2018

APSIPA

10th

Anniversary

The Era of Signal and Information Technologies and their Long-Term Prospective

 978-988-14768-4-5

CONTENTS

CONTENTS

Notes from Editors

Time flies. It is ten years now after the establishment of the Asia Pacific Signal and Information Association (APSIPA) in 2009, which is a non-profit making organization promoting broad spectrum of research and education activities in signal and information processing. The interest of APSIPA encompasses but not limited to signal and information processing, recognition, classification, communication, networking, computing, system design and implementation, security, and technology with applications to scientific, engineering, health, and social areas. We have to stress that APSIPA features with friendly collaborations and mutual fertilization of knowledge among members, and we work as members of a big family.

This is a magazine published for the celebration of the 10th Anniversary of APSIPA. Our design is to break away from the traditional thinking of nominal celebration magazines which mainly compose of congratulation statements and big headings of company names or individuals. We target at producing a magazine which reflects the body of knowledge of APSIPA, samples of technologies possessed by members and innovative ideas for short-term and long-term future development. The magazine should be long lasting with good impact to both academic and professional disciplines. We planned to make two to three rounds of Call-for-Articles. The beginning round was done by invitation only, and the last round is expected to be open to the public. However, the response is so overwhelming that the quota was filled soon after the 1st announcement. We would like to express our heartfelt thanks to all these support to the Magazine, and to members of the APSIPA as a whole. This shows our friendship, and the spirit of mutual fertilization of knowledge within the family of APSIPA.

We had specifically written in our submission guidelines asking authors (i) to write a short tutorial/introduction (in layman terms or with an overly simplified version of the theory) on one of the technologies possessed by the author in the hi-tech research area, or on a contemporary hi-tech topic with disputable argument, and (ii) to smartly predict the impact/effect/development of this technology in the coming 10 to 20 years.

We have now collected 42 articles written by APSIPA Presidents, Founders, Keynote Speakers, BoG Members, Annual Summit and Conference Organizers, Keynote Speakers, TC Chairs, Distinguished Lecturers and others. The articles include a wide range of current and new research topics, with critical comments, innovative ideas, creative suggestions and original contributions, which could be able to inspire readers with new ideas and research directions. Topics of the articles include artificial intelligence, learning, deep learning, convolutional neural networks, optimization, imaging, speech technology, videos, mobile technology, IoT and communications, robotics, mobile technology, DSP, filtering, automatic control, vehicle driving, air pollution, watermarking, DNA, data hiding, healthcare and monitoring. An electronic version of this magazine will also be available in the APSIPA web after the APSIPA-ASC 2018.

The Content of this Magazine starts from "Notes from Editors". This is followed by articles written by APSIPA (i) Presents and Founders, (ii) Current BoG Members, (iii) Past BoG Members, (iv) ASC Organizers & Keynote Speaker and (v) TC Chairs. The categorization of an article has been done mainly according to the role(s) indicated by the author in the article. For authors with multiple roles, either the 1st role or a popular one was chosen, quite arbitrarily. We are also glad to report that most authors of category (i) and (ii) have accepted our invitation to present their articles in three Overview Sessions of the APSIPA 10th Anniversary Magazine in APSIPA-ASC'2018, which are to be held on Wednesday, 14 November 2018, Hawaii.

The content is then followed by the 1st page of the Photo Gallery of each year's APSIPA-ASC. A link is provided at the end of each year allowing readers to access more photos from APSIPA website. An electronic version of the APSIPA 10th Anniversary Magazine is also available immediately after the APSIPA-ASC'2018. The magazine ends with samples of APSIPA Governance, such as APSIPA-By-Laws.

Editors:

Wan-Chi Siu,	President (2017-2018), APSIPA
Hitoshi Kiya,	President-Elect (2017-2018), APSIPA
Bonnie, N.F. Law	Secretary of EB & APSIPA Headquarters

Brief History of APSIPA:

APSIPA was formally established in 23 July 2009 as a non-profit making association, aiming at the promotion of research and education on signal processing, information technology and communications. Let us recall that the Association started with an initial meeting in Hawaii in April 2007, and was formally established in July 2009 in Hong Kong. The APSIPA-ASC comes back to Hawaii this year, 2018, which is the time that we celebrate the 10th Anniversary of APSIPA.

The best descriptions of the history APSIPA have been given by two of our founders: Prof. Sadaoki Furui and Prof. Ray Liu.

Prof. Furui said, *"During the 2007 IEEE ICASSP conference held in Hawaii, Prof. Ray Liu and I got the idea of creating an academic association encouraging research and education on signal and information processing in the Asia-Pacific region. The name APSIPA was decided at the 1stSteering Committee Meeting held at Tokyo Institute of Technology, Tokyo, Japan in December 2007, and after 3 more meetings, the inaugural APSIPA-ASC was held in Sapporo, Japan, in 2009."*

Prof. Liu said, *"It was April 2007, during IEEE ICASSP in Honolulu at the exact same locations of convention center and hotel, a group of us gathered together for a dinner to talk about if we might want to start an Asian/Pacific Association on signal and information processing for a very simple reason: The world's technology gravitation center has been shifting from Atlantic Ocean to Pacific Ocean. Yet there was no organized community to serve the specific needs for colleagues in Asian/Pacific regions.*

At the table were Sadaoki Furui, Lin-shan Lee, Soo-Chang Pei, Hideaki Sakai, Wan-Chi Siu, Min Wu, Meng-Hua Er, and myself. The consensus was unanimous that we shall take on this noble task, not because it is easy but because it is hard.

Sadaoki called for the first steering committee meeting in Tokyo in December 2007 to commence the activities of starting APSIPA. Afterward, we met in Las Vegas during ICASSP 2008, in Hong Kong in December 2008 hosted by Wan-Chi, and then in Taipei during ICASSP 2009, to lay out the groundwork to give birth to APSIPA and the hosting of the first Annual Summit and Conference (ASC) in Sapporo in October 2009."

After meetings in 2007 and 2008 in Tokyo, Las Vegas, Hong Kong and Taipei, more people joined APSIPA as founders or life-members. The inaugural APSIPA Annual Submit and Conference was held in Sapporo Japan in 2009. We subsequently held APSIPA Annual Summit and Conference in Singapore (2010), Xi'an, China (2011), Los Angeles, USA (2012), Kaohsiung, Taiwan (2013), Siem Reap, Cambodia (2014), Hong Kong SAR (2015), Jeju, Korea (2016), and Kuala Lumpur, Malaysia (2017), and in this year (2018) we hold it in Hawaii. The number of attendees increases from 240 in Sapporo to over 400 in the recent two years (Kuala Lumpur: 422, and Hawaii:450+).

Subsequent to running successfully a few APSIPA-ASCs, we launched of our open-access journal "APSIPA Transactions on Signal and Information Processing", published by the Cambridge University Press in 2012, with Prof. Antonio Ortega as the first EiC. In same year, 2012, both the Distinguished Lecturer Program and APSIPA Newsletters were lunched. The APSIPA Friend Labs Program was then launched in 2013, which is a means to increase knowledge exchange among members. In recent years, particularly in years 2016-2017, a few more awards were given to selected ASC papers for the ever increasing quality of our conferences. We have also enhanced the APSIPA activities with Summer Workshops, Sadaoki Furui Prize Paper Award, collaborations with IEEE, etc. In 2018, for the celebration of APSIPA 10th Anniversary, a Photo Gallery has been established on the APSIPA website, and the APSIPA 10th Anniversary Magazine is published and also available on the APSIPA website.

Editorial Board(EB) of APSIPA 10th Anniversary Magazine

Does AI Make People Happy?

Risk of Progress of Artificial Intelligence

In recent years, AI (artificial intelligence) technology has been making rapid progress. With the acceleration of computers and networks, the environment that collects, constructs, and uses big data has been developed, and effective learning methods of deep neural networks (DNNs), that is, deep learning has been realized. Along with that, studies on the impact of development of AI on employment, the relationship with ethics and human rights, and security risks are also being conducted.

Self-learning capability of AI

Why is recent AI effective? In conventional computer systems, information about contents of data and programs needed to be given by humans to recognize various information such as speech, image, video and language, and to process games such as Shogi and Go. On the other hand, in AI using DNNs, parameters for appropriately processing data can be automatically optimized by giving inputs and target outputs. In other words, the computer automatically understands and makes proper judgments.

Reasons why recent AI has become effective can be summarized as follows:

- *Advanced representation power*: With DNNs, almost any kind of complicated computation including nonlinear processing can be realized.
- *Learning capability*: Using the stochastic gradient descent method, network parameters for many complex computations can be effectively learned.
- *Distributed representation*: Symbols and symbol sequences, such as words and sentences, are expressed as points in a multidimensional vector space having a semantic distance relationship. For example, Paris - France + Italy ~ Rome, and king - male + female ~ queen.
- *Overall optimization*: By making all elements of the networks differentiable, the whole process going from input to output can be optimized. Conventionally, global optimization could not be achieved, since each element constituting complicated processing in voice and image recognition was individually optimized.
- *Representation learning*: Methods of extracting features of input data are automatically learned at the lower layer of the network.
- *Multi-level representation ability*: Even higher-order abstracted features are automatically learned in each layer of the network. Features common to various voices and various images are learned at layers closer to the input, and distinctions relevant to the name of a word or an object are learned at layers close to the output automatically.

With this background, DNNs of various structures are utilized in various areas of AI such as voice recognition, image recognition, automatic translation, information search, robots, and autonomous driving. Research on collecting and analyzing enormous literature and biological/genetic information to be used for preventive/diagnostic/early detection/treatment of diseases, and drug discovery is progressing. Even in marketing and investment, the use of DNNs is progressing.

Professor Sadaoki Furui

PhD, Life-FIEEE

A Founder and the 1st President of APSIPA (2009-2012)

President
Toyota Technological Institute at Chicago

Professor Emeritus
Tokyo Institute of Technology

Sadaoki Furui received the B.S., M.S., and Ph.D. degrees from the University of Tokyo, Japan in 1968, 1970, and 1978, respectively. After joining the Nippon Telegraph and Telephone Corporation (NTT) Labs in 1970, he has worked on speech analysis, speech recognition, speaker recognition, speech synthesis, speech perception, and multimodal human-computer interaction. From 1978 to 1979, he was a visiting researcher at AT&T Bell Laboratories, Murray Hill, New Jersey. He was a Research Fellow and the Director of Furui Research Laboratory at NTT Labs. He became a Professor at Tokyo Institute of Technology in 1997. He was Dean of Graduate School of Information Science and Engineering. He was given the title of Professor Emeritus in 2011. He is now serving as President of Toyota Technological Institute at Chicago (TTIC). He has authored or coauthored over 1,000 published papers and books. He was elected a Fellow of the IEEE, the Acoustical Society of America (ASA), the Institute of Electronics, Information and Communication Engineers of Japan (IEICE) and the International Speech Communication Association (ISCA). He received the Paper Award and the Achievement Award from the IEEE SP Society, the IEICE, and the Acoustical Society of Japan (ASJ). He received the ISCA Medal for Scientific Achievement, and the IEEE James L. Flanagan Speech and Audio Processing Award. He received the NHK (Japan Broadcasting Corporation) Broadcast Cultural Award and the Okawa Prize. He also received the Achievement Award from the Minister of Science and Technology and the Minister of Education, Japan, and the Purple Ribbon Medal from Japanese Emperor. He was accredited as Person of Cultural Merit by the Japanese Government in 2016.

What current AI cannot yet do or is not yet good at

Although AI has come to be widely used in recent years, there are still a lot of things that it cannot yet do or is not yet good at. They include:

- Explaining/interpreting the reasons for each judgment
- Logical thinking, abstraction, and reasoning
- Using hierarchical knowledge such as causality, inclusion relation, and context
- Having and using common sense
- Understanding the meaning of natural language and the flow of dialogue
- Dealing with new situations that were not included in the given learning data, except the cases when the target domain is “closed world” with clearly defined rules, such as Go and Shogi, where the computers can learn by themselves without using any data.
- General problem solving irrespective of domains, such as speech, image, and translation. (Artificial General Intelligence)
- Having a three-dimensional or four-dimensional time/space image (view of the world)
- Raising appropriate questions or problems
- Demonstrating creativity
- Having a mind/consciousness

DNN is, as it were, a black box, and it cannot explain why some specific output was obtained for a specific input. It cannot explain the reason. Therefore, it is difficult to pursue the cause in the event of an error, and it is unpredictable how it behaves for the target deviating from the learning data. It can even be said that it is unclear why the current DNNs do so well. This may be a problem when using AI for various real-life situations.

Even with these problems, AI progresses daily and many problems become closer to being solved.

The appearance of human-like AI

At an event held at the beginning of May 2018, there was a demo of the AI voice dialog system "Google Duplex," which became a hot topic. Google users can instruct the AI system with voice or keyboard to make a phone call to restaurants, hair salons, etc. on their behalf, to make reservations. The system calls a store with a natural human voice, talking to a store person while responding as if it is a human being. If it cannot make a reservation that is perfect for the user's condition, it can make a reservation close to the given condition through conversation with the store person.

This system uses speech recognition, dialogue understanding, and speech synthesis technology, developed using recursive neural network (RNN), etc. based on huge databases of human-human dialogues and human voices. Since the demonstrations show only successful cases, it is unclear how much general dialogue can be made and what variation of voices can be made, but it is certain that big technological progress is taking place.

It poses a problem when such an AI system resembling a human being is made. People who received calls from AI will respond and think that they are interacting with ordinary people, not knowing it is AI. Once he/she realizes that the opponent is actually AI by some chance, he/she feels tricked. In order to avoid such discomfort, as an ethical standard, AI probably needs to inform the other party in advance that it is AI.

In the case of the above demo, the AI should first report at the onset, "I am calling on behalf of my client." So, when that happens, will someone who received the call continue talking with AI happily? Does the AI need to respond like a human being? Would not rather simpler interaction be preferrable?

These days, not only the voice, but also a face image that looks exactly like someone, a natural face picture of a person who does not exist, and a video as if President Obama is talking, can be made. It gets harder to distinguish which is the natural one. It is getting harder for us to trust our eyes and ears.

Such a problem did not occur in a time when the robot apparently looked like a robot, but due to the rapid progress of technology, problems have come to occur. Boston Dynamics in U.S, which is leading the development of robots with motor skills such as humans and animals, has announced that it will release robots like animals with advanced abilities in the near future. They move around so naturally that, if they wear dog or donkey costumes, you will mistake them for real dogs or donkeys.

If AI can so effectively mimic a human, not only ethical problems will arise, but legal problems as well, such as the possibility of widespread cases of various fraud. Since we cannot stop advancing the technology, we will need to set and regulate appropriate standards on how to use it.

Does AI make people happy?

Advances in AI will benefit a large number of people, including the elderly and impaired, enrich all our lives, support our health, and contribute to our happiness. Although there is a possibility it may deprive some of work or make some jobs obsolete, new professions are also born. People will be released from hard labor.

However, some people point to the demo of "Google Duplex" is an example that only people who can use or access the technology (AI) will have a kind of privilege. Those who can use AI can leave the boring and hard work to AI, but those who may not have access will remain stuck to do the work themselves.

Also, some people are worried that communication with AI will become the dominant part of life, and communication between people will be lost. Even today, I see a case where people surrounding the same table communicate with each other online. Even among families, there are people who talk more online than speaking directly. When people start fighting via emails or online messages, it is sometimes difficult to stop. Communication with only text is producing extreme hate language. The

importance of talking is forgotten, and a socially rigid society is about to spread.

An even bigger problem is the application of AI to weapons. Unmanned aircraft weapons have already been created, but development of lethal autonomous weapon systems must be stopped. It is necessary to maintain a level of human control and not fully automate. Also, the use of AI weapons in potential terrorist activity must be stopped at all costs.

AI definitely will develop rapidly from now. It will be extremely important to demonstrate human wisdom and intelligence to control the usage and ensure responsible applications, avoiding the harmful or even unintended outcomes.

Acknowledgement: This article is based on discussion with Professor David McAllester at TTIC.

Photo Gallery: APSIPA-ASC'2011 in Xian

Wireless AI: Deciphering our World with a New Sixth Sense

What smart impact will future 5G and IoT bring to our lives? Many may wonder, and even speculate, but do we really know? With more and more bandwidth readily available for the next generation of wireless applications, many more smart applications/services unimaginable today may be possible. In this article, we will show that with more bandwidth, one can see many multi-paths, which can serve as hundreds of virtual antennas that can be leveraged as new degrees of freedom for smart life. Together with the fundamental physical principle of time reversal to focus energy to some specific positions and the use of machine learning, a revolutionary radio analytic platform can be built to enable many cutting-edge IoT applications that have been envisioned for a long time but have never been achieved. We will present the world's first ever centimeter-accuracy wireless indoor positioning systems that can offer indoor GPS-like capability to track human or any indoor objects without any infrastructure, as long as WiFi or LTE is available. Such a technology forms the core of a smart radios platform that can be applied to indoor tracking, home/office monitoring/security, radio human biometrics, and vital signs monitoring. In essence, in the future of wireless world, communication, as we see it, will be just a small component of what's possible. There are many more magic-like smart applications that can be made possible, allowing us to decipher our surrounding world with a new "sixth sense".

Multipaths, a new-found friend

In wireless communications, when a signal emitted from a transmitter gets reflected or scattered by a scatterer, an attenuated copy of the original signal is generated and reaches the receiver through a different path. The phenomenon is well known as the multipath propagation, which can cause destructive interference and degrades the performance of communication.

However, viewed from another perspective, the scatterers in the environment in fact act as virtual antennas/sensors that can be leveraged to offer some desirable outcomes. Just imagine that everyday human activities with motion and body movements affect wireless signal propagation surrounding us and thus change the channel profiles, and information about these activities is embedded in the signals. When signals get bounced back and forth by the scatterers, multiple "replicas" are generated, which contains enriched meaningful information about our activities. Each of such multipaths is in essence a degree of freedom naturally existing in our surrounding environment. They can be considered as tens or hundreds of virtual antennas ready to serve us on demand. The spatial resolution in resolving independent multipath components is determined by the transmission bandwidth. The larger the bandwidth, the better the spatial resolution and thus the more multipaths can be resolved, as shown in Fig. 1.

How to utilize the multipaths as virtual antennas/sensors? We find that a good starting point is to resort to the physics of time reversal (TR) and its focusing effect [1]. It has been well known that the convolution of the time-reversed waveform and the channel can generate a unique peak at the specific receiver's location, called the spatial focusing effect [2]. This indicates that the multipath channel profile works as a unique and location-specific signature, and the spatial focusing effect only happens when the channel can "match" the time-reversed waveform. By comparing the multipath CSI with a set of time-reversed CSI pre-collected at multiple known locations, one can infer the current location of a device, and this idea can be applied to assist positioning.

Since each multipath profile is in essence a focusing point on the "time-reversal logical space", if an event happens such as a door opens or closes that affects the

Professor Ray K.J. Liu

PhD, FIEEE

A Founder of APSIPA & Advisor, Advisory Board

Christine Kim Eminent Professor of Information Technology
Electrical and Computer Engineering Department
Distinguished Scholar-Teacher University of Maryland

Professor K. J. Ray Liu was named a Distinguished Scholar-Teacher of University of Maryland, College Park, in 2007, where he is Christine Kim Eminent Professor of Information Technology. He is the founder of Origin Wireless, Inc., a high-tech start-up developing smart radios for smart life. Professor Liu was a recipient of the 2016 IEEE Leon K. Kirchmayer Award on graduate teaching and mentoring, IEEE Signal Processing Society 2014 Society Award for "influential technical contributions and profound leadership impact", IEEE Signal Processing Society 2009 Technical Achievement Award, and more than a dozen best paper awards. Recognized by Web of Science as a Highly Cited Researcher, he is a Fellow of IEEE and AAAS. Professor Liu is IEEE Vice President, Technical Activities – Elect, He was Division IX Director of IEEE Board of Director, President of IEEE Signal Processing Society, where he has served as Vice President – Publications and Editor-in-Chief of IEEE Signal Processing Magazine. He also received teaching and research recognitions from University of Maryland including university-level Invention of the Year Award (three times), and college-level Poole and Kent Senior Faculty Teaching Award, Outstanding Faculty Research Award, and Outstanding Faculty Service Award, all from A. James Clark School of Engineering (each award honors one faculty per year from the entire college).

multipath, as a result the multipath profile is now mapped to another focusing point. If one can perform analytics or machine learning to distinguish both events, then one shall be able to infer what has happened. With this notion, one can further design various types of analytics based on the multipath CSI, which we refer to as the *radio analytic*. By fully exploiting the rich multipath CSI, radio analytic can decipher the propagation environment, reveal subtle information on various human activities, as if a new extended sixth human sense. Radio analytics can enable many cutting-edge IoT applications, such as accurate indoor positioning, tracking, wireless event detection, human recognition and vital signs monitoring, as we will illustrate in the sequel.

Centimeter-accuracy indoor positioning with Wi-Fi

Most of the existing indoor positioning systems (IPS) can only achieve a meter-level accuracy, and the performance becomes even worse in the non-line-of-sight (NLOS) condition because it is generally very difficult or even impossible to obtain precise measurements due to the rich scattering indoor environment. Since TR is able to focus the energy of the transmitted signal only onto the intended location, by utilizing a unique, location-specific CSI, the TR-based indoor positioning system (TRIPS) [3] can position a user by matching the CSI with the geographic allocation. Since spatial focusing is a half-wavelength focus spot, the TRIPS can achieve a 1- to 2-cm level positioning.

As discussed earlier, a large bandwidth is indispensable for resolving multipath CSI and high correlation of CSI values from different locations, because insufficient bandwidth can result in positioning ambiguity. To understand how bandwidth affects the accuracy of indoor positioning, we conducted extensive experiments in a typical indoor space. We deployed two channel sounders under an NLOS setting, with one of them placed on an experimental structure with a 5-mm measurement resolution. At each location of the experimental structure, we collected multiple CSI values as location-specific fingerprints under different bandwidths.

The experimental results in Fig.1 illustrate the TRRS distribution among the central location on the experimental structure and its nearby locations under different effective bandwidths [4]. The figure shows that 40 MHz of bandwidth is insufficient to distinguish nearby locations centimeters away. The ambiguity decreases significantly with an increasing bandwidth.

When the effective bandwidth reaches 360 MHz, the region of ambiguity shrinks to a ball with an approximately 1-cm radius, which indicates centimeter-level accuracy. Unfortunately, the bandwidths on mainstream 802.11nWi-Fi chips are merely 20 or 40 MHz, insufficient for centimeter-level indoor positioning. This motivates the formulation of a large effective bandwidth by exploiting the diversities on Wi-Fi devices, i.e., the frequency and spatial diversity.

Achieving centimeter accuracy via TRRS

Fig. 2 shows a generalized framework of diversity exploitation. More specifically, Fig. 2(a) shows an example of fusing CSIs from four different WiFi channels, while Fig. 2(b) demonstrates the way of fusing CSIs from four receiving antennas. Both diversities can be exploited at the same time as shown in Fig. 2(c), where CSIs on two WiFi channels and two receiving antennas are combined into one fingerprint. In WiFi, the frequency diversity is achieved by performing frequency hopping on different WiFi channels [5], and the spatial diversity can be achieved by collecting CSIs on multiple antenna links on multiple-input-multiple-output (MIMO) WiFi devices [6]. Denote the maximum frequency diversity by F, the maximum spatial diversity by S, and the physical bandwidth for each WiFi channel by B, the effective bandwidth is given as $S \times F \times B$. As one can see, diversity exploitation increases the effective bandwidth to $S \times F$ times compared to the physical bandwidth B.

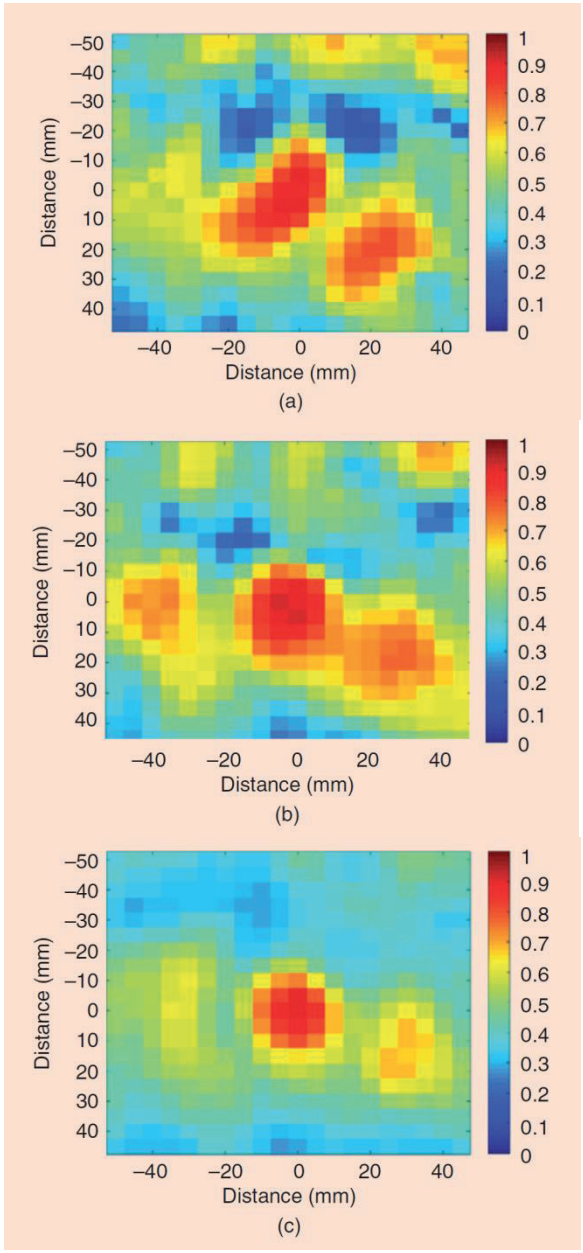


Fig. 1 Ambiguity among nearby locations under (a) 40 MHz bandwidth (b) 120 MHz effective bandwidth (c) 360 MHz effective bandwidth.

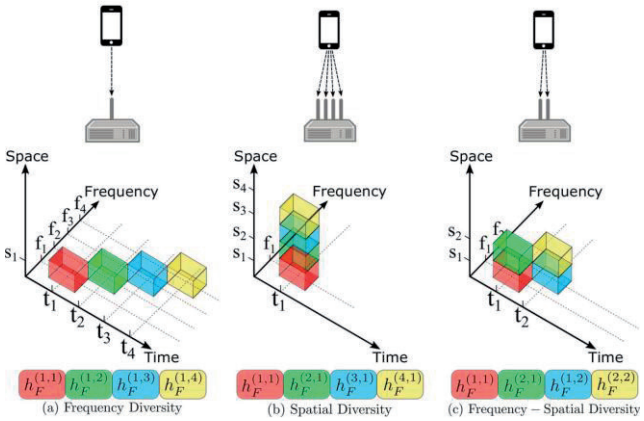


Fig. 2 Leveraging frequency and spatial diversities in WiFi to achieve large effective bandwidth.

Mapping-free indoor tracking with decimeter accuracy

The reliability of the centimeter-accuracy IPS discussed above depends on whether the CSI fingerprints in the offline database are up to date or not. If the environmental changes affect the CSI and thus degrade the positioning accuracy, the CSI database need to be updated that will increase the overhead of the IPS. To avoid the re-calibration of the CSI fingerprints, we analyze the TRRS spatial distribution and find it exhibits a periodic damping pattern that is only dependent on the moving distance [7][8][9]. Combining the moving distance estimation and moving direction estimation obtained from inertial measurement unit, as well as a map-based position correction to correct the accumulated error in tracking, we implement a low-complexity high-accuracy real-time indoor tracking system. Extensive experiments have verified that the system can achieve decimeter accuracy even under NLOS.

Wireless event detection

The variations in the multipath CSI due to different indoor events, such as opening or closing a door or window can also be characterized by TR technique. By treating each path of the multipath channel in arich-scattering environment as a distributed virtual antenna, a TR-based indoor event detection system [10][11] can be designed that takes the multipath CSI as the feature and determines the occurrence of an indoor event according to the current CSI in the propagation environment.

Human radio biometrics

Most of the current biometrics systems require special devices that capture human biometric traits in an extremely LOS environment, i.e., the subject should make contact with the devices. Considering the combination of all the physical characteristics and other biological features that affect the propagation of EM waves around the human body and how variable those features can be among different individuals, the chance for two humans to have the identical combinations is significantly small, no matter how similar those features are. Consequently, human radio biometrics, which record how the wireless signal interacts with a human body, can be viewed as unique among different individuals. Motivated by this finding, we have implemented a TR-based human identification system [12] that first refines the collected CSI to focus only on the human radio feature by removing the background information and then identifies the human by comparing the current CSI with those in a radio biometrics database.

Wireless vital signs monitoring

Since CSI can capture environmental perturbations due to human activities, the features of different human activities can be extracted by analyzing the CSI. Among the various human activities, breathing is an important one since it is a fundamental physiological function of human which can act as a vital indicator for their health status and predictor of medical conditions. Although breathing introduces very minor environmental perturbations and only affects CSIs slightly, the periodic pattern of breathing embedded in the CSI time series is very distinct and can be extracted with high fidelity by performing spectrum analysis on the CSI time series [4].

Long-term Impact

As a revolutionary platform that connects everything around the world, the IoT has dramatically changed our lifestyle and enabled us to measure and track everything connected to it. Because of the ubiquitous deployment of wireless devices, wireless sensing that can make many smart IoT applications possible has recently received a great deal of attention. As the next generation of wireless systems embraces a larger bandwidth, richer information can be revealed through wireless sensing, e.g., in the form of multipaths. As bandwidth increases, the number of multi paths that can be resolved also increases, allowing them to serve as hundreds of virtual antennas. Motivated by the physical principle of TR, we developed various radio analytics for smart IoT applications in indoor positioning, event detection, human recognition, and vital signs monitoring. Unlike conventional approaches for these applications, the proposed radio analytics approach can work well under NLOS and enjoys low implementation complexity, thus making it an ideal paradigm for smart IoT sensing, positioning, and tracking.

References

[1] B. Wang, Y. Wu, F. Han, Y.-H. Yang, and K. J. R. Liu, "Green wireless communications: A time-reversal paradigm," IEEE J. Sel. Areas Commun., vol. 29, no.8, pp. 1698–1710, Sept. 2011.

[2] Y. Chen, B. Wang, Y. Han, H. Q. Lai, Z. Safar, and K. J. R. Liu, "Why time reversal for future 5G wireless?" IEEE Signal Process. Mag., vol. 33, no. 2, pp.17–26, Mar. 2016.

[3] Z. H. Wu, Y. Han, Y. Chen, and K. J. R. Liu, "A time-reversal paradigm for indoor positioning system," IEEE Trans. Veh. Technol. (special section on Indoor localization, tracking, and mapping with heterogeneous technologies), vol. 64, no. 4, pp. 1331–1339, Apr. 2015.

[4] C. Chen, Y. Han, Y. Chen, F. Zhang, H. Q. Lai, B. Wang, and K.J. R. Liu, "TR-BREATH: Time-reversal breathing rate estimation and detection," IEEE Trans. Biomed. Eng, vol. PP, no.99, pp. 1-1.

[5] C. Chen, Y. Chen, Y. Han, H. Q. Lai, and K. J. R. Liu, "Achieving centimeter accuracy indoor localization on WiFi platforms: A frequency hopping approach," IEEE Internet Things J., vol. 4, no. 1, pp. 111–121, Feb. 2017.

[6] C. Chen, Y. Chen, Y. Han, H. Q. Lai, F. Zhang, and K. J. R. Liu, "Achieving centimeter accuracy indoor localization on WiFi platforms: A multi-antenna approach," IEEE Internet Things J., vol. 4, no. 1, pp. 122–134, Feb. 2017.

[7] F. Zhang, C. Chen, B. Wang, H. Q. Lai, and K. J. R. Liu, "A time-reversal spatial hardening effect for indoor speed estimation," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, New Orleans, Mar. 2017.

[8] F. Zhang, C. Chen, B. Wang, H. Q. Lai, and K. J. R. Liu, "WiBall: A time reversal focusing ball method for indoor tracking," unpublished, submitted to IEEE

Internet Things J., 2017.

[9] F. Zhang, C. Chen, B. Wang, H. Q. Lai, and K. J. R. Liu, "WiSpeed: A statistical electromagnetic approach for device-free indoor speed estimation," to appear, IEEE

Internet Things J., 2018.

[10] Q. Xu, Y. Chen, B. Wang, and K. J. R. Liu, "TRIEDS: Wireless events detection through the wall," IEEE Internet Things J., vol. 4, no. 3, pp. 723–735, June2017.

[11] Q. Xu, Z. Safar, Y. Han, B. Wang, and K. J. R. Liu, "Statistical learning overtime-reversal space for indoor monitoring system," IEEE Internet Things J., vol. PP, no. 99, pp. 1– 1, Jan. 2018.

[12] Q. Xu, Y. Chen, B. Wang, and K. J. R. Liu, "Radio biometrics: Human recognition through a wall," IEEE Trans. Inf. Forensics Security, vol. 12, no. 5, pp. 1141–1155, May 2017.

Photo Gallery: APSIPA-Pre-Establishment Summit in Tokyo, December 2007

Interpretable Convolutional Neural Networks

There is a resurging interest in developing a neural-network-based solution to the supervised machine learning problem. The convolutional neural network (CNN) will be our focus in this article. CNNs are widely used in the computer vision field nowadays. They offer the state-of-the-art solutions to quite a few vision and image processing problems such as object detection, scene classification, room layout estimation, semantic segmentation, image super-resolution, image restoration, object tracking, etc. Although CNNs are the main stream machine learning tool for big visual data analytics, they are used as black box tools by the great majority. Efforts have been made in the interpretability of CNNs based on various disciplines and tools. They include approximation theory, optimization theory and visualization techniques. Here, the CNN operating principle is explained using data clustering, rectification and transform, which are familiar to researchers and engineers in the signal and information processing community. As compared with other studies, this approach appears to be more direct and insightful. It is expected to contribute to further research advancement on CNNs.

Why Nonlinear Activation?

The need of nonlinear activation in CNNs is explained in [1],[2] using a RECOS (REctified COrrelation on a unit Sphere) model as shown in Fig. 1.

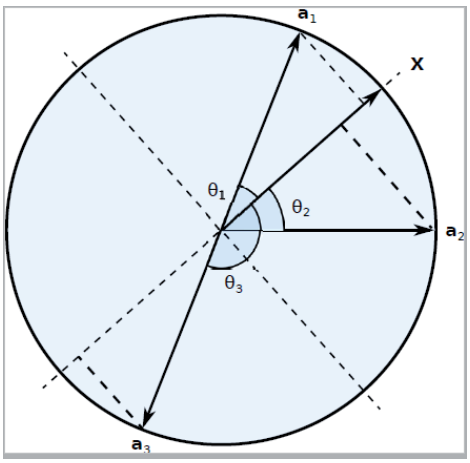


Figure 1: Illustration of the RECOS model, where $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$ denote three anchor vectors while $\mathbf{x}$ denotes an input vector on a high-dimension unit sphere. The convolution of $\mathbf{x}$ and $\mathbf{a}$ can be viewed as a projection. The projection values can be either positive or negative. There will be a sign confusion problem if two RECOS models are in cascade.

The convolution can be viewed as a correlation of an input vector and an anchor vector as shown in the figure. When two RECOS models are in cascade, we will have the following two confusion cases.

- A positive correlation followed by a link with a positive outgoing weight
- A negative correlation followed by a link with a negative outgoing weight.

Professor C.-C. Jay Kuo

PhD, FIEEE

The 2nd President of APSIPA (2013-2014)

Distinguished Professor of Electrical Engineering and Computer Science

University of Southern California

C.-C. Jay Kuo received the B.S. degree from the National Taiwan University and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology. He has been with the University of Southern California (USC) since January 1989. He is presently USC Distinguished Professor of Electrical Engineering and Computer Science and Director of the Multimedia Communication Laboratory. Dr. Kuo's research interests are in the areas of multimedia coding and communication, multimedia computing, deep learning and computer vision. Dr. Kuo is listed as the top advisor in the Mathematics Genealogy Project in the number of supervised PhD students. He has guided 145 students to their PhD degrees and supervised 26 postdoctoral research fellows. He is a co-author of about 272 journal papers, 916 conference papers, 39 patents, and 14 books. He delivered 36 keynote speeches and 700 invited lectures in conferences, research institutes, universities and companies. Dr. Kuo was Editor-in-Chief for the IEEE Transactions on Information Forensics and Security (2012-2014) and the Journal of Visual Communication and Image Representation (1997-2011). Dr. Kuo received the best paper awards from the Multimedia Communication Technical Committee of the IEEE Communication Society in 2005, from the IEEE Vehicular Technology Fall Conference (VTC-Fall) in 2006, and from IEEE Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP) in 2006. Dr. Kuo received the 1992 National Science Foundation Young Investigator (NYI) Award, the 1993 National Science Foundation Presidential Faculty Fellow (PFF) Award, the 1994 USC Northrop Junior Faculty Research Award, the 2007 Okawa Foundation Research Award, the 2010 Electronic Imaging Scientist of the Year Award, the 2010-11 Fulbright-Nokia Distinguished Chair in Information and Communications Technologies, the 2011 Pan Wen-Yuan Outstanding Research Award, the 2014 USC Northrop Grumman Excellence in Teaching Award, the 2016 USC Associates Award for Excellence in Teaching, the 2016 IEEE Computer Society Taylor L. Booth Education Award, the 2016 IEEE Circuits and Systems Society John Choma Education Award, the 2016 IS&T Raymond C. Bowman Award and the 2017 IEEE Leon K. Kirchmayer Graduate Teaching Award. Dr. Kuo is a Fellow of AAAS, IEEE and SPIE.

Similarly, we have the other two confusing cases:

- A positive correlation followed by a link with a negative outgoing weight
- A negative correlation followed by a link with a positive outgoing weight.

Because of these confusion cases, the CNN will encounter difficulty in resolving the white and the black cats in Fig. 2.

Figure 2: Two negatively correlated cat images.

The nonlinear activation unit imposes a constraint on the output so that only the positive correlation goes through while the negative correlation is blocked. This is especially obvious for the operation of the Rectified Linear Unit (ReLU). With such an interpretation, the nonlinearity of CNNs is less mysterious. It sheds light on the filter weight determination for the convolutional and the fully connected layers.

Filter Weights Selection or Initialization

In the training of CNNs, most work adopts random initialization for filter weights determination and, then, uses the backpropagation process to update them until the cost function converges. The work in [3] treats the filter weight selection process as a dimension reduction problem for the convolutional layers. Based on this new viewpoint, it adopts the principal component analysis (PCA) technique to determine the initial filter weights. The weights of incoming links of a node corresponds the kernels of a leading PCA component. The number of nodes indicates the number of selected PCA components.

Furthermore, our recent studies in [4] show that the filter weights in the fully connected (FC) layers can be selected efficiently by interpreting the operations in FC layers as a cascade of multiple linear-least-squared (LLS) regression processes.

Alternative Approach

The CNN provides an end-to-end solution to the supervised learning problem. The network architecture allows backpropagation, yet it has side effects such as adversarial attacks. The work in [3] presents a new transform, called the Subspace approximation with augmented kernels (Saak) transform, to extract features from the input data. Then, the extracted features can be fed into well-known classifiers such as the support vector machine (SVM), the random forest (RF) and the multi-layer perceptron (MLP) classifiers. It goes back to the traditional pattern recognition framework by dividing feature extraction and decision making into two separate modules.

If the number of object classes is fixed, the training and testing data are reasonably correlated, and the cost function is suitably defined, the end-to-end optimization approach taken by CNNs offers the state-of-the-art performance. However, this end-to-end optimization methodology has its weaknesses: 1) robustness against perturbation [5]; 2) scalability against the class number; and 3) portability among different datasets.

In contrast with CNN's end-to-end optimization methodology, the modular design such as the Saak transform approach is expected to be more robust against perturbation and less sensitive to the variation of object classes. The small perturbation will not affect leading last-stage Saak coefficients much due to the use of PCA. Kernels in earlier Saak transform stages should not change much if their covariance matrices do not change much. More comparison should be conducted to get a more conclusive result.

Long-Term Impacts

Data-driven signal and information processing is an emerging field because of the limitation of a mathematical or physical modeling approach. By extracting features from a large set of data such speeches, texts, images, videos, etc. via neural networks or successive PCA stages, one can derive a rich set of representations of data. To use images as an example, the representation can range from the pure spatial domain to the pure spectral domain as the two extremes. There is a family of joint spatial-spectral representations with various spatial-spectral tradeoffs between them.

Big data collection is easier than big data labeling. Thus, the trend is expected to move from supervised learning to weakly supervised (or unsupervised) learning. The latter is very challenging. Furthermore, statistics will play an important role in big data analytics. It is still not used as extensively as desired. By incorporating statistics, I believe that we can reduce a lot of computational burdens. Besides, it will make the black box of CNNs more transparent.

In my opinion, the recent success of many data understanding tasks is primarily attributed to big data and statistics. There is little intelligence such as inference in the neural-network-based solutions. In other words, we are still far away from human intelligence. This is still a long way to go to make machines as powerful as human brains.

References

[1] C.-C. Jay Kuo, "Understanding convolutional neural networks with a mathematical model," the Journal of Visual Communications and Image Representation, Vol. 41, pp. 406-413, November 2016.

[2] C.-C. Jay Kuo, "The CNN as a guided multi-layer RECOS transform," the IEEE Signal Processing Magazine, Vol. 34, No. 3, pp. 81-89, May 2017.

[3] C.-C. Jay Kuo and Yueru Chen, "On data-driven Saak transform," the Journal of Visual Communications and Image Representation, Vol. 50, pp. 237-246, January 2018.

[4] C.-C. Jay Kuo, Siyang Li and Min Zhang, "Unveiling convolutional neural networks (CNNs) and its application to effective filter weight initialization," in preparation.

[5] Yueru Chen, Zhuwei Xu, Shanshan Cai, Yujian Lang and C.-C. Jay Kuo, "A Saak transform approach to efficient, scalable and robust handwritten digits recognition," Picture Coding Symposium, San Francisco, California, USA, June 24-27, 2018.

Random Forests for Fast Image/Video Super-resolution

Image super-resolution (SR) usually refers to an increase of the resolution of a single low-resolution (LR) image to form a high-resolution (HR) image which should preserve the characteristics of natural image, such as sharp edges and rich texture. This is a difficult and ill-posed problem, since a number of unknown pixels have to be inferred from very limited information. However, due to the demand of hi-tech applications, including image/video upsizing for mobiles, ultra-high definition TV, visual surveillance, human and face recognition, this topic has recently arisen the attention of many researchers.

Let us also mention a few exceptional applications. Have you thought of using a single camera (perhaps with wide-angle) to capture the scene of a marriage ceremony and have the key persons clearly shown on the scene (or on a separate screen) as shown in fig.1A(a)? This is miracle! This can be done by using super-resolution techniques for the video. Since the number of key persons is very small, we can perform "content specific super-resolution", i.e. sample videos of the key persons can be used for training in our random forests/tree structure to form regression models, which are subsequently used for real-time video enlargement. This allows exceptionally high quality enlargement. Further examples include the use of learning based super-resolution for zooming the football matches, staff identification in company surveillance systems, etc.

Fig.1A(a): targeted super-resolution video for marriage ceremony

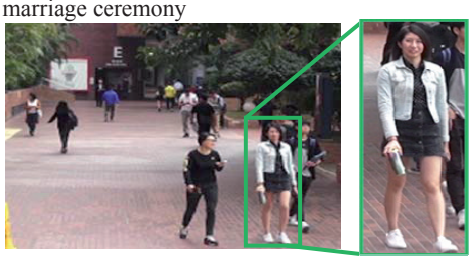


Fig.1A(b): Video super-resolution in Surveillance

Conventional approaches starting from simple bi-cubic interpolation to reconstruction-based approaches using geometric duality, nonlocal similarity, etc. had been investigated thoroughly. Machine learning by making use of k-nearest neighbors (k-NN), sparse coding and convolution neural network (CNN) appears to win and give substantial improvement over the conventional approaches. Most of these make use of deep learning with various neural network structures. However, the speed of realization is usually very slow. Inspired by its possible fast speed of realization, we have proposed to use random forests/trees for super-resolution. There is no study of using the random forests/trees for image/video super-resolution in the literature in early years. Our study has shown that the random forest structure can perform high-speed image super-resolution with resultant image quality comparable to or even better than the state-of-the-art super-resolution technologies, with a speed of 5 to more than 10 times faster, compared with conventional and CNN approaches.

There are many issues which make the random forests approach successful. For example, in our approach the super-resolution is done in two-phases: (i) Classification and then (ii) Regression as shown in fig.2. For Classification, a very simple split function is used, which provides a way to classify training blocks into an

Professor Wan-Chi Siu

PhD, DIC, Life-FIEEE

President (2017-2018) APSIPA

Emeritus Professor

(Former Chair Professor, HoD(EIE) and Dean of Engineering Faculty)

Department of Electronic & Information Engineering, Hong Kong Polytechnic University

WAN-CHI SIU *received the MPhil and PhD degrees from The Chinese University of Hong Kong in 1977 and Imperial College London in 1984. He is Life-Fellow of IEEE, Fellow of IET and HKIE, and President (2017-2018) of APSIPA (Asia-Pacific Signal and Information Processing Association). Prof. Siu is now Emeritus Professor, and was Chair Professor, HoD(EIE) and Dean of Engineering Faculty of The Hong Kong Polytechnic University. He is an expert in DSP, fast algorithms, super-resolution videos, and machine learning for object tracking. He has published 500 research papers (over 200 are international journal papers) and has 9 recent patents granted. Prof. Siu was also an independent non-executive director(2000-2015) of a publicly-listed video surveillance company and convenor of the First Engineering/IT Panel of the RAE(1992/93) in Hong Kong. He is an outstanding scholar, with many awards, including the Best Faculty Researcher Award and IEEE Third Millennium Medal (2000). Prof. Siu is/was Guest Editor/Subject Editor/AE for IEEE Transactions on CAS, IP, CSVT, and Electronics Letters, and organized very successfully over 20 international conferences including IEEE society-sponsored flagship conferences, such as TPC Chair of ISCAS1997 and General Chair of ICASSP2003 and ICIP2010. He was a VP(2012-2014) of the IEEE Signal Processing Society, is now a member of Fourier Award Committee and some other IEEE committees.*

extremely large number of classes. "Extremely large number classes" is good since it means blocks with really close features are grouped within the same class. A regression model is formed for each class. Hence simple split functions were formed and regression models (say C^1_4 or C^n_2 in fig.2) were built during the training stage. This is the reason why the super-resolution can be done very fast and efficiently. For Regression, a learned regression model is reached by a block very quickly by making use of the trained split functions, and high quality super-resolution results are then obtained. Note that using random trees/forests for super-resolution is novel, which opens up a new research direction for research in fast super-resolution.

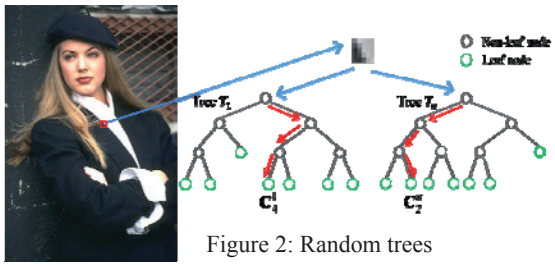


Figure 2: Random trees

Innovative ideas, such as the comparative splitting function, formation of feature pool with various filters, residual signal modeling, manifold formulation, etc. have also been proposed to push up the speed and quality of random forest super-resolution. Fig.3 shows a technology road map of interpolation in terms of PSNR improvement. It is seen that the quality improves gradually in these few years, and the newly improved results are mainly obtained by machine learning approaches. It is also true deeper learning (with more layers, see fig.4) gives better results. Interestingly all these approaches bear some similarity with the recent development of CNN for super-resolution.

Long-term Impact

Deep learning in NN (such as CNN) and other associated techniques have brought to a big jump in super-resolution quality, but the realization speed is usually very slow. Very often the deeper the learning and the larger the data set are, the better the quality is. To us, in simple terms, the system then covers more comprehensively all possible appearances, such that fine classifications are achieved. Hence this results in better quality. With the random forests structures, we have found another way to achieve the same quality, with possibility of even faster speed. We do not mean to beat the CNN. This just means to create a new direction for fast super-resolution. In the coming 10 to 20 years, there could be more creative systems other than the Neural Networks that can achieve the same goal. Should the structure again be "first-classification-then-regression"?

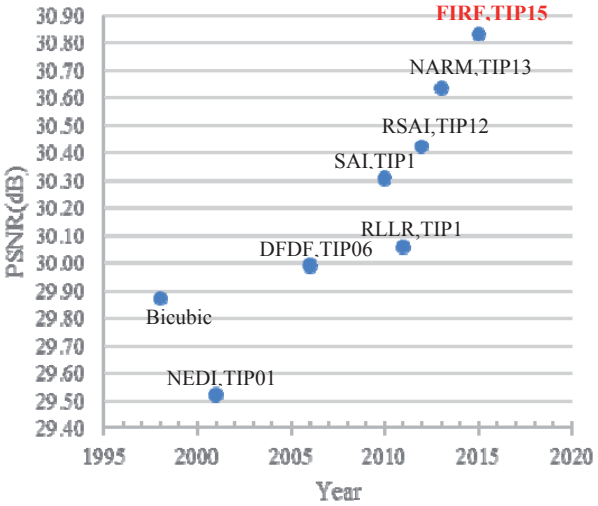


Figure 3: Technology road map of conventional Image interpolation development.

References

[1] Wing-Shan Tam, Chi-Wah Kok and Wan-Chi Siu, "A Modified Edge Directed Interpolation for Images", pp.13011_1-20, Journal of Electronic Imaging, Vol.19(1), 013011, Jan-March 2010.

[2] K.W. Hung and W.C. Siu, "Robust soft-decision interpolation using weighted least squares," *IEEE Trans. Image Process*., vol.21, no.3, pp.1061-1069, March 2012.

[3] He He and Wan-Chi Siu, "Single Image Super-Resolution using Gaussian Process Regression", Proceedings, pp.449-456, IEEE Computer Vision and Pattern Recognition Conference (CVPR 2011), June 20-24, 2011, Crowne Plaza, Colorado, USA.

[4] Kwok-Wai Hung and Wan-Chi Siu, "Novel DCT-based Image Up-sampling using Learning-based Adaptive k-NN MMSE Estimation ", pp.2018-2033, Vol.24, No.12, IEEE Transactions on Circuits and Systems for Video Technology, December 2014.

[5] Jun-jie Huang, Wan-Chi Siu and Tian-Rui Liu, "Fast Image Interpolation via Random Forests", pp.3232-3245, Vol.24 No.10, IEEE Transactions on Image Processing, October 2015.

[6] Jun-Jie Huang and Wan-Chi Siu, "Learning Hierarchical Decision Trees for Single Image Super-Resolution", pp. 937-950, Vol.27, No.5, IEEE Transactions on Circuits & System for Video Technology, May 2017.

[7] Zhi-Song Liu, Wan-Chi Siu and Yui-Lam Chan, "Fast Image Super-Resolution via Randomized Multi-Split Forests", Proceedings, pp.2747-2750, IEEE International Symposium on Circuits and Systems (ISCAS'2017), 28-31 May 2017, Baltimore, MD, USA.

[8] Yu-Zhu Zhang, Wan-Chi Siu, Zhi-Song Liu, Ngai-Fong Law, "Learning Via Decision Trees Approach for Video Super-Resolution", Proceedings, pp.558-562, 2017 International Conference on Computational Science and Computational Intelligence (CSCI'17), 14-16 December 2017, Las Vegas, USA.

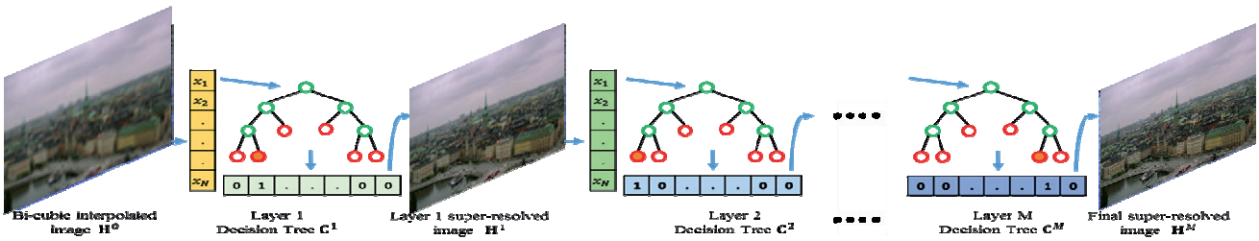


Figure 4: Deeper Learning using random trees structure

Compressible and Learnable Encryption for Untrusted Cloud Environments

With the wide/rapid spread of distributed systems for information processing, such as cloud computing and social networking, not only transmission but also processing is done on the internet. Therefore, a lot of studies on secure, efficient and flexible communications have been reported. Moreover, huge training data sets are required for machine learning and deep learning algorithms to obtain high performance. However, it requires large cost to collect enough training data while maintaining people's privacy. Nobody wants to include their personal data into datasets because providers can directly check the data. Full encryption with a state-of-the-art cipher (like RSA, or AES) is the most secure option for securing multimedia data. However, in cloud environments, data have to be computed/manipulated somewhere on the internet. Thus, many multimedia applications have been seeking a trade-off in security to enable other requirements, e.g., low processing demands, and processing and learning in the encrypted domain,

Accordingly, we first focus on compressible image encryption schemes, which have been proposed for encryption-then-compression (EtC) systems, although the traditional way for secure image transmission is to use a compression-then encryption (CtE) system. EtC systems allow us to close unencrypted images to network providers, because encrypted images can be directly compressed even when the images are multiply recompressed by providers. Next, we address the issue of learnable encryption. Cloud computing and machine learning are widely used in many fields. However, they have some serious issues for end users, such as unauthorized access, data leaks, and privacy compromise, due to unreliability of providers and some accidents.

Compressible Image Encryption, EtC System

A block scrambling-based image encryption scheme has been proposed for EtC systems with the assumption of the JPEG standard as a compressible image encryption scheme [1], in which a user wants to transmit image I securely to an audience, via a social networking Service (SNS) provider like Twitter or a cloud photo storage service (CPSS) such as Google Photos, as illustrated in Fig.1.

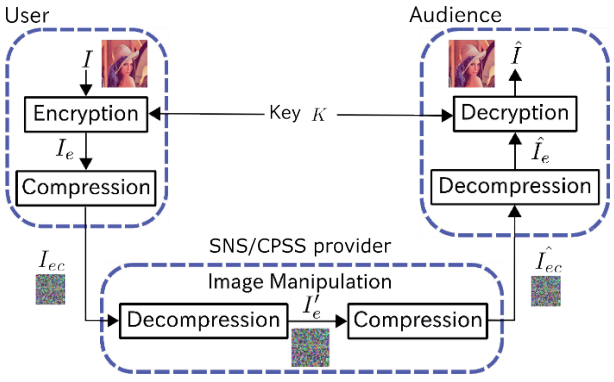


Fig. 1 EtC system

Professor Hitoshi Kiya

PhD, FIEEE

President-Elect (2017-2018) APSIPA

Professor

Department of Computer Science

Tokyo Metropolitan University

Hitoshi Kiya received his B.E and M.E. degrees from Nagaoka University of Technology, in 1980 and 1982 respectively, and his Dr. Eng. degree from Tokyo Metropolitan University in 1987. In 1982, he joined Tokyo Metropolitan University, where he became a Full Professor in 2000. From 1995 to 1996, he attended the University of Sydney, Australia as a Visiting Fellow. He is a Fellow of IEEE, IEICE and ITE. He currently serves as President-Elect of APSIPA, and he served as Inaugural Vice President (Technical Activities) of APSIPA from 2009 to 2013, and as Regional Director-at-Large for Region 10 of the IEEE Signal Processing Society from 2016 to 2017. He was also President of the IEICE Engineering Sciences Society from 2011 to 2012, and he served there as a Vice President and Editor-in-Chief for IEICE Society Magazine and Society Publications. He was Editorial Board Member of eight journals, including IEEE Trans. on Signal Processing, Image Processing, and Information Forensics and Security, Chair of two technical committees and Member of nine technical committees including APSIPA Image, Video, and Multimedia Technical Committee (TC), and IEEE Information Forensics and Security TC. He has organized a lot of international conferences, in such roles as TPC Chair of IEEE ICASSP 2012 and as General Co-Chair of IEEE ISCAS 2019. He has received numerous awards, including six best paper awards.

Because the user does not give the secret key K to the provider, the privacy of image I to be shared is under control of the user even when the provider recompresses image I. Therefore, the user is able to control image privacy for his own demand. However, in CtE systems, the user has to disclose unencrypted images to recompress them.

Figure 2 illustrates the procedure of the block scrambling-based encryption with the block size of 16 × 16, which consists of four encryption steps. An example of an encrypted image is shown in Fig.3(b); Fig.3(a) is the original one. This Image encryption scheme has been extended as a grayscale-based image encryption one to enhance the security of EtC systems [2]. An example of an encrypted image is shown in Fig. 3 (c), in which the block

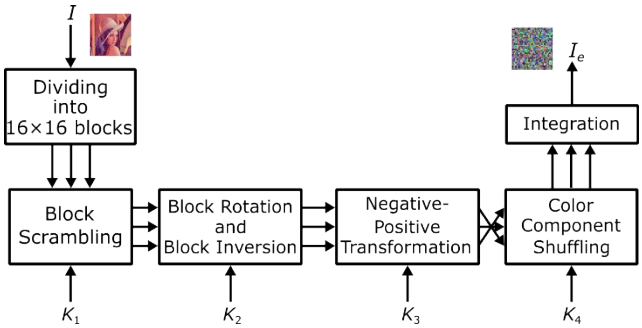


Fig. 2 Block scrambling-based image encryption

size is smaller, the number of blocks is larger, and the encrypted image includes less color information, than Fig.3 (b). Images encrypted using these schemes have the following properties.

(a) The compression efficiency for encrypted images is almost the same as that for the original ones under the use of JPEG compression.
(b) Robustness against various attacks has been demonstrated.

Figure 4 shows the rate-distortion (RD) curves of JPEG compressed images without any encryption and with the block scrambling-based encryption, where the average bitrate and PSNR values of 20 images are plotted, after decrypting the images. The encrypted images were found to not be affected by JPEG compression.

Application to SNS and CPSS

SNS providers like Twitter and Facebook, and CPSS providers like Google Photos are generally well known to manipulate images uploaded by users, and to support the JPEG standard, as one of the most widely used image compression standards. For this reason, images protected by almost all encryption schemes such as RSA are inapplicable to the providers, while images encrypted by the aforementioned schemes are applicable to most of the providers if encrypted images meet some conditions.

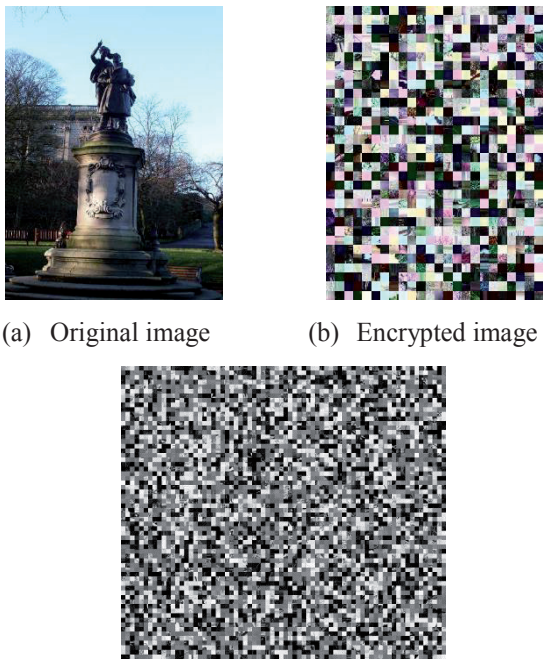
In [3], EtC systems with the block-based image encryption have been applied to SNS providers. In one experiment, encrypted and non-encrypted JPEG images were uploaded to various SNS providers to determine the robustness of the EtC systems under various conditions. Moreover, the paper investigated how each SNS provider manipulates uploaded images, and the encryption schemes used in the EtC systems were evaluated in terms of the robustness against image manipulation by SNS providers.

Security Evaluation against Jigsaw Puzzle Solver Attacks

Security analysis for EtC systems is needed, because the encryption schemes used in EtC systems do not have provable security. Therefore, safety has been evaluated first based on its key space assuming brute-force attacks, and the schemes have generally been shown to have enough key spaces for protecting against such attacks. However, each block in encrypted images has almost the same correlation as that of original images, which are needed to maintain a high compression performance. Jigsaw puzzle solvers, which utilize the correlation between pieces, have been actively studied in the area of computer vision, and they have succeeded in solving puzzles with a large number of pieces. We can regard the blocks of an encrypted image as pieces of a jigsaw puzzle. In [4]-[5], jigsaw puzzle solver attacks were discussed in addition to brute-force attacks, as a ciphertext-only attack (COA). Some solvers have been shown to be able to decrypt encrypted images even when the key space is large enough. However, assembling jigsaw puzzles becomes difficult under the following conditions.

(a) The number of blocks is high.
(b) The block size is small.
(c) The encrypted images include JPEG distortion.
(d) The images have no color information

Other attacking strategies such as known-plaintext attack (KPA) and chosen-plaintext attack (CPA) should be considered for the security. Block scrambling-based image encryption becomes robust against KPA through assigning a different key to each image for the encryption. In addition, the keys used for the encryption do not need to be disclosed because the encryption scheme is not public key cryptography. Therefore, the encryption can avoid the CPA unlike public key cryptography.


(a) Original image (b) Encrypted image

(c) Grayscale-based encrypted image

Fig. 3 Example of encrypted images

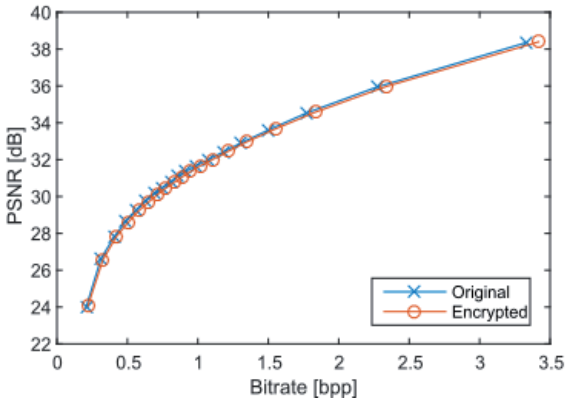
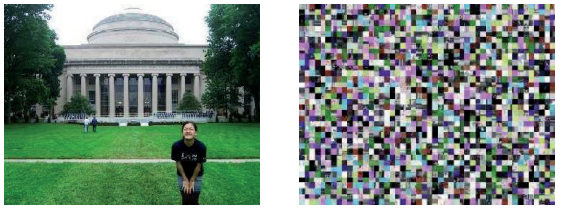
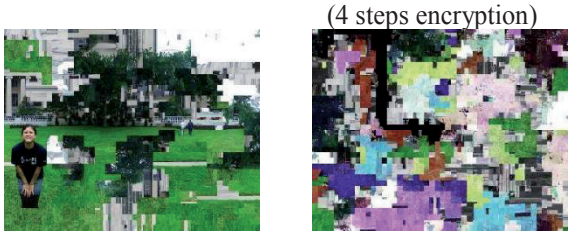


Fig. 4 RD curves of original images and encrypted ones



(a) Original image (b) Encrypted image (4 steps encryption)



(c) Assembled image (1 step encryption) ($D_c = 0$, $N_c = 0.3$, $L_c = 0.1$)

(d) Assembled image (4 step encryption) ($D_c = 0$, $N_c = 0$, $L_c = 0$)

Fig. 5 Examples of encrypted images and assembled images.

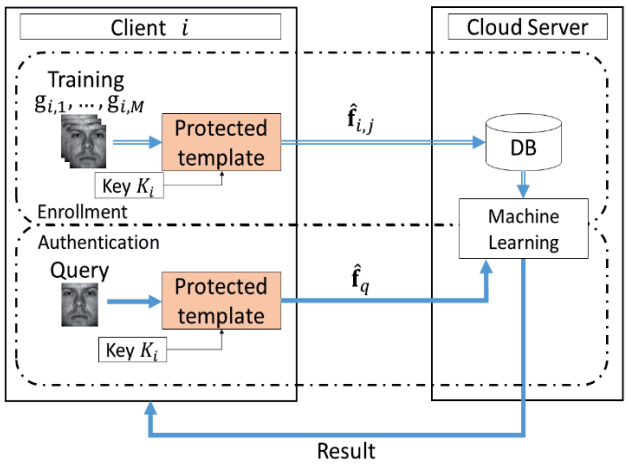


Fig. 6 Learnable encryption for machine learning

Figure 5 illustrates examples of an encrypted image and the assemble images, where three measures ware used to evaluate the results: direct comparison (D_c), neighbor comparison (N_c), largest component (L_c). In the measures, $D_c, N_c, L_c \in [0, 1]$, a larger value means a higher compatibility. Encryption with four steps in Fig. 1 was shown to make assembling images more difficult, than encryption with one step, i.e. with only block scrambling.

Learnable Image Encryption

Considerable efforts have been made in the fields of fully homomorphic encryption and multi-party computation [6]. However, these encryption schemes are still difficult to be applied to learning algorithms, although some attempts have been made to deep learning [7]-[8]. Moreover, the schemes require preparing algorithms specialized for computing encrypted data, and high computational complexity. In addition, the latest book on the subject [9] demonstrates the severity of the problem by providing a taxonomy of attacks and studies of adversarial learning. It also analyzes conventional attacks as well as the latest discovered weaknesses in deep learning systems.

Furthermore, privacy preserving computing schemes without homomorphic encryption and multi-party computation have also been considered for machine learning and deep learning [10]-[12]. They allow not only a light-weight computing cost, but also direct the computation of typical learning algorithms, without preparing any algorithms specialized for secure computing. In those methods, the block scrambling-based encryption in Fig.1 plays an important role as it does in EtC systems. Figure 6 illustrates the scenario of the privacy preserving computing. In the enrollment, client i, prepares training samples g_i, such as images, and a feature set $\mathbf{f}_{i,j}$, called a template, is extracted from the samples. Next the client creates a protected template set $\hat{\mathbf{f}}_{i,j}$ using a secret key p_i and sends the set to a cloud server. The server stores it and implements learning with the protected templates for a machine learning algorithm. In the authentication, Client i creates a protected template as a query and sends it to the server. The server carries out a classification problem with a learning model prepared in advance, and then it returns the result to Client i. Note that the cloud server has no secret keys and that the classification problem can be directly carried out using well-known algorithms.

Long-term Impact

Communications, computing and data storage environments have been dramatically changing. A variety of defenses have to be considered for learning systems and various attack types in the new environments, although considerable efforts have been made in the field of information security and forensics.

For example, the EtC systems described in this article are applicable to only still images, so EtC systems for video data have not been developed yet, due to the difficulty in performing motion estimation algorithms in the encrypted domain. Moreover, most signal processing-friendly encryption schemes such as block scrambling-based ones have no provable security, but like our house keys, they are absolutely necessary for our lives. Therefore, new evaluation measures of the safety should be discussed, because encryption schemes without provable security have a lot of attractive features. Robustness against jigsaw puzzle solver attacks is one of the measures.

Machine learning and deep learning systems are very powerful tools in many fields. However, huge training data sets are required for the learning, and the data, such as surveillance data are generally sensitive. Therefore, privacy-preserving computing schemes are required to utilize various learning systems safely for our lives. Unfortunately, those computing schemes have not been sufficiently developed yet. In other words, a lot of research subjects still need to be conducted in this field. That is very fortunate for researchers.

References

[1] K. Kurihara, M. Kikuchi, S. Imaizumi, S. Shiota, and H. Kiya, “An Encryption-then-Compression System for JPEG / Motion JPEG Standard,” IEICE Trans. Fundamentals, vol.E98-A, no.11, pp.2238–2245, November 2015.

[2] W. Sirichotedumrong, T. Chuman, S. Imaizumi, and H. Kiya,"Grayscale-based Block Scrambling Image Encryption for Social Networking Services," Proc. IEEE International Conference on Multimedia and Expo, July, 2018.

[3] T. Chuman, K. Iida, and H. Kiya, "Image Manipulation on Social Media for Encryption-then-Compression Systems," Proc. APSIPA Annual Summit and Conference, December, 2017.

[4] T. Chuman, K. Kurihara, and H. Kiya, "On the Security of Block Scrambling-based ETC Systems against Jigsaw Puzzle Solver Attacks," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp.2157–2161, March, 2017.

[5] T. Chuman, K. Kurihara, and H. Kiya, "On the Security of Block Scrambling-based EtC Systems against Extended Jigsaw Puzzle Solver Attacks," IEICE Trans. Inf. & Sys., vol.E101-D, no.1, pp.37–44, January 2018.

[6] M. Barni, G. Droandi, and R. Lazzeretti, "Privacy Protection in Biometric-based Recognition Systems: A Marriage between Cryptography and Signal Processing," IEEE Signal Processing Magazine, vol. 32, no. 5, pp. 66–67, 2015.

[7] N. Dowlin, R. Giladbachrach, K. Laine, K. E. Lauter, M. Naehrig, and J. Wemsing, "Cryptonets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy," MSR-TR-2016-3, Microsoft Tech Report, 2016.

[8]Y. Wang, J. Lin, and Z. Wang, "An Efficient Convolution Core Architecture for Privacy-Preserving Deep Learning," Proc. IEEE International Symposium on Circuits and Systems, April 2018

[9] D. Joseph, B. Nelson, Benjamin I. P. Rubinstein, and J.D. Tygar," Adversarial Machine Learning," Cambridge University Press, August 2018

[10]M.Tanaka, "Learnable Image Encryption", arXiv:1804.00490, March 2018.

[11] A. Kawamur, T. Maekawa, Y. Kinoshta, and H. Kiya, "SVM Computing using EtC Images in the Encrypted Domain," Technical Report of IEICE, SIS2018-1, pp.1-6, June, 2018.

[12] T. Nakachi and H. Kiya, " Practical Secure OMP Computation and Its Application to Image Modeling," Proc. International Conference on Information Hiding and Image Processing, September, 2018.

Photo Gallery: APSIPA ASC'2017 in Kuala Lumpur

Data Analytics for the Smart Grid

Introduction

In recent years there has been a proliferation of data that has appeared in numerous engineering applications. Massive amounts of heterogeneous data are becoming available with deployment of sensor networks and the Internet of Things (IoT).Engineering systems such as the electric power grid are currently modeled by systems of dynamical equations that describe the physical properties of the electric grid including current, voltage, and frequency.

Like many engineering applications, we are seeing an increasingly large influx of data for the electric power grid. At the transmission level, Phase Measurement Units (PMU)s provide a wealth of information about electrical grid measurements. Additional data is provided at the distribution level with sensors such as Advanced Metering Infrastructure (AMI), environmental data from weather sensors, and even social data about customer usage levels of electricity. This additional data needs to be processed to get useful information and then integrated with physical models of the electric power grid in order to make intelligent decisions about grid operations and security. This is combined with increasing penetration of distributed renewable energy generation (e.g. solar photovoltaic (PV) panels), more electric vehicles, more forms and usage of energy storage, and demand response programs. The decision making is enhanced through the use of control theory, optimization methods, signal processing, and machine learning.

This summary gives a high level discussion of some problems that have been formulated to enhance decision making along with activities in this area that have been presented in the APSIPA community. This consists of getting information about the electrical grid and environment via sensor networks, interpreting information received via signal processing and machine learning, and then using the information to make intelligent decisions about the grid using control and optimization algorithms. The focus is on the electrical grid beyond the last substation, the distribution grid. For the smart distribution grid there is an increasing amount of distributed renewable energy sources and possible distributed storage. This necessitates gathering more information about the electrical grid, environmental data, and building energy usage. With this information we can forecast distributed renewable energy sources and develop algorithms for distributed state estimation. We can then develop demand response algorithms to control loads (e.g. appliances, thermostats, air conditioners, hot water heaters).

In the second section we discuss some past activities in using signal and information processing for the electric power grid. This includes the activities that have occurred in the APSIPA community and some of the activities that have occurred in the IEEE and also the United States. The third section gives a high level summary of some activities that my group has worked on in this area. Finally, we summarize the research and suggest directions for further directions.

Professor Anthony Kuh

Ph.D., FIEEE

President-Elect, 2019-2020
VP-Technical Activities &
BoG Member 2009-2018
APSIPA

Program Director, National Science Foundation
Professor, University of Hawaii

Anthony Kuh received his B.S. in Electrical Engineering and Computer Science at the University of California, Berkeley in 1979, an M.S. in Electrical Engineering from Stanford University in 1980, and a Ph.D. in Electrical Engineering from Princeton University in 1987. He is currently a Professor at the University of Hawaii, also serving as director of the interdisciplinary renewable energy and island sustainability (REIS) group. Previously, he served as Department Chair of Electrical Engineering Dr. Kuh's research is in the area of neural networks and machine learning, adaptive signal processing, sensor networks, and renewable energy and smart grid applications. In January, 2017 he started service as a program director for the National Science Foundation(NSF). He is in the Electrical, Communications, and Cyber Systems (ECCS) division working in the Energy, Power, Control, and Networks (EPCN) group.

Dr. Kuh won a National Science Foundation Presidential Young Investigator Award and is an IEEE Fellow. He was also a recipient of the Boeing A. D. Welliver Fellowship and received a Distinguished Fulbright Scholar's Award working at Imperial College in London. Dr. Kuh was an Associate Editor for the IEEE Transactions on Circuits and Systems and was a Distinguished Lecturer for the IEEE Circuits and Systems Society. Dr. Kuh served as the technical co-chair for the 2007 IEEE ICASSP held in Honolulu. He served as the IEEE Signal Processing Society Regions 1-6 Director at Large and was a senior editor of the IEEE Journal of Selected Topics in Signal Processing. He currently serves on the Board of Governors of the Asia Pacific Signal and Information Processing Association as Vice President of Technical Activities.

Signal and Information Processing Activities for the Smart Grid

Artificial Intelligence (AI) and particularly machine learning is currently seeing a higher level of interest than ever before. This is due to many factors including the speed of current day computers (data centers), the massive amount of heterogeneous data that is available through sensor networks, social network media, and the IoT, and recent successes using deep learning architectures for image processing and speech recognition.

Where is research in this area headed to? Two illuminating articles discuss some of the possibilities, Jeannette Wing's discusses computational thinking in [1] and Michael Jordan's discusses the current state of AI in [2]. There has been much progress in the areas of machine learning in terms of algorithmic development (supervised learning, unsupervised learning, and reinforcement learning), hardware implementation, and applying machine learning to numerous engineering, scientific, social, and economic areas. However, in many ways the data science revolution is just beginning.

For many engineering applications data is received in an online manner and learning and decision making must be done in real-time. Applications include the electric power grid, autonomous systems, manufacturing plants, and chemical processing plants. Examples of decision making include detecting anomalies and bad data. Learning must be done in an online mode with real-time decision making.

Recognizing the need for using machine learning and data analytics, the IEEE Power and Energy Society (PES) formed a Big Data subcommittee to bring leaders from industry, academia, and government to define the architectural, computational, and practical challenges and opportunities brought by the emerging big data in smart grid at http://sites.ieee.org/pes-bdaps/ . The National Science Foundation (NSF) has also been active in this area organizing a workshop this past February on Real Time Learning and Decision Making in Dynamical Systems bringing together many engineering research communities (power systems, control systems, transportation, signal processing, computational intelligence) to discuss real time learning and decision making, [3]. Other NSF activities include community building through the NSF Smart Grids Big Data Spokes, http://smartgridsbigdataspoke.org and a Dear Colleague Letter for funding in real-time learning and decision making for engineering systems, [4].Here we focus on signal and information processing research and activities applying machine learning and data analytics to power systems and the smart grid.

In the past several years there has been significant attention given to applying signal and information processing to smart grid and energy problems. The *IEEE Signal Processing Magazine* had a special issue on "Technical challenges of the smart grid: from a signal processing perspective" that appeared in September, 2012. The articles in the special issue addressed some of the signal processing methodologies that are important in the design and operation of the future smart grid [5]. In December, 2014, the *IEEE Journal on Selected Topics of Signal Processing* had a special issue on "Signal processing in smart electrical power grid" with twelve articles discussing using signal processing for a variety of topics including state estimation, electric vehicle charging, demand-side management, fault detection, electricity market balancing, energy consumption models, and power balancing [6]. The *APSIPA Transactions on Signal and Information Processing* have also had special issue articles devoted to signal and information processing for the smart grid, [7]. There have also been special sessions at both the *IEEE International Conference on Acoustic Speech and Signal Processing(ICASSP)* and the *Asia Pacific Signal and Information Processing Annual Summit and Conference (APSIPA ASC)* devoted to applying signal processing methods to smart grid and energy problems. Other recently created IEEE Journals that focus on the smart grid and sustainable energy are the *IEEE Transactions on Smart Grid* and the *IEEE Transactions on Sustainable Energy*. These two journals and others from the IEEE Power and Energy Society often use algorithms and approaches from signal and information processing.

Signal and Information Processing for the Distribution Grid

Here we focus on the electrical grid beyond the last substation, the distribution grid. In the traditional legacy grid the power is passed through the last substation and the distribution grid steps down the voltage through transformers to eventually reach the customers (residential, commercial, industrial).This is changing with the introduction of distributed renewable energy sources located at customer premises, energy storage devices, and electric vehicles hooking up to the distribution grid. The distribution grid can be a residential community, military base, or a University campus. In fact, many Universities around the United States and globally are examining their distribution grid and looking at making changes to convert this to a smart microgrid. Smart microgrids are defined by the Galvin Electricity Initiative as modern smaller scale versions of today's electricity grids [8]. These can be distribution grids that not only generate, distribute, and regulate the flow of electricity, but can still function when separated from the main grid [8]. It is envisioned that the future electric grid may be decomposed into a hierarchical structure with microgrids playing an important role and each hierarchical entity have some degree of autonomy and independence [9]. This section summarizes some of the research in my group on using signal and information processing and machine learning applied to problems at the distribution grid level. Here we look at three problems for the distribution grid; sensing to get data, processing data, and decision making. We discuss some representative research in each of these areas with more details in [10].

At the distribution level there is an increasing amount of distributed renewable energy sources (e.g. solar PV panels), distributed storage options, and programs for demand response. This necessitates the need for both environmental sensors and electric grid sensors. For deploying sensors, a key question is where to place the sensors and how many sensors to deploy. It is desirable to place a set of sensors to form a sensor network so that relevant information can be gathered. Tools from signal processing, statistics, and machine learning are used. The problem is often framed as an optimization problem of

minimizing some error criterion or maximizing some information criterion.

In [11] we formulated the static sensor placement problem using a mean square error criterion and showed the problem reduced to an integer programming problem that becomes infeasible when the number of locations, n and sensors, m become large. We found families of greedy algorithms that run in polynomial time and also found upper and lower bounds to optimal performance. The problem is broadly applicable in a number of domains including placement of weather boxes at discrete locations, PMU placement, and placement of AMI on a distribution grid. In [11] we conducted a number of simulations on randomly generated data and also on an IEEE 57-bus test system. The simulations showed that greedy algorithms run in polynomial time, give good approximations to optimal algorithms, and that we could find reasonably tight lower and upper bounds to the optimal algorithms.

Once we have placed sensor networks for the electric grid, the environment, and gathered information about energy usage in buildings we then need to interpret data using signal and information processing. There is a variety of tasks that can be performed including spatial and temporal forecasts of distributed PV energy generation and distributed state estimation.

The electrical distribution grid beyond the last substation consists of feeder lines that distribute power to customers. It can usually be modeled as a radial network. Our goal is to measure both voltages and currents on the radial network. As we move towards a smart microgrid, the radial network will have smart meters deployed on the radial network and also have distributed renewable energy sources such as solar PV energy generation. A simple radial network is shown in Fig. 1. There may also be additional wireless communication capabilities on the radial network.

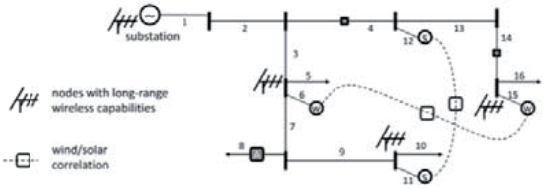


Fig. 1 . Model of radial network with distributed renewable energy sources and metering

We would like to perform state estimation for this radial network finding voltages at nodes and currents at branches. Here the circles represent distributed renewable energy generation (solar PV and wind) and the shaded squares represent meters taking measurements. This radial network can be modeled as a factor graph and then equations can be formulated to solve for node voltages and branch currents.

For real distribution grids the radial networks are often very large consisting of many nodes and connections in the factor graph. For most radial networks the factor graph representation is a tree if there are no distributed renewable energy sources. For these networks distributed state estimation algorithms are needed to efficiently find states. This is done using algorithms such as message passing and belief propagation which converge for radial networks. When we have distributed renewable energy generation such as solar PV, loops are created in the factor graph. The distributed solar PV energy sources are highly correlated, stochastic, and intermittent. We have performed belief propagation on factor graphs modeling simple distribution grids with distributed renewable energy generators and show that the algorithms converge to good solutions [12].

For accurate distributed state estimation we need good forecasts of distributed renewable energy sources. Distributed renewable energy generation such as solar PV energy is a stochastic and intermittent energy source. We conduct forecasting viewing solar energy sources as time series and perform time series prediction. Both batch and online learning algorithms are considered. We also examine the cost of making errors in solar energy forecasting. If we underestimate the amount of solar energy we can often curtail the excess solar energy. However, if we overestimate the amount of solar energy, then serious problems can occur as energy loads may need to be reduced through demand response or outages could occur. In most cases, prediction uses symmetric cost functions such as the mean squared error criterion. If underestimation is less costly than overestimation the estimates can be optimally biased to account for the different costs. However, in [13] we show that optimizing with respect to different asymmetric cost functions gives better performance. A convex piecewise linear cost function is used and an adaptive stochastic gradient descent algorithm such as Least Mean Square (LMS) algorithm are used.

Once we have processed data, then decision making can occur. Here we discuss two research problems that we have worked on; using reinforcement learning algorithms to perform demand response to control hot water heaters and developing online unsupervised kernel algorithms for outlier detection to detect bad electrical grid data.

Demand response can occur in homes by controlling appliance usage, setting thermostats and hot water heaters, and controlling when electric car batteries are charged. The optimization approach that is used depends on how the problem is defined. Things to consider include the cost of electricity, where energy is coming from (renewable sources or firm sources), storage options, user load profiles, and comfort indices. There have been many different approaches to demand response including using linear programming, non convex programming, dynamic programming, game theory, and approximate dynamic programming(ADP). Here we focus on ADP as this allows systems to learn load profiles and forecast renewable energy production. This involves both exploring and exploiting the state space to come up with good solutions. Conditions can also change with time due to load profile changes and seasonal weather changes.

Here we briefly discuss a research project of controlling a hot water heater using ADP. This work is based on [14]. The problem consists of minimizing a combination of costs to heat water in the hot water heater and the discomfort of the customers when they do not get enough hot water.

These are modeled as a Markov Decision Process (MDP). The MDP could be solved using a finite horizon dynamic programming model however, we do not know the state and action transition probabilities. A key is to determine when to turn on the water heater and when to turn off the water heater. Models are formulated for how the water in the water is heated when the water heater is turned on and how the water in the water heater cools when the water heater is turned off. For this problem we need to find information about how residents use hot water. This consists primarily of taking showers and using hot water for cleaning. This can be learned by observing households over a period of time. A finite horizon ADP problem is then formulated with the horizon being one day. Both time and temperature of the hot water heater are discretized. The ADP is shown to converge and give a significant improvement over water heaters that have fixed set points determined by time of day. Using ADP results for the hot water heater involves predicting when a user will take a hot shower. The ADP algorithm will typically turn on the hot water heater well before the anticipated shower takes place.

Finally, there is increasing interest in detecting bad data on the power grid. This could be due to some physical problem or some malicious cyber-attack. Real-time learning and decision making must occur. In [15] an online least squares one-class support vector machine classification algorithm is presented to detect outliers in a data stream. An approximate linear dependence criteria is used to obtain a sparse solution (by determining which data vectors are added as support vectors)with data vectors processed sequentially one vector at a time. Experiments were conducted for critical measurements for IEEE bus systems showing good performance of the proposed algorithm.

Summary and Further directions

Signal and information processing along with machine learning will play an increasingly more important role in decision making for the grid. This summary shows some activities in this area and a summary of some representative research in my lab.

As more heterogeneous data become available, more distributed computing is used, and changes occur in the electric power grid (integration of distributed renewable energy sources and storage methods) more sophisticated data analytic methods are needed. This involves advances in theory, algorithm design, and appropriate application of these algorithms for real-time learning and decision making.

References

[1] J. M. Wing, ``Computational Thinking'', Communications of the ACM, Vol 49, #3, 33-35, Mar. 2006.

[2] M. Jordan, ``Artificial Intelligence – The Revolution Hasn't Happened Yet'', https://rise.cs.berkeley.edu/blog/michael-i-jordan-artificial-intelligence%E2%80%8A-%E2%80%8Athe-revolution-hasnt-happened-yet/ , May, 2018.

[3] Website: https://sites.google.com/tamu.edu/nsf-workshop/home

[4] Website: https://www.nsf.gov/pubs/2018/nsf18063/nsf18063.jsp

[5] H. Gharavi, A. Scaglione, M. Dohler, and X. Guan: Guest Editors. Technical challenges of the smart grid: from a signal processing perspective. IEEE Signal Processing Magazine , 29, 5, 2012.

[6] Y. F. Huang, S. Kishore, V. Koivunen, D. Mandic, and L. Tong: Guest Editors. Signal processing in smart electrical power grid. IEEE Journal of Selected Topics in Signal Processing , 8, 6, 2014.

[7] A. Kuh, B. Chen, M. Ilic, and R. Yu: Guest Editors. Signal and information processingfor the smart grid. APSIPA Transactions on Signal and Information Processing, http://journals.cambridge.org/action/displayJournal?jid=SIP,2014,2015.

[8] ``What are smart microgrids'' Website report, Galvin Electricity Initiative, http://www.galvinpower.org/microgrids/.

[9] D. Bakken, A. Bose, K. Mani Chandy, P. Khargonekar, A. Kuh, S. Low, S. von Meier, K. Poolla, P.P. Varaiya, and F. Wu. Grip: Grid with intelligent periphery: A control architecture for grid 2050, 2011 IEEE SmartGridComm , Brussels, Belgium, Oct. 2011.

[10] A. Kuh. ``Signal and information processing applications for the smart grid'', 2015 APSIPA ASC. Hong Kong, Dec. 2015.

[11] M. Uddin, A. Kuh, A. Kavcic, and T. Tanaka. ``Nestedperformance bounds and approximate solutions for the sensor placement problem'', APSIPA Transactions onSignal and Information Processing, 3(4), 2014.

[12] Y. Hu, A. Kuh, T. Yang, and A. Kavcic. ``A belief propagation based power distribution system state estimator'', IEEE Computational Intelligence Magazine , 6(3):36–46, August 2011.

[13] S. Fatemi, A. Kuh, and M. Fripp. ``Online and batch methods for solar radiation forecasting under asymmetric cost functions'', Renewable Energy, Vol 91, 397-408, June 2016.

[14] M. Motoki, M. Umeda, M. Fripp, and A. Kuh. ``An approximate dynamic programming algorithm for control of a residential water heater'', 2015 IEEE International Joint Conference on Neural Networks , Killarney, Ireland, July 2015.

[15]M. Uddin and A. Kuh. ``Online least squares one-class support vector machine for outlier detection in power grid data'', IEEE ICASSP 2016, Shanghai, China, Mar. 2016.

Deep Spectral Mapping Models for Speech Signal Preprocessing

With a rapid increase of computing power and storage capacity in the past decade, many emerging speech-enabled applications on portable devices have drawn attentions from developers and users. Moreover, big data and deep learning paradigms have facilitated new algorithms in many speech areas, e.g., automatic speech recognition (ASR) [1]. However, the environmental robustness [2] is still a critical factor of user experience and wide technology deployments.

Though DNN-based solutions to classification problems have been wide spread, in order to tackle the environmental robustness issues, we instead had focused our attention on developing new high-dimensional regression approaches to classical speech signal preprocessing problems, including speech enhancement [3,4], speech dereverberation [5], speech separation [6], outperforming the performances of all state-of-the-art competing algorithms that we have tested and compared with in a wide range of noisy conditions.

A Unified View for Speech Preprocessing

All speech preprocessing approaches basically start from an explicit model of signal mixing in the time domain [2]:

$$y(t) = h(t) * x(t) + n(t), \qquad (1)$$

where $y(t)$, $x(t)$, and $n(t)$ represent the t-th sample of noisy speech, clean speech, and additive noise, respectively, and $h(t)$ is a time-varying convolutional distortion. In general, the objective of speech preprocessing is to recover the clean speech $x(t)$ given the noisy speech $y(t)$, which is obviously an underdetermined problem. For speech enhancement, most existing techniques focus on removing the additive noise $n(t)$ by ignoring $h(t)$. For speech dereverberation, on the contrary, we often treat $h(t)$ as a measured room impulse response without considering $n(t)$. For speech separation, the noise is still additive but with more confusing interfering speech from other speakers. From a research perspective, all these three problems are often tackled separately. However in realistic situations, different additive and convolutional distortions are usually mixed, making speech preprocessing extremely challenging.

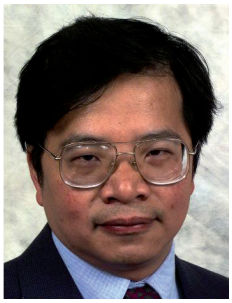
Conventionally, the methodologies to address different types of distortions varies considerably. However, a common practice is with mathematical assumptions about signals, distortions and their interactions. This often makes the performance of speech preprocessing severely degraded in adverse environments when assumption no longer hold.

With the powerful deep neural networks (DNNs), we have developed a unified framework totally different from the conventional approaches such that speech preprocessing is cast as a high-dimension nonlinear regression problem with the DNN directly approximating a special mapping function that converts spectral features from input noisy to output clean speech. To train the DNN, we need a large set of clean and noisy speech pairs. But it is often quite time consuming and expensive to collect time-synchronized stereo-data from realistic environments to have a full coverage of signals and distortions. Alternatively, we use the explicit model in Eq. (1) to generate a huge amount of simulation data. Given a set of clean speech signals and distortions, an unlimited amount of noisy speech data can be generated theoretically. There are three main advantages of DNN-based regression when compared with the conventional approaches. First, our framework makes almost no model signal assumptions, resulting in a more powerful capability when the distortion signal is quite hard to express, e.g., non-stationary. Second, the strong prior information about the clean speech signals, distortions and their interactions has been learned and stored in DNN. While in the conventional approaches, the prior information could not be well utilized. Third, in realistic environments, background noises, reverberations, and interfering speech often simultaneously exist. In such case, the conventional approaches could not make further inferences as the recovery of clean speech signal with so many unknown distortions in Eq. (1) is hard to tackle. In contrast, in our approach, we can use Eq. (1) to generate the diversified simulation data, including different types of distortions, which can well alleviate the inference problem in the conventional approaches.

As an illustration, we take speech enhancement [3] as an example. In Figure. 1, a block diagram of the proposed DNN-based speech enhancement system is depicted. In the

Prof. Chin-Hui Lee

PhD, Fellow-IEEE ***A Founder of APSIPA***



Professor

School of Electrical and Computer Engineering

Georgia Institute of Technology

Chin-Hui Lee is a professor at School of Electrical and Computer Engineering, Georgia Institute of Technology. Before joining academia in 2001, he had accumulated 20 years of industrial experience ending in Bell Laboratories, Murray Hill, as a Distinguished Member of Technical Staff and Director of the Dialogue Systems Research Department. Dr. Lee is a Fellow of the IEEE and a Fellow of ISCA. He has published over 500 papers and 309 patents, with more than 35,000 citations and an h-index of 80 on Google Scholar. He received numerous awards, including the Bell Labs President's Gold Award in 1998. He won the SPS's 2006 Technical Achievement Award for ``Exceptional Contributions to the Field of Automatic Speech Recognition''. In 2012 he gave an ICASSP plenary talk on the future of automatic speech recognition. In the same year he was awarded the ISCA Medal in scientific achievement for ``pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition''.

training stage in the upper panel, a regression DNN model is trained from a collection of stereo data, consisting of pairs of noisy and clean speech represented by the log-power spectrum features. In the enhancement stage in the bottom panels, the well-trained DNN model is fed with the features of noisy speech in order to generate the enhanced log-power spectra features. The needed phase information is calculated from the original noisy speech. The assumption is that the phase is not as sensitive for human auditory perception, so only an estimate of the magnitude spectrum is required. Finally, an overlap-add method is used to synthesize the waveform of estimated clean speech. A detailed description of feature extraction and waveform reconstruction modules can be found in [7]. As shown in Figure 2, the DNN parameters are first pre-trained by using a stack of restricted Boltzmann machines (RBMs) in a layer-by-layer manner. Then a minimum mean squared error object function between the target and enhanced log-power spectral features is used to perform fine-tune back-propagation.

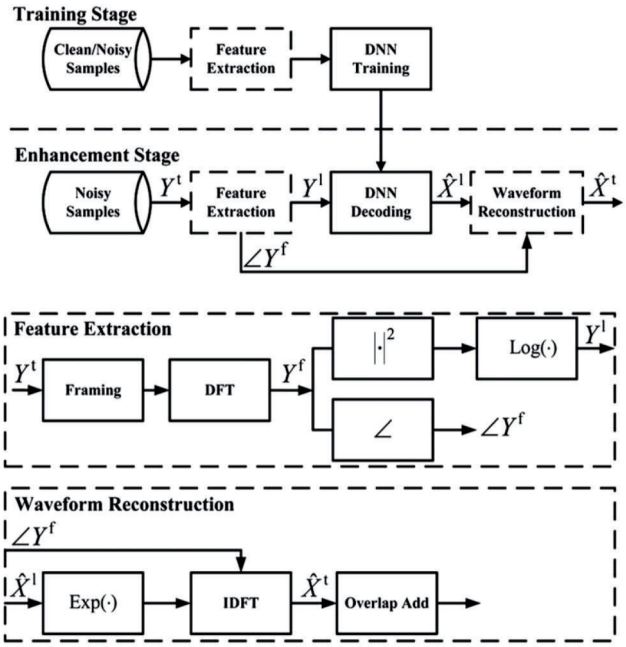


Fig. 1 DNN-based speech enhancement [3].

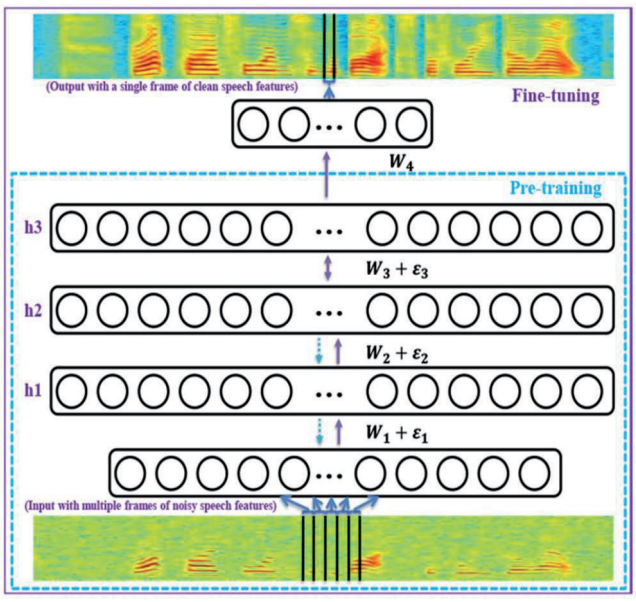


Fig. 2 Illustration of DNN training [4].

Applications

As a fundamental technique, our proposed DNN-based speech preprocessing, can be widely used to improve the environmental robustness of many speech-enabled systems. For speech communication, DNN-based approach was verified to significantly improve both speech quality and speech intelligibility with both subjective and objective experiments when compared with the conventional approaches [4]. For noise reduction in cochlear implant patients, DNN-based speech enhancement is shown to greatly improve speech intelligibility [8]. For voice activity detection, DNN-based speech preprocessing was used in a noise-universal detector with top performance in high noise, adverse acoustic environments [9]. For ASR, the proposed approach played a key role in achieving the lowest word error rates for several recent challenges, such as CHiME-2 for speech separation and recognition of mixed speech [6], CHiME-4 [10] for ASR of microphone array speech and REVERB challenge for ASR of reverberant speech [11]. For speaker diarization, DNN-based speech enhancement was demonstrated much more effective than the conventional approaches to reduce speaker diarization error rate [12], implying that enhanced speech had better discrimination among different speakers. It also turns out that many issues in diarization were caused by overlapping speech. Unsupervised source separation [13] is not only academically interesting but could also be practically useful in dealing with speakers talking at the same time. Compact DNN design had been tackled from a transfer learning perspective [14]. To reduce latency in DNN processing and still maintain high speech quality, a multi-objecting learning and ensembling approach has been recently explored [15]. DNN-based bandwidth expansion to convert narrowband into wideband speech has also been studied [16].

From the application perspective, one more advantage of DNN-based speech processing is that it is quite flexible to design customized models for specific scenarios, potentially yielding better performance than the general model. For example, we can only use the distortion signals of one noise environment (e.g., factory, tank, or helicopter) to build environment-dependent models. We can also use only the speech data of one speaker to build speaker-dependent models [17]. These could be easily implemented by only modifying the simulation data generation part.

Long-term Impact

The proposed DNN-based regression framework is a new paradigm shift in speech preprocessing. It demonstrates the strong approximation power and complementarity with the conventional approaches. Although it is not perfect at the current stage, we believe that great progresses will be made in the future from the following aspects. First, better training procedures and DNN architectures can be designed to leverage upon the unlimited simulation data to achieve better generalization capability for unseen speakers and unseen environments. Moreover, domain knowledge can be also incorporated into DNN architecture design, e.g., SNR [18], to better understand the conventional neural network as a black box. Second, the MMSE criterion for optimizing DNN often leads to an over-smoothing problem and does not guarantee the consistent improvements of measures in different applications, such as speech intelligibility and quality in speech communication, recognition accuracy in

ASR, and diarization error rate in speaker diarization. Accordingly, the optimization of speech preprocessing DNN should be guided by corresponding measures in the specific application by using some reinforcement learning techniques. Finally, we can combine the conventional and deep learning techniques to fully utilize the individual complementary strength in both approaches. For example, the online adaptivity of the conventional approach can be incorporated into the design of deep models. Mask estimation to improve speech quality [19] and reduce ASR error rate [20] are two recent efforts in this area.

Acknowledgment

First, the author would like to thank Prof. Jun Du at University of Science and Technology of China (USTC) for a great collaboration teamwork. The DNN-based regression framework was conceived at Georgia Tech and developed at USTC. Many colleagues and students also contribute to the systems, algorithms and applications discussed in this paper, including Prof. Marco Siniscalchi, Prof. Fengpei Ge, Dr. Y. Tsao, Dr. Zhen Huang, Dr. Bo Wu, Kehuang Li, and Sicheng Wang once at or from Georgia Tech, and Dr. Yong Xu, Dr. Yannan Wang, Dr. Qing Wang, Yanhui Tu, Tian Gao, and Lei Sun once at or from USTC.

References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 82-97, 2012.

[2] A. Acero, *Acoustic and Environment Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, 1993.

[3] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, Vol. 21, No. 1, pp. 65-68, Jan. 2014.

[4] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, Vol. 23, No. 1, pp. 7–19, Jan. 2015.

[5] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, Vol. 25, No. 1, pp. 102–111, Jan. 2017.

[6] J. Du, Y. Tu, L.-R. Dai, and C.-H. Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE/ACM Trans. Audio, Speech and Language Processing*, Vol. 24, No. 8, pp. 1424–1437, August 2016.

[7] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," *Proc. INTERSPEECH*, pp. 569–572, 2008.

[8] Y.-H. Lai, Y. Tsao, X. Lu, F. Chen, Y.-T. Su, J. K.-C. Chen, M.-J. Lien, H.-Y. Chen, L. P.-H. Li and C.-H. Lee, "A Noise Classification Based Deep Learning Noise Reduction Approach to Improving Speech Intelligibility for Cochlear Implant Recipients," *Ear and Hearing.* Vol. 39, No. 4, pp. 795-809, Jul/Aug 2018.

[9] Q. Wang, J. Du, X. Bao, Z.-R. Wang, L.-R. Dai and C.-H. Lee, "A universal VAD based on jointly trained deep neural networks," *Proc. INTERSPEECH*, pp. 2282-2286, 2015.

[10] Y.-H. Tu, J. Du, Q. Wang, X. Bao, L.-R. Dai and C.-H. Lee, "An information fusion framework with multi-channel feature concatenation and multi-perspective system combination for the deep-learning-based robust recognition of microphone array speech," *Computer Speech & Language*, Vol. 46, pp. 517-534, 2017.

[11] B. Wu, K. Li, F. Ge, Z. Huang, M. Yang, S. M. Siniscalchi, and C.-H. Lee, "An End-to-End Deep Learning Approach to Simultaneous Dereverberation and Acoustic Modeling for Robust Speech Recognition," *IEEE Journal on Selective Topics in Signal Processing*, Vol. 11, Issue 8, pp. 1932-1300, December 2017.

[12] L. Sun, J. Du, T. Gao, Y.-D. Lu, Y. Tsao, C.-H. Lee and N. Ryant, "A novel LSTM-based speech preprocessor for speaker diarization in realistic mismatch conditions," *Proc. ICASSP*, April 2018.

[13] Y. Wang, J. Du, L.-R. Dai and C.-H. Lee, "A Gender Mixture Detection Approach to Unsupervised Single-Channel Speech Separation Based on Deep Neural Networks," *IEEE/ACM Trans. Audio, Speech and Language Proc.,* Vol. 25, No. 7, pp. 1535-1546, July 2017.

[14] S. Wang, K. Li, Z. Huang, S. M. Siniscalchi, and C.-H. Lee, "A Transfer Learning and Progressive Stacking Approach to Reducing Deep Model Sized with An Application to Speech Enhancement," *Proc. ICASSP,* New Orleans, March 2017.

[15] Q. Wang. J. Du L.-R. Dai and C.-H. Lee, "A Multi-Objective Learning and Ensembling (MOLE) Approach to High-Performance Speech Enhancement with Compact Neural Network Architectures," Vol. 26, No. 7, pp. 1181-1193, *IEEE/ACM Trans. Audio, Speech and Language Proc.*, July 2018.

[16] K. Li and C.-H. Lee, "A Deep Neural Network Approach to Speech Bandwidth Expansion," *Proc. IEEE ICASSP,* pp. 4395-4399, Brisbane, Australia, April 2015.

[17] T. Gao, J. Du, L.-R. Dai and C.-H. Lee, "A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR environments," *Speech Communication*, vol. 95, pp. 28-39, 2017.

[18] T. Gao, J. Du, L.-R. Dai and C.-H. Lee, "Densely connected progressive learning for LSTM-based speech enhancement," *Proc. ICASSP*, April 2018.

[19] Y. Tu, J. Du, L. Sun and C.-H. Lee, "LSTM-Based Iterative Mask Estimation and Post-Processing for Multi-Channel Speech Enhancement," *Proc. APSIPA*, Kula Lumpur, Malaysia, Dec. 2017.

[20] Y. Tu, I. Tashev, S. Zarav and C.-H. Lee, "Combining Conventional and Deep Learning Techniques for Speech Enhancement and Recognition," *Proc. ICASSP,* Calgary, April 2018.

Unleashing the Intelligence in Signal & Data

Professor B.H. Juang

PhD, FIEEE ***A Founding Member of APSIPA***

Motorola Foundation Chair Professor, Georgia Institute of Technology

Biing-Hwang (Fred) Juang is the Motorola Foundation Chair Professor and a Georgia Research Alliance Eminent Scholar at Georgia Institute of Technology. He is also enlisted as Honorary Chair Professor at several renowned universities.

He received a Ph.D. degree from University of California, Santa Barbara. He conducted research at Speech Communications Research Laboratory (SCRL) and Signal Technology, Inc. (STI) in the late 1970s on a number of Government-sponsored research projects and at Bell Labs during the 80s and 90s until he joined Georgia Tech in 2002. He was Director of Acoustics and Speech Research at Bell Labs (1996-2001).

Prof. Juang is well known for his work in data and signal modeling for compression, recognition, verification, enhancement, physical and statistical analyses, secure communication, and synthesis. He is accredited with the original concept of signal modeling for discrimination that has served as an important guiding principle of deep learning. Prof. Juang has published extensively, including the book "Fundamentals of Speech Recognition", co-authored with L.R. Rabiner.

He received the Technical Achievement Award from the IEEE Signal Processing Society in 1998. He served as the Editor-in-Chief of the IEEE Trans on Speech and Audio Processing from 1996 to 2002. He was elected an IEEE Fellow (1991), a Bell Labs Fellow (1999), a member of the US National Academy of Engineering (2004), and an Academician of the Academia Sinica (2006). He was named recipient of the IEEE Field Award in Audio, Speech and Acoustics, the J.L. Flanagan Medal, and a Charter Fellow of the National Academy of Inventors (NAI), in 2014.

Promises of Artificial Intelligence

When one of the best GO players in the world lost the game to a computer in 2016, many people gasped in awe while others smiled at the inevitability. The subject of artificial intelligence (AI) once again became the talk of the town as it did in the 70s and 80s, enjoying its status as a celebrity science. Since its unofficial birth in 1956, nonetheless, AI has come in ebb and flow. Will it truly realize its promises this time?

While many insist that the answer is iffy at best, it's important that we understand the scientific progress that was accomplished and that led to this round of worldwide enthusiasm. Furthermore, it'd be enlightening to develop such an understanding in contrast to human intelligent capabilities as this comparison would bring to light appropriate aspirations for the modern discipline of artificial intelligence.

Recent Progress and Successes

Many attributed to the so-called Deep Neural Networks (DNN) [1] as the game changer; many also argue that Artificial Neural Networks (ANN) had been known for decades. Innovations that were presented after the turn of the millennia seem to have been largely misunderstood.

As machine learning started to be widely recognized as a field in computer science, the notion of supervised and unsupervised learning, subjects that have long been studied in other disciplines under different names, became ubiquitous among computer scientists.

Geoffrey Hinton's work in the Boltzmann machines, first as a stochastic associative memory in the 1980s to enhance the non-stochastic Hopfield Net [3] and later the restrictive Boltzmann machines (RBM) [4] with his contrastive divergence algorithm [5] for estimating the network connection parameters, is an interesting sequential development of how the idea of memory through association can be encapsulated by a simple connectionist's model. Indeed, while the content of a "device" memory is and must be deterministic, "human" memory is inherently probabilistic and becomes deterministic only after some mental decision is made. Hinton further proposed the "deep belief network" (DBN) based on the innovative paradigm of unrolled and stacked RBMs [6]. In many examples, DBN has been shown as a formidable tool for achieving the goal of unsupervised learning, particularly when the data dimensionality is large [4].

As the other main idea of ANN, the more traditional FNN (feed-forward neural networks, or multi-layer perceptron) has also long presented itself as a potentially useful tool for approximating a mathematical function [7]. Pattern recognition can be easily formulated as a function that maps an observed pattern to a discrete class label. Such a function is discriminative, in contrast to Bayes' teaching of decision theory based on idealized probability models [8]. Determination of an FNN configuration for a task remains an art and many early attempts to increase the number of network layers for performance improvement failed. The

reason for this disappointing fact was later found to be the so-called diminishing gradient problem [9].

The integration of RBM/DBN and FNN proposed around 2005 marks the most significant progress in this area of work [1]. Conceptually, this integration builds upon the notion of decomposing the ability of cognition (more specifically recognizing a pattern here, not the general human cognition) as consisting of association and discrimination. The former is accomplished through layers of feature transforms by RBM/DBN in the so-called "pre-training" procedure and the latter is carried out by the FNN stacked over the earlier layers of DBN and optimized by the usual "error backpropagation" algorithm. This leads to the practice of DNN, and the prevalent term of Deep Learning.

Since then, these studies (DNN and its variants) have flourished and been applied extensively to many applications with impressive results. As mentioned, a machine can now play the sophisticated GO game as good as any human master. In face recognition, a machine can recognize more faces more accurately and more speedily than any human being. In speech recognition, an automatic speech recognition system rivals a human listener in terms of word accuracy when DNN is embedded in the now conventional hidden Markov model framework [10].

What remains intriguing in spite of these impressive results is the lack of systematic understanding of DNN as it at times produces outcomes that are hard to interpret or explain. Such deficiency nonetheless has not hampered the widespread use of DNN and deep learning.

Brain-Inspired Task-Oriented AI

By now, it has become clear that the current AI that is being demonstrated is no longer the so-called strong or general AI but a computation system that is designed to accomplish an intelligent *task*. It is thus interesting to differentiate the human intelligence and an intelligent task that a computer can accomplish.

Human's intelligence encompasses many capabilities, generally described as cognition, planning, analysis, learning, adapting, memorizing, etc. However, these capabilities are intertwined and hardly isolated. For example, no planning can be accomplished without memory. In short, human intelligence is a holistic concept with many attributes that are still beyond the current realm of computational intelligence.

In contrast, computational algorithms tend to be atomistic, each with a particularly designed objective or result. Examples are: sorting, dynamic programming, matching, and memory recall, to name a few. Although we cannot claim that every brain function can be written as a math function (counter examples: a desire to drink, a hungry feeling), those operational ones that can be computed are accomplished by a proper computer much more effectively and efficiently, often by several orders of magnitude in performance, than the human brain itself.

With the above contrast (holistic vs. atomistic) as a backdrop, an extremely interesting development in recent AI is worth noting here. As mentioned, the success of DNN can be attributed to the integration of two computational models, i.e., RBM/DBN which performs association and feature transformation, and FNN which performs discrimination and discernment. Mathematically, they learn from data to perform nonlinear transformation and continuous-to-discrete function mapping, respectively. Nonetheless, the structure of DNN is essentially based on the McCulloch-Pitts neural calculus [10], which was totally inspired by the biological brain model. And yet, the performance of DNN (and all related variants) has in many cases exceeded the human performance in pattern recognition within the prescribed design scope. One can thus conclude that the purpose of modern AI, while may be inspired by the human brain and intelligence, should not dwell on duplicating the human performance but to exceed it in "intelligent tasks." An analogy can be drawn in aviation engineering. We do not fly by flapping wings like a bird; we fly by moving forward rapidly to create a lift with airfoil wings. As a result, we fly at a much higher speed than any bird.

So, how should scientists and engineers tackle the challenge of designing a machine or computational system to accomplish an intelligent task? An engineer would examine and decompose an otherwise holistic intelligent task into several subtasks, the individual design objective of which can be computationally achieved. The above mentioned task of cognition, or more precisely pattern recognition, demonstrates this principle of modern AI from a computational standpoint; it practically and possibly rightly assumes that the ability of cognition involves two tasks or capabilities at a lower level, i.e., association (to relate) and discrimination (to discern). The use of properly configured computational structures to accomplish the integrated task then ensues. The fact that these structures share the same architecture, i.e., an FNN, and can be directly coupled in the case of DNN is particularly remarkable. The seamless integrated structure also brings about convenient implementation achieved by today's powerful multi-processor computing systems.

Long Term Prospects

I. Worthy Aspirations of AI

We have witnessed the power of modern principles of AI taking advantages of sophisticated machines learning algorithms and massive amount of data. When these principles are at work and supported by sufficient amount of real data, they often accomplish a performance for the attempted intelligent tasks far better than humans do.

Coupled with advanced mechanization and robotics, modern AI will be able to do what a human *cannot*. That's the aspiration we engineers and scientists ought to embrace because that is where the real benefits of AI lie. *AI is not to replace human labor but to enhance human capabilities*.

II. The Power of AI Is in Human Data and Signal

AI intends to help with human intelligent tasks, the core of which involves human generated data, ranging from sensory data, i.e., sight, sound and touch, to behavioral data, such as opinions, preferences, choices, purchases, habits, etc.

The power of AI thus hinges on the sciences and technologies around data and signal; it can only be unleashed by advanced data and signal research. The disciplines of signal processing and systems, statistical modeling, data analytics and machine learning become therefore ever more important for us to meet the challenge and realize the above worthy aspirations – helping people to do what the natural intelligence can't.

References

[1] Y. LeCun, Y. Bengio, and G.E. Hinton, "Deep Learning," Nature, Vol. 521, pp 436-444, 2015.

[2] G. E. Hinton, and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks." Science 313.5786, 504-507, 2006.

[3] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," Proceedings of the National Academy of Science (PNAS), Biophysics, Vol. 79, pp. 2554-2558, April 1982.

[4] N. Le Roux and Y. Bengio, "Representational power of restricted Boltzmann machines and deep belief networks," Neural Computation, Jun., 20(6): 1631-49, 2008.

[5] G. E. Hinton, "Training Products of Experts by Minimizing Contrastive Divergence," Neural Computation 14, 1771–1800, 2002.

[6] B.H. Juang, "Deep neural networks – a developmental perspective", APSIPA Transactions on Signal and Information Processing, 5, E7. doi: 10.1017/ATSIP.2016.9/

[7] Kurt Hornik, Maxwell Stinchcombe, Halbert White, "Multilayer feedforward networks are universal approximators," Neural Networks, Volume 2, Issue 5, p.359–366, 1989.

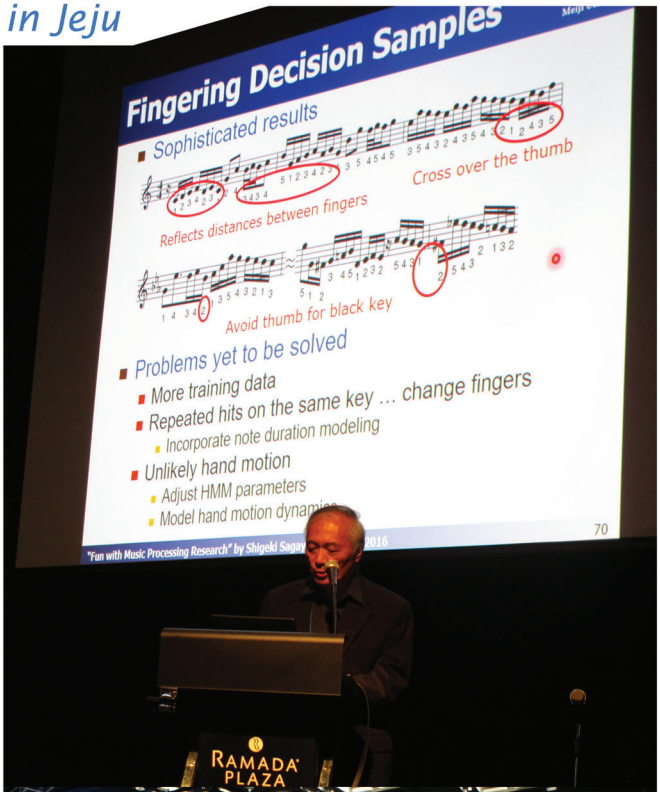
[8] B.H. Juang and S. Katagiri, "Discriminative Learning for Minimum Error Classification," IEEE Transactions on Signal Processing. VOL. 40, NO. 12, pp.3043-3054, Dec. 1992.

[9] S. Hochreiter. Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut f. Informatik, Technische Univ. Munich, 1991. Advisor: J. Schmidhuber.

[10] Warren McCulloch and Walter Pitts, "A Logical Calculus of Ideas Immanent in Nervous Activity," Bulletin of Mathematical Biophysics 5 (4): 115–133, 1943.

[11] L.R. Rabiner and B.H. Juang, "An introduction to hidden Markov models," IEEE ASSP Magazine, Vol. 3, No.1, p.4-16, Jan. 1986.

Photo Gallery: APSIPA ASC'2016 in Jeju



Speech Signal for Unsupervised Identity Authentication

Nowadays, with the rapid development of mobile phone applications in internet, the cyberspace which human beings depend on has been greatly extended, which is different from the traditional physical space. Not only the life styles of human beings have been changed significantly, but also enormous economic opportunities have been occurring as the result of the extensions. The huge demands for remote user authentication in mobile applications, such as mobile banking, mobile shopping, have drawn attentions from researchers and developers. In the physical space, the individual identities could be authenticated by authorized staffs, in another word, it's supervised. However, in case of self-services in certain physical space and the whole cyberspace, it's impractical for authorized staffs to authenticate in person, in another word, the authentication is unsupervised. It's referred as **unsupervised identity authentication** (**USIA**). In addition to security concerns, this has something to do with privacy protection.

Possible Technical Solutions to USIA

Traditional identity authentication methods are divided into two types. Type I is to check "*what you know*" like password and dynamic verification code. Type II is to check "*what you have*" such as IC cards and USB keys. However, both of them are easy to be forgotten, duplicated, or stolen, and cannot meet the needs to protect users' personal information and properties.

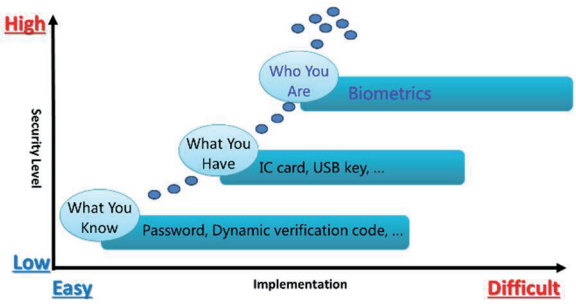


Fig.1 Possible solutions to **USIA**

Recently, biometric recognition has been regarded as an alternative authentication method with higher security and greater convenience. Biometric recognition is the state-of-the-art automatic technology for measuring and analyzing an individual's physiological or behavioral characteristics, and can be used to verify or identify an individual, which can be used to discern "*who you are*". Biometrics is instinctive and refers to metrics related to human characteristics, which can be applied in identity authentication anytime and anywhere. Obviously, with biometrics, people can '*prove yourself by yourself*'. Compared with the above mentioned traditional authentication methods, biometric authentication features convenience, efficiency and security to a bigger extent.

Prof. Thomas Fang Zheng

PhD, SMIEEE ***VP (2018-2019) APSIPA***

Professor

Center for Speech and Language Technologies

Tsinghua University

Dr. Thomas Fang Zheng is a research professor and Director of the Center for Speech and Language Technologies (CSLT), Tsinghua University. His research and development interests are on speech and language processing, having published more than 280 papers and 13 books, holding over 16 patents.

He is an IEEE Senior member, a CCF (China Computer Federation) Senior Member, a council member of Chinese Information Processing Society of China, a council member of the Acoustical Society of China, and the founding chair of steering committee of National Conference on Man-Machine Speech Communication (NCMMSC) of China.

He serves/d as VP (2013-2014, 2016-2017, & 2018-2019) of APSIPA, head of the Voiceprint Recognition special topic group of the Chinese Speech Interactive Technology Standard Group, Vice Director of Subcommittee 2 on Human Biometrics Application of Technical Committee 100 on Security Protection Alarm Systems of Standardization Administration of China (SAC/TC100/SC2).

He serves/d as an associate editor of IEEE Trans. on ASLP, an editorial board member of Speech Communication, an editorial board member of APSIPA Transactions on SIP, a series editor of SpringerBriefs in Signal Processing.

He ever served as General Co-Chair of Oriental COCOSDA 2003, General Co-Chair of APSIPA ASC 2011, APSIPA Distinguished Lecturer (2012-2013), General Co-Chair of IEEE ChinaSIP 2013, Area Chair of Interspeech 2014, General Co-Chair of ISCSLP 2014, Publication Chair of ICASSP 2016, and will serve as General Co-Chair of APSIPA ASC 2019, and General Co-Chair of Interspeech 2020.

Technical Requirements of Biometric Recognition for USIA

Biometrics represents the inherent characteristics of a person, and has the property of universality, uniqueness,

stability, and non-reproducibility. Yet in some practical applications, biometrics has its own limitations. In some cases, fingers and palms are exuvial, and the authentication will be hard to be implemented. Outlaws can conceal their real identities to escape from justice by wearing fingerprint caps. Iris recognition also requires expensive camera focus and appropriate light source. Retina recognition needs laser irradiation on the back of the eyeballs to obtain the characteristics of retina, which is inconvenient to users, even harmful sometimes. Some spoofing methods, such as real-time facial reenactments, make it easier to attack the face recognition system.

Considering these potential problems as mentioned above, some technical requirements for USIA should be seriously considered. Here the five essential requirements for USIA are summarized as below:

1. **Unification**: it should have uniqueness of the biometric and higher accuracy of the automatic recognition technology.
2. **Hardly forged**: it should be hard to spoof and attack.
3. **Reflecting real intention**: the user's real intention should be distinguished from passive action by external force or menace.
4. **Evidence traceability**: the action of authentication should be easily traced.
5. **Convenience and low-cost**: it should be convenient to use with low-cost for sampling, transmission, and computation.

Characteristics of Speech Signal

Spoken language is the most natural way that people communicate with each other. The information media of spoken language is speech signal, which contains rich information such as accent, language, content, emotion, gender, and identity, but in a simple one-dimension form, as shown in Fig.2, here we abbreviate this kind of characteristics as *rich information in simple form* (**RIiSF**). Different kinds of speech processing technologies are under study accordingly.

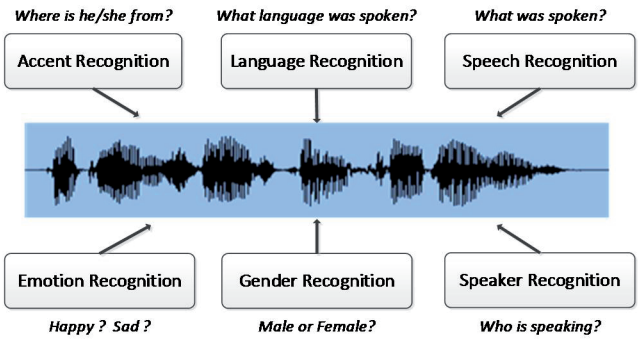


Fig.2 Speech signal as RIiSF

In addition, many researches have been making great efforts to describe how these different kinds of information are intermixed in speech signal. As early as in 1997, Prof. Fujisaki [2] proposed a definition that speech signal is a multi-layer medium and there are three levels of information delivered in speech interaction, as shown in Fig. 3. In this figure, the speech signal is assumed to be a hybrid of linguistic information, para-linguistic information, and non-linguistic information, and each information has its own representative properties.

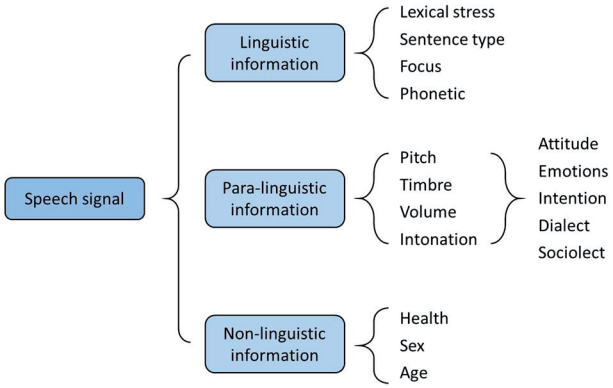


Fig.3 Multi-level information of speech signal

Speaker Recognition

Among the intermixed information in speech signal, voiceprint is a set of measurable information that uniquely symbolizes and identifies an individual. It is a kind of behavioral biometrics. Accordingly, voiceprint recognition (speaker recognition) is referred to as recognizing a person's identity from his/her voice. Research studies [3-7] have suggested that the voiceprint of a person has the property of uniqueness and stability, which can stay relatively stable and is not easy to change especially in adulthood. Furthermore, researchers believe that any two individuals do not sound identical because the shapes and sizes of their spoken organs, such as vocal tract, larynxes, lungs, nasal cavity and other parts of voice related organs, are quite different. In addition to these physiological differences, each speaker has his/her own characteristics or manner of speaking, such as the use of a particular accent, rhythm, intonation style and so on.

After decades' research, current speaker recognition systems have been achieving satisfactory performance. However, critical robustness as well as security issues still need to be addressed, especially in practical situations. Several additional technical focuses for speaker recognition systems are described as follows [1]:

1. **Ageing / time-varying**
   Ageing or time-varying represents a phenomenon that over a long period of time intervals, performance degradation would appear in speaker recognition systems. It could restrict promoted applications of speaker recognition.
2. **Speaking styles**
   Speaking styles usually come from speaker's spontaneous or unplanned speech and contains such abundant speaker individual information as emotion, speaking rate, volume, idiom, which enlarge the intra-speaker variations and reduce the system performance.
3. **Short utterance**
   In real scenarios, it is quite difficult to collect enough data (e.g. forensic applications) or users do not prefer to speak long enough (e.g. e-banking). So short utterance speaker verification is a great challenge.

4. **Anti-spoofing**
   Most biometric systems are vulnerable to spoofing, so is the speaker recognition system. How to prevent different kinds of spoofing attacks (e.g. impersonation, speech synthesis, voice conversion and replay) is a crucial issue of system security.

Desired Architecture for USIA

In view of the advantages of biometric recognition technologies and the RIiSF characteristics of speech signal, a voiceprint recognition based '3 + 2 + 2' architecture can be designed to satisfy USIA's five technical requirements, as shown in Fig. 4. It is based on the triple security mechanism.

1. **Three-biometrics**: three kinds of biometrics, the voiceprint plus lip print and face, are used to ensure the uniqueness and accuracy of identity authentication.
2. **Two-liveness detection**: both speech recognition and lip reading technologies are used to verify the content information of the user's utterance with an identical timing for both speech content and lip reading. In this way, it could implement the voice liveness detection and confirm whether the presented speech signals originated from a live user.
3. **Two-real-intention detection**: both the emotion recognition and the facial expression recognition technologies are used to capture the real intention of users so as to confirm if the user uses the system unconsciously and is under coercion.

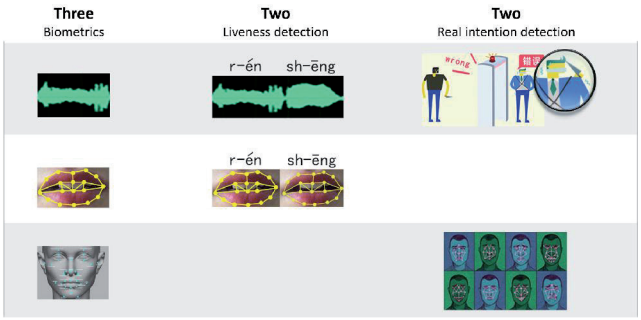


Fig.4 Voiceprint based **'3 + 2 + 2'** architecture for USIA

Long-term Impact

The demand of eID (electronic identity) is growing very rapidly. In 2006, the European Union launched the 2010 Pan European eID management (eIDM) framework roadmap, and in 2011 USA announced the National Strategy for Trusted Identities in Cyberspace (NSTIC) for setting up an identity ecosystem, and in 2013 China started to draft over 30 standards related to eID and in 2016 launched a Service Platform for Trusted Identity Authentication in Cyberspace. The eID will be widely used in more and more unsupervised scenarios such as mobile banking, mobile payment, entry/exit administration, and applications in need of remote user authentication.

In all these plans or regulations, it can be seen that the security, especially the anti-spoofing requirement, is a key factor. However, with the General Data Protection Regulation (GDPR) of European Union becoming effective on 25 May 2018, the personal data (privacy) protection becomes an additional hotspot.

Speech signal is one of the most suitable carriers for identity authentication, because the RIiSF characteristics helps not only to prevent spoofing easily, but also to detect real intention at a low cost.

Moreover, by using the content, intention, and speaker information contained in speech signal, the "*One-Sentence for Everything*" scenario will become more and more popular, as illustrated in Fig. 5.

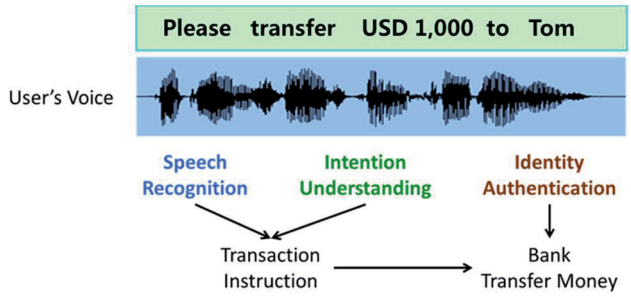


Fig. 5 'One-Sentence for Everything' scenario.

References

[1] Zheng T F, Li L-T. Robustness-Related Issues in Speaker Recognition [M]. Springer Singapore, 2017.

[2] Fujisaki H. Prosody, Models, and Spontaneous Speech [J]. Computing Prosody, 1997:27-42.

[3] Campbell J P J. Speaker recognition: a tutorial [J]. Proceedings of the IEEE, 1997, 85(9):1437-1462.

[4] Furui S. Recent advances in speaker recognition [J]. Pattern Recognition Letters, 1997, 18(9):859-872.

[5] Reynolds D A. An overview of automatic speaker recognition technology[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2011: 4072-4075.

[6] Kinnunen T, Li H. An overview of text-independent speaker recognition: From features to supervectors [J]. Speech Communication, 2010, 52(1):12-40.

[7] Beigi H. Fundamentals of Speaker Recognition [M]. Springer US, 2011.

Photo Gallery: APSIPA ASC'2011 in Xian

High Dynamic Range Video: Towards Extraordinary Visual Experience

Dr Guan-Ming Su

Ph.D. SMIEEE

VP of Industrial Relations and Development (2018-2019)

Senior Manager

Dolby Labs

Guan-Ming Su is with Dolby Labs, Sunnyvale, CA, USA. He is the core inventor of Dolby Vision Codec and the inventor of 80+ U.S. patents and pending applications. He is the co-author of 3D Visual Communications (John Wiley & Sons, 2013). He served as an associate editor of Journal of Communications; associate editor in APSIPA Transactions on Signal and Information Processing, and Director of review board and R-Letter in IEEE Multimedia Communications Technical Committee. He also serves as the Technical Program Track Co-Chair in ICCCN 2011, Theme Chair in ICME 2013, TPC Co-Chair in ICNC 2013, TPC Chair in ICNC 2014, Demo Chair in SMC 2014, General Chair in ICNC 2015, Area Co-Chair for Multimedia Applications in ISM 2015, Demo Co-Chair in ISM 2016, Industrial Program Co-chair in IEEE BigMM 2017, Industrial Expo Chair in ACMMM 2017, and TPC Co-Chair in IEEE MIPR 2019. He serves as chair of APSIPA Industrial Publication Committee 2014-2017 and VP of APSIPA Industrial Relations and Development starting 2018. He is a Senior member of IEEE. He obtained his Ph.D. degree from University of Maryland, College Park.

Reproducing the real-world scene truthfully on displays has been an important goal for image processing related research and product. With the new development of high dynamic range (HDR) technologies in recent years, this dream finally comes true. The HDR technologies bring in two new perspectives to significantly increase the end users' experience: expanded contrast ratio and wide color gamut. The enhanced contrast ratio can provide not only much lower dark level and higher peak brightness, but also sharper and more detailed image. The increased color gamut allows end users to watch all existing colors in the real world, not just the distorted/compressed one in traditional narrower color gamut.

The fundamental to build up such HDR video pipeline is to define the future-proof HDR signal format (Electro-Optical Transfer Function (EOTF)) to cover range of 0.0 ~10,000 cd/m^2. Knowing the limitation of the conventional EOTF (i.e. gamma), several formats have been proposed in the past few years, such as Perceptual Quantizer (PQ) defined in SMPTE ST.2084, ITU-R BT.2100, Hybrid Log Gamma (HLG), and SL-HDR defined in ETSI TS 103 433. In addition, new color spaces, such as R.2020 and ICtCp [1], are proposed and standardized to support HDR in wide color gamut. The new HDR signal formats allow us to digitize the captured signal, exchange between image processing modules, and apply image processing and compression on the full perceived dynamic range.

With those new signal formats, on the other hand, we also face new challenges in the HDR video pipeline. Firstly, the new signals exhibit different characteristics from traditional signals and require us to pay special attention to each module along the pipeline. Secondly, there are still lots of legacy devices, such as standard dynamic range (SDR) TVs, in the current market, and the backward-compatibility issue becomes very important for this SDR to HDR transition period. Those challenges are encountered during the entire video pipeline, namely, content creation stage, content distribution stage, and content consumption stage.

During the content creation stage, we need new content grading tools (both hardware and software) to achieve the best colorist/director/fidelity intent by fully utilizing this new dynamic range and color gamut. Different mindset is also needed to manage higher luminance range and richer color. Backward-compatibility and consistency support to enable content being viewed in both HDR and legacy SDR devices are important as well. Metadata becomes a critical issue to enable this content mapping and transform.

In the content distribution stage, the common solution is to reuse the existing mature video codec [2] without completely reinventing a brand new system to facilitate the time-to-market deployment. There are two different major categories depending on whether the SDR backward compactivity is needed. Without the backward compatibility limitation, the reshaping concept [3][4] is introduced to both change the characteristics of HDR signal and condense 12 bit HDR signal to 10 bit reshaped signal. The reshaped signal is then fed into legacy 10-bit video codec for compression. The HDR signal can be reconstructed via the inverse operation with metadata at the decoder side. With backward compatibility constraint, there are two further different approaches [5]: dual-layer architecture and single-layer architecture. In both architectures, the base layer is encoded for the SDR video such that any legacy device can playback the baseline content. Whenever a HDR device is available, the HDR signal can be constructed through the SDR to HDR prediction with the aid of metadata. The dual-layer architecture has another video bitstream as enhancement layer to further compensate the difference between the original HDR signal and predicted HDR signal. Though dual-layer system provides higher fidelity to source HDR

signal, the decoder requires two video decoding instances and consumes more memory and computation.

For the content consumption stage, advanced image processing can be introduced to process HDR for better perceptual performance. The common observed artifact in the HDR pipeline is the amplified video compression artifact owing to dynamic range prediction process and false contouring artifact owning to lower bit depth (8 or 10 bit) in base layer. Hardware efficient solution is proposed to alleviate the false contouring artifact [6][7]. Motivated by the high dynamic across in the luminance domain, luminance-driven adaptive image scaling is proposed in [8][9]. Advanced image processing technique to tackle the different characteristics of different EOTF and a universal statistics transfer framework are proposed in [10].

Long-term Impact

HDR display and capture devices are already in the market. The capability for those devices will keep growing to further extend the dynamic range and reach the limitation of human visual system. Applying conventional image processing techniques used in SDR does not provide satisfactory visual quality along this new HDR pipeline. Developing new image processing tools for the HDR pipeline becomes a very important research area for the next decades. The deployment of HDR technology has expanded from OTT service to broadcasting business, such as DVB and ATSC; and from 2D image product to more immersive viewing system. Applications of HDR technology are not limited to entertainment industry. For example, a much higher fidelity visual system with richer information can enhance the performance/accuracy for surveillance, advanced science exploration and analysis, and computer vision related products, such as robots and autonomous driving car. HDR technology will change how we observe, produce, and consume visual information in the next 20 years.

References

[1] Taoran Lu, Fangjun Pu, Peng Yin, Tao Chen, Walt Husak, Jaclyn Pytlarz, Robin Atkins, Jan Fröhlich, and Guan-Ming Su, ``IPT-PQ Colour Space and Its Compression Performance for High Dynamic Range and Wide Colour Gamut Video Distribution", ZTE Communications, special issue on Recent Progresses on Multimedia Coding, Analysis and Transmission, No.1 2016.

[2] Konstantinos Konstantinides, Guan-Ming Su, and Neeraj Gadgil, ``High Dynamic Range Video Coding," Handbook of Signal Processing Systems 3rd ed, Springer, editors: Shuvra Bhattacharyya, 2018.

[3] Jan Froehlich, Guan-Ming Su, Scott Daly, Andreas Schilling, and Bernd Eberhardt, ``Content Aware Qauntization: Requantization of High Dynamic Range Baseband Signals Based on Visual Masking by Noise and Texture", IEEE Inter. Conf. on Image Processing, Sep. 2016.

[4] Chau-Wai Wong, Guan-Ming Su, and Min Wu, ``Impact Analysis of Baseband Quantizer on Coding Efficiency for HDR Video", IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1354-1358, October 2016.

[5] Harshad Kadu, Qing Song, and Guan-Ming Su, ``Single Layer Progressive Coding for High Dynamic Range Videos", IEEE Picture Coding Symposium, June, 2018.

[6] Qing Song, Guan-Ming Su, and Pamela Cosman, ``Hardware-Efficient Debanding and Visual Enhancement Filter for Inverse Tone Mapped High Dynamic Range Images and Videos", IEEE Inter. Conf. on Image Processing, Sep. 2016.

[7] Subhayan Mukherjee, Guan-Ming Su, Irene Cheng, ``Adaptive Dithering using Curved Markov-Gaussian Noise in the Quantized Domain for Mapping SDR to HDR Image", International Conference on Smart Multimedia, August, 2018.

[8] Qian Chen, Guan-Ming Su, and Peng Yin, ``Spatial adaptive upsampling filter for HDR image based on multiple luminance range", IS&T/SPIE Electronic Imaging, 2014.

[9] Hossein Talebi, Guan-Ming Su, and Peng Yin, ``Fast HDR Image Upscaling Using Locally Adapted Linear Filters", IS&T/SPIE Electronic Imaging, 2015.

[10] Bihan Wen and Guan-Ming Su, ``TransIm: Transfer Image Local Statistics Across EOTFs for HDR Image Applications", IEEE International Conference on Multimedia & Expo, July, 2018.

Sample APSIPA Distinguished Lecture Activities (2015)

Scaling up the Deployment of Active Noise Control Systems

Professor Woon-Seng Gan

PhD, Fellow Audio Engg Society

Vice President, Institutional Relations & Education Program, Chairman of Speech, Language and Audio Technical Committee

Professor of Audio Engineering and Director of Infocomm Technology
School of Electrical & Electronic Engg.
Nanyang Technological University, SINGAPORE

Woon-Seng Gan is a Fellow of the Audio Engineering Society(AES), a Fellow of the Institute of Engineering and Technology(IET), and a Senior Member of the IEEE. He served as an Associate Editor of the IEEE/ACM Transaction on Audio, Speech, and Language Processing (TASLP; 2012-15) and was presented with an Outstanding TASLP Editorial Board Service Award in 2016. He also served as the Associate Editor for the IEICE transaction (2014-2016) on Fundamentals of Electronics, Communications and Computer Sciences(Japan). He is currently serving as Associate Editor of the IEEE Signal Processing Letters (2015); Associate Technical Editor of the Journal of Audio Engineering Society (JAES; 2013-); Editorial member of the Asia Pacific Signal and Information Processing Association (APSIPA; 2011-) Transaction on Signal and Information Processing; Associate Editor of the EURASIP Journal on Audio, Speech and Music Processing (2007-). He has published more than 300 international refereed journals and conferences, and has translated his research into 6 granted patents and has founded a company, Immersive Sound Technology Pte Ltd in 2015. He had co-authored three books on Subband Adaptive Filtering: Theory and Implementation (John Wiley, 2009); Embedded Signal Processing with the Micro Signal Architecture, (Wiley-IEEE, 2007); and Digital Signal Processors: Architectures, Implementations, and Applications (Prentice Hall, 2005).

Active Noise Control (ANC) techniques have been applied successfully in reducing noise in several key applications, such as in the noise cancelling headsets for airplane entertainment, reducing engine noise in automobile, and reducing airflow noise in heating, ventilation and air-condition (HVAC). Recently, we have seen a growth in research and industry applications in a larger scale deployment, such as in window openings, smart phones, big-size vehicle, factory, medical facilities, and construction machines and sites. All these diverse applications point to the robust and well-established theoretical studies and experiments of more than 85 years since a German inventor, Paul Leug, who submitted a patent accounting for the principle of the ANC in 1933. These growing activities to solve our daily noise problems bring many opportunities to the digital signal and information processing communities, like APSIPA. Not only are our proposed digital and analog techniques can be cleverly used in ANC systems to solve real-world noise problems, our skillset and knowledge can bring about a more efficient, low cost, and practical solution to meet long term deployment issues in new applications.

In this short article, I will do a quick introduction on the concepts of ANC, its history, highlight some of the recent research and experimentation activities on multi-channel ANC development that my team in NTU is working on, and points to some key long-term impacts in this exciting field.

Introduction to Active Noise Control

The principle of ANC can be best described from Leug's patent: *"The unwanted sound is picked up by one or more microphone, their electrical signals feed, after amplification, to one or more loudspeakers so that the sound wave produced is in "phase opposition" to the primary unwanted sound and cancels it."* This description, which appears to be a simple technique, is however more challenging in practical realization due to the physical limitation of sound wave propagation and the non-ideal responses from the electroacoustic system for producing the anti-noise. We will account for some of these challenges in the later sections.

Since the conception of ANC in 1933, ANC has been undergoing the Garnett hype cycle, as shown in Figure 1. In the beginning, several engineers, such as Conover began to build manual noise control system for canceling out transformer noise and also, Olsen worked on a noise reduction system for machine noise. Subsequently, there was a downturn in ANC activities due to the lack of tools and techniques at that time to make the noise control effective and automated. It is during the dormant period of 60s till 80s that two important works occurred. The first significant work is the invention of the renowned adaptive Least Mean Square (LMS) by Professor Bernard Widrow (Stanford University, USA), who opened the door for today's wide-ranging applications using adaptive signal processing for changing the transfer function of control filter; another significant work is the invention of digital signal processor (DSP) by Texas Instruments, which leads to the current advancement in digital technologies.

Gradually in the early 80s, there are a few encouraging experimentations in the digital implementation of active noise control system in air-ducts. Gradually, there are more activities in applying ANC to ear muffler, headphones, propeller airplane, automobile, head rest, and leading to several more recent applications of ANC pillow, MRI noise reduction, ANC for open window, and many others. The gradual rise of the ANC will not be possible without the excellent research effort and compilation of work by many researchers, in particularly, Professor Sen M Kuo (Northern

Illinois University, USA) and Professor Stephen Elliott (University of Southampton, ISVR in UK) who compiled two of the most influential books in the area of ANC in the early 90s. These new works had led to the most successful ANC commercial products, which is the noise canceling headphones. Recently, we have seen many interesting research and development activities in the field of ANC and in this short article, I am going to explain the work that is being carried out in the Digital Signal Processing Laboratory at the Nanyang Technological University in Singapore.

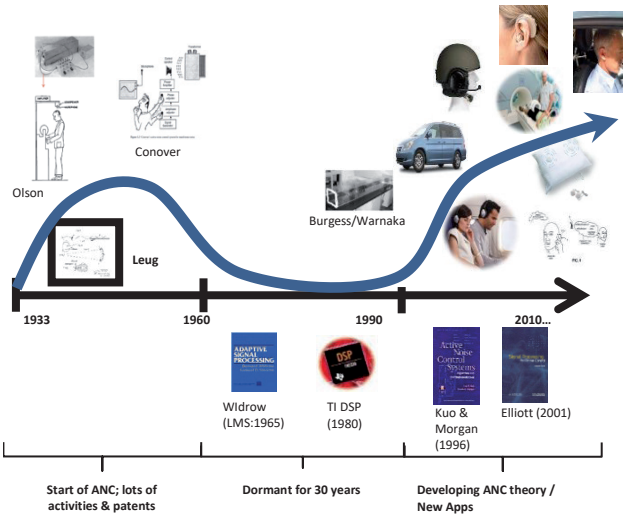


Figure 1: "Garnett" hype cycle for ANC development

Applying Multi-Channel Active Noise Control to Open Window

As explained, the technique of ANC is based on the principle of superposition, where a secondary source must be generated (mainly) by loudspeaker to create a sound field (anti-noise signal) that is of equal amplitude but opposite phase with the primary noise source. The resultant sound field captured at the error microphone provides an indication of how well the ANC system has generated the anti-noise signal. In the conventional feedfoward ANC system, a reference microphone is used to monitor the noise source and together with the error microphone, form the reference and error signals, respectively, to the controller. The controller is usually a digital processor (such as DSP and FPGA) that can be programmed with an adaptive algorithm to iteratively adjust the controller transfer function to arrive at producing the most optimum anti-noise signal through a loudspeaker (secondary source). The overall ANC setup is shown in Figure 2. Note that there are several analog-to-digital and digital-to-analog conversion taking place to perform digital adaptive noise control. The success application of the ANC is usually based on several important factors, such as (i) global vs local cancellation; (ii) placement of sensors and actuators; (iii) causality in secondary sound field generation; and (iv) coherence between the reference and error signals.

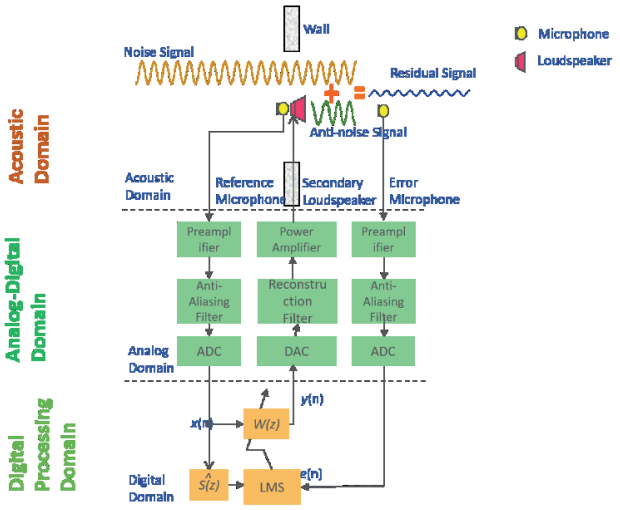


Figure 2: Overall block diagram of the ANC Implementation

A global noise control is often desired but hard to achieve. Global noise control leads to the overall spatial noise reduction without creating noise control only in certain region near the error microphone. In air duct experiment, global noise reduction can be achieved at the outlet of the duct, as noise flow is confined by the air duct. In other applications, secondary source can be positioned in close proximity to the primary noise source to achieve full cancellation of noise source.

The placement of sensors (i.e. microphone or accerometer) and actuators (i.e. loudspeakers) is one of the most critical factor in the success of ANC, but often a neglected consideration in academic papers. More sensors and actuators can provide better noise reduction, but also aggravated by the increase of computational complexity in the multiple channel ANC system.

The coherence (i.e. similarity) of the reference and error signals provides an important indicator on how well the ANC system can cancel out the noise source. If the coherence between reference and error signals is high (~0.9), an approximated reduction of -10 dB can be achieved. Below the coherence of 0.9, the noise reduction decreases in a logarithmic manner, reaching -3 dB with coherence of 0.5.

One of the major problems with the use of the electro-acoustic ANC system is the electrical latency introduced by different components (i.e. A/D, D/A, digital computation, analog filtering, amplification units, and sensors/actuator). This electric latency can be significant if the reference sensor is placed close to the error sensor. However, by increasing the distance between the reference and error sensors to gain longer processing latency, we sacrifice coherence and leads to poorer noise reduction.

Furthermore, to gain wider spatial region of noise reduction, there is a need to resort to multiple channel ANC system. Figure 3 shows the multi-channel ANC system, which consists of multiple-channel input signals picked up by J microphones, J secondary sources, and M error signals for adapting the multi-channel FXLMS algorithm:

$$\mathbf{w_{jj}}(n+1) = \mathbf{w_{jj}}(n) - \mu \sum_{m=1}^{M} \left[e_m(n) \mathbf{x'}_{\boldsymbol{mjj}}(\boldsymbol{n}) \right], \qquad (1)$$

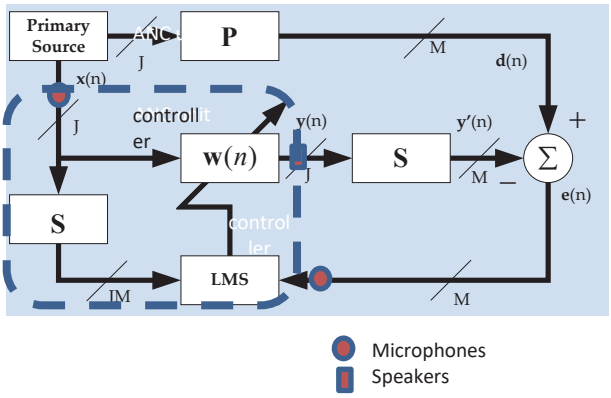


Figure 3 Multiple-channel (J input signals x J secondary sources x M error signals) FXLMS based ANC system

Figure 4 (a) External of the Experimentation Housing Setup of 2m x 2m x 2m; Fig 4(b) Interior of the Experimentation Housing Setup

However, the computational complexity in (1) is high and requires high-speed processor to achieve the necessary computation latency under certain sampling frequency. In this paper, we present a 24x24x24 multiple channel ANC system, as shown in Figure 4, for noise control in open window.

Here, external noise enters the experimental building (2m x 2m x 2m) through open window, where 24 reference microphones are used to pick up the noise source at the edge of the window opening and passes through the multiple channel ANC controller before sending the control signals to the secondary sources (i.e. 24 loudspeakers). The 24 error microphones, which are aligned to the 24 secondary sources are used to feedback to the multi-channel adaptive filters of the ANC system. In our experiments, broadband noise (400-1200 Hz) of 75 dBSPL can be reduced by approximately -10 dB with this multi-channel ANC system. However, high speed processor (such as the FPGA processor), sampling at 25 kHz must be used in the implementation of the above system. Currently, several works have been conducted to reduce the number of error channels, reduce adaptation complexity using partial updating or scanning techniques, collocating the reference sensors with the secondary sources, and exploiting geometry symmetry to reduce computation. This opens up new research areas in extending the noise reduction coverage in the open window and lead to new ANC applications.

In addition, to further extend the bandwidth of noise control, hybrid passive and active noise control techniques have been developed. One approach is to integrate active noise control system into a passive structure (i.e. silencer), which consists of absorbing material to attenuate high frequency noise, as shown in Figure 5. The performance of the hybrid

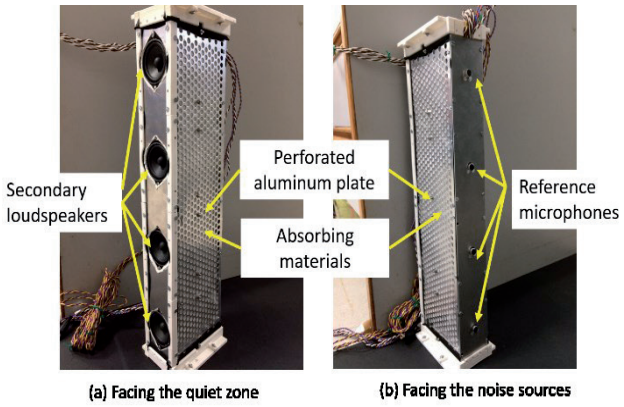


Figure 5: Hybrid Active Passive ANC Unit with (a) Speakers facing the quiet zone and (b) Microphones facing the noise sources

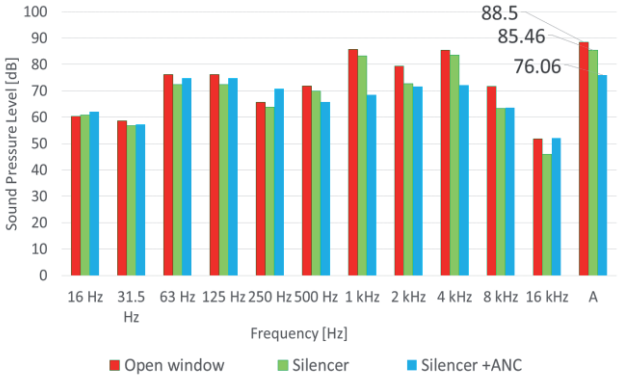


Figure 6: Noise Attenuation of the Hybrid Active and Passive Noise Control System

active and passive noise control extended the control bandwidth to 8 kHz. Therefore, we are pushing the effective frequency boundary of noise control by combining active and passive noise control techniques.

Long-term Impact on Multi-Channel ANC system for Scaled Up Deployment

With the advent of low-cost, low-power and high-computational processing power, we are seeing larger scale deployment of ANC to cover a wider spatial noise reduction. This current research and development have demonstrated the feasibility of extending the noise control region to a larger area under tight noise control constraints. Ultimately, we want to achieve a scalable and distributed noise control unit that can be easily installed and maintained to suit different applications' configurations and needs. In. my lab, we are taking a small step toward realizing the deployment of scalable multi-channel ANC system, but with more research findings across different research laboratories internationally, we are confident to see great progress in this field over the next five to ten years, and making our living and working environment more peaceful!

References

[1] B Lam, S Elliott, J Cheer, *W.S.Gan*, "Physical limits on the performance of active noise control through open windows," in *Applied Acoustics,* Vol 137, pp 9-17, 2018.

[2] S.J. Elliott, J Cheer, B Lam, C Shi, *W.S.Gan*, "A wavenumber approach to analysing the active control of plane waves with arrays of secondary sources," in *Journal of Sound and Vibration*, Vol 419, pp 405-419, 14 April 2018.

[3] T Murao, C Shi, *W.S. Gan*, M Nishimura, " Mixed-error approach for multi-channel active noise control of open windows," *Applied Acoustics*, Vol 127, pp 305-315, Dec 2017.

[4] J.Y. Hong, J.J. He, B. Lam, R Gupta, *W.S.Gan,* "Spatial Audio for Soundscape Design: Recording and Reproduction," in Special Issue on Spatial Audio, *Applied Sciences*, vol 7, 627, 2017, doi: 10.3390/app7060627.

[5] A. Agha, R. Ranjan, *W.S. Gan,* "Noisy Vehicle Surveillance Camera: A System to Deter Noisy Vehicle in Smart City," in the special issue on Acoustics in Smart Cities, *Applied Acoustics*, vol 117, Part B, 236–245, Feb 2017.

[6] Y. Kajikawa, *W.S. Gan*, S.M. Kuo, "Recent Advances on Active Noise Control: Open Issues and Innovative Applications," in the inauguration issue of *APSIPA Transaction on Signal and Information Processing* Volume 1, Dec 2012, e3, DOI: 10.1017/ATSIP.2012.4, Published online: 28 August 2012. *This paper has won the 2017 APSIPA Sadaoki Furui Prize Paper Award.*

[7] B. Lam, *W.S. Gan,* "Active Acoustic Window: Toward a Quieter Home," *IEEE Potential*, vol 35, no.1, pp 11-18, Jan/Feb 2016.

[8] DY Shi, *W.S. Gan*, "A Novel Selective Active Noise Control Algorithm to Overcome Practical Implementation Issue," in the *International Conference on Acoustic, Speech and Signal Processing*, 15-20 April 2018 Calgary, Canada.

[9] *W.S. Gan*, " Smart Audio Sensing for Environmental Deployment," invited distinguished speaker in the *IEEE 4th World Forum on Internet of Things*, 05-08 February 2018, Singapore.

[10] B Lam, C Shi, *W.S.Gan*, "Active Noise Control Systems for Open Windows: Current Updates and Future Perspectives," 24TH *International Congress on Sound and Vibration*, 23-27 July 2017, London, UK.

[11] D.Y Shi, C Shi, *W.S.Gan*, "Effect of the Audio Amplifier's Distortion on the Feedback Active Noise Control," in *Signal and Information Processing Association Annual Summit and Conference* (APSIPA), 12-15 Dec 2017, Kuala Lumpur, Malaysia

[12] V Belyi, *W.S.Gan*, "A New Psychoacoustic Subband Active Noise Control Algorithm," in *Signal and Information Processing Association Annual Summit and Conference* (APSIPA), 12-15 Dec 2017, Kuala Lumpur, Malaysia

[13] S Elliott, J Cheer, B Lam, C Shi, *W.S. Gan*, "Controlling Incident Sound Fields with Sourtce Arrays in Free Space and Through Apertures," 24TH *International Congress on Sound and Vibration*, 23-27 July 2017, London, UK.

[14] R Gupta, *W.S.Gan,* "3D Audio AR/VR Capture and Reproduction Setup for Auralization of Soundscape," 24TH *International Congress on Sound and Vibration*, 23-27 July 2017, London, UK.

[15] Z.T. Ong, *W.S.Gan,* "Effect of Masker Orientation on Masking Efficacy in Soundscape Applications," 46th *International Congress and Exposition on Noise Control Engineering*, INTER-NOISE 2017, 27-30 Aug 2017, Hong Kong.

Photo Gallery: APSIPA ASC'2017 in Kuala Lumpur

True Quiet Environment Creation with Active Noise Control

Professor Yoshinobu Kajikawa

PhD, SMIEEE

Vice President - Member Relations and Development (2018-2019) APSIPA

Professor

Department of Electrical, Electronic, and Information Engineering

Kansai University

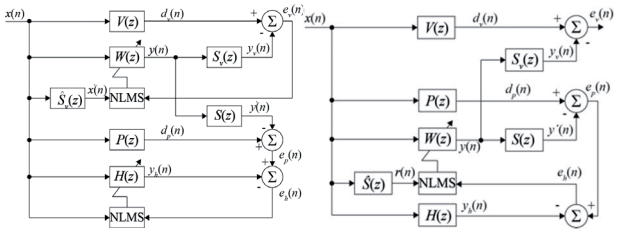
Brief Biography: Yoshinobu Kajikawa received his B.E. and M.E. degrees, both in Electronic Engineering from Kansai University, Japan in 1991 and 1993, respectively, and D.E. degree in Communication Engineering from Osaka University in 1997. He joined Fujitsu Ltd. in 1993. In 1994, he joined Department of Electronic Engineering, Faculty of Engineering in Kansai University, Japan. He is currently a Professor of Department of Electrical Electronic, and Information Engineering, Faculty of Engineering Science in Kansai University. His research interests are audio and acoustic signal processing including active noise control, adaptive signal processing, personal sound system, spatial sound processing, and acoustic transducer design. He received the Sato Prize Paper Award from the Acoustical Society of Japan and Sadaoki Furui Prize Paper Award from the APSIPA. He has published more than 200 international refereed journals and conferences, and has granted eight Japanese patents. He is currently a Senior Member of the IEEE and IEICE and is currently serving as a Vice President (2018-2019) in APSIPA and an Associate Editor of IET Signal Processing (2016-). He served as a vice president of IEICE Engineering Science Society (ESS), a chair of IEEE Signal Processing Society Kansai Chapter, an associate editor of IEICE Trans. Fundamentals and ASJ. He was also involved in organizing several international conferences, such as ICASSP, APSIPA ASC, and IWAENC.

Noise pollution is one of big issues in the industrial society. Active noise control (ANC) [1-3] is one of the solutions to reduce unwanted acoustic noise based on the wave superposition. ANC has very long history since this invention in 1936, but ANC still has many challenges we should improve in order to commercialize ANC more. What are challenges to commercialize ANC more? ANC usually requires physical microphones at the desired locations you want to obtain comfortable sound space (quiet zone). However, there are some difficulties to locate physical microphones there due to physical limitations, e.g. one of desired target areas is close to human eardrum, but it is too difficult to locate a microphone there. In such a situation, virtual sensing is one of powerful solutions to create quiet zones at the desired locations without locating any physical microphones. There are many varieties of virtual sensing techniques. The virtual sensing technique my team developed [4-9] is introduced later. This virtual sensing technique has superior features that the error microphones can be placed far from the desired place and the noise reduction performance is robust for environmental changes.

Another challenge is related to human hearing sensitivity. Even if unwanted acoustic noise is reduced by ANC, human may perceive residual noise because ANC cannot completely erase the noise. In this case, the noise components, which human did not care before ANC, are easily heard. This is due to human auditory system, that is psychoacoustic feature. Hence, ANC should consider not only noise reduction, but also human auditory features. One of solutions is a combination of ANC with masking technique where residual noise above absolute threshold of hearing would be masked by any natural sound making human comfortable. This idea should be always considered in practical noise problem situations, but some difficulties remain, e.g. which kind of natural sound is suitable for masking, how the system track the change of noise environment and noise characteristics, etc. Another approach is to use psychoacoustic features like A-weight and ITU-R 468 noise weighting in ANC structure [12]. These noise weighting filters modify error and reference signals and these modified signals are used to adjust the noise control filter in the ANC structure. This approach can improve the noise reduction in hearing sense.

ANC with Virtual Sensing

ANC creates quiet zone in the vicinity of the error microphone, which monitors residual noise (error signal). The error signal is iteratively minimized in the least mean square sense to obtain an optimal noise control filter in ANC. The size of the quiet zone is known to be approximated by 1/10 of the wavelength of the noise wave around the error microphone in the single-channel ANC system and relatively larger in the multi-channel ANC system. However, the error microphone may not be located closed to a desired location due to some physical limitations. In this case, the error microphone has to be located far from the desired place and adequate noise reduction may not be obtained around the desired place. Consumer ANC systems, such as the noise canceling headphones, head-mounted ANC systems, and ANC headrests, aim to reduce the noise level at the eardrum of the listener. If the noise contains noticeable power in the frequency range above 500 Hz, it is necessary to work out an ANC system that can realize the noise reduction at the desired location without locating the error microphone there. In this case, virtual sensing technique is a powerful solution to move the quiet zone to the desired location.



(a) Tuning stage (b) Control stage

Fig. 1 Block diagram of a single-channel feedforward ANC with virtual sensing.

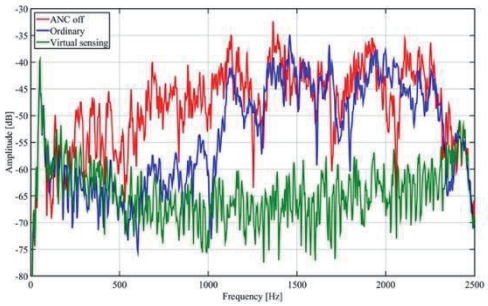


Fig. 2 Comparison of noise reduction performance between with and without virtual sensing at the desired place.

Figure 1 shows a block diagram of the feedforward ANC with virtual sensing technique we developed. In this system, an auxiliary adaptive filter, which maintains an in formation related to an optimal noise control filter for the desired location, is updated by locating a physical microphone to the desired location. This preliminary step is called "tuning stage". After the tuning stage, the physical microphone is removed from the desired location, and then the noise control filter is updated by an error signal modified by the auxiliary filter output. In this situation, the maximum noise reduction is obtained around the desired location. This step is called "control stage". Figure 2 shows a comparison of noise reduction performance between with and without virtual sensing. It can be seen from Fig. 2 that the higher noise reduction is obtained in the ANC with virtual sensing. This result indicates that the virtual sensing can move the quiet zone to the desired location. In the feedforward case, there is no limitation, that is, the error microphone can be placed far from the desired location (more than a few meters). Hence, the error microphone, which is implanted into a wall in a room, can realize noise reduction around any place in the room.

ANC Considering Psychoacoustic Features

The hearing sense has variety characteristics which are masking, loudness, pitch and so on. Loudness is the subjective perception. According to the definition, two sinusoidal waves with different frequencies are regarded to have equal loudness level if they are perceived as equally loud. This curve is called as equal-loudness contours standardized in ISO 226:2003. The noise weighting is suggested to quantify the hearing sensitivity of the human for the frequency. Two kinds of noise weighting filters, i.e. A-weighting filter and ITU-R 468 noise weighting filter, are treated for considering the human auditory properties in ANC. A-weighting is based on the equal-loudness contour for the 40 dB pure tone. On the other hand, ITU-R 468 noise weighting is defined to evaluate the random noise. These filters are integrated with ordinary ANC systems. The ANC with noise weighting filter can reduce noise over a specific frequency band effectively. Some subjective test results indicate that the ANC with noise weighting filter realizes true noise reduction for persons compared with ordinary ANC.

Long-term Impact

ANC realizes noise reduction in the desired location by virtual sensing techniques and true noise reduction by considering human auditory features. Of course, the development of hardware supports these state-of-art noise reduction techniques. However, ANC does not still erase the noise completely and can reduce the noise only in the limited space. True quiet environment should be created all over the room and enclosed space. Moreover, ANC setup strongly depends on the sound environments, this means ANC must be customized to fit the target environments. It is therefore difficult for non-expert engineers to implement ANC to any noisy environments. In the future, true quiet environments would be realized by noise vacuum cleaner which may be realized by true point source and easily implemented to any noisy environments. Hence, major breakthroughs in loudspeaker technology and signal processing technology are anticipated.

References

[1] S. M. Kuo and D. R. Morgan, Active noise control systems- algorithms and DSP implementations, New York: Wiley, 1996.

[2] S. J. Elliott, Signal processing for active control. CA: Academic Press, 2001.

[3] Y. Kajikawa, W. S. Gan, and S. M. Kuo, "Recent advances on active noise control: open issues and innovative applications," APSIPA Trans. Sign. Inf. Process., vol. 1, pp. 1–21, Aug. 2012.

[4] D. Moreau, B. Cazzolato, A. Zander and C. Petersen, "A Review of virtual sensing algorithms for active noise control," Algorithms., vol. 1, no. 2, pp. 69–99, Nov. 2008.

[5] N. Miyazaki, and Y. Kajikawa, "Head-mounted active noise control system with virtual sensing technique," J. Sound and Vibrations, vol. 339, pp. 65–83, Mar. 2015.

[6] S. Edamoto, C. Shi, and Y. Kajikawa, "Virtual Sensing Technique for Feedforward Active Noise Control," Proc. of Meetings on Acoustics, vol. 29, issue 1, 2017.

[7] S. Hirose and Y. Kajikawa, "Effectiveness of headrest ANC system with virtual sensing technique for factory noise," in Proc. APSIPA ASC 2017, Malaysia, Dec. 2017.

[8] R. Maeda, Y. Kajikawa, C. Y. Chang, and S. M. Kuo, "Helmet ANC system with virtual sensing technique," in Proc. International Congress on Sound and Vibrations, Hiroshima, Jul. 2018.

[9] R. Hasegawa, Y. Kajikawa, D. Y. Shi, and W-. S. Gan, "Window active noise control system with virtual sensing technique," in Proc. Internoise 2018, Chicago, Aug. 2018.

[10] R. Hasegawa, H. Yamashita, and Y. Kajikawa, "Study on the effectiveness of active noise control system for the nonstationary noise in consideration of psychoacoustic properties," in Proc. APSIPA ASC 2018, Hawaii, Nov. 2018.

Pore Features for High-resolution Facial Image Analysis

The human face is the most commonly seen object in images and videos. The analysis, detection, and recognition of face images lead to a wide range of applications. For face recognition, many different algorithms [1-8] have been proposed to tackle different challenges, such as the variations caused by illumination [9-12], pose [13], low resolution [14-16], facial expressions [17, 18], aging [19, 20], etc. However, when the resolution of face images increases, it has been found that these face recognition algorithms cannot improve their performance when the resolution has been sufficiently high. The reason for this is that all these methods are based on extracting features from facial appearances. When the resolution of an image is sufficiently high, further increasing the resolution will just mean increasing its redundancy, which is a waste of information. However, when we watch TV, we can easily see the pore patterns on the facial skin of actors. In other words, for high-resolution face images, we can exploit pore-based features for facial image analysis and recognition.

Pore-scale Facial Features

The pore-scale facial features include pores, fine winkles and hair, which are densely distributed over human faces, and which can be obtained non-intrusively using visible wavelength illumination. A few researchers have reported using the skin's appearance as a kind of texture surface for personal identification. The skin texture on the back of human hands was used for personal identification in [21]. Lin et al. [22] proposed to use skin-texture features, combined with global facial features, for face recognition. All of this research considers the skin as a texture, rather than identifying the pores in skin images. In addition, the recognition accuracy of these methods is not sufficiently high.

In this article we present a methodology capable of identifying people using the pore-scale facial features. This method fills an existing gap in the literature in terms of how to tackle the different problems and challenges for face recognition, i.e. pose variations, expression variations, illumination variations, misalignment, aging, identical twins identification, etc., within a new pore-scale framework. Based on this, a small skin region is shown to be sufficient for achieving highly accurate personal identification.

Pore Keypoint Detection

The pore-scale facial features are formed from pores, fine wrinkles, and hair. Due to small concavities where the incoming light is blocked, the pores are small, darker points, while the wrinkles are in the form of line structures. The hair also appears as small, darker points, and as lines. From a biological point of view, the quantity of pores on different faces should be similar even for people who have very different skin appearances; but the level of difficulty

Prof. Kenneth K.M. Lam

MSc, PhD, SrMIEEE

VP - Publications (2017-2019) APSIPA

Professor & Associate Dean

Department of Electronic and Information Engineering

The Hong Kong Polytechnic University, Hong Kong

Prof. Kin-Man Lam received his M.Sc. and Ph.D. degrees from Imperial College of Science, Technology and Medicine, in 1987 and University of Sydney, Australia, in 1996, respectively. In 1996, he joined the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University as an Assistant Professor, and has been a Professor since 2010. He is also an Associate Dean of the Engineering Faculty. He was actively involved in professional activities. In particular, he was a Technical Co-Chair of the 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2005 International Symposium on Intelligent Signal Processing and Communication System, and 2010 PCM. He was a General Co-Chair of the 2012 IEEE International Conference on Signal Processing, Communications, & Computing, APSIPA ASC 2015, and IEEE ICME2017. Prof. Lam was the Chairman of the IEEE Hong Kong Chapter of Signal Processing between 2006 and 2008. He received an Honorable Mention of the Annual Pattern Recognition Society Award for an outstanding contribution to the Pattern Recognition Journal in 2004.

Prof. Lam was the Director-Student Services and the Director-Membership Services of the IEEE Signal Processing Society, and was an Associate Editor of IEEE Trans. on Image Processing, an Area Editor of the IEEE Signal Processing Magazine. Currently, he is the VP-Publications of APSIPA, and also serves as an Associate Editor of DSP, APSIPA Trans. on Signal and Information Processing, and EURASIP International Journal on Image and Video Processing. He is also an Editor of HKIE Transactions. He has published more than 250 articles, with about 100 journal publications. His current research interests include human face recognition, image and video processing, and computer vision.

in detecting the pores varies. Therefore, a quantity-driven Difference of Gaussians (DoG) detector is employed to compute the location and the scale of the pore keypoints. The method uses an adaptive threshold to extract a certain number of keypoints on a skin region. Consider a specific face region shown in Fig. 1, Fig. 2 shows the corresponding pore keypoints and the scales detected from the region, on four subjects from the Bosphorus database. Different colors for the keypoints indicate their scale, as shown in the color bar.

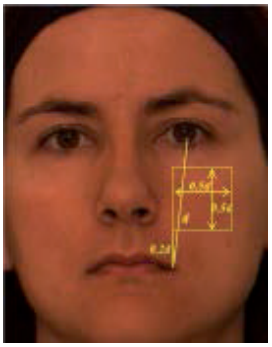


Fig. 1: The size of a cropped skin region, whose pore features are to be extracted.

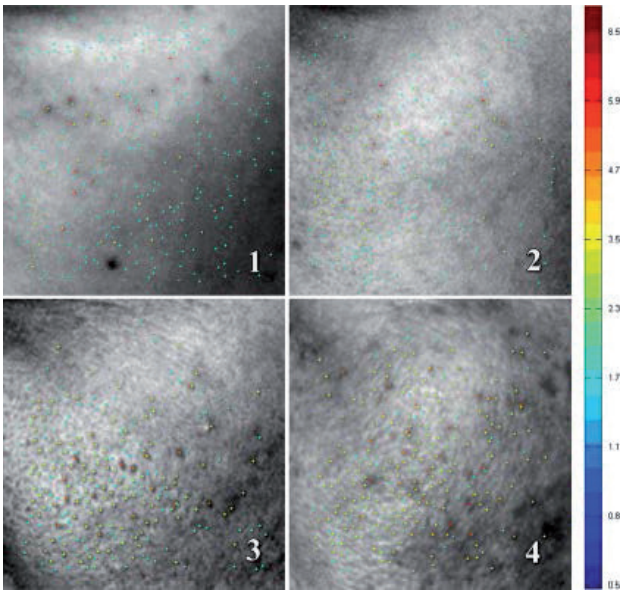


Fig. 2: Keypoint visualization based on four subjects from the Bosphorus database.

Pore Keypoint Feature Extraction

Pore-SIFT (PSIFT) detector [23], adapted from SIFT, was proposed using the gradient information of neighboring keypoints to describe the textures and the spatial distribution of pores. The dimension of the PSIFT feature vector is 512. Matching two keypoints using descriptors of 512 dimensions is too computationally expensive. To improve the efficiency, a more compact keypoint descriptor, namely Pore-PCA-SIFT (PPCASIFT) [24], was proposed, which uses principal component analysis (PCA) to reduce the dimensionality of the descriptor. The dimension of a PPCASIFT feature vector is 72 only.

With the PPCASIFT feature vector for each pore keypoint, the initial matching between the keypoints of two images is performed based on the distance between the feature vectors and the geometrical constraint. Then, the keypoint-matching results are converted to block-matching results using the constraint whereby the lines joining the matched keypoints are almost parallel. The number of matched blocks is used as a similarity measure for recognition.

Performances on Face Recognition

Experiments are conducted to show the effectiveness of the pore-scale facial feature in recognition performance. We first measure the recognition rates based on a single block, with different skin sizes. The gallery face images used in the experiments are of frontal view, neutral expression, and with uncontrolled illumination. The experiments are based on the Bosphorus database, which is a publicly available database containing 4,666 face images of 105 subjects. The face images in this database were captured with different poses and different expressions. The face images in the gallery set were constructed by a single image of each of the 105 subjects. The face images in the testing sets have different poses and expressions.

All the face images are cropped and resized to about 700×600 pixels. Using whole face images, a square region on one side of each face is located based on the corresponding eye corner and mouth corner, and has a size of $r \times r$ resolution, where r is a certain percentage of the length d of the line connecting the two outer corners, as shown in Fig. 3. d is about 360 pixels on average (about 8.1 cm according to [23]). By changing the percentage, squares of different sizes can be considered in the matching processing. We will show the performances of the pore-feature-based method for face recognition, with different values of r, i.e. different block sizes. The corresponding skin region used in the gallery set is equal to or larger than the one for the testing set so that location error between the two skin regions to be matched caused by misalignment and pose difference can be allowed. Then, the pore-scale keypoints from the query image are matched to a gallery face image.

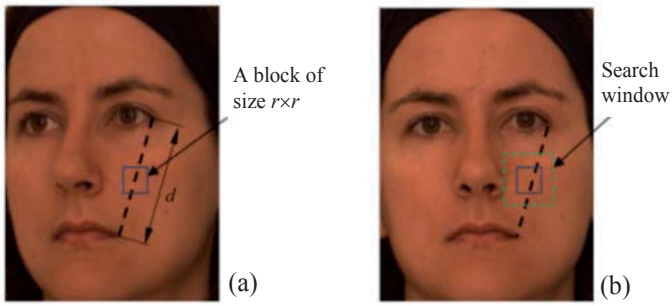


Fig. 3: (a) A testing face with a square skin region whose center is the mid-point of the line joining an outer eye corner and the corresponding outer mouth corner, and (b) a gallery face with a search window represented by a green, dashed box, which is larger than or the same size as the testing skin region.

We initially set r at 80% of d, and then reduce its value in steps. At each value of r, the recognition rate based on a testing set of face images is measured. When the region size is reducing, the number of pore-scale keypoints decreases, and may not be sufficient to make an accurate decision at a certain size.

We evaluate the performances when the face images are under different pose variations. Each testing set is formed by face images with four different poses, denoted as R10, R20, R30 and R45 for 10, 20, 30 and 40 degrees from frontal view, respectively. Table 1 shows the recognition rates under these different poses and with different window sizes used for recognition. We can see that under a small pose variation, using a small r, such as 72 pixels (equivalent to 2.62 cm^2 size of skin), can achieve a 100% recognition rate. The recognition performances degrade when the pose variation becomes larger. On the other hand, the whole-face images achieve a 100% recognition rate for all four poses.

	r (pixels)								Whole
Pose	72	90	108	144	180	216	252	288	face
R10	100	100	100	100	100	100	100	100	100
R20	98.1	97.1	98.1	99.0	99.0	100	100	100	100
R30	84.8	91.4	96.2	95.2	98.1	99.0	99.0	99.0	100
R45	48.6	56.2	60.0	68.6	78.1	82.9	82.9	93.3	100

Table 1: The recognition rates (%) for images under different poses.

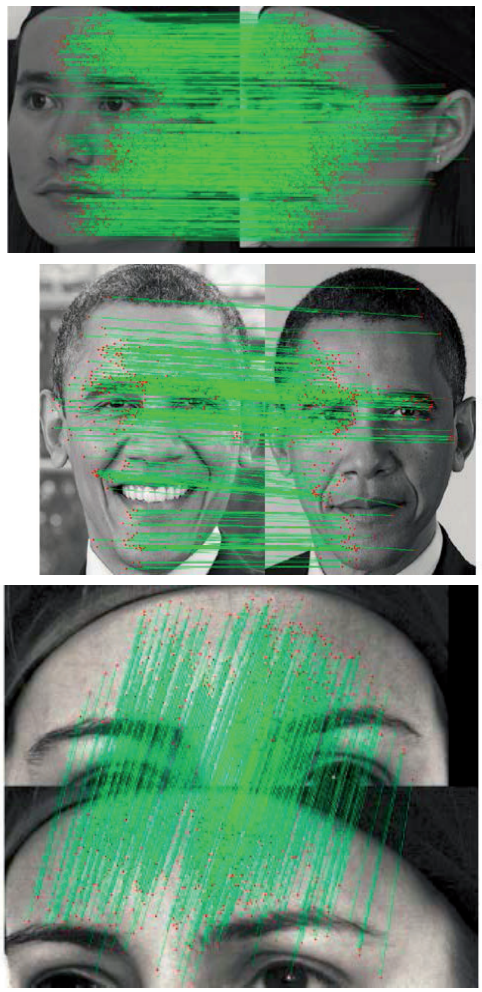
For testing the performance under different expressions, each testing set consists of faces expressing six emotions, namely anger, disgust, fear, happiness, sadness, and surprise. The recognition rates under different expressions and skin sizes are tabulated in Table 2. From the results, we can see that the anger, fear, and surprise expressions cause little distortion of the relative position of the pore keypoints, and hence have very little effect on the recognition performances. The other three expressions, i.e. disgust, happiness, and sadness, have degraded performances when the skin size becomes smaller. This is because the cheek skin suffers from different degrees of deformation with the different expressions. Nevertheless, the deformation affects only part of the faces, so using the whole-face images can still achieve a 100% recognition rate.

	r (pixels)								Whole
Pose	72	90	108	144	180	216	252	288	face
Anger	98.6	100	100	100	100	100	100	100	100
Disgust	91.3	97.1	98.6	98.6	100	100	100	100	100
Fear	100	100	100	100	100	100	100	100	100
Happiness	56.6	57.5	65.1	68.9	78.3	78.3	82.1	89.6	100
Sadness	95.5	98.5	98.5	100	100	100	100	100	100
Surprise	100	100	100	100	100	100	100	100	100

Table 2: The recognition rates (%) for images under different expressions.

Fig. 4 shows the pore keypoint matching results of face pairs of the same persons. In Fig. 4(a), the two faces are at two different poses, 10 and 45 degrees, respectively. In Fig. 4(b), the two images are Former US President Obama, before and after his presidentship. The two images have different facial expressions, lighting conditions, and ages. This result also shows that the pore features can be used for face recognition with age progression. Fig. 4(c) illustrates the recognition of two images using the forehead only. Actually, by using a region in face images, face recognition can be conducted.

Another great challenge for face recognition is the recognition of identical twins. This experiment is performed on the ND-TWINS-2009-2010 dataset. The testing set contains 1,856 images of 435 subjects with frontal view, neutral expression, and controlled illumination. This setting is similar to the R10 images tested previously. However, the identical twins have almost the same appearance, except at the pore scale. The distortions, such as those caused by the noise and blurring, have more influence on the recognition rate for identical-twin face images than that for other face images. In contrast, based on our observation, the quality of the images in the ND-TWINS-2009-2010 dataset is relatively lower than that of the Bosphorus database. Hence, the recognition rates degrade compared to the R10 testing results, i.e. the row of R10 in Table 1. Nevertheless, a 99.8% recognition rate with a 2.3% equal error rate can still be achieved. Fig. 5 shows the matching results for two identical twins.



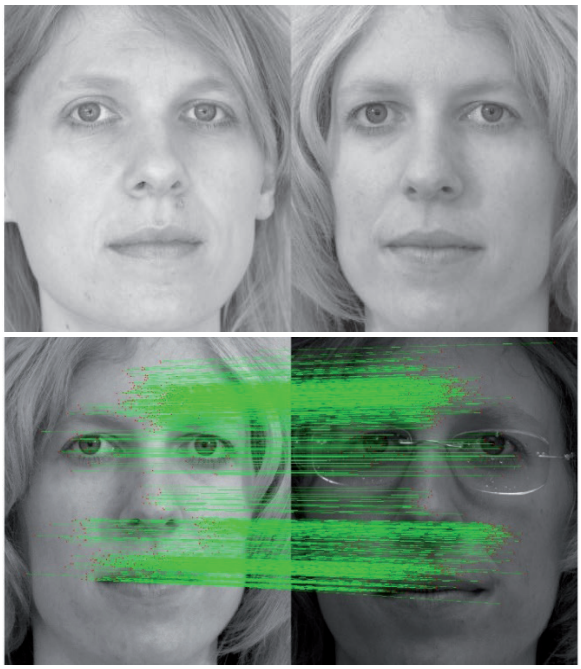
(a) (b) (c)

Fig. 4: Pore keypoint matching results.

Long-term Impact

Pore patterns on facial skin are different for different people, so they can be taken as a convenient biometric for people identification. If we can match the pore keypoints, within small corresponding regions of two people, we can accurately and confidently verify whether they are the same person. With the advancements of cameras in the future, higher quality and resolution could be achieved. This will make the detection and matching of the pore keypoints become a simple task. By matching the facial pore keypoints of a face image in a video sequence, it is possible to generate an accurate 3D face model. Not only

the depth of the keypoints at facial features, such as the eyes, nose, and mouth, are estimated, various locations on the facial skin surface, where pore keypoints are located, can also have their depth estimated. Currently, recognition of subtle expressions is a big challenge. However, by tracking the motion of the facial pores, subtle expressions can be detected and recognized. This technology can also be used for lie detection. We can foresee that there will be various applications based on pore keypoint analysis, and this research will attract a lot of attention in the near future.



(a)

(b)

Fig. 5: (a) The matching of two identical twins, where no pore keypoints are matched; and (b) the matching of the face images from one of the identical twins, where many pore keypoints are matched.

References

[1] Xudong Xie, Wentao Liu, and Kin-Man Lam, "Pseudo-Gabor wavelet for face recognition," *J. Electron. Imaging*, 22 (2), June 21, 2013.

[2] Zhan-Li Sun, Kin-Man Lam, Zhao-Yang Dong, Han Wang, Qing-Wei Gao and Chun-Hou Zheng, "Face Recognition with Multi-Resolution Spectral Feature Images," *PLOS ONE*, vol. 8, Issue 2, e55700, Feb. 2013.

[3] Wing-Pong Choi, Siu-Hong Tse, Kwok-Wai Wong, and Kin-Man Lam, "Simplified Gabor Wavelets for Human Face Recognition," *Pattern Recognition*, vol. 41, no. 3, pp. 1186-1199, March 2008.

[4] Xudong Xie and Kin-Man Lam, "Gabor-Based Kernel PCA with Doubly Nonlinear Mapping for Face Recognition with a Single Face Image," *IEEE Trans. on Image Processing*, vol. 15, no. 9, pp.2481-2492, 2006.

[5] Danghui Liu, Kin-Man Lam, and LanSun Shen, "Optimal Sampling of Gabor Features for Face Recognition", *Pattern Recognition Letters*, vol. 25, no. 2, pp. 267-276, January 2004.

[6] Kwan-Ho Lin, Kin-Man Lam and Wan-Chi Siu, "Spatially Eigen-Weighted Hausdorff Distances for Human Face Recognition", *Pattern Recognition*, vol. 36, no. 8, pp 1827 - 1834, 2003.

[7] Baofeng Guo, Kin-Man Lam, Wan-Chi Siu, and Shuyuan Yang, "Human Face Recognition Based on Spatially Weighted Hausdorff Distance," *Pattern Recognition Letters*, vol. 24, no. 1-3, pp. 499-507, 2003.

[8] Pengzhang Liu, Kin Man Lam, and Tingzhi Shen, "Face Recognition Using AdaBoost Modular Locality Preserving Projections," APSIPA ASC 2010, December 14-17, 2010, Singapore.

[9] Xudong Xie and Kin-Man Lam, "An Efficient Illumination Normalization Method for Face Recognition," *Pattern Recog. Letters*, vol. 27, no. 6, pp. 609-617, 2006.

[10] Danghui Liu, Kin-Man Lam, and LanSun Shen, "Illumination Invariant Face Recognition", *Pattern Recognition*, vol. 38, pp. 1705-1716, 2005.

[11] Xudong Xie and Kin-Man Lam, "Face Recognition under Varying Illumination Based on a 2D Face Shape Model", *Pattern Recog.*, vol. 38, no. 2, pp. 221-230, 2005.

[12] Muwei Jian, K.M. Lam and Junyu Dong, "Illumination Compensation and Enhancement for Face Recognition," Proceedings, APSIPA ASC 2011, October 2011, Xi'an, China.

[13] K.M. Lam and H. Yan, "An Analytical-to-holistic Approach for Face Recognition Based on a Single Frontal View, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 7, pp. 673-86, July 1998.

[14] Muwei Jian and Kin-Man Lam, "Simultaneous Hallucination and Recognition of Low-Resolution Faces Based on Singular Value Decomposition," *IEEE Trans. on Circuits and Systems for Video Tech.*, vol. 25, no. 11, 2015.

[15] Kuong-Hon Pong and Kin-Man Lam, "Multi-resolution feature fusion for face recognition," *Pattern Recognition*, vol. 47, no. 2, pp. 556-567, 2014.

[16] M. Saad Shakeel and Kin-Man-Lam, "Recognition of Low-Resolution Face Images using Sparse Coding of Local Features," Proceedings, APSIPA ASC 2016, December 2016, Jeyu, Korea.

[17] Xudong Xie and Kin-Man Lam, "Elastic Shape-Texture Matching for Human Face Recognition," *Pattern Recognition*, vol. 41, no. 1, pp. 396-405, January 2008.

[18] Xudong Xie and Kin-Man Lam, "Face Recognition Using Elastic Local Reconstruction Based on a Single Face Image," *Pattern Recog.*, vol. 41, no. 1, pp. 406-417, 2008.

[19] Huiling Zhou and Kin-Man Lam, "Age-invariant face recognition based on identity inference from appearance age," *Pattern Recognition*, 76, pp. 191-202, April 2018.

[20] Huiling Zhou, Kwok-Wai Wong, and Kin-Man Lam, "Feature-Aging for Age-Invariant Face Recognition," Proceedings, APSIPA ASC 2015, pp. 1161-1165 December 2015, Hong Kong.

[21] Jin Xie, Lei Zhang, Jane You, David Zhang, and Xiaofeng Qu, "A study of hand back skin texture patterns for personal identification and gender classification," Sensors, 12(7), pp. 8691–8709, 2012.

[22] Dahua Lin and Xiaoou Tang, "Recognize high resolution faces: From macrocosm to microcosm," Proceedings, pp. 1355–1362, IEEE Computer Vision and Pattern Recognition Conference, 2006.

[23] Dong Li and Kin-Man Lam, "Design and Learn Distinctive Features from Pore-scale Facial Keypoints," *Pattern Recognition*, vol. 48, no. 3, pp. 732-745, 2015.

[24] Dong Li, Huiling Zhou, and Kin-Man Lam, "High-Resolution Face Verification Using Pore-scale Facial Features," *IEEE Trans. on Image Processing*, vol. 142, pp. 117-123, 15 August 2015.

Human-like Conversational Robot

Impact of End-to-End Speech Recognition

In the last decade, we have observed a significant progress in speech and image recognition. It was surprising but predictable that deep learning and big data brought this improvement. But more surprising is end-to-end or seq2seq neural network model is replacing the statistical framework of speech recognition, which deemed very solid.

Conventionally, automatic speech recognition (ASR) has been a complex of acoustic model, phone model, lexical model and language model, each of which is formulated with a dedicated statistical model. Currently, a complex but single neural network is designed to conduct the entire process in an end-to-end framework. As illustrated in Figure 1, the direct mapping from a sequence of acoustic features into a sequence of words, called acoustic-to-word model, achieves comparable performance to the state-of-the-art system with an extremely simple architecture and an amazing decoding speed (RTF of 0.04).

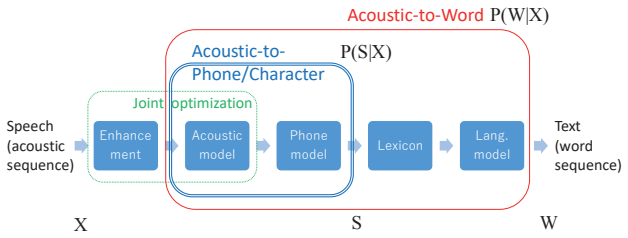


Figure 1 End-to-end speech recognition

End-to-End Dialogue System?

The next step will be end-to-end speech understanding, which directly converts speech into meaning or intentions of the utterance by combining the language understanding function. The question here is how to define a set of meaning and intentions, which may depend on the task domain.

Since the goal of intelligent machines is to respond to any user queries, an ultimate end-to-end model is to output proper responses given a user input. When we assume text for inputs and outputs, the model is called a neural conversational model, which has been intensively investigated in these years. However, we must be aware that dialogue is not a simple mapping from a user input to a system output, as shown in Figure 2. First, we need to take into account the context of the dialogue. The response to "How about you?" cannot be made without knowing the context. Second, we need an external database or knowledge base to respond to many queries. This is not limited to the conventional tasks of database query or question-answering, but even in chatting, we need personal profiles and common senses to make a consistent and relevant dialogue. What is needed is an open problem. Moreover, there are other factors that affects the responses in human dialogue. They include emotions, desire and characters. Moreover, the emotions and desire can change according to the input during the dialogue. These are not modeled in conventional dialogue systems, but would be required in human-like agents and robots.

Another serious problem in training dialog systems is there are many choices in responses given a user input and there is no ground-truth in chatting-style conversations. Modeling emotions, desire and characters will provide a solution. However, we need to define their objective measurement and the evaluation criteria.

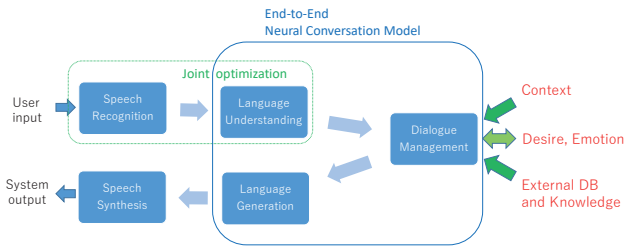


Figure 2 End-to-end spoken dialogue system

Tatsuya Kawahara

PhD, FIEEE

EiC APSIPA T-SIP (2018-), VP-Publications (2014-17)

Professor

School of Informatics

Kyoto University, Japan

Tatsuya Kawahara received B.E. in 1987, M.E. in 1989, and Ph.D. in 1995, all in information science, from Kyoto University, Kyoto, Japan. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a Professor in the School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR and NICT.

He has published more than 300 technical papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been conducting several projects including speech recognition software Julius *and the automatic transcription system for the Japanese Parliament (Diet).*

He was a General Chair of IEEE Automatic Speech Recognition and Understanding workshop (ASRU 2007). He also served as a Tutorial Chair of INTERSPEECH 2010 and a Local Arrangement Chair of ICASSP 2012.

He has been an editorial board member of Elsevier Journal of Computer Speech and Language *and* IEEE/ACM Transactions on Audio, Speech, and Language Processing. *He is a board member of APSIPA and ISCA, and a Fellow of IEEE.*

Android ERICA

Since 2014, we are conducting a project to develop an autonomous android ERICA. Our ultimate goal is to pass a **Total Turing Test**, convincing people that ERICA is indistinguishable from a human in terms of verbal and non-verbal communications including facial expressions and eye-gaze as well as spoken dialogue. This is apparently very challenging even in a 20-year span, as it is almost equivalent to make a human, but we hope this challenge would reveal what is missing in the current technology and what is critical in human communication.

Our realistic goal is to make the interaction with ERICA is as engaging as that with a human. Toward this goal, we set up several social tasks designated for ERICA. Unlike conventional conversational agents and robots, we focus on long and deep interactions, which are not a sequence of short query-response pairs. We choose tasks in which ERICA plays a relatively simple role, focusing on some aspects of interactions.

The first task is **attentive listening**, in which ERICA listens to senior people talking about a given topic such as memorable travels and recent activities. It is similar to counseling, in that she needs to encourage users to talk more by showing interest and empathy. We have investigated generation of natural backchannels in terms of timing, lexical tokens and prosody. We also incorporate partial repeats and elaborating questions based on focus word detection, which can be robustly applied to open-domain conversations.

The second task is **job interview**, in which ERICA asks questions to students applying for a job position in some company. In this scenario, dialogue is designed to be adaptive by generating questions on the fly, without assuming a particular job or a company.

You can watch some demonstration videos of dialogue with ERICA at:
https://www.youtube.com/channel/UCDjRgo5ecEw0Ou78-uJOssg (most of them are in Japanese).
We hope that ERICA can take a role of counselor and interviewer in the future, but the dialogue structure in these roles is relatively simple.

The next task we set up is **speed dating**, which involves asking and answering questions as well as listening and talking. This task calls for the dialogue modeling mentioned in the previous section. We design an emotion model which is affected by the way of talking and listening of the user. It is pre-trained with a set of questionnaires conducted in dialogue data collection, and fine-tuned with dialogue behaviors in an end-to-end manner. We are not sure speed dating with ERICA is pleasant or stressful, but would like to make it real.

Communication skill is one of the fundamental skills of human being and still very important in the current society. For example, face-to-face interview is essential in recruiting students and employees. ERICA is designed not only to replace some human roles, but also to help human practice the skill. This is the reason why we choose job interview and speed dating as target tasks.

Figure 3 Dialogue with ERICA

From Deep Learning to Wide Learning

One of the key concepts of deep learning is an integrated architecture and joint optimization of signal and information processing. This framework has been amazingly successful once we define input and output, and prepare a large amount of the paired data.

While this is regarded as a vertical connection of signal and information processing, we should also explore a horizontal connection that selects relevant information among many sources of signals. This capability is inherent in human being and needed for future AI.

References

[1] T.Kawahara. Spoken dialogue system for a human-like conversational robot ERICA. In Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS), (keynote speech), 2018.

[2] T.Kawahara, T.Yamaguchi, K.Inoue, K.Takanashi, and N.Ward. Prediction and generation of backchannel form for attentive listening systems. In Proc. INTERSPEECH, pp.2890--2894, 2016.

[3] S.Ueno, H.Inaguma, M.Mimura, and T.Kawahara. Acoustic-to-word attention-based model complemented with character-level CTC-based model. In Proc. IEEE-ICASSP, pp.5804--5808, 2018.

[4] ERICA channel: https://www.youtube.com/channel/UCDjRgo5ecEw0Ou78-uJOssg

Thermal Image Based Categorization and Estimation

Categorization and estimation are important basic functions in machine learning and artificial intelligence. These functions are normally performed based on visual information, which requires appropriate brightness. In some difficult situations such as nighttime, environment behind obstacles, and so on, visual information is mostly useless due to less or even none illumination and brightness. In these circumstances, information of heat inside solids, which is naturally varied due to different melting points, is possibly pondered to differentiate solids. These varieties of heat create interesting patterns, and the thermal information hence becomes possible to utilize in categorization and estimation. This article introduces thermal image based categorization and estimation by two cases studied on nighttime vehicle categorization and food calorie estimation.

1) Case study on nighttime vehicle categorization

Vehicle categorization is an important and useful function for an automatic tollgate on a free way. The function basically requires categorizing types of vehicle during daytime and nighttime. In analysis of vehicle heat, the heat is differentiated among windscreen, body, and engine as shown by a sample of sedan in Fig. 1. These heat patterns are actually varied by different types of vehicle. In the case of road without vehicle as shown in Fig. 2, the heat distribution has no unique pattern, and its distribution is seen as non-uniform as shown in the right side figure. Fortunately, heat distribution of vehicles such as sedan, van, and pickup seems like semi-Gaussian as shown by a couple samples in Fig. 3 and 4, while other objects like human, motorcycle, etc. are not the same as shown by a

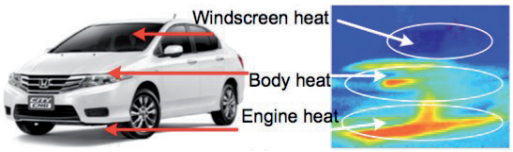


Fig. 1 heat pattern of a sedan

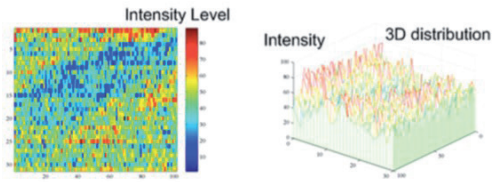


Fig. 2 heat distribution of road

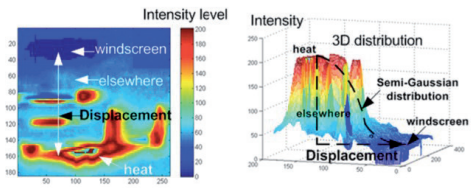


Fig. 3 heat distribution of van

Professor Kosin Chamnongthai

PhD, Professor

Member, Board of governors

Professor,
Electronic and Telecommunication Engineering Dept,

King Mongkut's University of Technology Thonburi

Kosin Chamnongthai currently works as professor at Electronic and Telecommunication Engineering Department, Faculty of Engineering, King Mongkut's University of Technology Thonburi (KMUTT), and also serves as president of ECTI Association (2018-2019). He served as editor of ECTI e-magazine during 2011-2015, associate editor of ECTI-CIT Trans during 2011-2016, associate editor of ECTI-EEC Trans during 2003-2010, associate editor of ELEX (IEICE Trans) during 2008-2010, and chairman of IEEE COMSOC Thailand during 2004-2007.

He has received B.Eng. in Applied Electronics from the University of Electro-communications in 1985, M.Eng. in Electrical Engineering from Nippon Institute of Technology in 1987, and Ph.D. in Electrical Engineering from Keio University in 1991. His research interests include computer vision, image processing, robot vision, and signal processing.

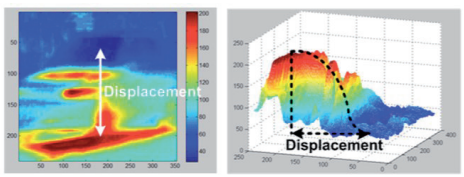


Fig. 4 heat distribution of sedan

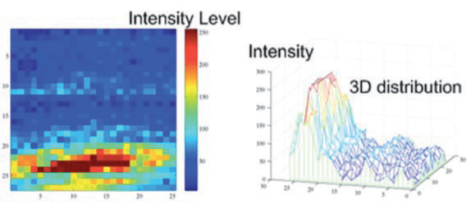


Fig. 5 heat distribution of an object

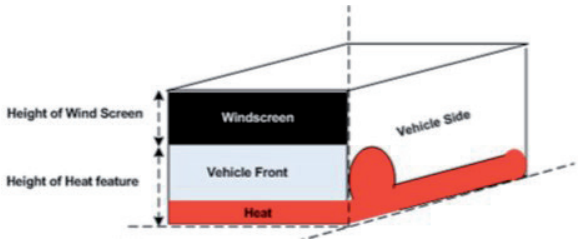


Fig. 6 heat from windscreen, body and engine

sample in Fig. 5. By these analysis results, existence of vehicle is able to confirm in the first step by their heat patterns, a boundary of vehicle can be found by heat pattern of road, and ratio of heat generated from windscreen, body, and engine as shown in Fig. 6 is possibly used in categorization of vehicle type by comparing with training data. The hardware system finally can be set up as shown in Fig. 7 under condition that the thermal camera as heat sensor has to be installed in the position that senses both front and side views of vehicle.

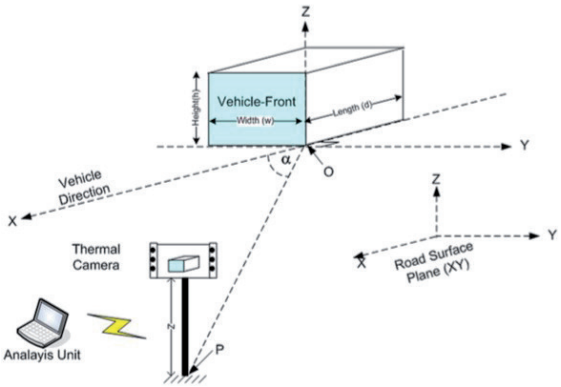


Fig. 7 system set up

2) Case study on food calorie estimation

To estimate calorie by real time in each dish is an essential function for food calorie control, because food in each dish has different calorie even same food. Food is basically cooked based on recipe, which are decided ingredients and their amounts. In analysis by heat in a bowl of Thai curry, which is assumed as a kind of complicated food, ingredients in the curry keep different temperature levels as shown in Fig. 8 due to their different

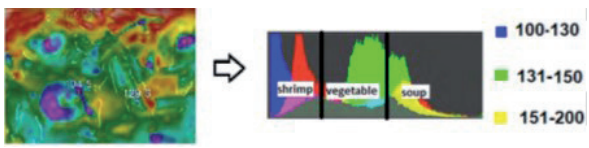


Fig. 8 different ingredient-temperature levels in a dish

melting points. Although temperature of each ingredient is always changed due to surrounding air, the order of ingredient temperatures may never be changed which can be considered as pattern for ingredient recognition. Suppose the food type in a dish is once recognized, the temperature pattern is possibly used to categorize ingredient boundaries based on its recipe, and volume of ingredient may be then used to calculate ingredient and food calories. The real-time food calorie estimation system is set up in a box as shown in Fig. 9 with a heat sensor and a brightness sensor on the ceiling, and a bowl containing food is located in the middle bottom. The system is assumed to be used in the future at canteen, kitchen, etc. for estimating calorie of food in a bowl by real time so that users immediately know exact food calorie before consuming.

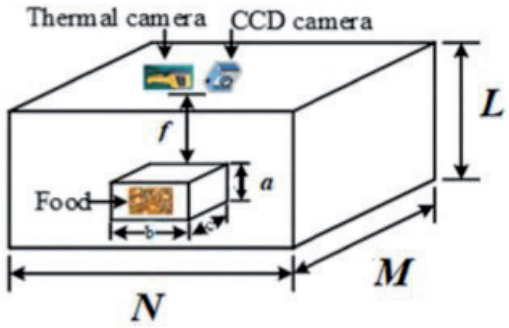


Fig. 9 food-calorie estimation system set up

Long-term Impact

The current thermal cameras functioning as heat sensor work excellently in some applications. Technology of heat sensing and sensors are expected to improve much more in the near future, and positive views of the sensor such as light weight, tiny size, low cost, high resolution, and so on will be realized soon. At that time, these may help providing better heat information, and make it closer to many applications in our daily life. Our studies prove the heat information should be used in some difficult circumstances, and can be fused with other sensors.

References

1. Apiwat Sangnoree and Kosin Chamnongthai, "Thermal-image processing and statistical analysis for vehicle category in nighttime traffic", journal of visual communication and image representation, April 2017, Vol.48, PP. 88-109
2. Sirichai Turmchokksam and Kosin Chamnongthai, "The design and implementation of an ingredient-based food calorie estimation system using nutrition knowledge and fusion of brightness and heat information", IEEE ACCESS, Volume 6, 2018, (DOI: *10.1109/ACCESS.2018.2837046*)

Sample APSIPA Seminar Activities (2011)

Giga-Pixel Mobile Imaging

Smartphone has become the primary imaging device for most people, and the improvement of image quality over the years has been remarkable. Imagine a smartphone that can capture giga-pixel photos. Then a person a hundred meter away can be identified. Such ultimate mobile imaging device is not available today, but the dream may come true in 10 years.

Enablers

The pursuit of high image quality never stops. Among the various image quality measures, resolution is often of the top priority. One can follow the conventional approach by optimizing the optics to increase image resolution. But it has cost and bottleneck issues including the thickness limit of a smartphone.

The advance of computers and computation allows the creation and application of mathematical, algorithmic, and learning tools that are more sophisticated than ever. Imaging becomes a process that involves the interplay between mathematical theory, physical models, and computational algorithms. Computational imaging is a promising approach which will bring us closer to and eventually reach the goal of ultimate mobile imaging.

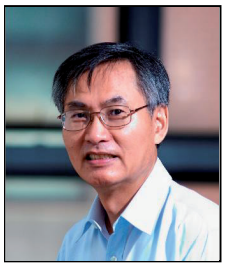
Multi-Aperture Imaging

A giga-pixel imaging system is a system that can generate high-resolution output at the level of 10^9 pixels per image [1]–[4]. The principle behind most existing giga-pixel imaging systems is similar. An array of telescopic cameras are used (hence a multi-aperture imaging system is resulted) and configured in such a way that each telescopic camera captures a part of the scene (Figure 1). Then the captured images are stitched together to generate a large image. The stitching has to take the alignment, rotation, illumination discrepancies between cameras into account.

An imaging system with multiple apertures can provide functions that are hard to achieve for a conventional camera such as scene depth inference, high dynamic range imaging, denoising, synthetic bokeh effect, and object-based segmentation. Particularly, an important application of multi-aperture imaging systems is light field photography, which has drawn worldwide attention in recent years. The most attractive functionality enabled by light field photography is perhaps refocusing, which allows a user to change the position of the focal plane by means of computation after a picture is taken. Refocusing is possible in light field photography because knowing the light field inside a camera is equivalent to knowing what is happening inside the camera at the moment the photographer presses the shutter from a geometrical optics point of view. It is also worthwhile to note that, with light field, 3-D perception of a visual world can be effected without the kind of discomfort due to vergence-accommodation conflict.

Professor Homer H. Chen

PhD, FIEEE **APSIPA BoG Member**



Distinguished Professor

Graduate Institute of Communication Engineering

National Taiwan University

Homer H. Chen received the Ph.D. degree in Electrical and Computer Engineering from University of Illinois at Urbana-Champaign. Dr. Chen's professional career has spanned academia and industry. Since August 2003, he has been with the College of Electrical Engineering and Computer Science, National Taiwan University. Prior to that, he held various R&D management and engineering positions with U.S. companies over a period of 17 years, including AT&T Bell Labs, Rockwell Science Center, iVast, and Digital Island (acquired by Cable & Wireless). His professional interests lie in the broad area of multimedia signal processing and communications, including music emotion recognition, computational imaging and display, perception-inspired image processing, video coding, P2P IPTV, multimedia retrieval, and content delivery. He was a U.S. delegate for ISO and ITU standards committees and contributed to the development of many new interactive multimedia technologies that are now part of the MPEG-4 and JPEG-2000 standards. Dr. Chen is an IEEE Fellow. He served on the IEEE Signal Processing Society Fourier Award Committee (2015-2017) and Fellow Reference Committee (2015-2017). He was an IEEE Circuits and Systems Society Distinguished Lecturer (2012-13) and an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology (2004-10), IEEE Transactions on Image Processing (1992-94), and Pattern Recognition (1989-1999). He was a Guest Editor of IEEE Journal of Selected Topics in Signal Processing (2014), IEEE Transactions on Multimedia (2011), and IEEE Transactions on Circuits and Systems for Video Technology (1999).

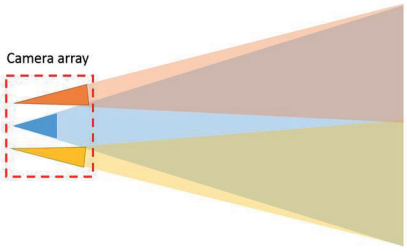


Figure 1. A simplistic sketch of camera configuration for high-resolution imaging.

Table 1: A summary of some existing multi-aperture cameras.

Name	Type	Sensor pixels	Image pixels	Output pixel ratio
Lytro camera	Lenslet Array	11M	1.1M	10%
Lytro ILLUM	Lenslet Array	40M	4.0M	10%
Raytrix R29	Lenslet Array	29M	7.3M	25%
Raytrix R11	Lenslet Array	11M	3.0M	27%
Stanford	Camera Array	355M	1.2M	0.3%
Pelican PiCam	Camera Array	12M	8.0M	67%

Non-overlapping FOVs

Despite the technical advancement, existing multi-aperture imaging systems and light field cameras are haunted by severe resolution reduction. The resolution of the output image is usually only a fraction of the total number of sensor pixels. The ratio between them is called output pixel ratio. For example, the resolution of the refocused image generated by an imaging system consisting of a 17×17 array of cameras is only slightly more than 0.3% of the number of sensor pixels of all cameras, and the output pixel ratio of the Lytro camera is only 10% (Table 1). The pixel ratio is generally unacceptably low for camera arrays configured with overlapping field of views (FOVs).

With overlapping FOVs (Figure 2), increasing the number of cameras or the sampling density does not necessarily improve the quality of super-resolved images [5], [6]. The quality ceiling is ultimately determined by the size of the point spread function (PSF). It is necessary that the cameras have a sufficiently small pixel size and a sharp PSF. The aliasing artifact caused by a sharp PSF greatly decreases the accuracy of image registration, which is necessary for super-resolution algorithms. It should be emphasized that the size of the PSF needs to be determined carefully. An overly sharp PSF can degrade the registration accuracy and hence output quality, while an overly flat PSF is guaranteed to produce blurry results. The tradeoff is a fundamental issue for reconstruction-based light field super-resolution.

We can overcome the shortcoming by keeping the FOVs of telescopic cameras non-overlap and, in addition, combining the array of telescopic cameras with one or more wide-angle cameras (Figure 1). The wide-angle camera is configured such that its FOV covers the entire scene and overlaps with the FOV of each telescopic camera. By doing so, the multiple images captured by a multi-aperture imaging system can be merged into one high-resolution sharp image without sacrificing functions such as depth inference and refocusing that are expected for a multi-aperture system [7].

Long-term Impact

A camera array with mixed focal lengths is capable of generating high-resolution images and disparity maps from the captured images. It can also produce the refocusing bokeh effect like what a light field camera does. To maximize the resolution of the output images, the camera configuration and the accompanying algorithm have to be optimized. Our experiences show that it is feasible to use the FOV coverage ratio to determine the tilt angle of the telescopic cameras. In addition, when merging the captured low-resolution images into a high-resolution output image, an image fusion algorithm with occlusion detection is useful. With occlusion detection incorporated in the loop, the resulting image is significantly sharper in regions with complex 3D structure, where dense image correspondence is expected to fail [8]. In conclusion, high-resolution imaging using a multi-aperture system offers a feasible path to giga-pixel mobile imaging.

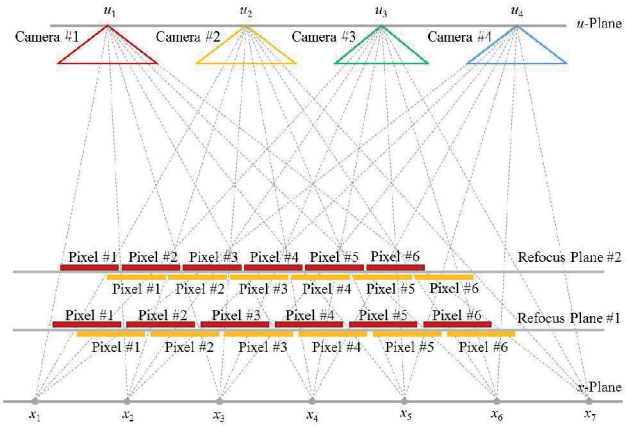


Figure 2. An illustration of generating subpixel refocused images at two different refocus planes. The low-resolution pixels of Cameras #1 and #2 are plotted to show the relative displacement between the cameras [7].

Acknowledgement: This article is based on a joint work with Dr. Kuang-Tsu Shih.

References

[1] J. Kopf, M. Uyttendaele, O. Deussen, and M. F. Cohen, "Capturing and viewing gigapixel images," *ACM Trans. Graph*., vol. 26, no. 3, pp. 93:1–93:10, 2007.

[2] D. J. Brady et al., "Multiscale gigapixel photography," Nature, vol. 486, pp. 386–389, 2012.

[3] O. S. Cossairt, D. Miau, and S. K. Nayar, "Gigapixel computational imaging," in *Proc. IEEE Int. Conf. Comput. Photography*, pp. 1–8, 2011.

[4] X. Yuan, L. Fang, Q. Dai, D. J. Brady, Y. Liu, "Multiscale gigapixel video: A cross resolution image matching and warping approach," *in Proc. IEEE Int. Conf. Comput. Photography*, pp. 1–9, 2017.

[5] K.-T. Shih and H. H. Chen, "Performance analysis of reconstruction-based super-resolution for camera arrays," in *Proc. IEEE Int. Conf. Image Processing* (ICIP), pp. 1162–1166, 2017.

[6] K.-T. Shih, C.-Y. Hsu, H. H. Chen, "Analysis of the effect of calibration error on light field super-resolution rendering," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process*., pp. 534–538, May 2014.

[7] K.-T. Shih, High Resolution Imaging and Depth Acquisition Using a Camera Array, Ph.D. Dissertation, National Taiwan University, 2017.

[8] K.-T. Shih and H. H. Chen, "Generating High-resolution image and depth map using a camera array with mixed focal lengths," *IEEE Trans. Comput. Imaging*, to appear.

Three-dimensional Video Capturing and Processing

With the emerging market of AR/VR imaging products, 3D video has become an active area of research and development in recent years. 3D video is the key to provide more realistic and immersive perceptual experiences than the existing 2D counterpart. In this article, we briefly describe main topics of 3D video capturing and processing.

Camera Calibration

The real world object exists in the 3D space. However, when we capture the 3D object by a camera, it is represented on the 2D plane. In order to reconstruct the 3D scene from the captured 2D images, we need to estimate the depth information of the scene. In our lab at GIST, we have built a multi-view color and depth camera system, as shown in Fig. 1, to generate 3D contents for AR/VR applications [1].

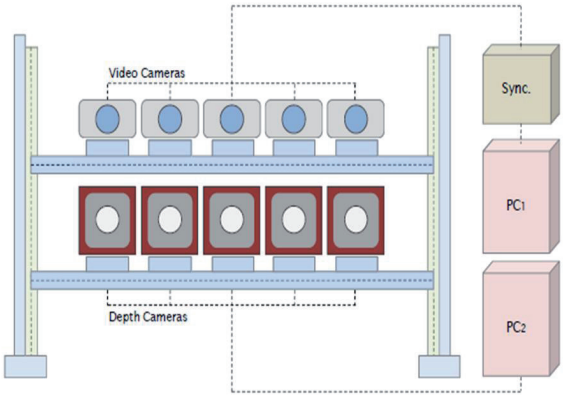


Fig. 1. Multi-view color and depth camera system

If we want to know where a point in the 3D space is projected onto the 2D plane, you should find the geometric relationship between them: rotation and position of the camera (extrinsic parameters), and relationship between the lens and image sensors in the camera (intrinsic parameters). Fig. 2 shows an example condition for stereo camera calibration. In order to obtain intrinsic and extrinsic camera parameters, we can capture the calibration patterns simultaneously from the left and right cameras [2, 3].

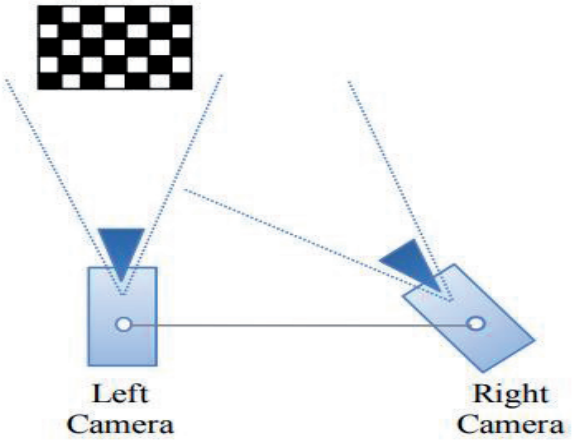


Fig. 2. Camera calibration

Professor Yo-Sung Ho

PhD, FIEEE

BOG Member of APSIPA (2018-2020)

Professor

School of Electrical Engineering and Computer Science

Gwangju Institute of Science and Technology

Yo-Sung Ho received the B.S. and M.S. degrees in electronic engineering from Seoul National University, Seoul, Korea, in 1981 and 1983, respectively, and the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, in 1990. He joined Electronics and Telecommunications Research Institute (ETRI), Daejon, Korea, in 1983. From 1990 to 1993, he was with Philips Laboratories, Briarcliff Manor, New York, USA. In 1993, he rejoined the technical staff of ETRI. Since September 1995, he has been with Gwangju Institute of Science and Technology (GIST), where he is currently Professor of School of Electrical Engineering and Computer Science. Since August 2003, he has been Director of Realistic Broadcasting Research Center (RBRC) at GIST in Korea. He has authored or coauthored over 1,000 published papers and books. He is a Fellow of IEEE. He is currently a BoG member of APSIPA, and he served as Founding Chair of Image, Video and Multimedia TC in 2009-2013, and Vice President of Institutional Relations and Education Program in 2016-2016. He was also President of the Korean Institute of Broadcast and Media Engineers (KIBME) in 2016. He has served as an Associate Editor of IEEE Transactions on Multimedia (T-MM) and IEEE Transactions on Circuits and Systems Video Technology (T-CSVT). He has organized several international conferences, including APSIPA ASC 2016. Because of his distinctive contributions, Dr. Ho received many research awards and paper awards, including the Outstanding Research Award from National IT Industry Promotion Agency in Korea.

The camera calibration step is necessary when we deal with the geometric relationship among multiple cameras, such as 3D warping or image rectification. 3D warping is a mapping to forward the depth information from one viewpoint image to another viewpoint image using camera parameters. We can forward a point located in the left image to a point located in the right image via a point placed in the world coordinate system. Image distortion caused by the hardware problem such as a radial distortion can be also corrected using the camera parameters.

Image Rectification

Even if multi-view cameras are installed close to an ideal arrangement, there are geometrical errors among the cameras. Those geometric errors existed in the multi-view color image can not only deteriorate the depth quality, but also interfere with the natural 3D view. Thus, we minimize the geometric errors by performing the multi-view image rectification to the color image [4, 5].

Fig. 3 shows the multi-view images before image rectification. Fig. 3 (c) and Fig. 3 (d) indicate that multi-view images are not properly corrected. Fig. 4 shows the multi-view images after rectification.

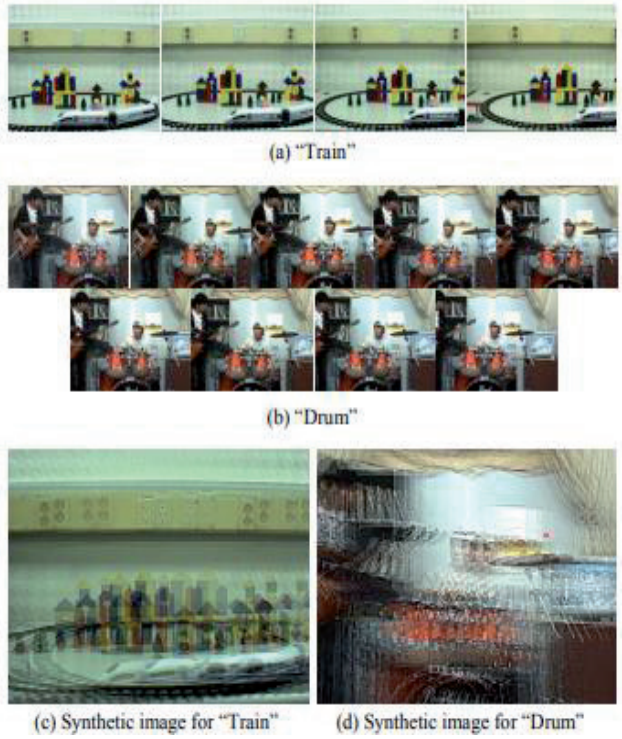


Fig. 3. Multi-view images before rectification

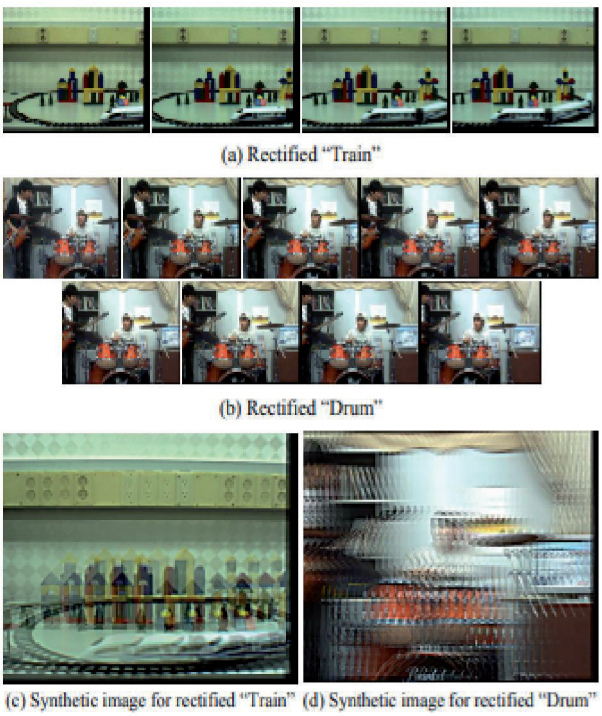


Fig. 4. Multi-view images after rectification

Color Correction

Color images captured by the multi-view camera system include illuminance variations and noises resulting from image sensors. Such color inconsistency may affect the performance of the stereo matching operation for depth estimation. Moreover, it can influence the quality of the reconstructed 3D scene and the virtual view synthesis.

In order to solve the color mismatch problem in multi-view color images, several methods have been proposed [6, 7]. Fig. 5 shows the results of different color correction methods [8, 9].

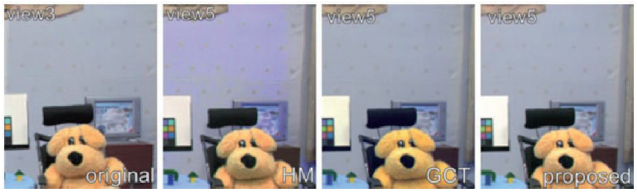


Fig. 5. Color correction

Depth Estimation

The depth information is very important for 3D image processing. Although the color information is relatively easy to capture using color cameras, it is still a big challenge to acquire an accurate depth map in real time.

Typical depth cameras measure the distance information of objects using structured lights or infrared rays. Depth measurement using the depth camera is fast and accurate in indoor environments, but it is not working well in outdoor environments due to hardware limitations and other environmental issues. A depth map captured by the depth camera usually has a lower resolution than the color image. Thus, we need to up-sample the depth map to have the same resolution as the color image. Depth warping based on camera parameters can be used for depth map up-sampling. However, it may generate many holes due to the resolution difference. Hence, various hole filling methods are used for the depth map up-sampling [10, 11].

The depth map can also be estimated from stereo images by calculating disparity values between two correspondence points in both images. The disparity estimation process from stereo images is called stereo matching. Since this disparity value has the same characteristics as the human binocular disparity, the depth information can be predicted from the disparity value. The disparity value increases as the object is closer to the camera, and it decreases as the object moves away from the camera. There are various methods for stereo matching [12, 13]. Fig. 6 shows the estimated depth maps from stereo image pairs.

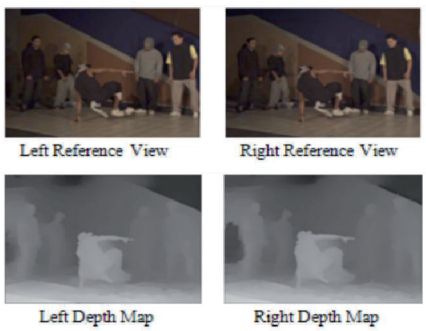


Fig.6. Depth estimation by stereo matching

By stereo matching, we can acquire the disparity map of the same resolution as the corresponding color image pairs. In addition, depth estimation using stereo matching methods is less affected by environmental factors. Stereo matching is basically a pattern matching operation to estimate disparity values by searching corresponding points between two images. Hence, it requires quite intensive computations. Although stereo matching provides quite accurate disparity values in textured regions, disparity values are relatively inaccurate in those areas of no texture or periodically repeated texture. Furthermore, stereo matching methods are weak in heavy occlusion regions.

Our multi-view color and depth camera system obtains more accurate depth information in real time by exploiting the advantages of the depth camera and stereo matching [1]. The depth value measured by the depth camera is warped into the corresponding position in the color image. The disparity value converted from the warped depth value can be used as the initial value for stereo matching to limit its search range. Thus, the hybrid depth estimation method saves the time significantly for disparity estimation and improves its accuracy as well.

Virtual View Synthesis

In order to generate immersive video contents from multi-view images, we can use depth image based rendering (DIBR). In other words, we can employ virtual view synthesis methods using the depth information to generate more viewpoint images than the captured ones [14].

In general, 3D image warping is applied using the depth information for viewpoint shifting. Since the depth map indicates the distance information between the camera and objects in the scene, we can map corresponding pixels between different viewpoints. When the viewpoints are changed, some background regions are disappeared or appeared because of foreground objects. Hence, the 3D image warping operation may create small holes near disoccluded object boundaries. Those holes can be filled by various interpolation methods [15]. Fig. 7 shows the overall operation of intermediate view synthesis.

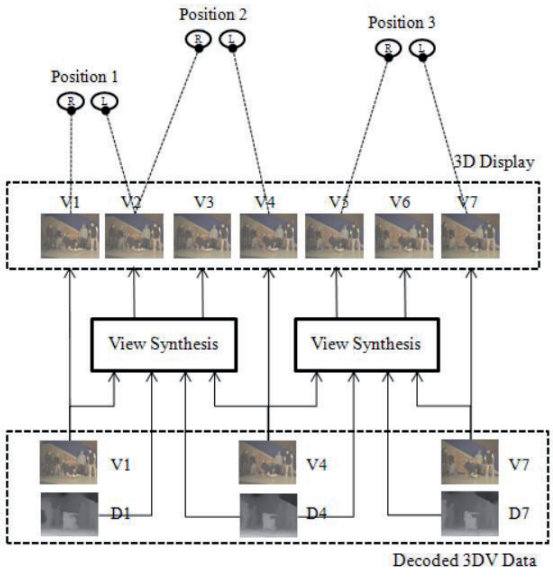


Fig. 7. Virtual view synthesis

Long-term Impact

There are many applications of 3D video, such as 3DTV, 3D movie, and AR/VR with 360˚ video, which are considered the main drive of the next-generation technical revolution. Stereoscopic display is the current mainstream technology for 3DTV, while auto-stereoscopic display is a more promising solution. The field of 3D image processing requires more research endeavors to resolve the associated technical problems.

References

[1] Y.S. Kang, E.K. Lee, and Y.S. Ho, "Multi-Depth Camera System for 3D Video Generation," International Workshop on Advanced Image Technology, pp. 44(1-6), Jan. 2010.

[2] Z. Zhang, "A Flexible New Technique for Camera Calibration," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 11, pp. 1330-1334, Nov. 2000.

[3] Y.S. Kang and Y.S. Ho, "Geometrical Calibration of Stereo Images in Convergent Camera Arrangement," IEEE-RIVF International Conference on Computing and Communication Technologies, pp. 44-47, Feb. 2012.

[4] D.V. Papadimitriou and T.J. Dennis, "Epipolar Line Estimation and Rectification for Stereo Image Pairs," IEEE Transactions on Image Processing, Vol. 5, No. 4, pp. 672-676, Apr. 1996.

[5] Y.S. Kang and Y.S. Ho, "Geometric Error Correction of Convergent Multiview Images," Electronic Letters, Vol. 51, No. 7, pp. 557-558, Apr. 2015.

[6] J.I. Jung and Y.S. Ho, "Color Correction for Multiview Images Using Relative Luminance and Chrominance Mapping Curves," Journal of Signal Processing Systems for Signal Image and Video Technology, Vol. 72, No. 2, pp. 107-117, Dec. 2012.

[7] J.I. Jung and Y.S. Ho, "Color Correction Algorithm Based on Camera Characteristics for Multiview Video Coding," Signal Image and Video Processing, Vol. 8, No. 5, pp. 955-966, Jul. 2014.

[8] U. Fecker, M. Barkowsky, and A. Kaup, "Histogram-based Prefiltering for Luminace and Chrominace Compensation of Multiview Video," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 18, No. 9, pp. 1258-1267, Jun. 2008.

[9] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color Transfer Between Images," IEEE Computer Graphics and Applications, Vol. 21, No. 5, pp. 34-41, Sep/Oct. 2001.

[10] S.B. Lee and Y.S. Ho, "Discontinuity Adaptive Depth Up-sampling for 3D Video Acquisition," Electronic Letters, Vol. 49, No. 25, pp. 1612-1614, Dec. 2013.

[11] Y.S. Kang, S.B. Lee, and Y.S. Ho, "Depth Map Up-sampling using Depth Local Features," Electronic Letters, Vol. 50, No, 3, pp. 170-171, Jan. 2014.

[12] Q. Yang, L. Wang, R. Yang, S. Wang, M. Liao, and D. Nister, "Real-time Global Stereo Matching using Hierarchical Belief Propagation," British Machine Vision Conference, pp. 989-998, Jan. 2006.

[13] Y.J. Chang and Y.S. Ho, "Disparity Map Enhancement in Pixel Based Stereo Matching Method Using Distance Transform," Journal of Visual Communication and Image Representation, Vol. 40, Part A, pp. 118-127, Oct. 2016.

[14] C. Lee and Y.S. Ho, "View Synthesis Using Depth Map for 3D Video," Asia-Pacific Signal and Information Processing Association, pp. 1-8, Oct. 2009.

[15] J.H. Mun and Y.S. Ho, "Multi-directional Hole Filling Method for Virtual View Synthesis," Journal of Signal Processing Systems for Signal Image and Video Technology, Vol. 85, No. 2, pp. 211-219, Nov. 2016.

Computational DNA

DNA contains genetic information of every living organism. Studying DNA sequences enables us to know about our genetic make-up, facilitate early identification of genetic related diseases as well as establish biological connection. Conventionally, experiments were carried out to gain this knowledge. With advancement in technologies, DNA sequencing now becomes faster and more affordable in which individual sequencing is getting popular. Through sequencing, DNA can be represented as a long string containing 4 bases: *A* (Adenine), *C* (Cytosine), *G* (Guanine) and *T* (Thymine). In this way, signal processing techniques can be applied conveniently to extract the knowledge embedded in the DNA sequences in a computational manner. In this article, we would summarize about how signal processing techniques are used in DNA-related study.

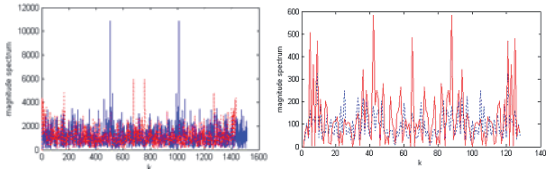
Periodicity Detection in Genes

There are about 3 billion bases in a human DNA sequence. Genes are embedded in the sequence and constitute about 5% of the total number of bases. It is estimated that a person contains between 20,000 and 30,000 genes. Genes are further divided into two regions, namely exon and intron in which exon codes for the proteins while intron may involve in regulation function. An important problem in DNA sequence analysis is to distinguish these two regions. Three-base periodicity is well known to be present in exons and absent in introns. A natural choice for detecting periodicity in signal processing area is the Discrete Fourier Transform (DFT). To obtain a numerical representation of the DNA sequence, we may consider rewrite the sequence using a binary indicator sequence $u_B[n]$ which takes the value of 1 or 0, depending on whether the base *B* exists at *n* or not, i.e.,

$$a\,u_A[n] + c\,u_C[n] + g\,u_G[n] + t\,u_T[n] \qquad (1)$$

where $\{a,c,g,t\}$ are weightings. In this way, DFT can be performed along the sequence for periodicity detection. By observing if a peak at frequency $\omega = 2\pi/3$ exists or not, region can be classified to be either exon or intron. Figure 1 (a) shows an example. The total number of bases (*N*) in the sequence is 1400. Peaks can clearly be seen at around positions 467 and 933 (*N*/3 and 2*N*/3) for exon (represented using blue color) while peak are absent for intron at those positions (represented using red color).

For human genome, the length of exons is rather short. Approximately 80% of exons contain less than 200 bases. The DFT works well for long exons containing more than 1000 bases as can be seen in Figure 1(a). However, the performance drops significantly for short exons containing less than 200 bases. Figure 1 (b) shows the magnitude spectrum of a short exon consisting of 120 bases (in blue color). The peak at around the 80th base is not obvious. To improve the robustness of DFT-based methods, biological properties should be incorporated in the periodicity detection. For example, distribution of bases of the pruine (A or G) and pyrimidine (C or T); distribution of the bases of amino (A or C) and keto (G or T) types; and distribution of the bases of the weak H-bond (A or T) and strong H-bond (G or C) types can be considered. This is commonly termed as Z-curve approach [1] in which the three-base periodicity becomes more prominent in these distributions as compared to the original binary indicator sequences. As shown in Figure 1 (b), applying the DFT to the distribution of bases of the pruine and pyrimidine, the peak is prominent. Through using the DFT on these distributions, the three-base periodicity can be better captured for the identification of the exon region.



(a) (b)

Figure 1: The magnitude spectra of (a) long exon and intron where the number of bases is more than 1000, and (b) short exon where the number of bases is only 120.

Dr Bonnie Ngai-Fong Law

PhD, SMIEEE

BoG member (2018-2020) and Secretary, APSIPA Headquarters

Associate Professor

Dept of Electronic & Information Engineering

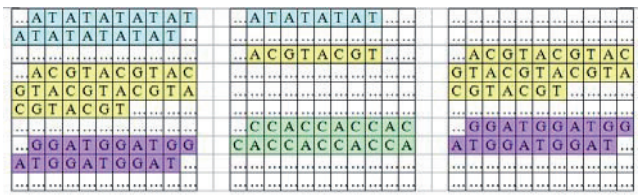
The Hong Kong Polytechnic University

Bonnie Ngai-Fong Law received the BEng degree with first class Honours from the University of Auckland, New Zealand, in 1993 and a PhD degree from the University of Tasmania, Australia, in 1997, both in Electrical and Electronic Engineering. She currently serves as a BoG member of APSIPA and secretary of APSIPA Headquarters. Her research interests include wavelets, image search, bioinformatics (gene expression data analysis and DNA sequence study) and image forensics. She has published over 90 research papers, including 40 of which published in prestigious international archival journals. She is editorial board member of two journals and members of a number of technical committees including APSIPA Biomedical Signal Processing and System. She has organized a number of international conferences and was a recipient of two conference paper awards.

Use of Repetitions for Establishing Biological Connection

In addition to the use of three-base periodicity for exons/introns identification, non-coding regions also contain repetitions which can be useful to provide our genetic fingerprint and establish biological connection among individuals. While 99.9% of DNA between two people is the same, introns vary among individuals. In particular, human genome contains a number of short repeating sequences in introns such as "*AT*" (repeating 8 times) and "*GGAT*" (repeating 4 times). However, the exact number of repeats is similar between biologically related individuals but vary among unrelated ones. This variation forms the basis of establishing biological connections among individuals.

Consider an example shown in Figure 2. We considered four artificial repeating sequences: *AT* (Locus A), *ACGT* (Locus B), *GGAT* (Locus C), and *CCA* (Locus D). There are three individuals. The first individual has *AT* repeated 10 times (A10), *ACGT* repeated 7 times (B7) and *GGAT* repeated 5 times (C5). The first individual does not have any repeats on *CCA* (i.e., D is absent). Hence the genetic fingerprint of the first individual is "A10, B7, C5". Using similar calculations, the genetic fingerprint of the second and the third individuals are "D7, A4, B2" and "B7, C5" respectively. Thus the first and the third individuals are likely to be biologically related while the second individual is likely not related to the first and the third individuals. This way of establishing biological connections will be statistically valid through using sufficient number of loci (e.g., more than 13) to form the genetic fingerprint.



(a) (b) (c)

Figure 2: DNA sequences for the (a) first, (b) second and (c) third individual.

Use of Repetition for Compression

The advancement of sequencing technologies introduced a new challenge. It is predicted that millions of individual human genome would be generated in the near future. Lossless compression of these sequences is thus needed to achieve an effective data storage, distribution and management. Often, repetition inside DNA sequences is used for compression. Identical subsequences within the DNA compression are found so that these subsequences can be encoded together. Different ways to identify and characterize intra-sequence similarities have been proposed. However, the average bit per base (bpb) is only reduced to around 1.73 for benchmark DNA sequences. As compared to the uncompressed 2 bpb, the compression gain is insufficient to deal with the exponential growth in the number of DNA sequences.

We have found that in addition to intra-sequence similarities, there are other kinds of similarities in DNA sequences. For example, for different chromosome sequences, there are partial sequence similarities in which sub-sequences within different chromosome sequences are similar to each other [2-4]. For DNA sequences from different individuals of the same species, their DNA sequences are highly similar to each other. A reference-based lossless compression scheme was developed to consider both intra-sequence and inter-sequence similarities [5, 8]. For a group of DNA sequences, the scheme works by first identifying similar sub-sequences within the group of DNA sequences to be compressed. After that, similar sub-sequences are encoded together to achieve compression. Using this reference-based compression, we found that the bpb can be significantly decreased. For example, for the dataset containing 33 sequences of E. coli with an average length of 5M, the bpb becomes 0.4550. As for the highly similar human mitochondrial dataset containing 3616 sequences with an average length of 16K, the bpb drops to 0.0386. As compared to the uncompressed 2 bpb, a great compression can be achieved.

Use of a reference sequence appears to provide an effective solution for compressing a group of highly similar sequences. Despite that, a key issue is the selection of the appropriate reference sequence. It needs to be a good representation of the other sequences to be compressed. Basically, the reference sequence can be selected as one of the DNA sequences to be compressed or constructed artificially. Besides, there is a premise that a single reference sequence is sufficiently representative. Unfortunately, sequences may have sub-structures even from the same species. For example for Han Chinese, southern group and northern group could have big variations. Thus, clustering can be used to group sequences into different clusters [6, 7]. In this way, sequences within one cluster can be well approximated by a single reference sequence in that cluster. Figure 3 shows the idea.

In order to minimize the computational cost of clustering, k-means clustering is used. In particular, a normalized local histogram is constructed to represent characteristics of each input sequence. Then feature-based clustering is applied to group similar DNA sequences together. After clustering, a representative sequence is derived for each cluster so that a reference-based compression can be performed [8]. Using this multiple reference-based compression for the highly similar human mitochondrial dataset containing 3616 sequences with an average length of 16K, the bpb drops to 0.0168, which is much smaller than 0.0386 achieved by using a single reference. For the dataset of E. coli containing 33 sequences with an average length of 5M, the bpb drops from 0.4550 (single reference sequence) to 0.2769 (multiple reference sequences).

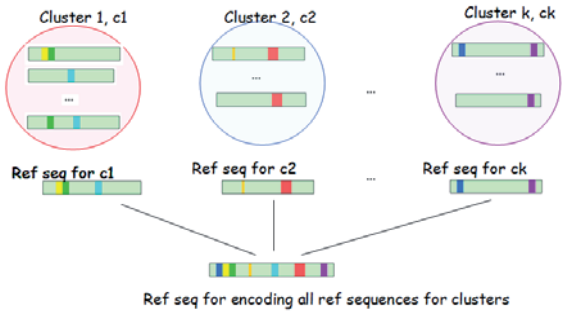


Figure 3: Clustering-based compression.

Long-term Impact

The advancement in the sequencing technologies has opened up new areas of study related to the DNA. Besides the issues discussed above, DNA can in fact store a vast amount of data. It has been demonstrated that artificial DNA sequence can be created so that messages such as text, image and video can be incorporated into these sequences for storage. In the future, DNA may be used for data storage, or even use for transmitting secret messages. Signal processing techniques are thus needed for encoding, decoding and detecting if there is any secret message.

References

[1] N.F. Law, K.O. Cheng and W.C. Siu, "On Relationship of Z-curve and Fourier Approaches for DNA Coding Sequence Classification", Bioinformation, Vol. 1 (7), 242-246, 2006.

[2] Paula Wu, N.F. Law and W.C. Siu, "Cross Chromosomal Similarity for DNA Sequence Compression", Bioinformation, Vol. 2 (9), 412-416, 2008.

[3] P. Wu, N.F. Law and W.C. Siu, "Analysis of Cross Sequence Similarities for Multiple DNA Sequences Compression", International Journal of Computer Aided Engineering and Technology, Vol. (4), 437-454, 2009.

[4] K.O. Cheng, N.F. Law and W.C. Siu, "A Novel DNA Sequence Compression Scheme using both Intra and Inter Sequences Correlation", APSIPA ASC 2015, 237-241, HK.

[5] K.O. Cheng, Paula Wu, N.F. Law and W.C. Siu, "Compression of Multiple DNA Sequences using Intra-sequence and Inter-sequence Similarities", IEEE/ACM Trans on Computational Biology and Bioinformatics, Vol. 12, No. 6, 1322-1332, Nov/Dec, 2015.

[6] K.O. Cheng, N.F. Law and W.C. Siu, "Clustering-based Compression for Population DNA Sequences", IEEE/ACM Trans on Computational Biology and Bioinformatics, (accepted), 2018.

[7] K.O. Cheng, N.F. Law and W.C. Siu, "Compressing population DNA Sequences using Multiple Reference Sequences", APSIPA ASC 2017, 760-764, 2017.

[8] DNA Compression Software:

http://www.eie.polyu.edu.hk/~nflaw/DNAComp/index.html and http://www.eie.polyu.edu.hk/~nflaw/RCC/index.html

Photo Gallery: APSIPA ASC'2016 in Jeju

Tactile Internet and Swarm Intelligence for Air Pollution Monitoring

Over the past decades, environmental awareness is becoming a vital issue in our societies. Owing to the industrialization of the countries and increased consumption of fossil fuels, the air pollution in the environment surpassed above the safety levels. Especially industrialization leads to pollution dispersion through plumes in the environment. The surpassed pollution levels can be controlled by searching the pollution source. An environmental monitoring UAVs can address this issue. However, there are many challenges including how do UAVs quickly navigate towards polluted area without collision with obstacles? How do UAVs collaboratively search for the pollution source? What is a reliable infrastructure to support the communication? How to integrate the Tactile Internet as a part of the process?

In the Creative Intelligent System Lab. of CSE department of National Sun Yat-sen University, a project that will integrate the Tactile Internet, intelligent UAV, swarm intelligent algorithm, 4G/5G wireless communication technology, human-computer interaction to carry out the air pollution source detection is under way. A 5G secure communication system architecture with network slicing to support the task is shown in Figure 1.

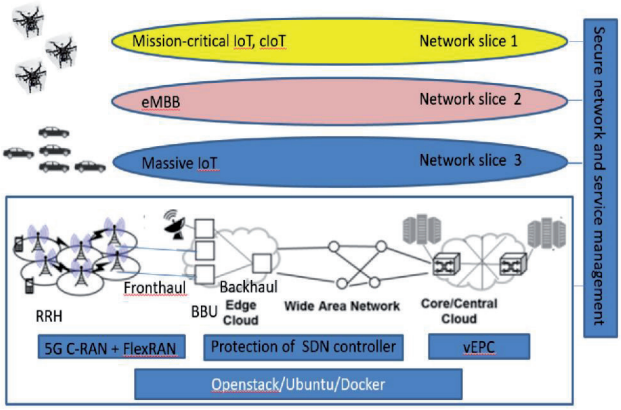


Figure 1 The 5G secure communication system architecture

The development of the fifth generation (5G) communication technology is in progress to satisfy the increasing demands of higher capacity, lower latency, and ubiquitous mobile access in the next generation mobile cellular networks. Furthermore, 5G networks shall be able to provide diverse and challenging requirements. Due to Cloud-Radio Access Network architecture has been identified as a promising approach to 5G, and network slicing offers a serviceable way to meet diverse requirements for 5G [5]. We realize a feasible solution for managing network slices in the C-RAN architecture in order to facilitate the deployment of the 5G prototype. We make use of OpenStack, Docker and OpenAirInterface to implement a C-RAN architecture with properties of flexibility, scalability and rapid deployment [4].

5G allows a large number of data transmission intensive applications including vehicle networking, IoT, wearable devices, virtual reality to operate smoothly. But the real advantage of 5G communication is that the tactile Internet, which will be the next wave of innovation, has a significant impact on human life. We can apply the Tactile Internet to help the pollution monitoring too.

We propose a novel methodology by using the collaborative intelligence learned from Golden shiners schooling fish [3]. We adopt shiners collective intelligence using the particle swarm optimization and use a Gaussian

Prof. Chung-Nan Lee

PhD, the outstanding engineering professor from Chinese Institute of Engineers, Taiwan

Member-at-Large of Board of Governors Advisory Board

Distinguished Professor

Department of Computer Science and Engineering

National Sun Yat-sen University

Chung-Nan Lee received the B.S. and the M.S. degrees, both in electrical engineering, from National Cheng Kung University, Tainan, Taiwan, in 1980 and 1982, respectively, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, in 1992. Since 1992, he has been with National Sun Yat-Sen University, Kaohsiung, Taiwan, where he was an Associate Professor in the Department of Computer Science and Engineering from 1992 to 1999; he was the Chairman of the Department of Computer Science and Engineering from August 1999 to July, 2001; Director of Wireless Broadband Communication Protocol Cross-Campus Research Center from 2013 to 2017; President of Taiwan Association of Cloud Computing from 2015 to 2017; currently, he is a Distinguished Professor and Director of Cloud Computing Research Center; Board of Governors of APSIPA from 2017. In 2016 he received the outstanding engineering professor from Chinese Institute of Engineers, Taiwan. His current research interests include multimedia over wireless networks, cloud computing, and evolutionary computing.

plume model for depicting the pollution distribution in the UAV search environment. Furthermore, our proposed method incorporates path planning and collision-avoidance for UAV group navigation.

To allow the communication of UAV group navigation, we use the Message Queuing Telemetry Transport communication protocol as the communication mechanism. This mechanism allows UAV to automatically share flight parameters or sensed values and perform further complex operations.

Long-term Impact

Though the UAVs has been applied to many areas such as in military, power lines inspection, crop monitoring, drama shooting, rescue exploration, monument maintenance, tourism assistance, environmental monitoring, traffic monitoring, intelligent monitoring and even land conservation, there are still many to explore. The UAVs not only can autonomously navigate alone, but also to cooperate to achieve a certain task like the pollution source identification.

Due to the limited battery capacity, UAVs have a limitation on range. Hence, the charging stations either in wireless or wire power transfer should be built just like gas stations and UAVs should be able to find a charging station themselves to recharge their power automatically too.

The system should allow the video captured from the cameras and the tactile sensors to the edge cloud for further image processing to recognize scene and avoid of collision. At the same time via a cockpit built in the user side, users can manipulate the virtual reality to present and interact with UAV in the first person view, let the users see obstacles and air scene, expand the view of bird-like overlooking and subjective patrol to guide the UAVs reach the target as soon as possible.

The challenging aspect of the Tactile Tnternet is to ensure that the underlying systems and infrastructure are ready to work, such as 5G network, UAVs, control systems, sensors and AR/VR/MR. The future application may not limit to air pollution, they should be able to detect the water pollution by carrying some specific water pollution sensors. All together it will make the world a better place to live.

References

[1] K. D. Hutchinson, "Applications of MODIS satellite data and products for monitoring air quality in the state of Texas," Atmospheric Environment, Vol. 37, 2003, PP. 2403-2412.

[2] F. Lopez-Pena, G. Varela, A. Paz-Lopez, R. J. Duro and F. J. G. Castano, "Public transportation based dynamic urban pollution monitoring system," Sensors & Transducers, Vol. 8, 2010, PP. 13-25.

[3] H. Liang and G. Gao, "Navigating robot swarms using collective intelligence learned from golden shiner fish," Proceedings of Collective Intelligence Conference, (CI2014), Cambridge, MA, Jun 2014

[4] C. N. Lee, M. F. Lee, J. M. Wu and W. C. Chang, "A Feasible 5G cloud-RAN architecture with network slicing functionality," APSIPA 2018, Hawaii, USA, 12-15, Nov. 2018

[5] H. J. Son and C. Yoo, "E2E Network Slicing—Key 5G technology." NETMANIAS TECHBLOG
Available at
https://www.netmanias.com/en/post/blog/8325/5g-iot-network-slicing-sdn-nfv/e2e-network-slicing-key-5g-technology-what-is-it-why-do-we-need-it-how-do-we-implement-it, November 27, 2015

Photo Gallery: APSIPA ASC' 2013 in Kaohsung

Sample ASIPA Activities (2016)

Signal Processing of Human Driving

Modeling human intelligence by means of computer technology has recently become a practical issue for industries. Among the many domains in which such research efforts are going on, one of the most impactful is in automotive applications. How does human learn how to drive? What is the essential knowledge for driving? How human perception can be replaced or enhanced by sensing technologies? etc…In simplest terms, driving can be modeled by a filtering process of a signal flow, i.e., taking a set of sensor signals of the traffic context (surrounding environment) as input, generating output signals of steering wheel angle, gas and brake pedal positions. Obviously, this is within the target field of APSIPA community [1,2].

Signal processing technologies such as computer vision and telecommunication have been already central issues in the context of intelligent vehicles, however, not much about modeling human driving behavior.For building human-cooperative mobility systems, further researchesare needed in modeling human driving behavior by means of signal processing.

Characterizing Drivers' Personality

One of important challenges of modeling driving behavior is to characterize the difference among drivers by means of signal processing. This technology can be regarded as natural extension of the speaker recognition in speech research. In general, human behavior signals, i.e., signals obtained through observing human behaviors, reflecting two types of personalities. One type includes static characteristics that are associated with human biological nature. The other type includes dynamic characteristics, in other words, how they react to a certain situation;the signals associated with this personality usually depends on the context [3].

In the speaker recognition, the static or biological characteristics are estimated using long-term signal such as average F0 frequency, whereas cepstrum analysis are often used for calculating the latter/behavioral characteristics. One interesting approach in characterizing driving behavior is to apply cepstrum analysis so that the driving behavior, such as pedal signals are blindly deconvolved and frequency response of the pedal operation can be extracted (Fig.1). This idea has been tested using the real-world driving data of 276 drivers. By introducing the cepstrum analysis, the driver identity can be correctly identified with more than 75% accuracy based on the gas/brake pedal signals of 5 minutes driving. The accuracy improvement from the raw signals was more than 30 points or 55% error reduction.

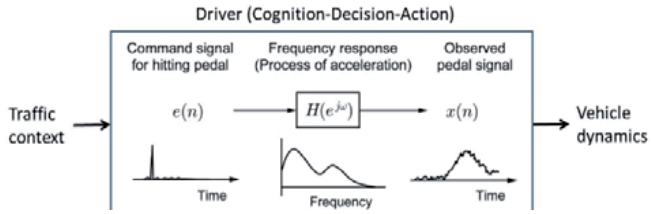


Fig. 1: Cepstrum analysis of driving signal.

Professor Kazuya Takeda

Dr.Eng, FIEICE, SrMIEEE

BoG Member (2016-2018) APSIPA

Professor, Institute of Innovation for Future Society/Graduate School of Informatics, Nagoya University

Prof. Kazuya Takeda is working in the field of signal processing technology research for acoustic, speech and vehicle applications. Understanding human behavior through data centric approaches utilizing real world signal corpus has been his main interest.

Prof. Takeda is a Professor at the Nagoya University, Japan. He received his B.E.E., M.E.E. and Ph.D. in 1983, 1985 and 1995, respectively from Nagoya University. After graduating from the university, he worked for ATR and KDD R&D Lab. He visited MIT as a visiting scientist before joined in Nagoya University in 1995. He is a fellow of IEICE (the Institute of Electronics, Information and Communications Engineers) and a senior member of IEEE.

Prof. Takeda has served as an academic leader in various signal processing fields. Currently, he is a BoG (Board of Governors) member of IEEE ITS Society and AISIPA. He is also a vice president of Acoustical Society Japan. He served as a general chair of FAST-zero 2017, Universal Village 2016 and program chair of IEEE ICVES 2009, IEEE ITSC 2017 and other scientific meetings. He is also working as a director of a university startup company Tier IV.

HMMs for Path Generation

One of the central applications of the driving behavior analysis is reproducing the human driving under the given conditions, i.e., head distance, relative location/speed to the surrounding vehicles. Such driver models can be used for traffic simulation, vehicle safety evaluation and automotive driving. In terms of the controlling theory, predicting instantaneous acceleration of the vehicle in 2D space, i.e., $A = (\ddot{x}, \ddot{y})$is the main goal. On the other hands, predicting the trajectory $T = \{(x(t), y(t)) | 0 \le t \le t_0\}$in advance under the realistic driving scenario is an interesting challenge for autonomous driving applications. Particularly, reproducing the trajectory that is familiar to the driver is important.

Such 'trainable' autonomous driving can be built based on the stochastic framework of sequential signal. Particularly, this problem can be regarded as a parallel problem of speech synthesis and Hidden Markov model (HMM) can

be used as a stochastic model of the vehicle trajectory. In general, lane change (LC) driving can be subdivided into three states, preparation, transition and adjusting, therefore,three-state HMM is used for the modeling. As shown in Fig.2, first an HMM is trained to fit a set of LC trajectories, T, of a driver. Then,generating trajectories by sampling the trained HMM can be performed by setting state durations with random values. Generatedprobable LC trajectories arethen evaluated against to the given driving context through the probability density of the relative locations to the surrounding vehicles, which is also trained by the data (Safe Map).

Based on the experimental evaluation, it was confirmed that HMM driver model can generate the lane change trajectory reflecting the personal preference [4].

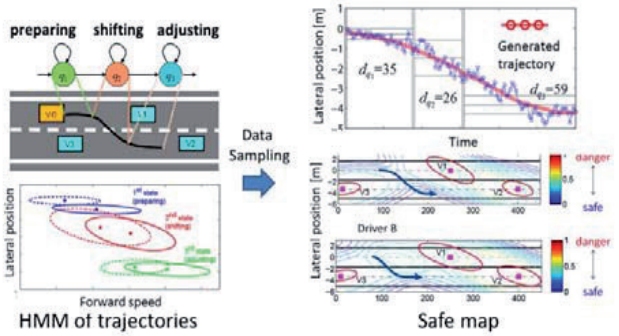


Fig. 2. Trainable algorithm for LC path generation.

Integrating Visual and Driving Behaviors

Visual behavior is closely related to the attentiveness while driving. In general, inattentive driver keeps looking at the road center only and distracted driver does not look at the road center at all. Proper visual behavior should be reflecting the surrounding traffic context, therefore, should be highly correlated with driving behavior. Analyzing the visual and driving behavior is also important research topic for detecting inattentive driving [5].

An experimental study was performed using the driving signals collected through 1000 lane changes under the real traffic. The subjective risk scores are given to all LC samples by 10 taggers. For the experiment, multi-stream HMM was used for modeling the correlations among signals corresponding to visual, driving and vehicle behaviors. Two HMMs were trained using 5% risky and 5% safety samples of the data and likelihood ratio of the test sample against to the two models were used to detect the risky lane changes.

As shown in Fig 3., by combining visual behaviors the detection accuracy of the risky LC can be improved by 15%, at 0.1-0.3 false alarm level [6].

Current Research Trends

Autonomous driving research is an emerging and rapidly growing research area and large body of them are using or trying to use signal processing and machine learning. IT giants' on-road tests bring huge amout of data including new signals such as 3D point cloud. The signal models discussed above will soon replaced with deep learning and the current performances are expected to become much higher.

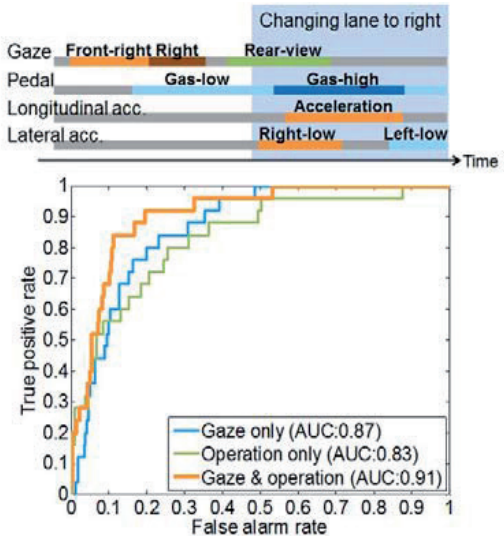


Fig 3. Integrating visual and driving behaviors for risky lane change detection.

Long-term Impact

Autonomous driving is becoming an industrial reality. However, most of technologies used in the current autonomous driving are algorithmic, with the exception of perception.In order to achieve autonomous driving systems with a driving behavior that is comfortable and optimal to the driver/society/vehicle under various conditions, the 'trainable' driving intelligence is inevitable. Driving behavior signal processing researches play a key role for building such flexible intelligence.

References

[1] Huseyin Abut, John H.L. Hansen, Kazuya Takeda, Gerhard Schmidt, Hanseok Ko (Eds.), "Vehicle Systems and Driver Modeling," De Gruyter, 2017.

[2] Kazuya Takeda, John Hansen, Pinar Boyraz, Lucas Malta, Chiyomi Miyajima and Huseyin Abut,"International Large-Scale Vehicle Corpora for Research on Driver Behavior on the Road," IEEE Transactions on Intelligent Transportation Systems, Vol.12 (4), pp. 1609-1623, 2012

[3] Chiyomi Miyajima and Kazuya Takeda, "Driver Behavior Modeling using on-road Driving Data," IEEE Signal Processing Magazine, VOl.33(6), 14-21, 2016

[4] Yoshihiro Nishiwaki, Chiyomi Miyajima, Norihide Kitaoka, Ryuta Terashima, Toshihiro Wakita, Kazuya Takeda, "Generating lane-change trajectories of individual drivers," Proc. of IEEE International Conference on Vehicle Electronics and Safety (ICVES 2008), 2008.

[5] Takatsugu Hirayama, Kenji Mase, Chiyomi Miyajima, Kazuya Takeda, "Classifications of driver's neutral and cognitive distraction states based on peripheral vehicle behavior in driver's gaze transition," IEEE Trans. on Intelligent Vehicle, Vol.1(2), 148-157, 2016

[6]Masataka Mori, Chiyomi Miyajima, Pongtep Angkititrakul, Takatsugu Hirayama, et al., "Measuring driver awareness based on correlation between gaze behavior and risks of surrounding vehicles," Proc. of IEEE International Conference on Intelligent Transportation Systems (ITSC2012), 2012

The ideal noise characteristics of the driving signal for driving OLED displays

Dr. Yuk-Hee Chan

PhD, MIEEE

Treasurer, APSIPA Headquarters

Associate Professor

Dept of Electronic and Information Engineering

The Hong Kong Polytechnic University

YUK-HEE CHAN received his BSc degree in electronics from The Chinese University of Hong Kong in 1987, and his PhD degree in signal processing from The Hong Kong Polytechnic University in 1992. Between 1987 and 1989, he worked as an R&D engineer at Elec & Eltek Group, Hong Kong. He joined The Hong Kong Polytechnic University in 1992. Dr. Chan has published over 150 research papers in various international journals and conferences. His research interests include image and video compression, image restoration, halftoning and fast computational algorithms in DSP. Dr. Chan was the Chair of IEEE Hong Kong Section in 2015 and is the Treasurer of Asia-Pacific Signal and Information Processing Association (APSIPA) Headquarters since it was founded.

Organic light-emitting diode (OLED) is currently the most advanced display technology [1]. It has obvious advantages over the LCD technology due to its better power efficiency, faster response time, wider viewing angles, improved brightness, lighter weight, more flexible substrates and lower cost [2].

There are different driving methods for driving OLED displays. Currently, most of the digital driving methods are based on pulse width modulation (PWM). The fast response time of OLED allows us to consider some other modulation schemes such as delta-sigma modulation (DSM). As compared with PWM, DSM helps to remove the dynamic false contour, reduce the harmonic tones of the frame frequency and release the gate scan time constraint to make a larger display realizable [3]. In general, these conventional digital driving methods share a common tactic that pixels are modulated independently. The spatial correlation among pixels is not considered.

No matter PWM or DSM is exploited to drive a full-color OLED display, in practice the OLED display behaves as a binary video display that plays a time sequence of oversampled binary image frames to render a gray-level image/video frame as shown in Fig. 1. The sequence of binary frames actually defines the signal used to drive the display. Obviously, its noise characteristics determine the visual quality of the displayed content. Then we have two questions as follows: 1. What are the ideal noise characteristics of this binary video sequence (i.e. the driving signal)? 2. How to generate the ideal binary video sequence effectively?

In theory, our human visual system (HVS) behaves as a low pass filter that can remove the high frequency noise of the displayed content to improve its perceptual quality. Hence, from visual quality point of view, each binary frame of the ideal driving signal should be a good binary representation of the original gray-level image and contains only high frequency spatial and temporal noise.

Figs. 2 and 3 show, respectively, the different binary frames for rendering an 8-bit gray level image 'Lena' when PWM [4] and DSM [3] are separately used. (Fig.3 shows the case when the over sampling rate (OSR) is 13.) In both cases, most of the binary frames are not good binary representation of the original image and all binary frames have strong low frequency noise. From that perspective, none of them is ideal for driving an OLED display.

Halftoning is a process that turns a gray-level image into its binary representation (a.k.a. halftone). Ulichney proposed a noise model that describes the ideal noise characteristics of a halftone [5]. Over the years, two better noise models have been developed [6,7]. These models can serve as guidelines for one to judge the visual quality of a halftone and to develop advanced halftoning algorithms. However, though many halftoning algorithms aim at producing halftones having the noise characteristics specified by the noise models, only a few of them (e.g. [7-11]) can flexibly control the noise characteristics and exactly achieve the goal. The tone-dependent error diffusion algorithm recently proposed in [11] is a low complexity solution among those few of them, but it is still not fast enough to produce binary frames continuously for driving an OLED display at video rate. Besides, halftoning focuses on how to generate a good binary frame instead of a binary frame sequence for one gray-level image/video frame. In other words, it does not take the inter-frame correlation of the binary frame sequence into account.

A recent modulation proposal referred to as STSDSM [12] takes both the inter- and intra-frame correlation of the binary frames into account and successfully shifts the noise to the high frequency bands in both spatial and temporal domains. Fig. 4 shows the ANPS performance of different modulation schemes. Each ANPS plot shows the average of the noise power spectra of individual binary frames of the binary sequence produced with a particular modulation scheme for testing video sequence 'Akiyo'. One can see that the overall noise level and the low frequency noise level of STSDSM are both low. As compared with the

pixel-oriented DSM [3], STSDSM can improve the visual quality of the played video content, solve the idle tone and flickering problems for low gray level constant inputs, achieve a higher effective bit resolution (3 more bits) and further release the gate scan time constraint. The performance can be achieved at a complexity cost of 3 shift-additions per pixel per frame and a pipeline realization is feasible to reduce the buffer size and the hardware cost.

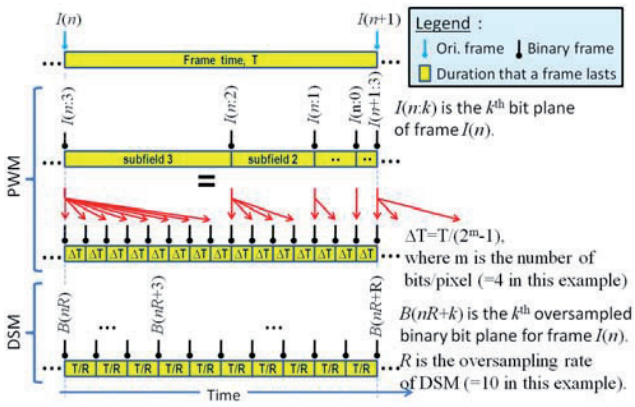


Fig.1: A simplified example showing how an OLED display plays a 4-bit gray level video with PWM and DSM. The oversampling rate exploited in DSM is 10.

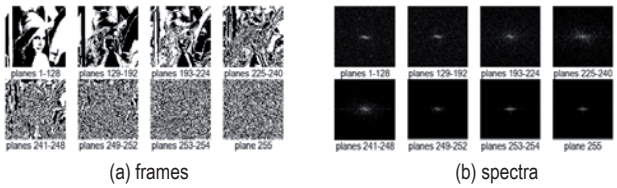


Fig.2: Individual binary planes produced by pixel-oriented PWM [4] to display image 'Lena' and their noise spectra.

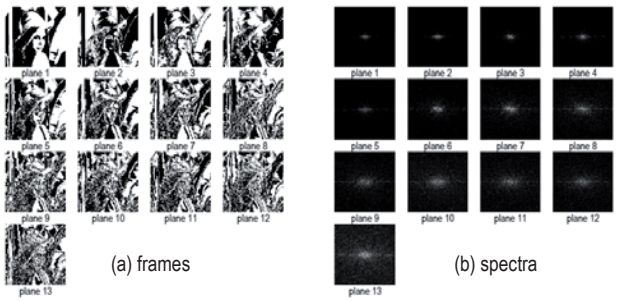


Fig.3: Individual binary planes produced by pixel-oriented DSM [3] to display image 'Lena' and their noise spectra

Long-term Impact

It is expected that in the near future OLED displays will be widely used in various products in various domains including education, entertainment and communication. By considering the huge production volume, a cost-effective driving module that can boost the display quality is always welcome and appreciated by manufacturers and users. The aforementioned progress in this research area creates a new direction for us to rethink the design of the driving module. It is expected that in the coming 20 years more and more advanced driving techniques will be developed along this direction and applied in display units of different grades.

References

[1] J. Chen, W. Cranton and M. Fihn (Ed.), Handbook of Visual Display Technology, Springer, Bristol UK, 2012.

[2] T. Tsujimura, OLED Display Fundamentals and Applications, Wiley, June 2012.

[3] J.H. Jang, M. Kwon, K. Lee and B. Jung, "A PDM-Based Digital Driving Technique Using Delta-Sigma Modulation for QVGA Full-Color AMOLED Display Applications," Journal of Display Technology, 6(7), July 2010, pp.269-278

[4] H. Murakami and R. Toyonaga, "A Pulse Discharge Panel Display for Producing a Color TV Picture with High Luminance and Luminous Efficiency," IEEE Trans. Electron Devices, 29, June 1982, pp. 988-994

[5] R. A. Ulichney, "Dithering with blue noise," Proc. IEEE, vol. 76, pp. 56–79, Jan. 1988.

[6] D. L. Lau and R. A. Ulichney, "Blue-Noise Halftoning for Hexagonal Grids," IEEE Trans. Image Process, vol. 5, no. 5, pp. 1270-1284, May, 2006.

[7] Y. H. Fung and Y.H. Chan, "Tone-dependent noise model for high-quality halftones," Journal of Electronic Imaging, 22 (2), 023004, Apr 12, 2013.

[8]. Y.H. Fung and Y.H. Chan, "Optimizing the error diffusion filter for blue noise halftoning with multiscale error diffusion," IEEE Transaction of Image Processing, 2012 Vol. 22, No.1, Jan 2013, pp.413-417

[9] Y.H. Fung, K.C. Lui and Y.H. Chan, "low-complexity high-performance multiscale error diffusion technique for digital halftoning," Journal of Electronic Imaging, 16 (1), 013010, Jan- Mar 2007.

[10] Y.H. Fung and Y.H. Chan, "Green Noise Digital Halftoning with Multiscale Error Diffusion," IEEE trans. on Image Processing, Vol.19, Jul 2010, pp.1808-1823

[11] Y.H. Fung and Y.H. Chan, "Tone dependent error diffusion based on an updated blue noise model," Journal of Electronic Imaging, 25(1), 013013, Jan 25, 2016.

[12] Y.H. Fung and Y.H. Chan, "Shaping the spatial and temporal noise characteristics of driving signals for driving AMOLED Display," IEEE Journal of Display Technology, Vol. 12, Dec 2016, pp.1652-1663

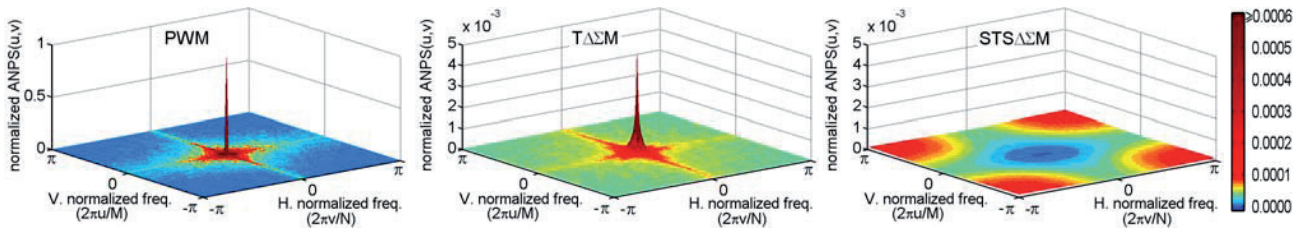


Fig.4: The $ANPS(u,v)$ plots of the driving signals generated for the testing video 'Akiyo' by (a) PWM [4], (b) pixel-oriented DSM [3] and (c) STSDSM [12]. The magnitudes of the components in all plots are normalized with respect to the strongest component in the 3 plots for easier comparison. The color bar shows the shared color mapping scheme used to map a range of normalized magnitude values to a color in the plots.

Audio Watermarking for Media Copyrights Protection

Introduction

On-line distribution of digital multimedia including audio, video, images and documents has proliferated rapidly in recent years. Along with the easy access to duplicating and editing, ownership infringement claims more and more attention. As a result, techniques have to be developed for copyright protection [1]. Different from traditional steganography and cryptography, watermarking systems do not restrict the right to access, but embed one or more watermarks with specific meanings into the host media as permanent signs [2]. When proprietorial disputes happen, the watermark(s) could be extracted as reliable proofs for assuring the authenticity.

Compared with images and video, inserting watermark(s) into digital audio files presents special challenge, since the human auditory system (HAS) is much more sensitive than the human visual system (HVS) and consequently the room for embedding information is limited [3]. In this short article, we introduce the fundamentals and framework of audio watermarking.

Audio watermarking is a technique providing a promising solution to copyrights protection for digital audio and multimedia products. Using this technique, hidden information, called *watermark*, containing copyrights information is imperceptibly embedded into the audio track of a host media. This watermark can be extracted later on from a suspected media to verify the authenticity.

Waleed H. Abdulla

PhD

Former VP-MRD and BoG Member, APSIPA

Associate Professor

Electrical, Computer, and Software Engineering

The University of Auckland
New Zealand

Waleed Abdulla received his PhD Degree from Otago University in New Zealand in 2002. He served as Vice President- Member Relations and Development in APSIPA for two terms followed by 2-year Board of Governors. He is one of the steering committee members who established APSIPA in 2009. He is APSIAP Newsletter founder and was Editor-in-Chief. He visited and delivered talks in several universities and conferences as presenter and keynote speaker. He has been serving as editorial board member of 5 journals. He supervised over 30 PhD and Master Degrees students. He was awarded APSIPA Distinguished Lecturer Award, Faculty Best Teachers of the Year in 2005 and 2012. His research is in "Signal Processing, Analysis, and Recognition", which encompasses multidisciplinary topics. He focuses on fundamental and applied research in domains with direct communal relevance including, human health and well-being, Human Biometrics, Big Data, and economic impact. His papers have been cited over 1300 times. He won two best paper awards in 2012 and 2016 in two major conferences. He co-authored a book in Audio Watermarking, which has been downloaded over 7500 times. It includes many topics in psychology of hearing and audio processing.

Audio watermark systems requirements

Audio watermarking must satisfy certain attributes to be effective in securing ownerships. These are imperceptibility, robustness, security, and data payload requirements. In the following points we shade lights on these attributes:

- **Imperceptibility** is an essential attribute. The process of audio watermarking is considered to be imperceptible or transparent if no differences between the host and watermarked signals are perceivable. In a nonprofessional term, watermarked signal must appear as if there is nothing added to the host media. To preserve the perceptual quality of the watermarked data, a psychoacoustic model derived from the auditory masking phenomenon is adopted to deceive the human perception [3].
- **Robustness** is a measure of reliability and refers to the capability of extracting the watermark from the attacked watermarked signal. Examples of attacks on audio watermarking include noise addition, resampling, compression, random samples cropping, time and pitch scaling, and more.
- **Security** is a prerequisite. It must be guaranteed that the watermarks cannot be extracted from the watermarked signal by reversing the embedding process or performing statistical detection [4, 5]. Secret keys (usually pseudorandom sequences) and/or scrambling operations can be adopted to add randomness into the embedding and detection processes, so that the digital watermarking system is self-secured.
- **Data payload** refers the number of bits that can be embedded in one-second audio fraction, expressed in bit per second (bit/s or bps) [6]. Watermarking system varies greatly, depending on the embedded parameters and the embedding algorithm. Audio watermarking

normally does not have high data payload, only 2~ 4 bits/s on average [7].

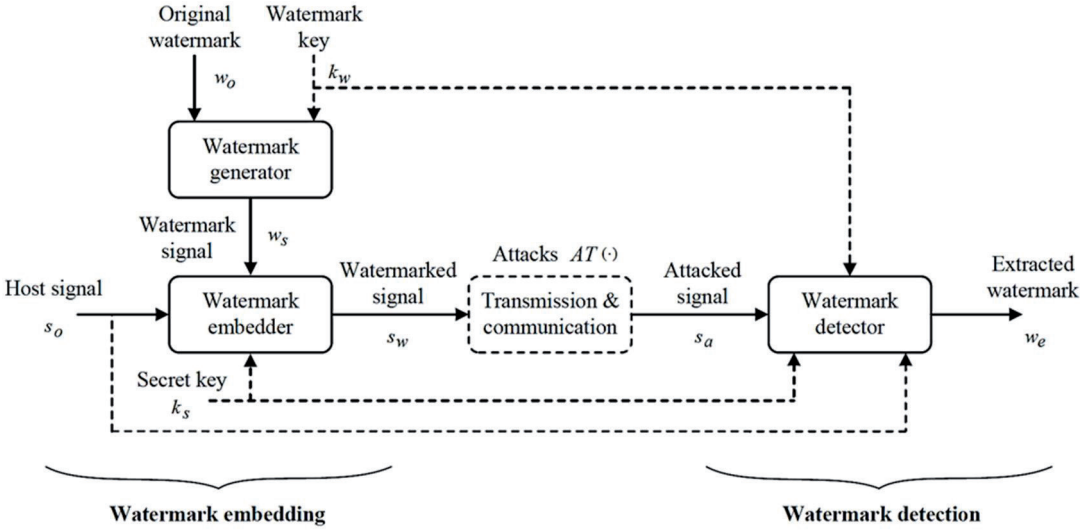


Figure 1. General block diagram of watermark

In practice, not a single system can fully satisfy all the requirements and some tradeoffs always exist among criteria. Typically, an audio watermarking system can operate with either excellent imperceptibility or strong robustness, but not both. In order to ensure the robustness, we embed the watermark(s) into perceptually important regions or increase the strength of the watermarking. However, such strategies are liable to cause perceivable distortion to the host signal, which is against the property of imperceptibility. Moreover, both of them are in close connection with data payload. If we embed more bits into an audio signal, the imperceptibility would become worse and the robustness would be stronger [8]. Similar compromises also occur between imperceptibility, robustness, and security.

Framework of the Watermarking Systems

A digital watermarking system consists of three fundamental parts, namely a watermark generator, an embedder, and a detector, as illustrated in Figure 1. The digital media to be protected is called host signal, s_0, where the original watermark, w_0, is embedded. The original watermark is diverse; possibly an image, a sequence of letters, or a simple series of bits. w_0 can be scrambled by a security key k_w to generate a more secure watermark w_s. Then, the watermark embedder incorporates the watermark signal into the host signal. An extra secret key k_s can be employed to provide further security to the generated watermarked signal s_w. The embedding process is mathematically described as follows:

$s_w = Embedding\ (s_0, w_0, k_w, k_s)$,

where s_w is perceptually similar to s_0. Then the watermarked signal s_w is ready for communication. s_w is likely to be modified, attacked, by either being processed by coding or tampered with by malicious attempts such as noise addition, MP3 compression, random samples cropping, etc. to remove the watermark.

In the detection, the watermark detector extracts the watermark from the received signal. The input to the watermark detector is called the attacked signal, s_a, which could be an identical or distorted version of s_w.

The detection process is defined by:

$w_e = Detection(s_a, s_0, k_w, k_s)$,

where w_e is the extracted watermark. For robust systems w_e must be very similar, within acceptable threshold, to w_0.

Human auditory system is very sensitive and it can sense any distortion in sound, in contrary to the human visual system that can easily be deceived. The intuitive question if the human auditory system is so sensitive, why we can't hear the distorted signal by the watermarking signal? To understand this we have to look at one important property of HAS.

Human auditory system has certain properties that we can invest to embed imperceptible signals in the host signal. There is a certain threshold, indicated by the dotted line shown in Figure 2, where we cannot hear the signal below it. This threshold is changeable and can be modified by loud signals, such as S_0 called masker. This masker can increase the threshold to become higher than that of the quiet environment and it can mask any signal below it. S_1 and S_2 are completely masked by S_0, thus we can embed the watermark signal below this threshold without noticeable distortion. We can track this moderation of the masking threshold by using psychoacoustic modelling [1].

Benchmarking on Audio Watermarking

There is a necessity for benchmarking various algorithms is imperative. Imperceptibility, robustness, and security are key principles in designing any audio watermarking scheme [9][10]. Perceptual quality assessment is usually classified into two categories: subjective listening tests by human acoustic perception and objective evaluation tests by perception modelling or quality measures. In subjective

listening tests, the subjects are asked to discern the watermarked and host audio clips.

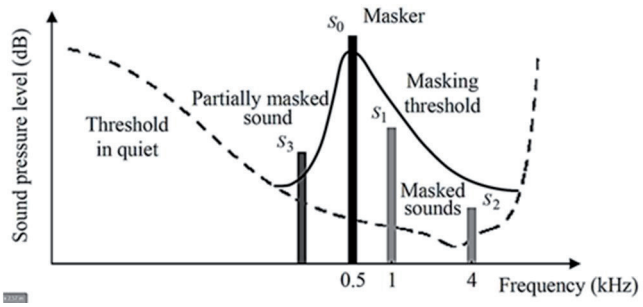


Figure 2. General block diagram of watermark

However, such audibility tests are not only costly and time-consuming, but also heavily depend on the subjects and surrounding conditions. Therefore, the industry desires the use of objective evaluation tests to achieve automatic perceptual measurement. Currently, the most commonly used objective evaluation is perception modelling, i.e., assessing the perceptual quality of audio data via a stimulant ear, such as Evaluation of Audio Quality (EAQUAL) [11], Perceptual Evaluation of Audio Quality (PEAQ) [12], and Perceptual Model-Quality Assessment (PEMO-Q) [13].

The goal of the robustness test is to test the ability of a watermarking system resistant to signal modifications in real applications. In the robustness test, various attacks are applied to the watermarked signal and produce a number of attacked signals. Then, watermark detection is performed on each attacked signal to check whether the embedded watermark survives or not. Examples of attacks are: noise addition, resampling, requantization, amplitude scaling, low-pass filtering, echo addition, reverberation, MP3 compression, DA/AD conversion, random samples cropping, jittering, zeros inserting, time-scale modification and pitch-scale modification [14]

Security analysis is performed to evaluate the characteristics of security for audio watermarking systems. Since security is attributed to the randomness merged by sequences of pseudorandom numbers (PRN) and/or scrambling operations, an intuitive method of security analysis is to calculate the number of possible embedding ways. If there are more possible ways of embedding, it would be difficult for unauthorized detection to ascertain the embedded watermark. This indicates that the system has a high level of security [15].

What is next?

Well, we believe that most valuable digital media are currently watermarked and this will grow in future in coherence with the growing use of digital media. Music and movie industries are very keen to secure their copyrights on their products through suitable watermarking techniques. The growing market of audio books will also be adopting such technology. This technology would save media industries billions of dollars through preventing or at least limiting illegal copy proliferation.

References

[1] Y.Q. Lin, W.H. Abdulla, Audio Watermark: A Comprehensive Foundation Using MATLAB, Springer (2015).

[2] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, Techniques for Data Hiding, IBM Systems Journal, vol.35, no. 3&4, pp. 313-336 (1996).

[3] L. Boney, A.H. Tewfik, K.N. Hamdy, Digital watermarks for audio signals, in *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, pp. 473–480 (1996).

[4] C.-P. Wu, P.-C. Su, C.-C.J. Kuo, Robust and efficient digital audio watermarking using audio content analysis, in *Proceedings of SPIE Security and Watermarking of Multimedia Contents II*, vol. 3971, pp. 382–392 (2000)

[5] W.-N. Lie, L.-C. Chang, Robust and high-quality time-domain audio watermarking subject to psychoacoustic masking, in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 45–48 (2001)

[6] I.J. Cox, M.L. Miller, J.A. Bloom, J. Fridrich, T. Kalker, *Digital Watermarking and Steganography,* Morgan Kaufmann Publishers, San Francisco (2008)

[7] S.J. Xiang, J.W. Huang, Histogram-based audio watermarking against time-scale modification and cropping attacks. IEEE Trans. Multimedia. 9(7), 1357–1372 (2007)

[8] W. Bender, D. Gruhl, N. Morimoto, A. Lu, Techniques for data hiding. IBM Syst. J. 35(3 & 4), 313–336 (1996)

[9] Y.Q. Lin, W.H. Abdulla, Perceptual evaluation of audio watermarking using objective quality measures, in *Proceedings of ICASSP*, pp. 1745–1748 (2008)

[10] Y. Lin, W. Abdulla, Objective quality measures for perceptual evaluation in digital audio watermarking. IET - Signal Process. 5(7), 623–631 (2011)

[11] A. Lerch, Software: EAQUAL - Evaluation of Audio Quality, v.0.1.3alpha ed. (2002)

[12] P. Kabal, An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality. Technical Report, TSP Lab, McGill University (2003)

[13] R. Huber, B. Kollmeier, PEMO-Q: a new method for objective audio quality assessment using a model of auditory perception. IEEE Trans. Audio Speech Lang. Process. 14(6), 1902–1911 (2006)

[14] Y.Q. Lin, W.H. Abdulla, Y. Ma, Audio watermarking detection resistant to time and pitch scale modification, in *Proceedings of IEEE International Conference on Signal Processing and Communications (ICSPC)*, pp. 1379–1382 (2007)

[15] Y.Q. Lin, W.H. Abdulla, A secure and robust audio watermarking scheme using multiple scrambling and adaptive synchronization, in *Proceedings of the 6th International Conference on Information, Communications and Signal Processing (ICICS)* (2007)

Sparsity Aware Adaptation – a New Challenge in Today's Adaptive Filters

Background

In practice, one often encounters systems that have a very long impulse response (IR), where most of the IR coefficients are either zero or near zero which are called inactive taps, and only few coefficients, usually present in cluster(s), have significant magnitudes which are called active taps. Figure 1 shows an example of a typical sparse impulse response. It has one cluster of active coefficients while the remaining part of the impulse response has zero or near zero values. One example of sparse systems is given by multipath cellular communication channels which, as shown in Fig. 2, have various alternate paths to reach the target with different delay profiles. Another example is the underwater acoustic channel as shown in Fig. 3, where acoustic waves from the underwater transmitter reach the target after getting reflected from the seabed as well from the water surface (total internal reflection) following various paths with different path delays. Others examples include room acoustic impulse response, echo paths in voice and data networks, HDTV channels etc.

Adaptive filters have been traditionally used to identify and track unknown systems. Most of the well known adaptive filters developed so far like the LMS, NLMS, affine projection, RLS etc. are, however, ``sparsity unaware", as they do not have any mechanism to incorporate the a priori information about the sparse nature of the system in the body of the algorithm. It has, however, been known that appropriate utilization of the a priori sparseness information can improve the performance of the adaptive filter substantially. For example, it may lead to much lesser complexity, faster convergence rate, lesser estimation error, better tracking capability etc. This has prompted a flurry of research activities in last decade or so to develop sparsity aware adaptive filters (also called sparse adaptive filters).

One of the most prominent class of adaptive filters in this category is the proportionate-NLMS (PNLMS) [1] and its family. In PNLMS algorithm, each coefficient is updated with an independent step size that is made proportional to the magnitude of the particular filter coefficient estimate, resulting in a combination of faster initial convergence (caused by active coefficients) and lesser steady state excess mean square error (EMSE) (caused by inactive coefficients) for sparse systems. The rate of convergence, however, slows down afterwards, being largely controlled by the inactive taps that outnumber the active coefficients, sometimes resulting in overall slower convergence than the NLMS algorithm. This problem, has been addressed somewhat in [2]-[4].

Prof. Mrityunjoy Chakraborty

PhD

BoG Member, APSIPA (2013-16)

Mrityunjoy Chakraborty obtained Bachelor of Engg. from Jadavpur university, Calcutta, Master of Technology from IIT, Kanpur and Ph.D. from IIT, Delhi. He joined IIT, Kharagpur as a faculty member in 1994, where he currently holds the position of a professor in Electronics and Electrical Communication Engg. The teaching and research interests of Prof. Chakraborty are in Digital and Adaptive Signal Processing, VLSI Signal Processing, Linear Algebra and Compressive Sensing.

Prof. Chakraborty is currently a senior editorial board member of the IEEE Signal Processing Magazine and also of the IEEE journal of Emerging Techniques in Circuits and Systems. Earlier, he had been an Associate Editor of the IEEE Transactions on Circuits and Systems, part I (2004-2007, 2010-2012) and part II (2008-2009), chair of the DSP Technical Committee (TC) of the IEEE Circuits and Systems Society, a guest editor of the EURASIP JASP and of special issues of TCAS-II, track co-chair (DSP track) of ISCAS 2015-2018, TPC co-chair of IEEE SIPS-2018, Special Session Co-Chair of DSP-18, Gabor track chair of DSP-15, and a TPC member of ISCAS (2011-2014), ICC (2007-2011) and Globecom (2008-2011). He is a co-founder of the APSIPA and also has been the founding chair of the APSIPA TC on Signal and Information Processing Theory and Methods (SIPTM).

Prof. Chakraborty is a fellow of the National Academy of Science, India, and also of the Indian National Academy of Engineering (INAE). During 2012-2013, he was selected as a distinguished lecturer of the APSIPA.

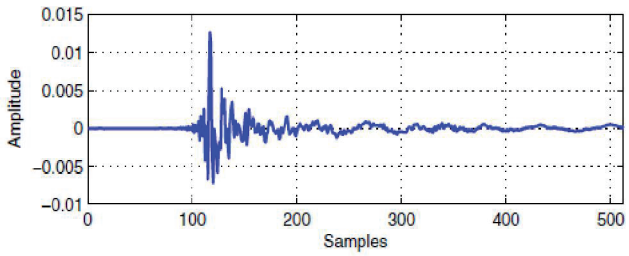


Figure 1. An example of a sparse impulse response

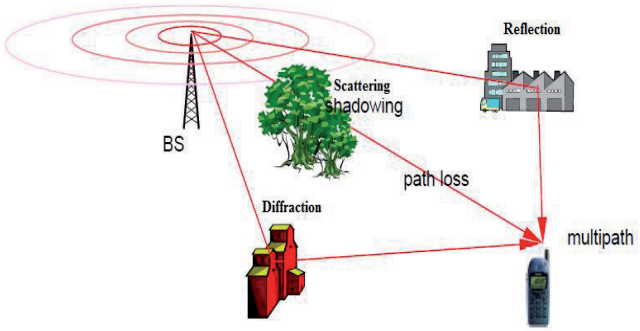


Figure 2. Wireless multipath channel – one example of a sparse system.

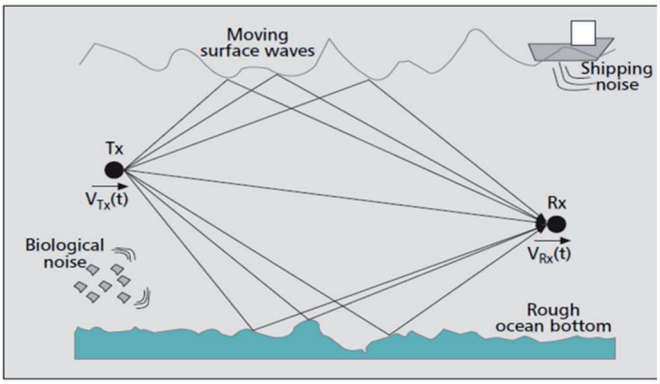


Figure 3. Wireless multipath channel – one example of a sparse system.

Separately, emergence of compressed sensing has given a new dynamism to the topic of sparse adaptive filters in last few years. In [5], Gu et al developed a new sparse adaptive filter by introducing a l_1 norm penalty (of the filter weight vector) in the LMS cost function. Minimization of the cost function introduces certain zero-attracting terms in the weight update equation which pulls the coefficients to zero value. The resulting algorithm, called zero-attracting LMS (ZA-LMS) enjoys both superior convergence rate and lesser steady state EMSE as compared to standard LMS, for highly sparse systems. The ZA-LMS has influenced design of sparse adaptive filters in last few years in a big way and several algorithms based on this came up, including its extension to NLMS, RLS etc. The review article [6] provides a coverage of the major developments in this area.

Robustness against time-varying sparsity: Often one comes across sparse systems that show time-varying sparsity, with the number of inactive taps varying from very large (highly sparse) to very few (highly non-sparse). The task then boils down to not only identifying the system, but also to track its time variation. Unfortunately, conventional sparse adaptive filters are not very robust against variable sparsity. For example, the PNLMS convergence slows down enormously as the system changes from being highly sparse to highly non-sparse. In the case of zero-attracting family, a reweighted ZA-LMS (RZA-LMS) was proposed in [5] that restricts the shrinkage to inactive taps only. Parameter selection for maintaining such shrinkage is, however, a tricky issue in the RZA-LMS algorithm especially for systems that have time varying sparseness with the active taps taking values over a wide range.

In [7], we have presented an elegant solution to the above problem, where we convexly combine the outputs of two adaptive filters – one sparsity aware (ZA-LMS) with coefficient vector $\mathbf{w}_1(n)$, producing output $y_1(n)$ and the other sparsity unaware (LMS) with coefficient vector $\mathbf{w}_2(n)$, producing output $y_2(n)$. Both the filters try to identify the same N×1 system impulse reponse vector $\mathbf{w}_0$ (supposed to be sparse with variable sparseness) by processing the system input x(n) and taking the system output $y_d(n) = \mathbf{w}_0^t\mathbf{x}(n)+v(n)$ as the desired response, where $\mathbf{x}(n) = [x(n), x(n-1), \ldots, x(n-N+1)]^t$ and v(n) is an observation noise idependent of x(n). The final output y(n) is formed by convexly combing $y_1(n)$ and $y_2(n)$ as $y(n) = \lambda(n)\ y_1(n)+(1-\lambda(n))\ y_2(n)$, where $\lambda(n)$ is a mixing parameter ($0 \le \lambda(n) \le 1$), which is updated in time by a gradient descent search on $e^2(n)$ where $e(n) = y_d(n) - y(n)$ is the combined filter output error. However, direct adaptation does not guarantee that $\lambda(n)$ will lie between 0 and 1. For this, $\lambda(n)$ is expressed monotonically in terms of another variable a(n) which is free to take any value from $-\infty$ to $+\infty$, and the above gradient descent search is carried out with respect to a(n). The proposed combination which follows the general adaptive convex combination philosophy of [8] is shown in Fig. 4 below.

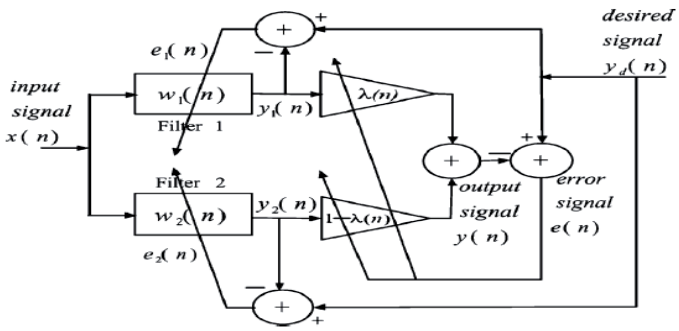


Figure 4. An adaptive convex combination of a sparsity aware (Filter1 : ZA-LMS) and a sparsity-unaware (Filter2 : LMS) adaptive filter.

A detailed convergence analysis, carried out by us in [7], shows that when the system is highly sparse, $\lambda(n) \to 1$, while, for a highly non-sparse system, $\lambda(n) \to 0$, meaning, in the case of former, the combination acts mainly like a ZA-LMS based adaptive filter, while for the latter, it acts more like a LMS-based adaptive filter. More interestingly, the analysis shows that for a wide range of system sparsity, from sufficiently non-sparse to sufficiently sparse, $\lambda(n)$

takes some intermediate values between 0 and 1, such that the overall combination acts better than both its constituent adaptive filters (i.e., LMS and ZA-LMS). The above conjectures were verified later by extensive simulation studies. In [7], the treatment was restricted to the basic form of adaptive filters like LMS and ZA-LMS. This was later extend to NLMS and ZA-NLMS in [9] and to affine projection algorithm (APA) and ZA-APA in [10].

Separately, in [11], we have addressed the well known problem of slowing down of the convergence of PNLMS algorithm, by developing a ZA-PNLMS algorithm. For this, we introduced a carefully constructed l_1 norm (of the coefficients) penalty in the PNLMS cost function which favours sparsity. The resulting zero attractors in the PNLMS weight update equation help in the shrinkage of the coefficients, especially the inactive taps, thereby arresting the slowing down of convergence and also producing lesser steady state excess mean square error (EMSE). A rigorous convergence analysis of the proposed algorithm is presented using the approach of [12] that expresses the steady state mean square deviation of the filter coefficients in terms of a zero attracting coefficient of the algorithm. The analysis reveals that further reduction of the EMSE is possible by means of a variable step size (VSS) simultaneously with a variable zero attracting coefficient in the weight update process. Simulation results confirmed that the proposed algorithm enjoys superior performance vis-à-vis existing sparse adaptive filters.

Lastly, we have recently developed a new sparse RLS-based adaptive filter [13], which employs a novel approximation of the l_0 norm of the filter coefficient vector for regularising the RLS cost function. The proposed algorithm overcomes some of the shortcomings of the existing algorithms as demonstrated via numerical simulations.

References

[1]. D.L. Duttweiler, ``Proportionate normalized least-mean-squares adaptation in echo cancelers", *IEEE Trans. Speech Audio Process*, vol. 8, no. 5, pp. 508-518, September 2000.

[2]. S.L. Gay, ``An efficient, fast converging adaptive filter for network echo cancellation", *Proc. Asilomar Conf. Signals, Systems, Comput.*, pp. 394-398, Nov., 1998.

[3]. J. Benesty and S.L. Gay,``An improved PNLMS algorithm", *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 1881-1884, 2002, Orlando, Florida, USA.

[4]. H. Deng and M. Doroslovacki, ``Improving convergence of the PNLMS algorithm for sparse impulse response identification", *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 181-184, 2005.

[5]. Y. Gu, Y. Chen and A.O. Hero,``Sparse LMS for system identification" , *Proc. IEEE ICASSP-2009*, pp. 3125-3128, Apr. 2009, Taipei, Taiwan.

[6]. R. L. Das and M. Chakraborty, ``Sparse Adaptive Filters - an Overview and Some New Results", Proc. ISCAS-2012, May, 2012, Seoul, Korea.

[7]. B. K. Das and M. Chakraborty, ``Sparse Adaptive Filtering by an Adaptive Convex Combination of the LMS and the ZA-LMS Algorithms", *IEEE Transactions on Circuits and Systems, Part I*, pp. 1499-1507, May, 2014.

[8]. J. Arenas-Garcia, A. R. Figueiras-Vidal, and A. H. Sayed, ``Mean-square Performance of a Convex Combination of Two Adaptive Filters", *IEEE Trans. Signal Process*., pp. 1078-1090, March, 2006.

[9]. B. K. Das, V. Chakravarthi and M. Chakraborty, ``A Convex Combination of NLMS and ZA-NLMS for Identifying Systems with Variable Sparsity", *IEEE Transactions on Circuits and Systems, Part II*, pp. 1112-1116, Sept., 2017.

[10]. V. Chakravarthi Gogineni, B. K. Das and M. Chakraborty, ``An adaptive convex combination of APA and ZA-APA for identifying systems having variable sparsity and correlated input", *Digital Signal Processing, Elsevier*, vol. 82, pp. 118-132, Nov., 2018 (to appear).

[11]. R. L. Das and M. Chakraborty, ``Improving the Convergence of the PNLMS Algorithm via *L1* Norm Regularization", *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1280-1290, July, 2016.

[12]. R. L. Das and M. Chakraborty, ``On Convergence of Proportionate-Type Normalized Least Mean Square Algorithms", *IEEE Transactions on Circuits and Systems, Part II*, pp. 491-495, May, 2015.

[13]. B. K. Das and M. Chakraborty, `` Improved l_0 -RLS Adaptive Filter", *Electronic Letters,* pp. 1650-1651, Dec., 2017.

Photo Gallery: APSIPA ASC'2012 in Hollywood

Artificial Intelligence Paradigms and Algorithms

Dr. Li Deng

Fellow of the IEEE

Former VP and member of Board of Governors, APSIPA

Chief Artificial Intelligence Officer, AI Research
Citadel America,
Seattle WA USA

Dr. Li Deng has been the Chief Artificial Intelligence Officer of Citadel since May 2017. Prior to Citadel, he was the Chief Scientist of AI, the founder of Deep Learning Technology Center, and Partner Research Manager at Microsoft. Prior to Microsoft, he was a tenured full professor at the University of Waterloo in Ontario, Canada as well as holding teaching and research positions at MIT (Cambridge), ATR (Kyoto, Japan), and HKUST (Hong Kong). He is a Fellow of the IEEE (since 2004), a Fellow of the Acoustical Society of America (since 1993), and a Fellow of the ISCA (since 2011). He has also been an Affiliate Professor at University of Washington, Seattle (since 2000). He was Editors-in-Chief of IEEE Signal Processing Magazine and of IEEE/ACM Transactions on Audio, Speech, and Language Processing. In recognition of the pioneering work on disrupting speech recognition industry using large-scale deep learning, he received the 2015 IEEE SPS Technical Achievement Award for "Outstanding Contributions to Automatic Speech Recognition and Deep Learning". He is an author or co-author of six technical books on deep learning, speech processing, pattern recognition and machine learning, and, the latest, natural language processing (Springer, June 2018).

Artificial Intelligence (AI) is a branch of computer science and a technology aimed at developing the theories, methods, algorithms, and applications for simulating and extending human intelligence. AI enables us to go from an old world where people give computers rules to solve problems to a new world where people give computers problems directly and the machines learn how to solve them on their own using a set of algorithms. An algorithm is a self-contained sequence of instructions and actions to be performed by a computational machine. AI consists of a rich set of algorithms used to perform AI tasks, notably those involving learning from data and experiences.

First Rise of AI --- Expert Systems

In the first rising wave of AI, starting in 70's and based on expert knowledge engineering, domain experts devised computer programs according to the knowledge about the (very narrow) application domains they have. The typical approach during this wave is exemplified by the expert system, a computer system that emulates the decision-making ability of a human expert. Such systems were designed to solve complex problems by reasoning about knowledge. The main "algorithm" used was the inference rules in the form of "if-then-else**".**

The early speech recognition research and system design, a long-standing AI challenge in machine perception, were based on the AI paradigm of expert knowledge engineering during the first rising wave of AI. During 70's and early 80's, the expert-system approach to speech recognition was quite popular. However, the lack of abilities to learn from data and to handle uncertainty in reasoning was acutely recognized by researchers contributing to the second rise of AI in late 80's. The author was part of the research community then and contributed to the transition from knowledge-based speech recognition to data-driven one [24-30] powered by machine learning methods described next.

Second Rise of AI --- Data-driven (Shallow) Learning

The second rising wave of AI arrived in 1980's (for speech, and somewhat later for other AI areas) after clear evidence that learning and perception capabilities are crucial for complex AI systems but missing in expert systems [6]. This is not just for speech recognition as discussed above, but also for vision and other AI systems. Much like speech recognition, the autonomous driving and vision researchers immediately realized the limitation of the first-generation AI paradigm due to the need for machine learning with uncertainty and with generalization capabilities.

This second-generation AI paradigm was based on machine learning, which we now call "shallow" due to the lack of abstractions constructed by many-layer or "deep" representations of data which would come in the third rise of AI.

In speech recognition, over more than 20 years from 1980's to 2010, the AI paradigm had been completely switched from earlier expert systems to the (shallow) machine learning paradigm using a statistical generative model called the Hidden Markov Model (HMM) integrated with Gaussian mixture models, along with various versions of its generalization [4,5,20,22]. The main algorithms and methods include Viterbi algorithm [50], Baum-Welch algorithm (which is a special case of EM when applied to HMM), and extended Baum-Welch (which includes learning algorithms for maximizing mutual information, minimizing classification errors, and minimizing phone errors) [9,36]. Among many versions of the generalized HMMs were statistical and neural-net based hidden dynamic models [7,19-21]. The former adopted EM and extended Kalman filter algorithms for learning model parameters [23,45], and the latter used backpropagation [49]. Both made extensive use of multiple latent layers of representations for the generative process of speech

waveforms following the long-standing framework of analysis-by-synthesis in human speech perception. More significantly, inverting this "deep" generative process to its counterpart of an end-to-end discriminative process gave rise to the very first industrial success of deep learning [17, 18,39,42], which is the foundation of the third rise of AI to be described next.

Third, Current Rise of AI --- Deep Learning

While the second-generation of AI systems performed a lot better than the previous generation, they were far from human-level performance. With a few exceptions, the machine learning models often did not have the capacity sufficiently large to absorb the large amounts of training data. Further, the learning algorithms, methods, and intra-structures were not powerful enough. All this changed about one decade ago, giving rise to the third rise of AI, propelled by the new paradigm of deep-structured machine learning or Deep Learning [35,42]**.**

In traditional machine learning approaches, features are designed by humans and feature engineering is a bottleneck requiring significant human expertise. Concurrently, the models lack the representation power and hence the ability to form levels of decomposable abstractions that automatically disentangles complex factors in shaping the observed data. Deep learning breaks away the above difficulties by the use of deep, layered model structure, often in the form of neural networks, and the associated end-to-end learning algorithms. The advances in deep learning are one major driving force behind the current AI inflection point and the resurgence of neural networks.

Speech recognition is the first real-world AI application impacted strongly by deep learning. Industrial applications of deep learning to large-scale speech recognition started around 2010. In late 2009 and also 2010, the author invited his academic collaborator Prof. Geoffrey Hinton (and later his students) to work with him and colleagues at Microsoft Research to apply deep learning to speech recognition. They co-organized the 2009 NIPS Workshop on Deep Learning for Speech Recognition and Related Applications. The workshop was motivated by the limitations of deep generative models of speech, and the possibility that the big-compute, big-data era warrants a serious exploration of deep neural nets (DNN). It was believed that pre-training DNNs using generative models of deep belief nets based on the contrastive-divergence learning algorithm would overcome the main difficulties of neural nets encountered in the 1990s. However, early into this research at Microsoft, it was discovered that without contrastive-divergence pre-training, but with the use of large amounts of training data together with the DNNs designed with corresponding large, context-dependent output layers, dramatically lower recognition errors were possible than then-state-of-the-art (shallow) machine learning system based on the second-generation AI paradigm and algorithms. This finding was quickly verified by several other major speech recognition research groups. Further, the nature of recognition errors produced by the two types of systems were found to be characteristically different, offering technical insights into how to integrate deep learning into the existing highly efficient, run-time speech decoding system deployed by major players in speech recognition industry today [1,10,16,39,57]. More recent advances in deep learning for speech recognition can be found in [2,51,54]. Backpropagation algorithm is uniformly used in all sorts of deep neural networks of all current speech recognition systems.

Large-scale speech recognition is the first and most convincing successful case of deep learning in the recent history, embraced by both industry and academia across the board. Since 2010, the two major conferences on signal processing and speech recognition, IEEE-ICASSP and Interspeech, have seen a huge increase in the numbers of accepted papers in their respective annual conference papers on the topic of deep learning for speech recognition. More importantly, all major commercial speech recognition systems (e.g., Microsoft Cortana, Xbox, Skype Translator, Amazon Alexa, Google Now, Apple Siri, Baidu and iFlyTek voice search, and a range of Nuance speech products, etc.) are all based on deep learning methods. The most cited paper per year in the speech recognition history was published on deep learning in 2012 in IEEE Signal Processing Magazine [39].

Quickly following the striking success of speech recognition in 2010 heralding the firm arrival of the third AI wave, two other important AI application areas --- computer vision [41] and machine translation [3] --- were completely taken over by the same deep learning paradigm. In addition, a large number of other real-world applications have been made successful due to deep learning, including image captioning [31,33,34,37,38], visual question answering [44,56], web search [40], speech and text understanding [53,55], dialogue systems [8,15], drug discovery and toxicology, customer relationship management, recommendation systems, medical informatics, advertisement, medical image analysis, robotics, self-driving vehicles, board games (e.g. AlphaGo, and Poker), etc.

Setting aside their huge empirical successes, models of neural-network based deep learning are often simpler and easier to design than the traditional machine learning models developed in the second-generation AI. In many applications, deep learning is performed simultaneously for all parts of the model, from feature extraction all the way to prediction, in an end-to-end manner. Another factor contributing to the simplicity of neural network models is that the same model building blocks (i.e. the different types of layers) are generally used in many different applications. Using the same building blocks for a large variety of tasks makes the adaptation of models used for one task or data to another task or data relatively easy. In addition, software toolkits have been developed to allow faster and more efficient implementation of these models. For these reasons, deep neural networks are nowadays a prominent method of choice for a wide variety of machine learning tasks over large datasets.

Summary

This article analyzes the paradigms and algorithms of the AI technology, organized in terms of three rises (and two falls) based on the historical path of the AI development. Instead of covering the extensive scope of AI, this article draws representative and most relevant examples mainly from notable applications in the machine perception aspect of AI. In particular, an attempt has been made to connect

the cycles of ups-downs of AI to their counterparts in speech recognition, which, as a core application area of AI, parallels the historical path of AI more closely than most other areas of AI.

Applications of AI are vast, from speech and image to natural language processing [12], information retrieval [48], compressed sensing [46,47], robotics, healthcare, agriculture, financial services and more. Due to the space limit, readers are referred to [11] for details. An outlook of future AI technology development can also be found in [11] as well as in [13,14,32,43,52].

References

[1] O Abdel-Hamid et al. "Convolutional neural networks for speech recognition," IEEE/ACM Trans. Audio, Speech and Language Processing, Vol. 22, Pg. 1533-1545, 2014.

[2] D. Amodei et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, ICML, pp. 173–182, 2016.

[3] D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. ICLR, 2015.

[4] J. Baker et al. Research developments and directions in speech recognition and understanding. IEEE Signal Processing Magazine, vol. 26, no. 3, pp. 75-80, 2009.

[5] J. Baker et al. Updated MINS report on speech recognition and understanding IEEE Signal Processing Magazine, vol. 26, no. 4, July 2009a.

[6] Chris Bishop. Pattern Recognition and Machine Learning, Springer, 2006.

[7] Bridle, J., et al. An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition," Final Report for 1998 Workshop on Language Engineering, CLSP, Johns Hopkins, 1998.

[8] A Celikyilmaz et al. Deep Learning in Spoken and Text-Based Dialog Systems. Chapter 3 in book: Deep Learning in Natural Language Processing, Springer, 2018.

[9] Chengalvarayan R. and Deng, L. "HMM-based speech recognition using state-dependent, discriminatively derived transforms on Mel-warped DFT features," IEEE Transactions on Speech and Audio Processing, pp. 243-256, 1997.

[10] G. Dahl, et al. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," IEEE Trans. Audio, Speech, and Language Processing, vol. 20, pp. 30-42, 2012.

[11] L. Deng. Artificial Intelligence in the Rising Wave of Deep Learning: The Historical Path and Future Outlook [Perspectives]. IEEE Signal Processing Magazine 35 (1), pp. 180-177.

[12] L. Deng and Y. Liu (Eds.). Deep Learning in Natural Language Processing, Springer, 2018.

[13] L. Deng and Y. Liu. A Joint Introduction to Natural Language Processing and to Deep Learning, Chapter 1 in book: Deep Learning in Natural Language Processing, Springer, 2018.

[14] L. Deng and Y. Liu. Frontiers of NLP in the Deep Learning Era, Chapter 11 in book: Deep Learning in Natural Language Processing, Springer, 2018.

[15] L. Deng. How deep reinforcement learning can help Chatbot, Venturebeat, August, 2016.

[16] L. Deng and D. Yu. Deep Learning: Methods and Applications, NOW Publishers, 2014.

[17] L. Deng, G. Hinton, and B. Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. ICASSP, 2013.

[18] L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton. Binary coding of speech spectrograms using a deep autoencoder. Interspeech, 2010.

[19] L. Deng and D. Yu. Use of differential cepstra as acoustic features in hidden trajectory modeling for phonetic recognition, ICASSP, 2007.

[20] L. Deng. Dynamic Speech Models – Theory, Algorithm, and Application, Morgan & Claypool, 2006.

[21] L. Deng, D. Yu, and A. Acero. "Structured speech modeling," IEEE Trans. on Audio, Speech and Language Processing, vol. 14, no. 5, pp. 1492-1504, 2006

[22] L. Deng and D. O'Shaughnessy. SPEECH PROCESSING – A Dynamic and Optimization-Oriented Approach, Marcel Dekker, 2003.

[23] L. Deng. Switching dynamic system models for speech articulation and acoustics. in Mathematical Foundations of Speech and Language Processing, pp. 115–134. Springer-Verlag, New York, 2003.

[24] L. Deng and K. Erler. "Structural design of a hidden Markov model based speech recognizer using multi-valued phonetic features: Comparison with segmental speech units," J. Acoust. Soc. Am. vol. 92, pp. 3058-3067, 1992.

[25] L Deng, M Lennig, V Gupta, and P Mermelstein. Modeling microsegments of stop consonants in a hidden Markov model based word recognizer. J. Acoustical Soc. Am., vol. 87(6), pp. 2738-2747, 1990.

[26] L. Deng, P Kenny, M Lennig, P Mermelstein. Modeling acoustic transitions in speech by state-interpolation hidden Markov models. IEEE Transactions on Signal Processing, vol. 40 (2), pp. 265-271, 1990a.

[27] L Deng, V Gupta, M Lennig, P Kenny, P Mermelstein. Acoustic recognition component of an 86000-word speech recognizer, ICASSP, 1990b.

[28] L Deng, P Kenny, M Lennig, V Gupta, P Mermelstein. A Locus model of coarticulation in an HMM speech recognizer, ICASSP, 1989.

[29] L Deng, M Lennig, P Mermelstein. Use of vowel duration information in a large vocabulary word recognizer, J. Acoustical Soc. America, vol. 86(2), pp. 540-548, 1989a.

[30] L Deng, M Lennig, V Gupta, P Mermelstein. Modeling acoustic-phonetic detail in an HMM-based large vocabulary speech recognizer, ICASSP, 1988.

[31] J. Devlin, et al. Language models for image captioning: The quirks and what works. ACL, 2015.

[32] Y. Eldar et al. Challenges and Open Problems in Signal Processing: Panel Discussion Summary from ICASSP 2017 [Panel and Forum], IEEE SIgnal ProcESSIng Magazine. Vol.34 (6), pp. 8-23.

[33] H. Fang, et al. From captions to visual concepts and back. CVPR, 2015.

[34] Z. Gan et al. Semantic compositional networks for visual captioning. CVPR, 2017.

[35] I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. MIT Press, 2016.

[36] X. He, L. Deng, and W. Chou. Discriminative learning in sequential pattern recognition – A unifying review for optimization-oriented speech recognition. IEEE Signal Processing Magazine, vol. 25, pp. 14-36, 2008.

[37] X. He and L. Deng. Deep Learning in Natural Language Generation from Images. Chapter 10 in book: Deep Learning in Natural Language Processing, Springer, 2018.

[38] X. He and L. Deng. Deep learning for image-to-text generation: a technical overview, IEEE Signal Processing Magazine vol. 34 (6), pp. 109-116.

[39] G. Hinton et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, vol. 29, pp. 82-97, 2012.

[40] P. Huang et al. Learning Deep Structured Semantic Models for Web Search using Clickthrough Data, CIKM, 2013.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. NIPS, 2012.
[42] Y. LeCun, Y. Bengio, and G. Hinton, Deep Learning. Nature. vol. 521: 436–444.
[43] Y. Liu, J. Chen, L. Deng. Unsupervised Sequence Classification using Sequential Output Statistics", NIPS, 2017.
[44] M. Lee et al. Reasoning in Vector Space: An Exploratory Study of Question Answering, ICLR, 2016.
[45] J. Ma and L. Deng. Target-Directed Mixture Dynamic Models for Spontaneous Speech Recognition. IEEE Trans. Speech and Audio Processing, vol. 12, pp. 47-58, 2004.
[46] H Palangi et al. Convolutional deep stacking networks for distributed compressive sensing, Signal Processing. Vol. 131, 181-189. 2017.
[47] H Palangi et al. Distributed Compressive Sensing: A Deep Learning Approach. IEEE Trans. Signal Processing. Vol. 64 (17), pp. 4504-4518, 2016.
[48] H Palangi et al. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval, IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 24(4), pp. 694-707, 2016.
[49] J. Picone, et al. Initial evaluation of hidden dynamic models on conversational speech. ICASSP, 1999.
[50] L. Rabiner and B.-H. Juang. Fundamentals of Speech Recognition, Prentice-Hall, 1993.
[51] G. Saon et al. English Conversational Telephone Speech Recognition by Humans and Machines, ICASSP, 2017.
[52] P. Smolensky et al. Basic Reasoning with Tensor Product Representations arXiv, January 2016.
[53] G. Tur et al. Deep Learning in Conversational Language Understanding. Chapter 2 in book: Deep Learning in Natural Language Processing, Springer, 2018.
[54] W. Xiong et al. Achieving Human Parity in Conversational Speech Recognition, Interspeech, 2016.
[55] X. Yang et al. End-to-end joint learning of natural language understanding and dialogue manager, ICASSP, 2017.
[56] Z. Yang et al. Stacked Attention Networks for Image Question Answering. CVPR 2016.
[57] D. Yu and L. Deng. Automatic Speech Recognition: A Deep Learning Approach. Springer, 2015.

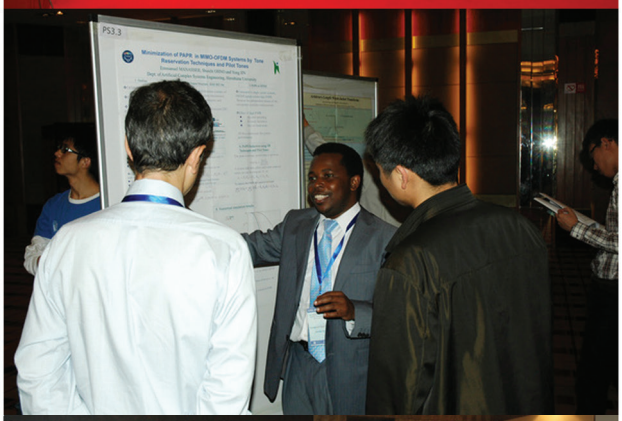


Photo Gallery: APSIPA-ASC'2011 in Xian

Adaptive Filtering with Selective Updates

Professor Yih-Fang Huang

PhD, Fellow IEEE

General Co-Chair, APSIPA ASC 2018

Professor of Electrical Engineering and Senior Associate Dean of College of Engineering

University of Notre Dame, Notre Dame, IN U.S.A.

Dr. Huang has been on the faculty at the University of Notre Dame since 1982 upon receiving his Ph.D. degree from Princeton University. He also served as chair of University of Notre Dame's Electrical Engineering department from 1998 to 2006. His research work employs principles in mathematical statistics to solve signal detection and estimation problems that arise in various applications that include wireless communications, distributed sensor networks and smart electric power grid.

Dr. Huang is a Fellow of the Institute of Electrical and Electronic Engineers (IEEE) ('95). He received the Golden Jubilee Medal of the IEEE Circuits and Systems Society in 1999, served as Vice President in 1997-98 and was a Distinguished Lecturer for the same society in 2000-2001. At the University of Notre Dame, he received Presidential Award in 2003, the Electrical Engineering department's Outstanding Teacher Award in 1994 and in 2011, the Rev. Edmund P. Joyce, CSC Award for Excellence in Undergraduate Teaching in 2011, and the Engineering College's Outstanding Teacher of the Year Award in 2013.

Dr. Huang was Toshiba Visiting Professor at Waseda University, Tokyo, Japan in Spring 1993, a visiting professor at the Munich University of Technology, Germany, April -July, 2007. In Fall, 2007, he was awarded the Fulbright-Nokia scholarship for lectures/research at Helsinki University of Technology in Finland. Dr. Huang was appointed Honorary Professor in the College of Electrical Engineering and Computer Science at National Chiao-Tung University, Hsinchu, Taiwan, in 2014.

In the past several decades, much research has been done in adaptive filtering, resulting in many great and impactful results. Among others, the recursive adaptive filtering algorithms such as Kalman filter (KF), recursive least squares (RLS) and least mean squares (LMS) have been well developed and found many useful applications.

While some may opine that those adaptive algorithms have reached maturity, a critical issue that has not been addressed much (relatively speaking) is the effective and discerning use of data in the implementation of adaptive algorithms. The conventional mechanism to implement those recursive filters is to update the parameter estimates at every epoch, i.e., every time a new datum (or measurement) is received. In reality, however, consecutive data often contain redundant information. Updating continually with redundant information is likely not productive, and can even be detrimental in some cases, e.g., causing numerical instabilities, causing estimates to drift, wasting resources, etc. Thus, an important question here is "when and how frequently should the parameter estimates be updated?". If we do not update the parameter estimates continually, i.e., at every epoch, then how do we decide when to update?

In this article, we present an adaptive filtering paradigm, namely, set-membership adaptive filtering, that enables a systematic mechanism to update parameter estimates *selectively*, depending on the *innovative information* contained in the received data.

Set-Membership Adaptive Filtering

The set-membership adaptive filtering (SMAF) is an unconventional filtering paradigm whose performance criterion is distinctly different from those of the conventional algorithms such as RLS, LMS, KF, etc. In particular, it requires that the filtering errors are bounded in magnitude at all time instants. The SMAF algorithms, see, e.g., [1,2,3,6,7,8,11,12], were initially developed to estimate the coefficients of linear regression models such as auto-regressive with exogeneous inputs (ARX), and auto-regressive moving average with exogeneous inputs (ARMAX) models. They are thus directly applicable to adaptive finite impulse response (FIR) and infinite impulse response (IIR) filters, and, more generally, system identification. A key feature of all SMAF algorithms, which results from the bounded-error criterion, is the *data-dependent selective update* of parameter estimates. Thus SMAF algorithms are viable alternatives to conventional algorithms that update the estimates continually, regardless of the benefits of those updates.

To illustrate the basic ideas of SMAF, consider a given data space $\mathcal{D}$ that comprises all possible input-desired output pairs $\{\mathbf{x}, d\}$, where $\mathcal{D} \subseteq \mathcal{C}^n \times \mathcal{C}$, and $\mathcal{C}^n$ is the n-dimensional real (or complex) Euclidean space. Given $\mathcal{D}$ and a designer-specified positive real number γ, the objective of SMAF is to design a filter whose output error $e = d - y$ is bounded in magnitude by γ for all $\{\mathbf{x}, d\} \in \mathcal{D}$, where $y = f_\theta(\mathbf{x}, d)$ is the filter output, and f_θ is a function that represents the adaptive filter with θ being the filter parameter that is to be estimated. If we only consider linear filters, then the filter output is $y = \theta^T\mathbf{x}$, and the objective of SMAF is to find a parameter θ such that

$$|e|^2 = |d - \theta^T\mathbf{x}|^2 \leq \gamma^2 \qquad (1)$$

In the case of regression models or adaptive FIR and IIR filters, $\mathbf{x}$ would be a *vector* that consists of input and prior output measurements, and d is the current output of the regression, or the filter. In system identification, $\mathbf{x}$ would be the input *vector* and d would be the desired output of the system at a particular time instant.

In the framework of recursive adaptive filtering, with the performance criterion specified in (1), we can define, for each time instant k, a *constraint set* H_{k}. The constraint set is the set of all parameter vectors θ that satisfy (1) with a given input-desired output pair $\{\mathbf{x}_{\mathrm{k}}, d_{\mathrm{k}}\}$, namely,

$$H_{\mathrm{k}} := \{\theta: |d_{\mathrm{k}} - \theta^T \mathbf{x}_{\mathrm{k}}|^2 \leq \gamma^2\} \tag{2}$$

Clearly, H_{k} is the region between two parallel hyperplanes in the parameter space $\mathcal{S}$. If we assume that the true parameter remains unchanged through all time instants, then the true parameter must lie in the intersection of the constraint sets at all time instants. We refer to such intersection set as the *exact membership set*, which is defined to be

$$\Omega_{\mathrm{k}} = \bigcap_{i=1}^{\mathrm{k}} H_i \tag{3}$$

for all time instants k, where i represents those time instants preceding k. It is clear that Ω_{k} is a polygon in $\mathcal{S}$. If the only assumptions made here are the bounded-error assumption and that the filter is linear, then every point in this polygon is a legitimate estimate for the parameter θ. As such, at every time instant, the SMAF renders a set of estimates, as opposed to a single point estimate rendered by the conventional algorithms. In essence, the objective of SMAF is to find a set of feasible filter coefficients such that the resulting output estimation errors are always bounded in magnitude.

If the parameter to be estimated, i.e., θ, is time-invariant, properly chosen error bounds would result in a non-empty Ω_{k} for all k. On the other hand, if θ changes at some point in time, Ω_{k} may become an empty set, for the inequality (1) may no longer hold. One way to circumvent the occurrence of an empty Ω_{k} is to adaptively change the error bound γ.

From (3), we see that Ω_{k} is a monotone non-increasing sequence of sets, i.e., $\Omega_{\mathrm{k}} \supseteq \Omega_{\mathrm{k+1}} \supseteq \Omega_{\mathrm{k+2}} \supseteq \cdots$, for all k. Thus, it is likely that, after some initial instants, Ω_{k_o} is a subset of H_{k_o+1}, for some k_o, as illustrated in Figure 1. Thus, no update of the exact membership set is necessary at time k_o. In general, however, it is difficult to find an analytical expression for Ω_k at every time instant k. In comparison, finding some analytically tractable outer bounding sets for Ω_k is usually more convenient. It is important to note, however, that the sequence of the outer bounding sets may not be monotone non-increasing. Instead, some set measure can be introduced that forms a monotone non-increasing sequence and that ensures convergence of the algorithm. Over the last few decades, various SMF algorithms have been proposed which differ in the manner that they define the *optimal* outer bounding sets, and the way that they define the set measure. Two such algorithms will be presented in the next section.

SMAF Algorithms

There are basically two types of outer bounding sets used in the derivations of SMAF algorithms - bounding ellipsoids and bounding spheroids. Use of bounding ellipsoids leads to a number of optimal bounding ellipsoid (OBE) algorithms, see, e.g., [1,2,8], while use of spheroids leads to a set-membership normalized least mean squares (SM-NLMS) algorithm [7]. In this section, we present some details of one of the OBE algorithms, namely, DH-

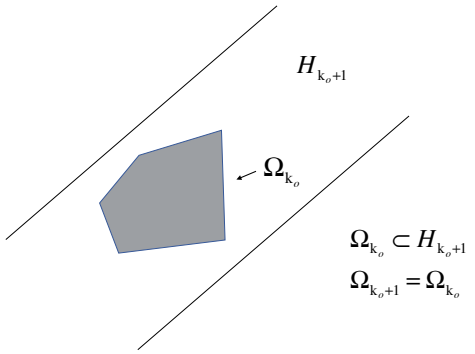


Fig. 1. No innovation rendered by the new constraint set.

OBE [2], and the SM-NLMS algorithm [7], to illustrate how the selective update of parameter estimates works.

To present DH-OBE, let the bounding ellipsoid at time k-1 be

$$E_{\mathrm{k-1}} = \{\theta \in R^N: (\theta - \theta_{\mathrm{k-1}})^T P_{\mathrm{k-1}}^{-1} (\theta - \theta_{\mathrm{k-1}}) \leq \sigma_{\mathrm{k-1}}^2\} \tag{4}$$

Then, at time k, with the new data pair $\{\mathbf{x}_{\mathrm{k}}, d_{\mathrm{k}}\}$, which defines the constraint set H_{k}, (2), an ellipsoid that *tightly* bounds the intersection of $E_{\mathrm{k-1}}$ and H_{k} is given by

$$\begin{aligned} E_{\mathrm{k}} = \{\theta \in R^N: & \\ & (1-\lambda_{\mathrm{k}})(\theta - \theta_{\mathrm{k-1}})^T P_{\mathrm{k-1}}^{-1} (\theta - \theta_{\mathrm{k-1}}) \\ & + \lambda_{\mathrm{k}} (d_{\mathrm{k}} - \theta^T \mathbf{x}_{\mathrm{k}})^2 \\ & \leq (1-\lambda_{\mathrm{k}}) \sigma_{\mathrm{k-1}}^2 + \lambda_{\mathrm{k}} \gamma^2\} \end{aligned} \tag{5}$$

One can re-formulate E_{k} in (5) with a standard ellipsoidal equation with updated P_{k}, σ_{k}, and θ_{k} as follows:

$$E_{\mathrm{k}} = \{\theta \in R^N: (\theta - \theta_{\mathrm{k}})^T P_{\mathrm{k}}^{-1} (\theta - \theta_{\mathrm{k}}) \leq \sigma_{\mathrm{k}}^2\} \tag{6}$$

where

$$P_{\mathrm{k}}^{-1} = (1-\lambda_{\mathrm{k}}) P_{\mathrm{k-1}}^{-1} + \lambda_{\mathrm{k}} \mathbf{x}_{\mathrm{k}} \mathbf{x}_{\mathrm{k}}^T \tag{7}$$

$$\sigma_{\mathrm{k}}^2 = \sigma_{\mathrm{k-1}}^2 + \lambda_{\mathrm{k}} \gamma^2 - \frac{\lambda_{\mathrm{k}} (1-\lambda_{\mathrm{k}}) (d_{\mathrm{k}} - \theta_{\mathrm{k-1}}^T \mathbf{x}_{\mathrm{k}})^2}{1 - \lambda_{\mathrm{k}} + \lambda_{\mathrm{k}} \mathbf{x}_{\mathrm{k}}^T P_{\mathrm{k-1}} \mathbf{x}_{\mathrm{k}}} \tag{8}$$

$$\theta_{\mathrm{k}} = \theta_{\mathrm{k-1}} + \lambda_{\mathrm{k}} P_{\mathrm{k}} \mathbf{x}_{\mathrm{k}} (d_{\mathrm{k}} - \theta_{\mathrm{k-1}}^T \mathbf{x}_{\mathrm{k}}) \tag{9}$$

$$P_{\mathrm{k}} = \frac{1}{1-\lambda_{\mathrm{k}}} \left(P_{\mathrm{k-1}} - \frac{\lambda_{\mathrm{k}} P_{\mathrm{k-1}} \mathbf{x}_{\mathrm{k}} \mathbf{x}_{\mathrm{k}}^T P_{\mathrm{k-1}}}{1 - \lambda_{\mathrm{k}} + \lambda_{\mathrm{k}} \mathbf{x}_{\mathrm{k}}^T P_{\mathrm{k-1}} \mathbf{x}_{\mathrm{k}}}\right) \tag{10}$$

Since there are infinitely many such tight bounding ellipsoids, the ellipsoid E_{k} formulated in (6) needs to be optimized in some sense (with respect to λ_{k}), rendering an *optimal bounding ellipsoid*. To optimize E_{k}, DH-OBE seeks to minimize σ_{k}^2 to render an optimal value of λ_{k}, thus an optimal bounding ellipsoid in the sense of minimum σ_{k}^2. The optimal value of λ_{k} in DH-OBE is obtained by

$$\lambda_{\mathrm{k}} = \begin{cases} \min(\xi, \nu_{\mathrm{k}}) & \text{if } \gamma^2 < \sigma_{\mathrm{k-1}}^2 + e_{\mathrm{k|k-1}}^2 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where ξ is a design parameter, $e_{k|k-1} = d_k - \theta_{k-1}^T \mathbf{x}_k$ and ν_k is determined as descried in [2]. Note that when $\lambda_k = 0$, there is *no update* of parameter estimate.

Minimizing σ_k^2 can be viewed as minimizing an upper bound on the normalized estimation error when θ_k is taken as the parameter estimate at time k. Since, at any time instant k, every SMAF algorithm renders a set of legitimate estimates (every point in the exact membership set is a legitimate estimate), the geometric center of the ellipsoid can be accepted as a natural point estimate. Furthermore, σ_k^2, when normalized with P_k^{-1}, is directly related to the size of the ellipsoid.

Using the center of the ellipsoid, θ_k, as the estimate of the parameter at every k reveals that formulations of the DH-OBE algorithm, (7) – (10), resemble those of RLS with a forgetting factor λ_k. The key difference here is that in RLS, λ_k is usually set *a priori*, and remains a constant through the entire operation of the filter. In contrast, in DH-OBE, the value of λ_k is optimized at every k in accordance with the received data. Since some optimal λ_k may be zero, optimization of λ_k, thus the bounding ellipsoid, offers a systematic mechanism for selective updates.

The formulation for SM-NLMS [7] is given below. Define the bounding spheroid at time k by

$$S_k = \{\theta \in R^N : ||\theta - \theta_k||^2 \le \sigma_k^2\}. \quad (12)$$

The parameter estimate at time k is given by

$$\theta_k = \theta_{k-1} + \lambda_k \frac{\delta_k \mathbf{x}_k}{\mathbf{x}_k^T \mathbf{x}_k} \quad (13)$$

$$\sigma_k^2 = \sigma_{k-1}^2 - \lambda_k \frac{|e_{k|k-1}|^2}{\mathbf{x}_k^T \mathbf{x}_k} \quad (14)$$

where

$$e_{k|k-1} = d_k - \theta_{k-1}^T \mathbf{x}_k \quad (15)$$

$$\lambda_k = \begin{cases} 1 - \frac{\gamma}{|e_{k|k-1}|} & \text{if } |e_{k|k-1}| > \gamma \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

Again, when $\lambda_k = 0$, there is *no update* of parameter estimate. Also, the formulations of SM-NLMS resemble those of NLMS, except for a data-dependent optimized step size, λ_k.

While the performance criteria for SMAF algorithms and for RLS and NLMS are different, due to the data-dependent optimized weight (or step size), the SMAF algorithms generally perform comparably, measured by mean square errors, to their counter parts, namely, RLS or NLMS. In fact, in some cases, SMAF algorithms perform even better than RLS or NLMS.

In general, a tighter error bound, i.e., a smaller value for γ, yields more accurate estimation results. However, a smaller value for γ usually calls for more frequent updates which implies more costs in computation and communications. Thus, choice of γ amounts to a performance-cost tradeoff, which is usually not conveniently available with RLS or LMS algorithms.

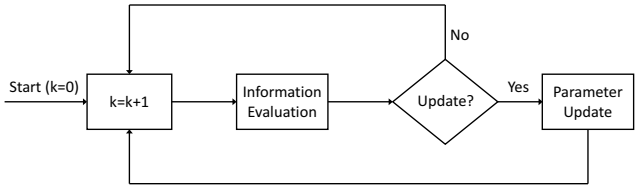


Fig. 2. Adaptive filtering with selective update.

In summary, regardless of the types of outer bounding set chosen, the SMAF approach offers a modular adaptive filtering architecture that consists of an *information evaluator* followed by an *updating processor*, as shown in Fig. 2. This leads to the unique feature that allows for selective update of the parameter estimates, which can be exploited to a great advantage in many applications. It is also worth noting that, in both DH-OBE and SM-NLMS, the filter output error at time k given the estimate at k-1, i.e., $e_{k|k-1} = d_k - \theta_{k-1}^T \mathbf{x}_k$, serves as a measure for *innovation*. The algorithms update parameter estimates only when the innovation exceeds certain threshold, as seen in (11) and (16).

Long-term Impact

Generally speaking, the main objective of adaptive filtering is to extract information from the received data, and transform the information into some sort of actionable intelligence. The goal of machine learning is similar. In fact, one may consider adaptive filtering as a rudimentary form of machine learning. In an era that a deluge of data can easily flow through the internet and be stored in relatively small devices, one may not think much about having a myriad of data. While storage of huge data sets is usually not a problem, the process of extracting information from data and, more importantly, transforming information into intelligence can be rather onerous. The complexity of such tasks grows exponentially as the size of data sets grows. Furthermore, in many practical situations, decisions and actions need to be made in a timely manner, or as quickly as possible. One way to expedite reaching the decision and action is to afford the ability of systematically separating relevant data from irrelevant data, thus enabling use of data in a more discerning way. The principle of SMAF that provides a systematic way of deciding whether or not the received data contains sufficient innovation, hence makes possible selective updates of parameter estimates, is a viable approach to developing more discerning and effective way of using received data. This can be informative in the development of more effective machine learning algorithms.

References

1. E. Fogel and Y.F. Huang, "On the Value of Information in System Identification - Bounded Noise Case," *Automatica*, Vol. 18, No. 2, pp. 229-238, March 1982.
2. S. Dasgupta and Y.F. Huang, "Asymptotically Convergent Modified Recursive Least Squares with Data-Dependent Updating and Forgetting Factor for Systems with Bounded Noise," *IEEE Transactions on Information Theory*, Vol. IT-33, No. 3, pp. 383-392, May 1987.

3. A.K. Rao, Y.F. Huang and S. Dasgupta, "ARMA Parameter Estimation Using a Novel Recursive Estimation Algorithm with Selective Updating," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-38, No. 3, pp. 447-457, March 1990.
4. A.K. Rao and Y.F. Huang, "Analysis of Finite Precision Effects on a Recursive Set Membership Parameter Estimation Algorithm," *IEEE Transactions on Signal Processing*, Vol. 40, No. 12, pp. 3081-3085, December 1992.
5. A.K. Rao and Y.F. Huang, "Tracking Characteristics of An OBE Parameter Estimation Algorithm," *IEEE Transactions on Signal Processing*, Vol. 41, No. 3, pp. 1140-1148, March 1993.
6. S. Gollamudi, S. Kapoor, S. Nagaraj and Y.F. Huang, "Set-Membership Adaptive Equalization and an Updator-Shared Implementation for Multiple Channel Communications Systems," *IEEE Transactions on Signal Processing*, Vol. 46, No. 9, pp. 2372-2385, September 1998.
7. S. Gollamudi, S. Nagaraj, S. Kapoor and Y.F. Huang, "Set-Membership Filtering and a Set-Membership Normalized LMS Algorithm with an Adaptive Step Size," *IEEE Signal Processing Letters*, Vol. 5, No. 5, pp. 111-114, May 1998.
8. S. Nagaraj, S. Gollamudi, S. Kapoor and Y.F. Huang, "BEACON: An Adaptive Set-Membership Filtering Technique with Sparse Updates," *IEEE Transactions on Signal Processing*, Vol. 47, No. 11, pp. 2928-2941, November 1999.
9. Nagaraj, S. Gollamudi, S. Kapoor, Y.F. Huang, and J.R. Deller, "Linear and Adaptive Interference Suppression in CDMA with a Worst Case Error Estimation Criterion," *IEEE Transactions on Signal Processing*, Vol. 48, No.1, pp. 284-289, January 2000.
10. E.W. Bai and Y.F. Huang, "Variable Gain Parameter Estimation Algorithms for Fast Tracking and Smooth Steady State," *Automatica*, Vol. 36, pp. 1001-1008, 2000.
11. J.R. Deller and Y.F. Huang, "Set-Membership Identification and Filtering for Signal Processing Applications," *Circuits, Systems, and Signal Processing*, Vol. 21, No. 1, pp. 69-82, January-February 2002.
12. L. Guo and Y.F. Huang, "Frequency-Domain Set-Membership Filtering and Its Applications," *IEEE Transactions on Signal Processing*, Vol. 55, No. 4, pp. 1326-1338, April 2007.
13. J.R. Deller, S. Gollamudi, S. Nagaraj, D. Joachim, and Y. F. Huang, "Convergence Analysis of the Quasi-OBE Algorithm and Related Performance Issues," *International Journal of Adaptive Control and Signal Processing*, Vol. 21, No. 6, pp. 499-527, August 2007.
14. S. Werner, M. Mohammed, Y.F. Huang and V. Koivunen, "Decentralized Set-Membership Adaptive Estimation for Clustered Sensor Networks," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3573-3576, Las Vegas, March 31-April 4, 2008.
15. S. Werner, Y.F. Huang, and M.L.R. de Campos, and V. Koivunen, "Distributed Parameter Estimation with Selective Cooperation," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 2849-2852, Taipei, Taiwan, April 19-24, 2009.

Photo Gallery: APSIPA Pre-Establishment Summit in Hong Kong, December 2008

Video Understanding with Depth Information

Professor Ming-Ting Sun

PhD, Fellow IEEE

Technical Program Co-Chair, APSIPA ASC 2018. BoG, APSIPA, 2011-2017.

Professor, Department of Electrical Engineering, University of Washington

Ming-Ting Sun *received the B.S. degree from National Taiwan University and the Ph.D. degree from UCLA, all in electrical engineering. He joined the University of Washington in 1996 where he is a Professor. Previously, he was the Director of Video Signal Processing Research at Bellcore. His main research interest is video and multimedia signal processing.*

Professor Sun holds 13 patents and has published about 300 technical papers, including 17 book chapters. He co-edited a book "Compressed Video over Networks." He has guest-edited 12 special issues for journals and given keynotes in several international conferences. He was an Editor-in-Chief of the Journal of Visual Communication and Image Representation (JVCI) from 2012 to 2016, the Editor-in-Chief of the IEEE Transactions on Multimedia (TMM) from 2000 to 2002, and the Editor-in-Chief of the IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) from 1995 to 1997. He was a Distinguished Lecturer of the Circuits and Systems Society from 2000 to 2001. He received an IEEE CASS Golden Jubilee Medal in 2000. He served as a General Chair of ICME 2016 and VCIP 2000, and an Honorary Chair of MMM 2018 and VCIP 2015. He was the Chair of the Visual Signal Processing and Communications (VSPC) Technical Committee of IEEE Circuits and Systems (CAS) Society from 1993 to 1994. He received the TCSVT Best Paper Award in 1993. From 1988 to 1991, he was the chairman of the IEEE CAS Standards Committee and established the IEEE Inverse Discrete Cosine Transform Standard.

Video signal processing is a practical field. In a practical field, the success of a technology depends on the success of its applications. In the early days, we saw active development of technologies for video compression and communications due to their mass market applications such as digital TV, video conferencing, and video streaming. Recently, many new mass market applications related to video understanding are emerging, such as video surveillance, self-driving cars, intelligent robots, and home entertainment with human computer interactions. These emerging mass market applications power development in video understanding technologies.

Video understanding tries to automatically recognize objects or activities in a video. Many traditional video understanding problems are hard problems. They try to recognize 3D objects or activities from a 2D video. When 3D objects and activities are projected into 2D video, depth information is lost. Since object and activity recognition are 3D problems, it is much easier to solve them in 3D with depth information if it is available.

With recent progress in technology, depth information is becoming much more accessible. It can be obtained in several different ways, including using laser scanners, low-cost depth sensors such as Kinect or RealSense, stereo matching with stereo videos, structure from motion with multiple images, or radar. With the depth information, we know the 3D coordinate of each pixel. We can also obtain much more reliable object boundaries. After normalizing with the depth information, we can compute the physical size of the object. Furthermore, depth information often is not affected by the lighting conditions and shadows. These properties allow us to resolve ambiguities in the 2D video and make the solutions much easier and more robust.

Depth information is usually relatively noisy. One challenge is how to denoise the depth image or how to utilize the relatively noisy depth information. We have been working on several video understanding related technologies and applications using the depth information. In the following, we give a few examples. Interested readers can obtain more detailed information from the references.

Automatic Powerline Detection

Powerlines present hazardous operating conditions for helicopters, especially when the pilot's vision is obscured by dust, smoke, fog, rain, snow, or darkness. For most RGB cameras, powerlines are usually subtle and difficult to detected, especially under poor lighting conditions, in the evenings, and when the helicopter is moving at relatively high speeds.

Radar can work under poor visibility conditions and in the evenings. The magnitude and phase of the returned radar signal in radar images are related to the depth of the objects in the radar video. A powerline will appear differently from the background in the radar images since it has a different depth compared to its background. Figure 1 shows a radar system on the helicopter and a radar image with powerlines.

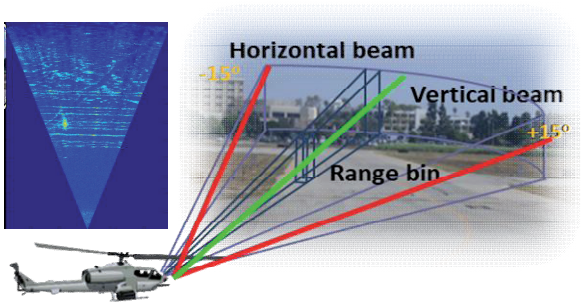


Figure 1. A radar system on the helicopter and a radar image with powerlines.

We proposed an automatic powerline detection framework principally based on computer vision and machine learning using millimeter wave radar videos. In the algorithm, we first perform Hough Transform to identify candidate powerlines. A Support Vector Machine (SVM) classifier is trained to differentiate true powerlines from false powerlines created by noise. A particle filter is used to track powerlines in the video sequence. Phase difference caused by different depths is used to improve the robustness of powerline detection under extremely noisy conditions. Experiments show that our algorithm can achieve very high accuracy and robust detection.

Semantic Segmentation

For research in self-driving car and intelligent robot applications, street view video datasets with dense semantic ground truth labels are very useful. Semantic segmentation which segments and labels objects in images is vital for training models for object recognition or scene understanding.

The difficulty in developing such a dataset is that performing manual segmentation and labeling is very labor intensive and time consuming. It is desirable to develop automatic methods to accurately perform segmentation and generate semantic labels. We first captured a large-scale street-view suburban video dataset, comprising over 400k images, using a laser scanner, stereo cameras, and a fisheye camera mounted on a car that was driven around town. By registering the video frames from all of the sensors and using the depth information from the laser scanner, the stereo camera, and structure from motion, we built a 3D model of the scene. We then developed a simple user interface to allow users to perform semantic labeling using simple bounding boxes on the 3D scene. The 3D labels are then projected into 2D video frames. A fully connected Markov Random Field (MRF) model with the depth information is used to correct incorrect labels. A result of the semantic segmentation is shown in Figure 2.

Using this 3D-to-2D label transfer method, we were able to obtain more accurate semantic labels than those obtained using existing 2D label transfer methods, and save a significant amount of time compared to using the manual labeling approach.

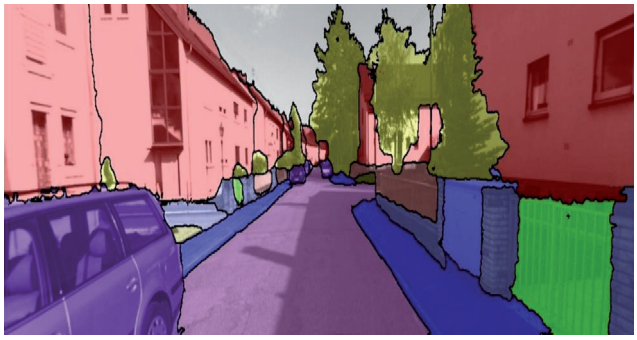


Figure 2. A result of the semantic segmentation using our proposed 3D-to-2D label transfer approach using the depth information.

Gaze Estimation

It is useful to know what a person is looking at on a screen, for example, for determining which part catches people's attention in a webpage layout. It would also enable applications where people can use their eyes instead of their hands for human-computer interaction.

We developed a real-time 3D gaze estimation system that can estimate the gaze direction using a Kinect camera, as shown in Figure 3. A 3D geometric eye model is constructed with help from the depth information as shown in Figure 4. The parameters of the 3D geometric eye model are determined through a simple calibration process. From the RGB image, we locate the iris center. From the iris center location, the eye corner location, and the 3D geometric eye model, we can estimate the gaze direction. Experimental results show that the system can achieve an accuracy of 1.4~2.7 degrees and is robust against large head movements.

Two real-time applications have been implemented to demonstrate the potential for practical applications. In one application, a person uses his eyes to type. In another real-time demo, players play chess games using their eyes. These demonstrate the potential to allow a completely paralyzed person to perform tasks using eye movements which was previously impossible. The core of the technology is the depth information which allows the 3D geometric eye model to be built relatively easily.

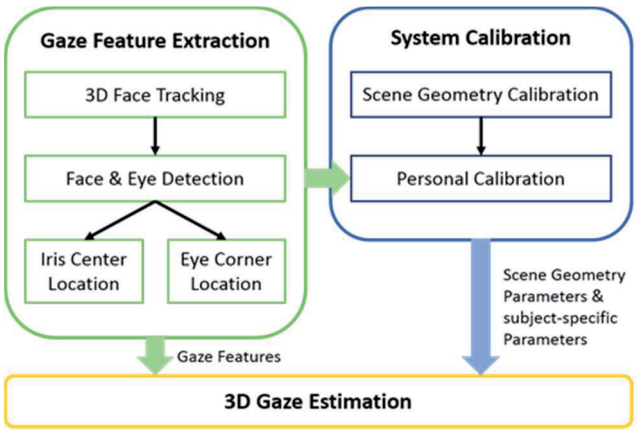


Figure 3. Our proposed 3D gaze estimation system.

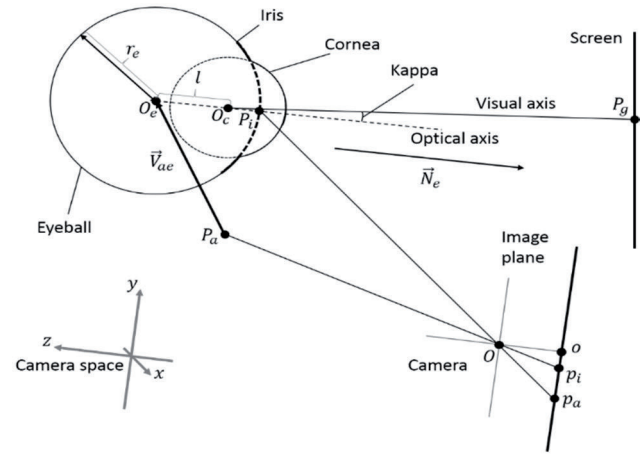


Figure 4. Our proposed 3D geometric eye model. The parameters are determined from a simple calibration process.

Elder Care

Many elderly people live alone. It is desirable to develop a system that not only monitors possible accidental falls, but also their diet, so that caretakers know whether or not they are eating healthily and if they have followed directions to take prescribed pills regularly. This is important since older people often forget if they have taken medicine as instructed. For this purpose, we developed a system to monitor activities in the kitchen, focusing on what people are eating. Some typical kitchen scenes are shown in Figure 5.

Figure 5. Common kitchen scenes for developing our diet monitoring system.

We developed a system, consisting of several algorithms, to recognize human activities that involve manipulating objects using a Kinect camera. This can be used for monitoring the diet of an elder person. From the RGB and corresponding depth images, we were able to segment out the person relatively easily even under poor lighting conditions, with the help of depth information from the Kinect. The depth information is useful, since in this case, it allows us to calculate the physical size of the object so that we can identify the torso, arm, and hand of the person. It also provides the boundary of the body and the background without being affected by shadows. The size and depth information allow us to resolve occlusions and locate the hand of the person to identify the objects being manipulated. With the identified object and the action being performed on the object, we trained a Hidden Markov Model (HMM) to infer the high-level task being performed. We evaluate our approach on a challenging dataset of 12 kitchen tasks involving 24 objects performed by 2 subjects. The entire system yields 82%/84% precision (74%/83% recall) for task/object recognition.

Long-term Impact

Many mass market applications (e.g., self-driving cars, intelligent robots, large scale video surveillance, home entertainment with human computer interaction, education, etc.) involve automatic object and activity recognition. Object and activity recognition are fundamentally hard problems that use traditional computer vision techniques. Since we live in a 3D world, object and activity recognition are inherently 3D problems. Projecting 3D objects and activities into 2D space creates ambiguities which are difficult to resolve using traditional 2D techniques. A 3D problem is best solved in 3D space utilizing depth information. Here we described a few technologies and systems using depth information which can achieve more accurate and robust results compared to traditional 2D approaches. With depth information becoming more easily accessible, we expect practical object and activity recognition applications utilizing depth information to become more prevalent and make a larger impact in our daily life.

References

1. Ma, D. Goshi, Y.C. Shih, and M.T. Sun, “An Algorithm for Power Line Detection and Warning based on a Millimeter-Wave Radar Video,” IEEE Transaction on Image Processing, vol.20, no.12, December 2011.
2. Q. Ma, D.S. Goshi, L. Bui, and M.T. Sun, “Robust Power Line Detection with Tracking in Radar Video,” APSIPA Transactions on Signal and Information Processing, vol. 4, September 2015.
3. J. Xie, M. Kiefel, M.T. Sun, and A. Geiger, “Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer, CVPR 2016.
4. L. Sun, M.L. Song, Z. Liu, M.T. Sun, “Realtime Gaze Estimation with Online Calibration,” IEEE MultiMedia, vol. 21, no. 4, pp. 28-37, Oct.-Dec. 2014.
5. L. Sun, Mingli Song, Z. Liu, M.T. Sun, ”Realtime Gaze Estimation with Online Calibration,” IEEE ICME 2014.
6. H. Liu, M. Philipose, and M.T. Sun, “Automatic Object Segmentation with RGB-D Cameras,” Special issue on 3D video processing, Journal of Visual Communication and Image Representation, vol. 25, Issue 4, pp. 709-718, May 2014.
7. H. Liu, M. Philipose, M. Patterson, and M.T. Sun, “Recognizing Object Manipulation Activities Using Depth and Visual Cues,” Special issue on 3D video processing, Journal of Visual Communication and Image Representation, vol. 25, Issue 4, pp. 719-726, May 2014.

Co-Evolution of Artificial Intelligence and Human Intelligence

Intelligence is the deciding factor of how human beings become the most dominant life form on earth. Throughout history, human beings have developed tools and technologies which help civilizations evolve and grow. Computers, and by extension, artificial intelligence (AI), has played important roles in that continuum of technologies. Recently artificial intelligence has garnered much interest and discussion. As artificial intelligence are tools that can enhance human capability, a sound understanding of what the technology can and cannot do is also necessary to ensure their appropriate use. While developing artificial intelligence, we also found out the definition and understanding of our own human intelligence continue evolving. The debates of the race between artificial and human intelligence (HI) have been ever growing. Although bold predictions are made for the future including some catastrophes ranging from machines replacing 99% of human jobs to machines ruling the human world, I would like to illustrate the future is a harmonic one where AI and HI co-evolve with each other.

AI vs. HI

With the combination of abundant computation, big data and recent advancement of deep neural networks (DNN's), computers are able to exhibit very intelligent behaviors such as accurately recognizing speech, image, video and language, and masterfully play complicated games like Chess and Go, sometimes even better than humans. The worry that computers will one day completely exceed human intelligence and even dominate human race has been ever greater. From historical perspective, this type of worry has been observed in Jan. 23rd, 1950's issue of Time Magazine [1], when computers barely existed in our life.

HIERARCHY OF INTELLIGENCE

Figure 1. illustrates a hierarchy of intelligence with the most fundamental intelligence at the bottom of the pyramid and the higher level of intelligence like "*creativity*" and "*wisdom*" at the top. Based on this simple hierarchy, let's examine how artificial intelligence measures against human intelligence.

Computation/Memory – Computation and memory were long considered important aspects of human intelligence in any culture. Mathematics/arithmetic is considered as the foundation for all science and engineering where calculation is pervasively required in all science and engineering discovery. Throughout human history, there are numerous attempts of inventing computing machines, including abacus, Pascal's mechanical calculator, Thomas' Arithmometer, Charles Babbage's difference engine, electronic calculator, …etc.

The invention of modern computers followed the pursuit of man-made computing machines. Computation and memory are exactly the two core components for the mathematic abstract model of general computation - Turing machine and the architecture of modern computers - Von Neumann machine. With modern computers in any form factor, humans clearly have no chance to compete with machines in computation and memory capability.

Dr. Hsiao-Wuen Hon

PhD, FIEEE

Keynote speaker of APSIPA ASC 2018

Corporate Vice President
Microsoft Corporation

Chairman
Microsoft's Asia-Pacific R&D Group

Managing Director
Microsoft Research Asia

At Microsoft, Dr. Hon drives Microsoft's strategy for research and development activities in the Asia-Pacific region, as well as collaborations with academia. Dr. Hon has been with Microsoft since 1995. He joined Microsoft Research Asia in 2004 as deputy managing director, stepping into the role of managing director in 2007. He founded and managed Microsoft Search Technology Center from 2005 to 2007 and led development of Microsoft's search products (Bing) in Asia-Pacific. Dr. Hon is an internationally recognized expert in speech technology. Dr. Hon has published more than 100 technical papers in international journals and at conferences. Dr. Hon holds more than 40 patents in several technical areas. Dr. Hon received a Ph.D. in Computer Science from Carnegie Mellon University and a B.S. in Electrical Engineering from National Taiwan University.

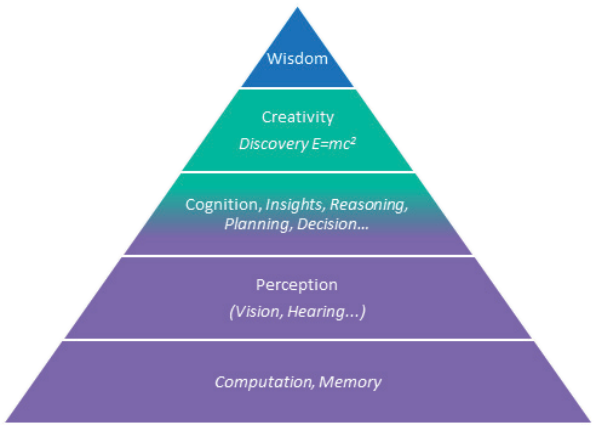


Figure 1 Hierarchy of Intelligence

Perception - It is not long ago the ability to identify and interpret the sensory information through eyes (vision) and ears (speech) is considered a unique intelligence of humans or advanced animals. Thus, the pursuit of automatic speech recognition (ASR) and vision recognition by computers has been the core component of artificial intelligence since its inception.

Perception is also the area which took most advantage of recent progress of DNN's. Recent progress in ASR and

computer vision recognition demonstrated human-parity or better performance in standard benchmark tasks like Switchboard [2] and ImageNet [3]. ASR and vision-based functions have been standard in mainstream products in security, autonomous driving, medical imaging and personal assistants, like Apple's Siri, Microsoft's Cortana, Amazon's Echo, and Google Assistant.

In many ways, DNN's could be considered the best technique to tackle complicated pattern recognition problems if ample labeled (supervised) training data are available because its ability to model any sophisticated non-linear representation with few or no artificial assumptions. While humans can recognize normal speech and objects with little training data, it is reasonable to conclude that DNN's technologies can perform better recognition for any arbitrary set of audio or visual objects with sufficient training data.

Cognition (Insights, Reasoning, Planning, and Decisions) - Recent progress in AI also enable machine to exhibit excellent cognition capability for some useful tasks. For instance, recent results on standard benchmark in **S**tanford **Qu**estion **A**nswering **D**ataset (SQuAD) for reading comprehension [4] and WMT17 shared newstest2017 tasks [6] for machine translation [5]; both considered substantial human cognitive tasks; have demonstrated human parity performance.

Figure 2 illustrates the general cognitive system being realized as a data-flow feedback loop. Humans analyze all observations and data they can sense to gain insights & reasoning and eventually make decision & planning of what to do next by manipulating the actuators to change the physical world. With recent AI progress, it is feasible to build a close-loop system for certain complicated industrial tasks where the cognition process of the system (Analysis and Decision parts) can be made automatically via AI technologies. In this view, the cognition process is modeled as a pattern recognition process where DNN's shine. The majority of emerging AI applications in the industry belong to this category, e.g. predictive maintenance [7]

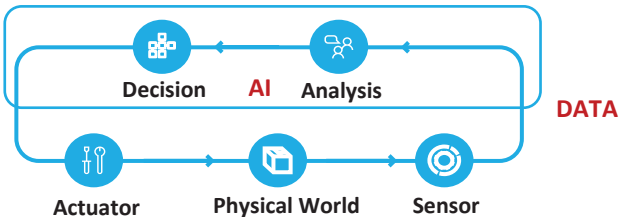


Figure 2 An illustration of the cognition system being realized as data-driven AI of a feedback loop.

The feasibility of such an automatic system shares the same characteristics of pattern recognition by DNN's – only effective for a closed system (where it isn't affected by any factor whose source is external to the system) with plenty of labeled training data. This type of cognition is often referred as black-box system where only the behavior of the stimulus/response will be accounted for, without any knowledge of its internal workings. Such black-box systems can then only answer "*what*", but not "*why*"; thus, is often referred to "*unexplainable*" AI. As AI-based dynamic systems becomes commonly used in our lives, clearer accountability will be required for decision making processes to ensure trust and transparency. Thus, the call for "*Explainable AI*" (XAI) whose actions can be easily understood by humans, has been gaining more momentum.

Humans often perform white-box cognition where inner components or logic are available for inspection and interpretation. Causality (cause and effect) analysis [8] is extensively conducted across components in different white box systems, so reasoning across different white box systems is even possible.

The philosopher John Searle presented a famous challenge to the Turing test, called the Chinese room experiment [9], in 1980. Searle contended that: 1) computers mindlessly manipulate symbols without understanding their meaning and 2) computers are not really thinking as we humans do. He called the former simulating intelligence as *weak* AI and the latter with understanding as *strong* AI. It is very interesting that today's mainstream AI cognition dominated by DNN's is still mostly resembling mechanical pattern recognition style of weak AI without much understanding. On the contrary, human's perception often leverages multi-layer and multimodal understanding of diverse semantics to compensate for imperfect memory or incomplete data. One illustration is the famous cocktail effect [10], where the brain's ability to focus one's auditory attention on a particular stimulus while filtering out a range of other stimuli, as a partygoer can focus on a single conversation in a noisy room.

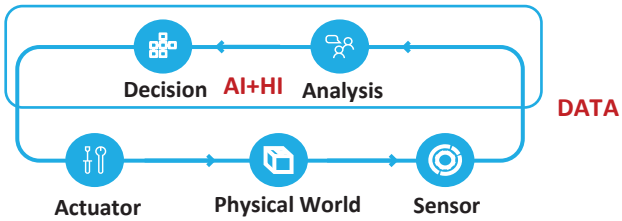


Figure 3 An illustration of the cognition system where "AI+HI" making analysis & decision together

Before a fully automatic white box cognition system can be attained, AI and many big data analytics tools can still be very useful to help humans for reasoning (cause-and-effect analysis), gaining insights, deriving planning & decision. Figure 3 shows such a collaborative system where AI and HI partner together in the critical "Analysis" and "Decision" process. Such a system is also very useful in dealing with open system interacting with external environment which cannot be fully observed.

Creativity - Recently some researchers use autoencoder DNN's (encoder-decoder model) to create interesting artistic works like poems, lyrics, music melodies [11], paintings and videos [12] that seems to demonstrate some creativity from artificial intelligence.

There isn't a precise definition of "*Creativity*" and there are many myths surrounding it, e.g. "*People are born with it*", "*It falls into your laps or god told me*". I generally like more actionable description like "Creativity is characterized by the ability to perceive the world in new ways, to find hidden patterns, to make connections between seemingly unrelated phenomena, and to generate solutions or artifacts." I take deliberation to give a narrow problem-solving definition of creativity to be "the ability to come up a new algorithm to solve a previously unsolved problem or existing solvable problem with a better solution".

Taking this definition, people would find computers are no match to humans since computers still cannot program itself or come up with their own algorithms. All the algorithms, including all AI algorithms, are from humans so far. Solving any problem involves both an algorithm (how to solve) and the real computation (the actual solving/compute). Let's do a simple exercise. There are two algorithms to solve the sum of "n" consecutive integers. One is the clever algorithm $n(n+1)/2$ thought to be invented by Carl Gauss while the other is straightforward n additions. Given the incredible

computation speed of modern computers (>teraflops), a human using clever Gauss algorithm would still lose to a computer using straightforward additions. Following this logic, humans should not be too worried about men losing to machine in Go or Chess games because (1) human's game playing algorithms still have a lot of merits given there is no comparison for the computation part; and (2) all the algorithms for playing computer game all come from humans anyway.

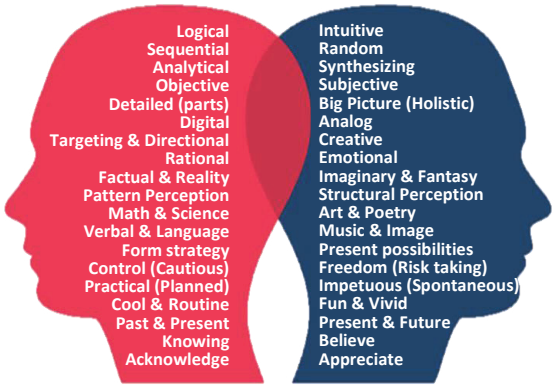


Figure 4 The theory of left brain vs. right brain [13].

Figure 4 shows the theory of left brain vs. right brain. One can find computer's close resemblance of human's left brain. The close collaboration relationship between humans providing creative algorithms and computers providing accurate & fast computation is just like our right and left brain working together to solve problems. This is yet another astounding example of "AI+HI".

Wisdom - Wisdom is probably harder to define than creativity. It is related to consciousness, empathy, sapience and mind. "Consciousness" belong exclusively to advanced animals. The mind-body problem is a longstanding philosophical problem concerning the relationship between the human mind and body. Many computer scientists believe the human brain is just like a computer powered by 1s and 0s and the human mind to be the equivalent of software running on the brain-computer while treating the human body as an irrelevant matter to be ignored.

In [14], Gelernter argues that the mind is in a particular body, and consciousness is the work of the whole body. The mind is clearly affected by the body's age and physical status. The mind operates in different ways through the course of each day. It works one way if the body is on high alert, another on the low focus (edge of sleep or hallucination). At high alert, the mind works exactly like a computer; thinking on purpose. It calls on the memory for data and patterns and instructions necessary to perform the jobs at hand. As we move toward low focus, the mental activity changes from thinking on purpose to thought wandering off on its own. It tends to pursue meaning by inventing stories as we try to do when we dream. Surprisingly or not surprisingly, many of our most creative ideas come from minds operating in low spectrum. German Chemist Friedrich August Kekulé said that he had discovered the ring shape of the benzene molecule after having a day-dream of a snake seizing its own tail. Pioneering neuroscientist Otto Loewi envisioned an experiment that could test his theory that the brain transmits some signals chemically while sleeping in 1921 eventually led him to a Nobel Prize. This is why Gelernter jokes about "We can't have artificial intelligence until a computer can hallucinate."

Co-evolution of AI & HI

The deep analysis of AI vs. HI across the hierarchy of intelligence shows that state-of-the-art AI follows different path than HI although both seem to share some high-level promises. DNN's, leveraging the incredible computation and memory capability of computers empower AI to solve a wide range of intelligent tasks with data-driven pattern recognition. HI on the other hand, still excels in white-box causality reasoning and creativity. The initial pursuit of AI to imitate HI does not actually happen. Instead, AI shines in pattern recognition with big data which can greatly augment HI with data-driven knowledge enhancement. The discovery and use of data-driven knowledge can be regarded as the biggest contribution of AI thus far which is destined to transform entire industries. Applications will continue drive the future development of AI technologies in areas like explainable AI, white-box causality analysis and integration of symbolic & common-sense knowledge.

We should feel very lucky that we are the first generation to live together with AI because AI can enhance our intelligence/mental capability. There is no need to be worried about AI taking over human race scenarios depicted by science fictions because all AI innovations are made possible by clever algorithms from humans. It is always AI+HI all along. Humans can run but cannot run faster and longer. Thus, we invented cars to take us fast and far. Similarly, we can compute and memorize things, but we cannot compute faster and will fatigue. Thus, we invented computers to help us to implement all imaginative algorithms we can come up with. Humans will learn and adapt new power offered by AI, just as humans have leveraged previous technologies. It becomes ever clear human intelligence co-evolves with artificial intelligence. Machines enables humans to become supermen - achieving beyond what we ever dreamed of while AI also helps us to re-think about "intelligence" and eventually what we are.

REFERENCES

[1] "*Science: The Thinking Machine*", Time Magazine, Jan. 23, 1950, p54-56.

[2] Xiong, W., et al. "*Toward human parity in conversational speech recognition*", IEEE/ACM Transactions on Audio, Speech, and Language Processing 25, 12 (2017), p2410–2423.

[3] He, K., Zhang, X., Ren, S., and Sun, J. "*Deep Residual Learning for Image Recognition*", *CVPR 2016*.

[4] Wang, W., et al. "*Gated Self-Matching Networks for Reading Comprehension and Question Answering*". Proc. ACL 2017.

[5] Hassan, H.. "*Achieving Human Parity on Automatic Chinese to English News Translation*". arXiv:1803.05567, *15 Mar 2018*

[6] http://statmt.org/wmt17/translation-task.html

[7] Wang, K. and Wang, Y. "*How AI Affects the Future Predictive Maintenance: A Primer of Deep Learning*". in Advanced Manufacturing and Automation VII, p1-10. Springer; 1st ed. 2018

[8] Pearl, J. and Mackenzie, D. "*The Book of Why: The New Science of Cause and Effect*". Basic Books; 1st ed. 2018

[9] Searle, J. "*Minds, Brains, and Programs*". in Behavioral and Brain Sciences, Volume 3, Issue 3 September 1980, pp. 417-424

[10] Arons, B. "*A review of the cocktail party effect*". in Journal of the American Voice I/O Society, 1992.

[11] Zhu H.. et al. "*XiaoIce Band: A Melody and Arrangement Generation Framework for Pop Music*". KDD 2018.

[12] Chen, D., et al. "*StyleBank: An Explicit Representation for Neural Image Style Transfer*". *CVPR 2017*.

[13] http://pulpbits.net/category/brain/

[14] Gelernter, D. "*The Tides of Mind: Uncovering the Spectrum of Consciousness*". Liveright; 1st edition, 2017.

Unified Information Hiding in Compressed Domain

Assoc. Prof. KokSheik Wong

PhD, SMIEEE

General Co-Chair of APSIPA ASC 2017

Associate Professor

School of Information Technology

Monash University Malaysia

KokSheik Wong received the B.S. and M.S. degrees in computer science and mathematics from Utah State University, USA, in 2002 and 2005, respectively, and the Ph.D. degree in engineering from Shinshu University, Japan, in 2009, under the scholarship of Monbukagakusho. From 2010 to 2016, he was with University of Malaya. In 2017, he joined School of Information Technology, Monash University Malaysia, where he leads the Multimedia Signal Processing and Information Hiding Group. His current research interests include information hiding, steganography, watermarking, multimedia perceptual encryption, multimedia signal processing, and their applications. He is a member of APSIPA.

Information hiding (IH) usually refers to two seemingly unrelated fields with conflicting requirements, namely inserting data into a content, and masking the perceptual meaning of a content. However, due to recent challenges in multimedia security, both fields are jointly researched and deployed to better manage contents. This article aims to give a brief overview of unified information hiding in compressed domain, its applications and future.

IH as Data Insertion

Information hiding (IH) can be treated as a process that inserts some data into a content. Its applications include watermarking, fingerprinting, secret communication, hyper-linking related contents, tagging, etc. [1]. One common requirement of IH is to maintain the quality of the output content so that it is perceptually similar to its original counterpart. In certain applications such as crime scene / forensic photo or medical imaging, *reversibility* is also required, where the inserted data can be removed to completely restore the original content.

IH as Perceptual Masking

Another way to perceive IH is a process that hides the perceptual meaning of a content, or in other words – encryption [2]. The main requirement is to ensure that the output is completely imperceptible, and there should not be any trace of the original content. Depending on the application, an approximation or the exact version of the original content should be recoverable when the required information (viz., key) is available.

Unified Information Hiding

Both data insertion and perceptual masking are put together, hence *unified*, to complement each other to better manage a content. Take content delivery for example, buyer's fingerprint and ownership information are inserted into the data, while the processed content is encrypted prior to transmission to avoid unauthorized viewing. The same framework can be applied for cloud storage where users want to keep their contents private while the administrator needs some information about the uploaded contents to better manage them.

Unified information hiding (UIH) can be achieved in 4 ways as summarized in Fig. 1. The main challenge here is that encryption reduces redundancy in the content, which is exploited by the traditional data insertion techniques to insert data. Given a content such as image, its perceptual semantic is first masked by means of encryption. Here, the encryption algorithms are usually tailor-made to prepare for data insertion. Selected pixels are usually predicted with high accuracy, and the rooms *reserved* from prediction are exploited for data insertion. To achieve encryption, [3] overwrites the pixel value at selected positions by the data to be inserted. This achieves perceptual masking and data insertion simultaneously. The precision of the stored prediction errors is manipulated to control the quality of the reconstructed image. To generalize, [4] maps any encrypted signal to the codewords in the universal domain, and the codewords are manipulated to insert data.

Researchers then aim for more features on UIH, including separability and commutative. Here, separability refers to the flexibility to extract the inserted data before and after decryption, while commutative refers to the flexibility in achieving the same output regardless of the order of operations. A quick solution to such requirements is to divide the content into two independent groups, where one group is reserved for data insertion, while the other group is responsible for masking the image. In [5], after encrypting all pixels, the 3 least significant bit planes are manipulated to insert data by exploiting uniformity in the plaintext image. When the image is decrypted, a fairly high quality image can be obtained while remaining LSB groups still hold the inserted.

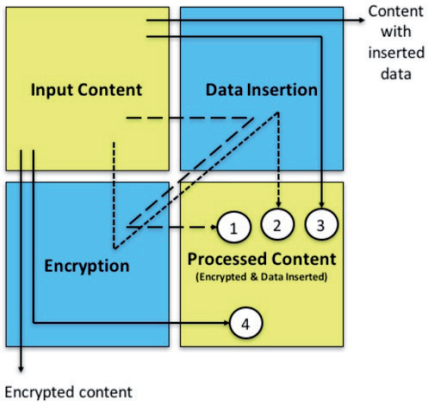


Fig. 1. Four approaches in achieving UIH

UIH and Compression

Since most content communicated nowadays are compressed by using some international standards such as JPEG for image and H.264-AVC for video, it is natural to apply UIH in compressed content. However, it adds another layer of complication because UIH can take place (a) before, (b) during, and (c) after compression. In this context, there are at least 3 performance issues to negotiate, namely, security, compressibility, and payload.

For (a), it is a known fact that encryption significantly diminishes the compression efficiency. [6] proposed an image encryption technique, where the properties exploited by JPEG compression standard are preserved. [7] then exploits the natural state of an image to insert data without compromising on security and compression efficiency with respect to JPEG. [7] is both commutative and separable.

For (b), it is also called in-the-loop processing and most video contents are handled under such scenario due to the challenges in handling context adaptive entropy coding techniques (e.g., CAVLC and CABAC). Different syntax elements are manipulated for different purposes. For example, in [8], during the video encoding under the HEVC standard, the Sign Bins, Transform Skip Bins and Suffix Bins are randomized to distort the video, while coding unit size is manipulated to insert data. Since the non-optimal parameters are considered for encoding, the compression efficiency deteriorates slightly.

For (c), the compressed content is partially decoded (up to entropy decoding) so that the syntax elements can be accessed and manipulated to achieve UIH. [9] masks the general appearance of a JPEG image by rearranging its DC values based on regions, and exploit the natural property of AC coefficients (i.e., short run of zeros and larger magnitude for lower frequency component) to insert data. While file size is mostly maintained, the technique is neither commutative nor separable. Similar to other predictive based UIH techniques in uncompressed domain, given a JPEG image [10] removes the DC coefficients and the first few low frequency AC coefficients in each block then manipulates the remaining coefficients to achieve content masking. New coefficients are then introduced to for data insertion. To recover the removed coefficients linear programming is deployed where the solution space is bounded by some constrains based on the properties of natural image.

Long-term Impact

The wide acceptance of cloud storage facility, online media store and on-demand services will further drive innovations in UIH to combat piracy, privacy infringement and authenticity related issues. Furthermore, data is being collected ubiquitously at all time (for example, array of cameras and sensors in smart city), where secure and robust UIH techniques are sought for to provide the three pillars of security, namely, confidentiality, integrity, and availability of the contents, which are particularly crucial when these contents are presented in court as evidences.

While many UIH techniques are put forward in the literature, most of them operate in the uncompressed domain. It is inevitable that most contents communicated nowadays are encoded in the compressed form, hence many most conventional techniques are impractical. Therefore, it is expected to see more UIH techniques in the compressed domain in near future.

For viability and long term development, the design of compression standards should also incorporate the mechanisms to manage content, such as perceptual masking and data insertion. The current compression standards (e.g., H.264-AVC and HEVC) assumes the traditional block-transform pipeline. Although some researchers manage to deploy UIH in the aforementioned standards, the rigid structure of such standards can be exploited by attacker to infer some perceptual information about the masked content as shown in [11]. Even worse, when the resolution is high, a rather detailed image can be sketched from each frame / slice. To achieve the desired properties in compression, data insertion and encryption, a revolutionary change in the encoding pipeline is needed. For example, machine learning techniques can be exploited to discover better basis vectors in representing signals. The remaining challenges include the deployment of the new codec so that they co-exist alongside with the legacy ones (e.g., JPEG2000 and JPEG), as well as managing the complexity of the techniques for adaptation to smart devices or sensors in performing UIH operations and compression. In additional, with the introduction of block-chain technology, information about the operations performed on a content can be recorded and verified to keep track of the current stage of processing, as well as the number of processes a content has underwent. This way, the completed history can be recorded, tracked, and verified. Imaging keeping track of every single content produced by a person throughout his / her life? This may lead to other interesting research problems, including how the ledger can be handled (i.e., externally vs internally stored via data insertion technology), the level of details to be recorded, the role of miner, compression of the ledger, privacy, to name a few.

References

[1] Y. Tew and K. Wong, "An Overview of Information Hiding in H.264/AVC Compressed Video," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 24, no. 2, pp. 305-319, Feb. 2014.

[2] K. Ratnavelu, M. Kalpana, P. Balasubramaniam, K. Wong, P. Raveendran, "Image encryption method based on chaotic fuzzy cellular neural networks," Signal Processing, vol.140, pp 87-96, Nov. 2017.

[3] R. M. Rad, K. Wong and J. M. Guo, "A Unified Data Embedding and Scrambling Method," in IEEE Transactions on Image Processing, vol. 23, no. 4, pp. 1463-1475, April 2014.

[4] Mustafa S. Abdul Karim, KokSheik Wong, "Universal data embedding in encrypted domain", Signal Processing, Vol. 94, pp. 174-182, Jan. 2014.

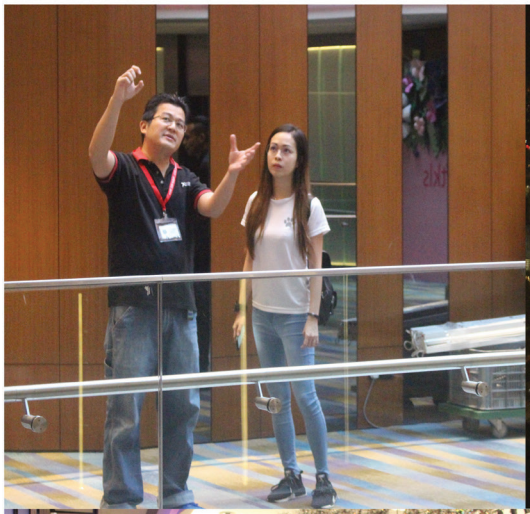
[5] X. Zhang, "Reversible data hiding in encrypted image", IEEE Signal Processing Letters, Vol. 18, no. 4, pp. 2011, 255–258.

[6] K. Kurihara, M. Kikuchi, S. Imaizumi, S. Shiota, H. Kiya, "An encryption-then-compression system for JPEG / motion JPEG standard", IEICE Transactions, vol. 98-A, no. 11, pp. 2238-2245, 2015.

[7] K. Wong and H. Kiya, "Reversible data hiding for compression-friendly image encryption method," 2017 Asia-Pacific Signal and Information Processing Association Annual

Summit and Conference (APSIPA ASC), Kuala Lumpur, 2017, pp. 1205-1209.
[8] Y. Tew, K. Wong, R. Phan, K. Ngan, "Separable Authentication in Encrypted HEVC Video," Multimedia Tools and Applications, Accepted.
[9] S. Ong, K. Wong and K. Tanaka, "Scrambling-Embedding for JPEG Compressed Image," Signal Processing, Vol., April 2015, pp. 56-68.
[10] K. Tan, K. Wong, S. Ong and K. Tanaka, "Rewritable data insertion in encrypted JPEG using coefficient prediction method," 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, 2017, pp. 214-219.
[11] K. Minemura, K. Wong, R. Phan, K. Tanaka, "A Novel Sketch Attack Framework for H.264/AVC Format-Compliant Encrypted Video," IEEE Transactions on Circuit and System for Video, Vol. 27, No. 11, Nov. 2017, pp. 2309 – 2321

Photo Gallery: APSIPA ASC'2017 in Kuala Lumpur



Robust Non-Contact Three-dimensional Measurement

There are many occasions when three-dimensional (3D) information of an object is needed. Assume that we are inspecting an internal component of an airplane turbine and a crack is found on its surface. At that time, what we are interested may not only be the color or shape of the crack, but also how deep the crack is, since it has strong implication to the reliability of the component. Indeed 3D measurement is needed in many applications such as 3D intra-oral dental measurement, human body shape measurement for shape guided surgery, facial shape 3D measurement for cosmetic surgery, endoscopic internal organ 3D inspection, 3D microscopy, etc. There are a few common requirements in these applications. First, the measurement system cannot (or preferably not to) have direct contact to the object. Second, the measurement needs to be high resolution and have high accuracy. And in some of these applications, the measurements need to be carried out quickly since the object can be moving during the course of measurement. Also, the measurement system may need to be installed on a handheld device and requires real-time performance.

In recent years, non-contact 3D scanning systems such as stereo vision camera and time-of-flight camera can be easily found even in mobile devices. Despite of the popularity of these systems, they are yet to be used in serious 3D measurement applications due to their relatively low resolution and low accuracy. Among the various non-contact 3D scanning techniques, the fringe pattern profilometry (FPP) can provide high resolution and relatively accurate 3D measurement. It is fast and allows full-field 3D scanning. With a projector-camera pair, the projector projects fringe patterns unto the target object and the deformed fringe patterns due to the object's 3D profile are captured by a camera as illustrated in Fig. 1. The 3D profile of the object can thus be obtained by measuring the amount of deformation of the fringe, which can be achieved by analyzing the phase shift of the fringes against a reference obtained in the initial calibration.

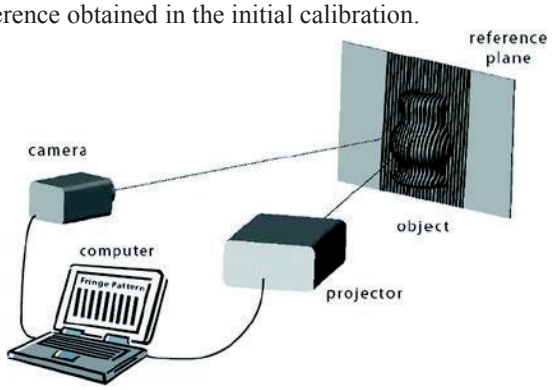


Fig. 1. A conceptual illustration of the setup of a fringe projection profilometry system.

There were many FPP methods developed in the last few decades to realize such analysis. However, the robustness of these methods is always a concern when using in serious applications (such as the ones mentioned in the first paragraph). Since many FPP systems cannot guarantee both the measurement device and the measuring objects to be totally static, it is required that a 3D measurement can be achieved with a single projection and single image capture. That is, we need single-shot FPP methods. However, most single-shot FPP methods have problem in their robustness. The three common sources of problem include,

(i) texture and color of the object;
(ii) highlights on the object due to illumination; and
(iii) discontinuities of the fringes due to multiple objects or occlusion.

For FPP, we only need the phase shift of the fringes to obtain the 3D profile of the object. If the object has vivd texture or color on its surface, bias of rapidly changing magnitude will be added to the fringes and makes the estimation of the phase shift far more difficult. Fig. 2 shows a slice of a fringe image obtained from an object with color stripes on its surface. One can see the rapidly changing *dc* level of the fringe signal. Also, for some objects made of highly reflective material, highlights are often found on their surface. They are the reflection of the global illumination made to the object. Such highlights overwhelm the fringes

Dr Daniel P.K. Lun

PhD, FIET, SMIEEE

Program Co-Chair, APSIPA ASC 2015

Associate Professor and Interim Head

Department of Electronic and Information Engineering, The Hong Kong Polytechnic University

Daniel P.K. Lun *received his B.Sc.(Hons.) degree from the University of Essex, U.K., and Ph.D. degree from the Hong Kong Polytechnic University (formerly called Hong Kong Polytechnic) in 1988 and 1991, respectively. He is now an Associate Professor and Interim Head of the Department of Electronic and Information Engineering of the Hong Kong Polytechnic University. Dr Lun is active in research activities. He has published more than 130 international journals and conference papers. His research interest includes signal and image enhancement, sparse representation and applications, and 3D data measurement and processing. He was the Chairman of the IEEE Hong Kong Chapter of Signal Processing in 1999-00. He was the General Co-Chair of DSP 2014, and Technical Co-Chair of APSIPA ASC 2015. He is an organizing committee member of a number of international conferences, including ICASSP 2003, ICIP 2010 and ICME 2017. Dr Lun is a member of the Digital Signal Processing and Visual Signal Processing and Communications Technical Committees of IEEE Circuits and Systems Society. Dr Lun is the recipient of a number of awards, including three best paper awards. He is an associate editor of IEEE Signal Processing Letters. He is a Chartered Engineer, a Fellow of IET, a corporate member of HKIE and a senior member of IEEE.*

such that only a bright spot can be seen. An example of the highlight in a fringe image can be found in Fig. 3. As the objects' 3D profile is estimated from the phase shift of the fringe, the regions with only highlight but no fringe will have great difficulty in estimating their 3D profile. Another source of problem comes from the discontinuities in the fringes. Due to the possibly irregular shape of the object, it is inevitable that some parts of the fringe projection are blocked or occluded by the object itself or by other objects in the scene. An example is shown in Fig. 4. Due to the position of the camera, the fringes in region B are blocked by region C. The missing fringes in region B certainly will introduce error to the measurement result.

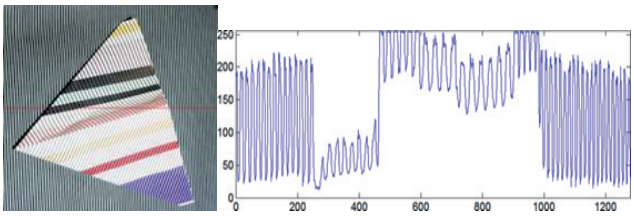


Fig. 2. Bias in the fringe signal due to the color stripes on the surface of the object. The fringe signal is taken along the red line of the image.

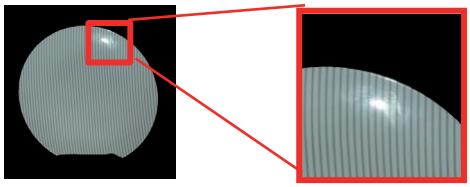


Fig. 3. Highlight in a fringe image. The fringes are overwhelmed by the bright spot.

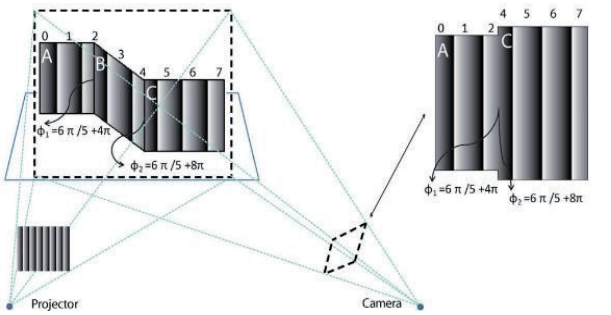


Fig. 4. Fringe discontinuity problem. Some fringes are blocked by the object thus cannot be found in the captured fringe image.

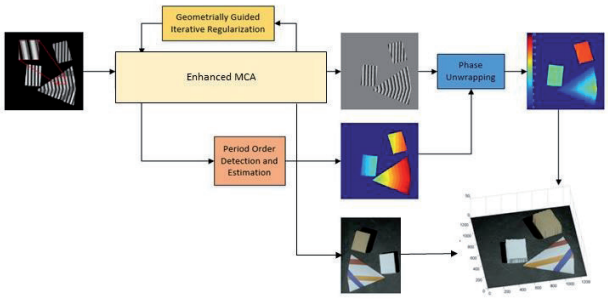


Fig. 5. Unified framework for robust single-shot FPP.

Our team has been working on FPP based 3D measurement for some years. With respect to the above problems, a unified framework was developed as shown in Fig. 5. The framework starts with a sub-system that extracts the deformed fringe pattern (thus separates the bias) using an enhanced morphological component analysis (MCA) technique. To estimate the fringes in the highlight regions, a geometrically guided iterative regularization sub-system is added. Finally, we introduce a period order detection and estimation sub-system to make sure the missing fringes due to occlusion will not affect the 3D measurement result. The details of these sub-systems can be found in our papers [1-11]. Here we would like to have a brief introduction of them.

In an FPP process, the captured fringe images can be modeled as the superimposition of the projected fringe patterns on the texture of the object. As explained above, the object's texture can introduce great difficulty to the analysis of the fringe's phase shift. Traditional solutions try to remove the texture in the fringe image but without success particularly when the texture has drastic changes in pattern (an example is shown in Fig. 2). In fact, in many diagnostic applications, the texture is also very important for different inspection functions. Rather removing the texture, a better strategy is to separate the texture from the fringe pattern such that both of them are available for subsequent uses. To do so, we proposed in [10, 11] an effective single-shot FPP algorithm that allows the object texture and fringe pattern to be estimated simultaneously. The heart of the proposed algorithm is an enhanced MCA tailored for FPP problems. Conventional MCA methods which use a uniform threshold in an iterative optimization process are inefficient to separate fringe-like patterns from image texture. We extend the conventional MCA by taking advantage of the low-rank structure of the fringe's sparse representation to enable an adaptive thresholding process. It ends up with a robust single-shot FPP algorithm that can extract the fringe pattern even if the object has a highly textured surface. It has a side benefit that the object texture can be simultaneously obtained in the fringe pattern estimation process. Fig. 6 shows an experimental result which demonstrates the improved performance of the proposed algorithm over the conventional single-shot FPP approach.

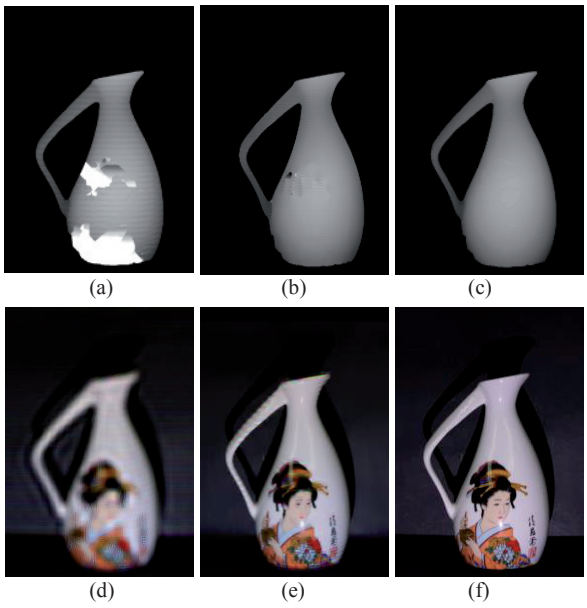


Fig. 6. Measured 3D profile and extracted texture image. (a) and (d) Results of the conventional approach; (b) and (e) results of the proposed approach; (c) and (f) the ground truth.

As to the highlight problem, we proposed in [6, 7] a novel inpainting algorithm to restore the fringe images in the presence of highlights. The proposed method first detects

the highlight regions based on a Gaussian Mixture model (GMM). Then an automatic geometric sketching of the missing fringes is carried out and the result is used as the initial guess of an iterative regularization procedure for regenerating the missing fringes. Simulation and experimental results show that the proposed algorithm can accurately reconstruct the 3D model of objects even when their fringe images have large highlight regions. It significantly outperforms the traditional approaches in both quantitative and qualitative evaluations. Fig. 7 shows that iterative inpainting process of the fringes in a region that is originally overwhelmed by highlight. Fig. 8 shows the 3D profile measured by the proposed approach as compared with the conventional one. Note that the error due to the highlight is largely recovered by the proposed method.

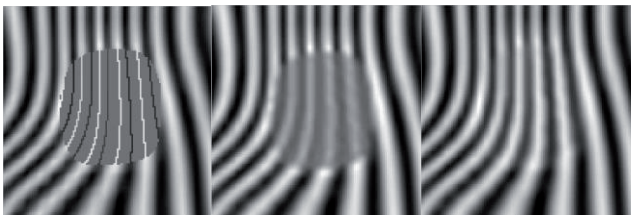


Fig. 7. The iterative estimation of the fringes inside the highlight region.

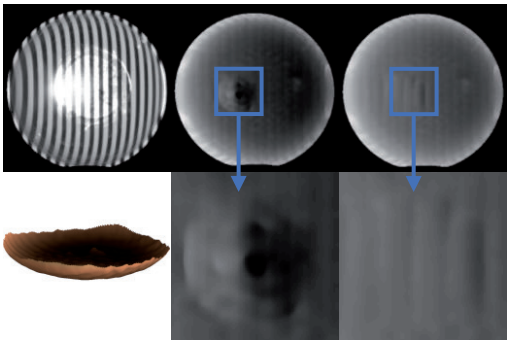


Fig. 8. Performance of the proposed inpainting method on a melamine plate with large highlights. (1st column) The fringe image with highlight and the 3D model reconstructed using the proposed method; (2nd column) the 3D profile recovered using the conventional approach; and (3rd column) the proposed approach.

As mentioned above, the FPP method achieves the 3D measurement by evaluating the amount of phase shift of the fringe pattern. Similar to many optical measurement methods, FPP can only provide a module-2π estimation of the phase, which is the so-called wrapped phase. Additional phase unwrapping procedure is needed to recover the true phase shift from the wrapped phase. For conventional FPP method, the phase unwrapping procedure is performed by assuming that the *Itoh* condition is held true. It suggests that the true phase can be estimated by integrating the wrapped phase differences. However, such assumption is invalid in the case that there is discontinuity in the fringe due to occlusion for instance. An example is shown in Fig. 4. As shown in the figure, only period 0, 1, 2, 4, 5, 6 and 7 are obtained; period 3 is missing. So the integration from period 2 to 4 must have error and the error will propagate to the integration of later periods. To solve this problem, we proposed in [5, 8, 9] two approaches to embed the period order numbers into the fringes to facilitate phase unwrapping even when some part of the fringes is missing due to occlusion. In the first approach [5], markers in the form of impulses are embedded in the sinusoidal fringe. The positions of the impulses encode the period order number of the fringe. While the impulse markers are easy to detect, misdetection can occur when the captured image is noisy or the object itself also has similar impulsive pattern. A more robust marker is needed. In [8, 9], we proposed a novel approach which encodes the period order numbers with some texture patterns and embeds them into the fringe patterns. When the encoded fringe image is captured, a modified MCA procedure and a sparse classification procedure are performed to decode and identify the embedded period order information. It is then used to assist the phase unwrapping process to deal with the different artifacts in the fringe image. Experimental results show that the proposed algorithm can significantly improve the robustness of an FPP system. It performs equally well no matter the fringe image has a simple or complex scene, or are affected due to the ambient lighting of the working environment. Fig. 9 shows the texture patterns and how they are embedded into a fringe pattern. Fig. 10 shows the performance of the proposed method when applying to a complex scene containing multiple objects. Note that, in Fig. 10, the handle of the jar is occluded by the head sculpture. Since the period order numbers are correctly decoded from the fringes, the 3D measurement around the handle has no error at all. More comparison results of the proposed algorithm can be found in [8, 9].

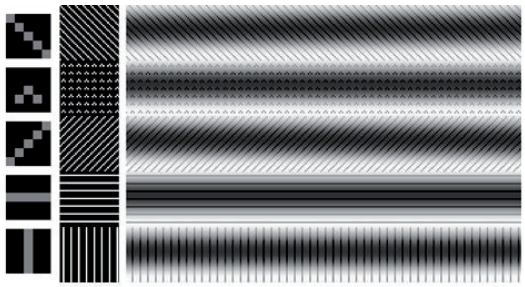


Fig. 9. 5×5 pixel binary textons (left); the code pattern generated by the corresponding texton (2nd column); and the encoded fringes (3rd column).

Fig. 10. A scene of complex objects (left); result of 3D measurement using the proposed algorithm (middle); and the top view (right).

Long-term Impact

The abovementioned approaches are effective to facilitate robust and high accuracy 3D measurements. However, many of these algorithms require time consuming iterative optimizations which hinder their application to practical systems, which often require real-time performance. Recently, we are working on replacing these optimization processes by some learning approaches and realizing them using deep neural networks (DNN). DNN has achieved great success in different signal and image processing applications in recent years. Particularly due to its regular structure, it can be easily implemented with graphics processing unit (GPU) and achieve massive parallelization. With the use of different DNN techniques, our initial experimental results [12] show that we can complete a 3D measurement with the abovementioned algorithms within a few hundred milliseconds. It opens up the possibility of high

resolution and high accuracy 3D measurement of moving objects, which is critical for many applications that involve live subjects, such as medical applications. The proposed algorithms provide the much needed robustness when working under undesired working conditions, such as when the target object has vivid textures, highlights due to global illumination, or occlusion by another object. The proposed approaches will make a great impact to the design of future 3D measurement devices.

References

[1] Tai-Chiu Hsung, Daniel P.K. Lun, and William W.L. Ng, "Zero spectrum removal using joint bilateral filter for Fourier transform profilometry", Proceedings, 2011 IEEE Visual Communications and Image Processing (VCIP), Tainan, Taiwan, Nov 2011, pp. 1-4. (DOI: 10.1109/VCIP.2011.6115948)

[2] W. W.-L. Ng and D. P.-K. Lun, "Effective bias removal for fringe projection profilometry using the dual-tree complex wavelet transform," *Appl. Opt.,* vol. 51, no. 24, pp. 5909-5916, 2012.

[3] T.-C. Hsung, D. P.-K Lun, and W. W.-L. Ng, "Efficient fringe image enhancement based on dual-tree complex wavelet transform," *Appl. Opt.,* vol. 50, no. 21, pp. 3973-3986, 2011.

[4] William W.L. Ng and Daniel P.K. Lun, "Image enhancement for fringe projection profilometry," Proceedings, 2013 IEEE International Symposium on Circuits and Systems (ISCAS), Beijing, China, pp. 729-732, 2013.

[5] B. Budianto, D. P.-K. Lun, and T.-C. Hsung, "Marker encoded fringe projection profilometry for efficient 3D model acquisition," *Appl. Opt.,* vol. 53, no. 31, pp. 7442-7453, 2014.

[6] B. Budianto and Daniel P.K. Lun, "Inpainting For Fringe Projection Profilometry Based on Iterative Regularization," Proceedings, 19th International Conference on Digital Signal Processing (DSP2014), Hong Kong, pp. 668-671, 2014.

[7] Budianto and Daniel Pak-Kong Lun, "Inpainting for Fringe Projection Profilometry Based on Geometrically Guided Iterative Regularization", *IEEE Trans. Image Process.,* vol.24, no.12, pp. 5531-5542, 2015.

[8] B. Budianto and Daniel P.K. Lun, "Gabor Feature Based Discriminative Dictionary Learning for Period Order Detection in Fringe Projection Profilometry", Proceedings, 2015 Asia Pacific Signal and Information Processing Summit and Conference, Hong Kong, pp.283-288, 2015.

[9] Budianto and Daniel Pak-Kong Lun, "Robust Fringe Projection Profilometry via Sparse Representation," *IEEE Trans. Image Process.,* vol. 25, no. 4, pp. 1726-1739, 2016.

[10] B. Budianto, D.P.K. Lun and W.P. Zhu, "Robust Single-Shot Fringe Pattern Projection for Three-dimensional Measurements", 2017 International Conference on Digital Signal Processing, London, U.K., pp. 1-4, August 2017.

[11] Budianto, Daniel P.K. Lun and Yuk-Hee Chan, "Robust Single-shot Fringe Projection Profilometry Based on Morphological Component Analysis", *IEEE Trans. Image Process., (accepted)*

[12] B. Budianto and D. P.-K. Lun, "Deep Learning Based Period Order Detection in Fringe Projection Profilometry," 2018. Available: https://doi.org/10.1109/APSIPA.2015.7415521.

Photo Gallery: APSIPA-ASC'2015 in Hong Kong

Integrating Adaptive Auditory Models into Deep Learning

The human auditory system operates on speech over a wide range of signal-to-noise ratios and a large dynamic range. Speech processing systems have attempted to emulate this; however high fidelity adaptive auditory models have generally been too computationally complex and led to limited success. The emergence of deep learning in speech processing has led to significant advances, but these advances are now plateauing and the lack of interpretability is impeding progress. The confluence of the auditory modelling and deep learning paradigms represents a major opportunity to develop a continuously adaptive and interpretable deep-learning front-end that incorporates desirable properties of the human auditory system.

The Peripheral Auditory System

The peripheral auditory system and in particular the cochlea can be viewed as a real-time spectrum analyser. The primary role of the cochlea is to transform the incoming complex sound wave at the ear drum into electrical signals. The cochlea is divided along its length by the basilar membrane which partitions the cochlear into two fluid canals and when unrolled, it would be around 3.5cm long (See Figure 1). The basilar membrane is a resonant structure that vibrates, vertically in sympathy with pressure variations in the cochlear fluid. Vibrations evoked by a single tone appear as travelling waves that propagate down the cochlea and reach maximum amplitude at a particular point before slowing down and decaying rapidly. Each point along the basilar membrane has a characteristic frequency to which it is most responsive, with the maximum membrane displacement occurring at the basal end for high frequencies and at the apical end for low frequencies [1]. The whole transformation can be seen as a process of a real time spectral decomposition of the acoustic signal to produce a spatial frequency map in the cochlea.

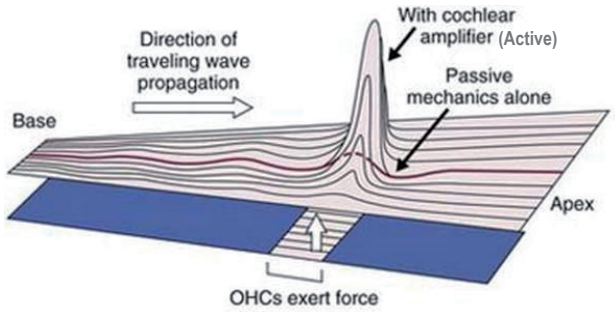


Figure 1: Uncoiled basilar membrane showing a travelling wave along it [1]

Running along the length of the membrane is the Organ of Corti, which contains one row of inner hair cells (IHC) and three rows of outer hair cells (OHC). The deflections of the basilar membrane cause the cilia of the inner hair cells to bend, which in turn triggers the hair cells to send nerve impulses to the auditory path way along afferent nerve fibres (Figure 2). The outer hair cells can act as motor units that amplify the movement of the basilar membrane, as shown in Figure 1, in response to a stimulus from the higher auditory processing system carried down efferent nerve fibres.

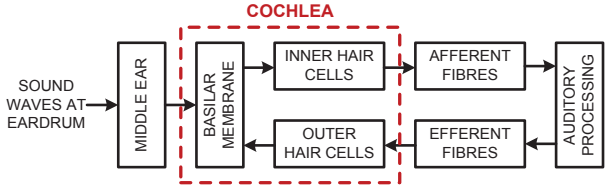


Figure 2: Block diagram of human auditory system including active feedback via outer hair cells

Professor Eliathamby Ambikairajah

PhD, FIET, FIEAust

APSIPA Distinguished Lecturer (2013-14), Advisory BoardMember (2015-2016)

Head of School

Electrical Engineering and Telecommunications

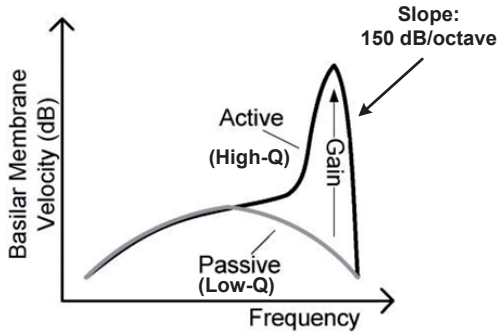
University of New South Wales (UNSW), Sydney, Australia.

Professor Eliathamby Ambikairajah received his BSc(Eng) (Hons) degree from the University of Sri Lanka and received his PhD degree in Signal Processing from Keele University, UK. He was appointed as Head of Electronic Engineering and later Dean of Engineering at the Athlone Institute of Technology in the Republic of Ireland from 1982 to 1999. He was an Invited Research Fellow with the British Telecom Laboratories, U.K., for ten years from 1989 to 1999.

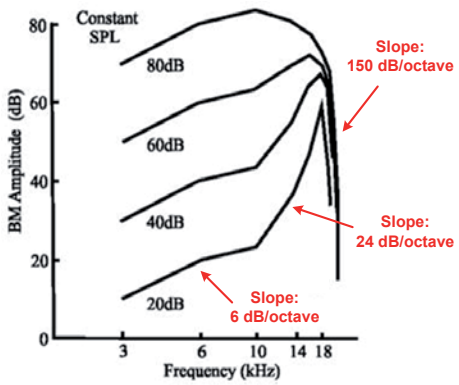
*His research interests include speaker and language recognition, emotion detection and biomedical signal processing. He has authored and co-authored approximately 300 journal and conference papers and is the recipient of many competitive research grants and a number of best paper awards. He was invited as a Visiting Scientist to the Institute of Infocomms Research (A*STAR), Singapore in 2009, where he is currently a Faculty Associate. He is currently an Associate editor for IEEE Education.*

He received the Vice-Chancellor's Award for Teaching Excellence in 2004 for his innovative use of educational technology, the School Awards for Teaching Excellence in 2003, Academic Management in 2001 and in 2014 he received the UNSW Excellence in Senior Leadership Award.

The auditory system has evolved to carry out this spectral mapping under a very wide range of acoustically adverse conditions. For instance, the human auditory system can process a vast range of sounds spanning some twelve orders of magnitude of input pressure intensity. Research has revealed [2] that in order to achieve this, the cochlea makes use of both passive and active systems. The passive system is operational at normal stimulus amplitudes, while the active system comes into play when the ear is presented with a low-amplitude stimulus. Specifically, the basilar membrane within the cochlea is normally in a passive state (low-Q spectral decomposition), but upon stimulation by a frequency of low amplitude, the section of the basilar membrane corresponding to that frequency transitions to an active state (adaptively higher-Q spectral decomposition) (Figure 3a). In this state, the experimentally-observed increased sensitivity is provided by some active mechanism feeding energy into the basilar membrane [3].



(a)



(b)

Figure 3: (a) Adaptive frequency selectivity in a cochlea with an active gain around 40dB [4]; (b) frequency response of the BM displacement, for various levels of acoustic excitation at one point measured by BM Johnstone et al (1986) [9]

The outer hair cells (OHC) provide this active mechanism and while details of their operation is still subject to active research [5], it is clear that they amplify the motion picked up by the IHC at low input sound levels at that frequency [3]. There are about 12,000 OHCs in the human cochlea and they each act through this mechanism as local feedback controllers of vibration. It is surprising how this large number of locally acting feedback loops can act together to give a large and uniform amplification of the global response of the BM. It is also remarkable how quickly the OHCs can act, since they can respond at up to 20 kHz in humans and 200 kHz in dolphins and bats [6]. This continuously adaptive operation of the cochlea helps ensure that the neuronal representation of the sounds is relatively invariant across a much greater dynamic range of input sounds than anything state-of-the-art machine based speech processing systems are capable of handling. In addition to the ability to handle input sounds in an extremely large dynamic range, measurements on human and other mammalian auditory systems have shown that the cochlea also has extraordinary frequency sensitivity and selectivity (Figure 3) as a result of this active feedback loops [7] and this has been confirmed by models of the mechanics of the cochlea that incorporate excitation by the OHCs [8].

Ideally, models of the cochlea should exhibit level-dependence, sharp auditory tuning curves and fast adaptation to changes in the input stimuli. These properties of the actual cochlea can be seen in Figure 3b which shows high frequency selectivity (High-Q) for low-level input stimuli and broad selectivity for high-level input stimuli as measured by Johnstone et al. [9]. However, modelling these characteristics is computationally expensive, and almost universally, speech processing systems model only the non-linear frequency scaling of the cochlea and ignore these adaptive characteristics [1].

Spectral decomposition modelled on the human auditory system (typically implemented as a simple parallel filterbank) has formed the front-end of the vast majority of speech processing systems [10]. State-of-the-art end-to-end deep learning systems have also been developed along similar lines, with the first layer typically comprising convolutional units that implement a bank of filters, with these filters being learnt from the training data [11]. However, these banks of fixed filters are time-invariant and lack the adaptive properties of the human peripheral auditory systems, which provide a number of desirable properties that are beneficial under adverse noise conditions, specifically level-dependent frequency sensitivity and selectivity that is adaptive to changes in the input stimuli [12].

It would be desirable to have an active model of the cochlea that incorporates the level-dependent adaptive gain and adaptive frequency selectivity properties, including adaptive non-linear compression of up to 3:1, into the deep learning paradigm that can serve as the single flexible future front-end for all speech processing systems.

Integrating Auditory Models into a Deep Learning System

How can one design a front-end that can provide a truly robust speech representation that can be used by any speech processing system under all realistic conditions? Keeping in mind that the human auditory system accomplishes this goal very well and that state-of-the-art deep-learning systems provide the means to train various kinds of speech processing systems, and there is a need to integrate the desirable properties of the human cochlea and auditory system into the deep-learning framework.

Current deep learning systems are trained to handle adverse conditions and noise by including data from

adverse conditions in the training set. The challenge here is to introduce novel *knowledge-driven* deep architectures that emulate the traits of the human auditory system and combine them with the *data-driven* learning capabilities of current deep learning systems. From a model estimation point of view, this will address the problems of training very large and complex models to cater to previously unseen adverse conditions by imposing prior knowledge about the human auditory system on the model space. This in turn will allow for the training of systems with a large number of hidden layers, which in turn can help provide greater levels of abstraction, and thus robustness, in the speech representation.

Conceptually this combines the strengths of auditory models, namely high fidelity and robustness to noise, with the strengths of the deep learning approaches, namely the ability to learn complex relationships and relative computational efficiency.This integration of active auditory modelling with end-to-end systems within a deep learning framework can involve three elements as suggested in Figure 4. Namely,

- A feed-forward adaptive spectral decomposition model based on the cochlea;
- A back-end dependent feedback path to improve the adaptive spectral decomposition;
- Extending the end-to-end system to learn a channel-invariant speech signal representation.

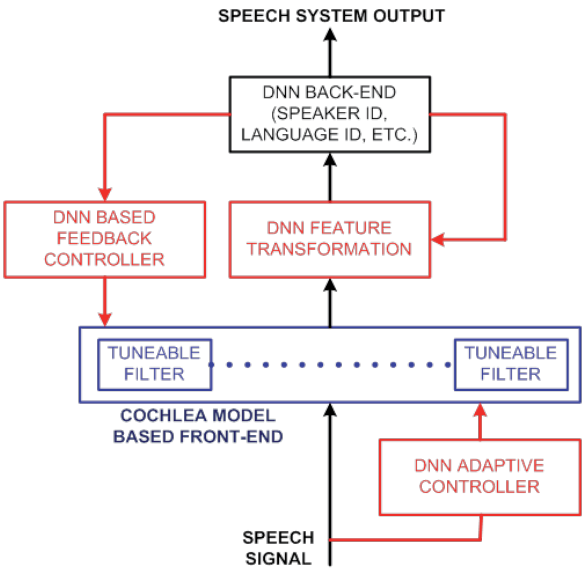


Figure 4: Overview of the three elements involved in integrating auditory models with deep learning based systems

Long-term Impact

With the rapid growth and uptake of AI and machine learning systems, speech interfaces between people and machines will be increasingly adopted. This can already be seen in the widespread adoption of smart home devices and smartphone assistants such as Google Home, Amazon Alexa, Apple Siri, etc. The speech processing capabilities of these systems are however still orders of magnitude worse than human capabilities, particularly under noisy environments and other adverse conditions (such as reverberation). Integrating adaptive auditory models into these speech systems will provide them with the adaptive signal conditioning capabilities, driven by suitable machine learning back-ends, emulating the human peripheral auditory system to operate under similar adverse conditions. It is also expected that in the future, insights obtained from these systems can be transferred to the next generation of cochlear sound processing hardware such as cochlear implants, which have to date relied on passive spectral analysers in their front-ends. For example, a major area of speech processing where the integration of adaptive auditory models may provide significant benefits is voice biometrics, and in particular the analyses and detection of spoofing attacks on voice biometric systems.

References

[1] J.McGee and E. J. Walsh, "Cochlear Transduction and the Molecular Basis of Auditory Pathology," in Cummings Otolaryngology-Head and Neck Surgery E-Book, P. W. Flint et al., Eds.: Elsevier Health Sciences, 2014.

[2] G. Ni, S. J. Elliott, and J. Baumgart, "Finite-element model of the active organ of Corti", Journal of The Royal Society Interface, vol. 13, no. 115, 2016.

[3] S. J. Elliott and C. A. Shera, "The cochlea as a smart structure", Smart Materials and Structures, vol. 21, no. 6, 2012.

[4]C. J. Plack, “The sense of hearing”, Psychology Press, 2013.

[5] G. Zweig, "Linear cochlear mechanics," The Journal of the Acoustical Society of America, vol. 138, no. 2, pp. 1102-1121, 2015.

[6] G. Ni, S. J. Elliott, M. Ayat, and P. D. Teal, "Modelling cochlear mechanics," BioMed research international, vol. 2014, 2014.

[7] S. Camalet, T. Duke, F. Jülicher, and J. Prost, "Auditory sensitivity provided by self-tuned critical oscillations of hair cells," Proceedings of the National Academy of Sciences, vol. 97, no. 7, pp. 3183-3188, 2000.

[8] M. LeMasurier and P. G. Gillespie, "Hair-cell mechano-transduction and cochlear amplification," Neuron, vol. 48, no. 3, pp. 403-415, 2005

[9] B. Johnstone, R. Patuzzi, and G. Yates, "Basilar membrane measurements and the travelling wave," Hearing research, vol. 22, no. 1-3, pp. 147-153, 1986.

[10] R. M. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition", *IEEE signal processing magazine,* vol. 29, no. 6, pp. 34-43, 2012.

[11] H. B. Sailor and H. A. Patil, "Novel unsupervised auditory filterbank learning using convolutional RBM for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 12, pp. 2341-2353, 2016.

[12] D. S. Kim, S.Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," IEEE Transactions on speech and audio processing, vol. 7, no. 1, pp. 55-69, 1999

Blind Bandwidth Extension of Audio Signals

Professor Changchun Bao

PhD, SrMIEEE

Chair of APSIPA SLA TC (2015-2016)

Professor of Speech and Audio Signal Processing
Faculty of Information Technology
Beijing University of Technology

Changchun Bao *received the Ph.D. degrees in communication and electronic system from JiLin University (former JiLin University of Technology) in 1995. From 1995 to 1997, he was a Postdoctoral Research Fellow in Xidian University. He joined Beijing University of Technology as an Associate Professor at the end of 1997, and was promoted to a full professor in 1999. He is a senior member of IEEE and was the Chair of APSIPA SLA TC from 2015 to 2016. His research interests are in the areas of speech & audio signal processing, speech coding, speech enhancement, speech transcoding, audio coding, audio enhancement, bandwidth extending for speech and audio signals and 3D audio signal processing. He is the co-author of over 290 papers in journals and conferences and holds 10 patents. He was ever an Associate Editor for the Journal on Communications, and currently an Editor for Signal Processing and an Editor for Journal of Data Acquisition & Processing. Prof. Bao is also a Board and Senior Member of Chinese Institute of Electronics (CIE), a Board member of the Acoustical Society of China (ASC), a Board Member of Signal Processing Academy of CIE and a member of International Speech Communication Association (ISCA).*

The frequency range of full-band audio signal generally varies from 20 Hz to 20 kHz, which is commonly adopted for high-fidelity sound reproduction. However, due to the limitations of network transmission rate and data processing ability, existing communication systems usually limit the bandwidth of the transmitted audio signals and only reproduce the low-frequency (LF) components in order to increase coding efficiency. For speech communication, wideband (WB) signals, which are sampled at 16 kHz with the bandwidth of 50~7000 Hz, can provide excellent intelligibility and subjective quality. But when music signals are transmitted on wireless channel, the auditory quality of WB audio signals is not satisfactory, and people are looking forward to enjoying the higher-quality audio service. A significant enhancement in audio quality can be gained by transmitting super-wideband (SWB) audio signals with the bandwidth of 50 to 14000 Hz or more.

To minimize the difference of audio quality between WB and SWB, bandwidth extension (BWE) could artificially recover the high-frequency (HF) components at the receiver from the decoded WB audio signals. For non-blind BWE methods, the time-frequency energy of audio signals is first extracted at the encoder. Then, the proper spectral patching in the light of correlations between LF and HF spectra is determined. Finally, the energy and control parameters are transmitted to the decoder as side information for reconstructing the HF components. Non-blind BWE methods can provide reproduced signals with a high quality, but all the audio transmission system is required to support side information for BWE. Whereas, blind BWE could artificially restore the truncated HF components at the decoder by building a statistical relationship between the LF and HF components. The advantage of blind BWE is that it can avoid any modifications inside the source coding and the network transmission process. In recent decades, many blind BWE solutions have been developed for speech and audio signals, and these typical BWE methods can be summarized to perform two main tasks, namely, the estimation of the spectral envelope and the extension of the fine spectrum. Informal listening test results indicate that the estimation accuracy of the HF spectral envelope is crucial to the improvement of auditory quality for the reproduced signals. Therefore, most BWE approaches have concentrated on modelling the mapping relationship between LF and HF spectral coefficients based on statistical learning methods and further estimated the HF spectral envelope under some error criterions without any auxiliary information. Based on this mapping relationship, we proposed several efficient blind BWE methods for audio signals.

Blind BWE Based on Non-linear Prediction and HMM

In this method, we used nonlinear prediction to implement audio bandwidth extension from WB to SWB in the frequency domain [1]. Firstly, the LF fine spectrum separated from WB audio was converted into a multi-dimensional space using a state space reconstruction (SSR). According to dynamical system theory, the trajectory in the reconstructed state space is completely equivalent to the original audio system in terms of diffeomorphism, and the point in state space shares the similar evolving behaviors with its nearest neighbors. Inspired by these, a nonlinear prediction based on nearest-neighbor mapping (NNM) was employed to restore the fine spectrum of the high frequencies. The nearest neighbor of the given state point was selected from the state points of the LF components, and the evolving trajectory of the nearest neighbor was used to substitute the evolution of the given point for further predicting the unknown HF fine spectrum. Moreover, the hidden Markov model (HMM) was applied in the spectral envelope extension of the high frequencies. By exploiting the state transition process of HMM, the temporal correlation between adjacent frames can be captured to make the spectral envelope of the extended audio signals smoother over time and better-matched to the original ones.

This is beneficial to the auditory quality of the extended audio signals. Then, a minimum mean square error (MMSE) estimator based on HMM was utilized to estimate the spectral envelope of the HF components. Finally, the HF components were regenerated by appropriately shaping a recovered fine spectrum and were combined with the original wideband audio to form a SWB audio signal with a bandwidth of 14 kHz. The proposed method has been applied to the ITU-T G.722.1WB audio codec for comparison with the ITU-T G.722.1C super WB audio codec. Objective quality evaluation results indicated that the proposed method is preferred over the reference bandwidth extension methods. Subjective listening results showed that the proposed method has a comparable audio quality with G.722.1C and improves the extension performance compared with the reference methods.

Blind BWE Based on Temporal Smoothing Cepstral Coefficients

In this method, we embed the extraction of the temporal relationship of audio signals into cepstral coefficients and proposed a temporal smoothing cepstral coefficient (TSCC) based scheme for BWE of audio signals [2]. It improved the temporal smoothness of the extended HF spectrum, without increasing the dimension of input feature and without burdening the storage space and computational complexity for enhanced/more sophisticated statistical models. In this work, firstly, a Gammatone filter bank was adopted to decompose the audio signals, and the audio signal energy of each frequency band was long-term smoothed by means of minima controlled recursive averaging (MCRA) to suppress transient signal components. Secondly, the resulting "steady-state" spectrum was processed by frequency weighting, and TSCCs were extracted by means of the power-law loudness function and cepstral normalization. Finally, the extracted cepstral coefficients were applied into a GMM-based BWE scheme to restore the HF components. The proposed method suppressed the transient components existing in the WB audio signal in the BWE frontend, and provided higher mutual information between the LF and HF parameters. Informal listening tests showed that, for most audio signals, TSCC played an important role on the improvement of spectral distortion and the temporal smoothness of the reconstructed audio signals, while for some rock music with the accompaniment of strong percussion music temporal envelope modification needs to be applied in order to maintain a good extension performance. The proposed bandwidth extension method has been applied into the ITU-T G.729.1 wideband codec and outperformed the Mel frequency cepstral coefficient (MFCC)-based method in terms of log spectral distortion (LSD), cosh measure, and differential log spectral distortion. Further, the proposed method improved the smoothness of the reconstructed spectrum over time and also gained a good performance in the subjective listening tests

Blind BWE Based on Phase Space Reconstruction

In this work, a blind high-frequency reconstruction method of audio signals based on phase space reconstruction (PSR) and non-linear prediction was proposed [3], where PSR was used to convert the LF modified discrete cosine transform (MDCT) coefficients of wideband audio to a multi-dimensional space. A non-linear prediction model was built up in the phase space to simulate the hidden relationship between the given phase points and the unknown MDCT coefficients. By using this model, we can restore the audio spectrum of the HF components from the LF components in the phase space by building a multidimensional phase space from a one-dimensional audio signal. In order to improve the performance, the energy and harmonic components of the reconstructed HF spectrum were further adjusted according to listening perception. The objective and subjective evaluations showed that the proposed method achieves a better performance than linear extrapolation (LE) method and was comparable to the efficient high-frequency bandwidth extension (EHBE) method.

Blind BWE using ensemble of recurrent neural networks

In this method, we introduced a continuous state space model into spectral envelope extension, and proposed a blind BWE method based on ensemble of recurrent neural networks [4]. In this work, the feature space of wideband audio was firstly divided into different regions through clustering. For each region in the feature space, a specific recurrent neural network with sparsely connected hidden layer, referred as the echo state network (ESN), was employed to dynamically model the mapping relationship between wideband audio features and high-frequency spectral envelope. Different from the traditional fully connected recurrent neural networks, ESN adopted the sparsely connected hidden layer in which the connectivity and weights of hidden neuron nodes were fixed and randomly assigned. The recurrent nature of the connections turned the time varying WB audio features into specific temporal patterns in high dimension. Then, a simple linear transformation was used to map the state of hidden neuron nodes to the desired HF spectral envelope. Given a rich collection of recurrently connected nodes, ESN was able to effectively fit the nonlinear mapping function between WB audio features and HF spectral envelope. Next, the outputs of the parallel ESNs for different regions in the feature space were weighted and fused by means of network ensemble techniques, so as to form ensemble echo state networks (EESN) for further estimating the HF spectral envelope. Finally, combining with the HF fine spectrum extended by spectral translation, the proposed method effectively extended the bandwidth of WB audio to SWB, and upgraded the subjective and objective quality of WB audio signals. Objective evaluation results showed that the proposed method outperformed the hidden Markov model-based bandwidth extension method on the average in terms of both static and dynamic distortions. In subjective listening tests, the results indicated that the proposed method was able to improve the auditory quality of the wideband audio signals and outperformed the reference methods.

Blind BWE Based on Ensemble Echo State Networks with Temporal Evolution

In this blind BWE work, we estimated the HF spectral envelope based on an ensemble echo state network with

temporal evolution (TE-EESN) [5]. First the feature space of audio signals was divided into several regions through clustering similar to [4]. Within each region the dynamic mapping relationship between WB audio features and HF spectral envelope was modeled by using one specific echo state network (ESN) as well, which is a particular realization of neural networks with recurrent structure. Additionally, the transition process among regions over time was approximated by an HMM. Then, by using network ensemble techniques, the outputs of multiple ESNs were weighted and fused according to the posterior probabilities derived from an HMM so as to obtain the estimation of the HF spectral envelope. Finally, by combining the HF fine spectrum extended by spectral translation, the proposed method effectively extended the bandwidth of WB audio to SWB. In the proposed BWE method, ESNs and HMM were separately used to build up two independent dynamic processes to describe the temporal evolution of audio signals. It is beneficial to the improvement of subjective and objective quality for the extended audio. The proposed extension method has been applied to the ITU-T G.729.1 wideband audio codec and was further evaluated in comparison with the ITU-T G.729.1 Annex E super-wideband audio codec and HMM-based reference bandwidth extension method. Objective quality evaluation results indicated that the proposed method was preferred over the hidden Markov model-based reference bandwidth extension method in terms of log spectral distortion, cosh measure, and differential log spectral distortion. Further, the proposed method improves the auditory quality of the wideband audio and also gained a good performance in the subjective listening tests.

Long-term Impact

As an efficient approach that increases coding efficiency of audio signals, the BWE will intensively improve our life quality and enabled us to enjoy high-quality speech and music anywhere by wireless devices. Because of the constraint of frequency resource, current wireless communication systems limited the bandwidth of audio signals for transmission. The bandwidth of 14 kHz is far from meeting the demands for high-fidelity audio signals, especially for music signals. Artificial BWE that can make WB audio codec has the similar quality to the full-band audio codec has recently received a great deal of attention, such as audio coding used for cloud computing and Bluetooth. Motivated by the non-linear relationship between LF and HF components, we developed several blind BWE methods from 7 kHz to 14 kHz. Although audio quality has been improved compared to the conventional methods, the mechanism of mapping between LF and HF bands has not been excavated completely due to the complexity of audio signals, especially its chaotic characteristics.

References

[1] Xin Liu, Changchun Bao. Blind bandwidth extension of audio signals based on nonlinear prediction and hidden Markov model. APSIPA Transactions on Signal and Information Processing, 2014, Vol. 3, e8

[2] Xin Liu, Changchun Bao. Audio bandwidth extension based on temporal smoothing cepstral coefficients. EURASIP Journal on Audio, Speech, and Music Processing. 2014:41.

[3] Changchun Bao, Xin Liu, Yongtao Sha, Xingtao Zhang. A blind bandwidth extension method for audio signals based on phase space reconstruction. EURASIP Journal on Audio, Speech, and Music Processing. 2014:1 .

[4] Xin Liu, Changchun Bao. Audio Bandwidth Extension Using Ensemble of Recurrent Neural Networks. EURASIP Journal on Audio, Speech, and Music Processing. 2016:12.

[5] Xin Liu and Changchun Bao. Audio Bandwidth Extension Based on Ensemble Echo State Networks with Temporal Evolution. IEEE/ACM Transactions on audio, Speech, and Language Processing, Vol. 24, No. 3, March 2016, 594-607

Photo Gallery: APSIPA-ASC'2015 in Hong Kong

The AI from data to edge product

Fast labeling tool and Embedded AI-based ADAS Technology

For deep learning, computing is its essence and data is its foundation. At beginning, how to create valid deep learning database and how to generate a good deep learning model are high time-consuming problems. We start our AI journey from data and target to edge AI product. A fast automatic labeling tool is designed to save the time needed to develop deep learning product and allow those who are not familiar with deep learning algorithm to generate an AI data easily. Collecting, labelling, augmenting, analyzing, and digging are five important processes to make budding data to become mature one. ezLabel[1], a web-based platform, provides an intelligent way to label data, collect data, analyze data, and dig data with high diversity. Then, we are able to figure out innovative and productive deep learning algorithms and applications efficiently. A light deep-learning-based object and its behavior detection algorithm, which is able to be ported on various embedded platform, e.g., Nvidia Jetson TX2, Renesas R-car H3, and Hikey 960, and a light deep-learning-based multipath convolutional neural networks and conditional back-propagation mechanism object detection algorithm, which is good at detecting far object, are proposed and are verified. These algorithms make a big step to realize the future filled with edge AI product.

The fast automatic labeling tool, ezLabel

Fig. 1 The user interface of ezLabel

There are three main pain points: inefficiency, insufficiency, and numerous applications waiting for the solutions. Collecting data is definitely fast; however, transferring those data to be understood by deep leaning algorithm is unbelievably slow. This boring, exhausted, annoying, tedious, hand-made process cannot be skipped. The worse thing is that the quality is not guaranteed after spending a lot of time. Even the data can be labeled smoothly, data diversity is another important issue, which is the key to commercialization. Details, details, details… There are always many details to be annotated for different product purposes. Thus, ezLabel is born to solve these annoying and unavoidable problems. It features automatic route prediction and fitting algorithm, which will significantly reduce the time needed to label manually and ensure the quality of the samples. ezLabel is an online platform opening to deep learning experts and ordinary people worldwide, thereby gathering data throughout the world to increase data diversity and facilitating the development of deep learning model.

Professor Jiun-In Guo

PhD ***APSIPA SPS TC Chair***

Distinguished Professor,
Dept. of Electronics Engineering and Institute of Electronics, National Chiao-Tung University (NCTU)

Associate Dean,
Collage of Electrical and Computer Engineering, NCTU

Director,
SoC Research Center, NCTU

Jiun-In Guo received the B.S. and Ph.D. degrees in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1989 and 1993, respectively. He is currently a full Professor of the Dept. of Electronics Engineering, National Chiao-Tung University. His research interests include DSP, VLSI design, SOC design, and intelligent vision processing. He is the author of over 200 technical papers on the research areas of low-power algorithm, architecture, and system design for DSP/Multimedia/Vision processing applications.

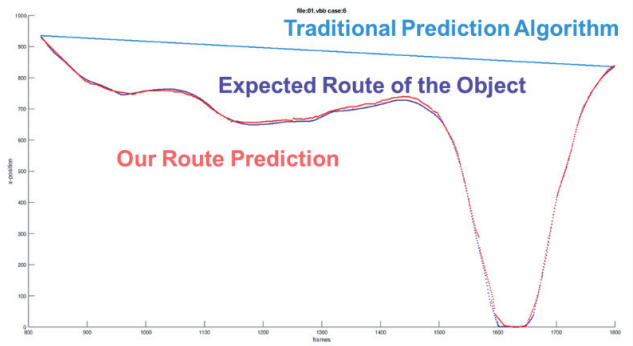


Fig. 2 A real case of object route including expected one, traditional one, our result

Unique route prediction facilitates users to label target object in only two frames instead of labeling in all frames. Traditional labeling tool only supports linear route prediction, that is, a deviate route result as the blue line shown in Fig. 2 is gained, which means user needs to adjust the bounding box frame by frame. In the case shown in Fig. 2, user has to check 1000 frames to guarantee that each target object is located in the bounding box. Instead, our

route prediction allows user to label target object with only two frames. A perfect route as the red line shown in Fig. 2 is delivered well by our unique algorithm.

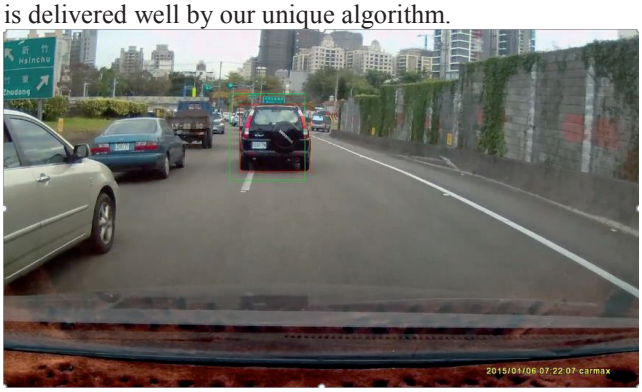


Fig. 3 The result of our fitting algorithm (green: original bounding box, red: result one)

Fitting not only speeds up the labeling process, but also guarantees the quality of labeled data. Current AI is not as smart as human beings. In order to recognize an object, deep learning needs thousands of or millions of data samples to teach the algorithm. A good data implies the good recognition model can be formed. Thus, not only efficiency but also the quality should be seriously considered. Our unique fitting algorithm decreases the unrelated background information and fits the bounding box to the target object. As shown in Fig. 3, a real example, the original bounding box in green includes many unrelated background information; moreover, the position of this bounding box is incorrect. It does not matter. Our fitting algorithm provides correct and precise bounding box automatically.

A web-based platform provides the comprehensive service for the world. We offer a web-based platform. Users are not limited by time and by place. With the internet and an ezLabel account, they can label data efficiently at anytime and anywhere. The diversity of data sample is hard to imagine. ezLabel gathers the power of everyone from worldwide in order to complement the deep learning database. ezLabel collects data and makes those data useful. We keep upgrading our service at every moment.

One-stage Faster RCNN and Heatmap-layer-added C3D

2D convolution object detection network and 3D convolution behavior recognition network are combined to implement rear overtake warning system. On a 2D convolution object detection network, object detection network is modified to let a network with a small amount of parameters generally have a better detection effect. Additionally, the original open source is revised to make this architecture more platform portable than other architectures. On the 3D convolution behavior recognition network, the original architecture is improved to make behavior recognition network with low resolution input have the ability of object localization, which utilizes the last layer of convolution layer to learn the overtaking object location in the latest image.

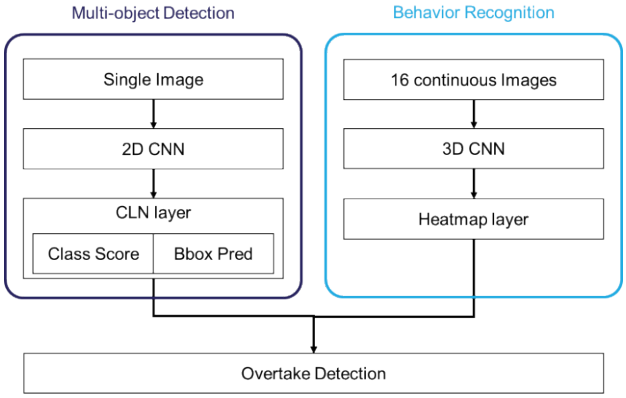


Fig. 4 Flowchart of the proposed rear overtake warning system

The original Faster RCNN[2] architecture is a two stage object detector. Compared to one stage object detector, it has Region Proposal Networks (RPN) Layer. RPN Layer connected after 2D CNN processes bounding box prediction, and computes object score. This layer focuses on the binary classification; as a result, it performs bounding box on the object without any diverse object attribute tag in the result image. With the characteristic that two stage object detector is good at detection quality, we try to keep this advantage and to make the proposed algorithm can be implemented on embedded platforms, i.e., original Faster RCNN is modified from two stage object detector architecture to one stage object detector one as shown in Fig. 4. This proposed one-stage-object-detector-based Faster RCNN, which eliminates ROI polling layer and fully connected layer, is named one stage Faster RCNN. One stage Faster RCNN consists of 2D CNN and proposed Classification and Localization Network (CLN) layer, which supports diverse classification.

C3D[3] network has faster speed than other state-of-the-art architectures have, which fits our goal that porting on the embedded systems, and its accuracy is good enough. Based on C3D network, which uses eight 3D convolution layers and two Fully-Connected layers, we propose a light C3D network, which has only five 3D convolution layers and two Fully-Connected layers. After combining 2D convolution object detection network, i.e., One-stage Faster RCNN, and 3D convolution behavior recognition network, i.e., Heatmap-layer-added C3D, it can recognizes and locates the objects and overtaking targets as shown in Fig. 5.

Fig. 5 The result of rear overtake warning system with object detection and behavior recognition

Multipath CNNs and conditional back-propagation mechanism for salient objects detection on the road

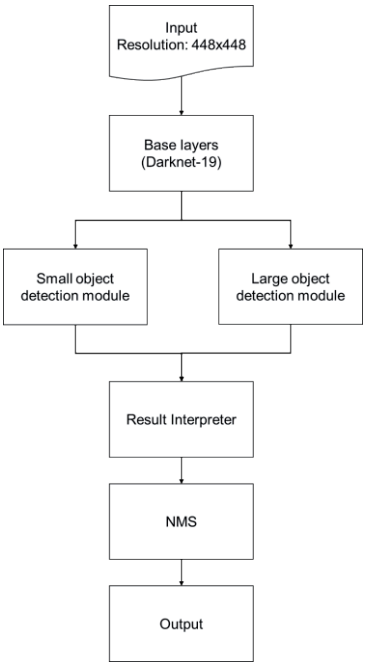


Fig. 6 System flow of objects detection with multipath CNNs and conditional back-propagation mechanism

We propose the design and verification of common object detection and warning systems on the road. This design considerations are actually applied to the implementation of low-power systems. The system separates the object detection networks of different sizes and applies the corresponding algorithm. For larger objects, a neural network with a large visual receptive field is used; for the detection of small objects, the network of smaller receptive field and fine grain features are used for precise predictions. Conditional Back Propagation mechanism makes different types of networks perform data driven learning for the set criteria and learn the representation of different object size without the degradation of each other. The design of multiple paths objects bounding boxes regressors can simultaneously detect objects in various kinds of scales and aspect ratios. The framework and algorithm make the models able to extract robust features under the same training data, which can be applied to different weather, e.g., sunny, night and rainy days, and different countries. Only single inference of neural network is needed for each frame to support the detection of multiple types of objects, such as bicycles, motorbikes, cars, buses, trucks, and pedestrians, and find out their exact positions.

Fig. 7 Robust far object detection results

Long-term Impact

Deep learning becomes more and more popular, which facilitates human beings in various fields, e.g., autonomous driving, surveillance, factory, medical, and finance. Data plays an important role as a translator between application and deep learning algorithm, that is, how to transfer the domain know-how via data dominates whether the deep learning performs well or not. ezLabel provides not only a fast labeling functions but also data sample management ones, which generates a shortcut to create more innovative and productive deep learning applications and keep enhance the quality of deep learning models efficiently. Some new ideas like mentioned in this article, i.e., rear overtaking warning system implemented by deep-learning-based multi-object detection and behavior recognition and multipath convolutional neural networks and conditional back-propagation mechanism for salient objects detection on the road, are designed and are verified in short time. A real AI future filled with edge AI products is much more closed. It is exciting that AI can truly be adopted in every single equipment.

References

[1] ezLabel of creDa, from https://www.aicreda.com/

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in proceeding of conference on Neural Information Processing Systems (NIPS), 2015.

[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in proceeding of IEEE International Conference on Computer Vision (ICCV), 2015.

Stochastic Signal Processing in Large MIMO Channels

Professor Shinsuke Ibi

PhD, MIEEE ***APSIPA WCN Chair***

Associate Professor

Department of Information and Communications Technology

Osaka University

Shinsuke Ibi received the B.E. degree in advanced engineering from Suzuka College of Technology, Japan, in 2002, and the M.E. and Ph.D. degrees in communication engineering from Osaka University, Japan, in 2004 and 2006, respectively. From 2005 to 2006, he was a visiting researcher at the Centre for Wireless Communications, University of Oulu, Finland. In 2006, he joined the Graduate School of Engineering, Osaka University, and he is currently an Associate Professor in the Department of Information and Communications Technology, Osaka University. From 2010 to 2011, he was a visiting researcher at the University of Southampton, UK. His research interests include EXIT-based coding theory, iterative detection, digital signal processing, and communication theory. He received the 64th and 71st Best Paper Awards from IEICE, and the 24th Telecom System Technology Award from the Telecommunication Advancement Foundation. He is members of IEEE, IEICE, and ION.

This paper introduces the fact that the large (or massive) multi-input multi-output (MIMO) is not only a technique for capacity enhancements, but also for cost-efficiency. The expensive hardware unit at the base station can be replaced by massive utilization of parallel low-cost units in the uplink large MIMO channels. In terms of the massive growth of wireless traffic in the near future, wireless engineers should think of cost-efficiency more seriously. To deal with the requirement, stochastic signal processing will play a key role.

Large MIMO detection

Large multi-input multi-output (MIMO) systems, in which a large number of antennas are equipped on both the transmitter and receiver sides, are gaining attention as a potential technique for meeting the ever-growing demand on wireless communication systems [1]. We consider uplink multi-user detection (MUD), where the BS has N receive (RX) antennas and M individual user equipments (UEs) have a single transmit (TX) antenna. The m-th UE conveys a modulated symbol x_m, which represents one of Q constellation points $\mathcal{X} = \{\chi_1, \ldots, \chi_q, \ldots, \chi_Q\}$. The average energy of the constellations in the set $\mathcal{X}$ is denoted by E_s. Let $\boldsymbol{x} = [x_1, \ldots, x_m, \ldots, x_M]^{\mathrm{T}} \in \mathcal{X}^{M\times 1}$ and $\boldsymbol{y} = [y_1, \ldots, y_n, \ldots, y_N]^{\mathrm{T}} \in \mathbb{C}^{N\times 1}$ denote the TX and RX symbol vectors, respectively. Assuming frequency flat fading environments, the RX vector $\boldsymbol{y}$ is given by

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{z}, \tag{1}$$

where $\boldsymbol{H} \in \mathbb{C}^{N\times M}$ is an $N \times M$ channel matrix. The vector $\boldsymbol{z} = [z_1, \ldots, z_n, \ldots, z_N]^{\mathrm{T}} \in \mathbb{C}^{N\times 1}$ is a complex additive white Gaussian noise (AWGN) vector, whose entries z_n obey independent identically distributed (i.i.d.) complex Gaussian distribution with zero mean and N_0 variance $\mathcal{CN}(0, N_0)$, where N_0 is the noise spectral density and the covariance matrix is given by $\mathbb{E}\{\boldsymbol{z}\boldsymbol{z}^{\mathrm{H}}\} = N_0\boldsymbol{I}_N$. Under the assumption that the channel state $\boldsymbol{H}$ is known at the RX with the aid of a pilot sequence, the optimal maximum likelihood (ML) detector explores the most-likely TX vector $\boldsymbol{x}$ as follows

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{\chi}\in\mathcal{X}^{M\times 1}} p(\boldsymbol{y}\,|\,\boldsymbol{\chi}) = \arg\max_{\boldsymbol{\chi}\in\mathcal{X}^{M\times 1}} \|\,\boldsymbol{y} - \boldsymbol{H}\boldsymbol{\chi}\,\|^2, \tag{2}$$

where the conditional PDF of RX vector $\boldsymbol{y}$ is given by

$$p(\boldsymbol{y}\,|\,\boldsymbol{\chi}) = \frac{1}{(\pi N_0)^N}\exp\left[-\frac{\|\,\boldsymbol{y} - \boldsymbol{H}\boldsymbol{\chi}\,\|^2}{N_0}\right]. \tag{3}$$

Large MIMO channels make large-scaled MUD difficult in uplink scenarios, as MUD becomes computationally expensive owing to the increased dimensions of the MIMO channels.

Channel Hardening Effects in Large MIMO

Linear spatial filters such as matched filter (MF), zero forcing (ZF) filter, and minimum mean square error (MMSE) filters trade detection capability for low computational cost relative to optimal ML detector. Denoting the weight matrix of the spatial filter by $\boldsymbol{W}$, the filter output is given by

$$\tilde{\boldsymbol{x}} = \boldsymbol{W}^{\mathrm{H}}\boldsymbol{y} = \boldsymbol{W}^{\mathrm{H}}\boldsymbol{H}\boldsymbol{x} + \boldsymbol{W}^{\mathrm{H}}\boldsymbol{z} \triangleq \boldsymbol{G}\boldsymbol{x} + \boldsymbol{v}. \tag{4}$$

The weight matrices $\boldsymbol{W}^{\mathrm{H}}$ of MF, ZF, and MMSE are $\boldsymbol{H}^{\mathrm{H}}$, $[\boldsymbol{H}^{\mathrm{H}}\ \boldsymbol{H}]^{-1}\ \boldsymbol{H}^{\mathrm{H}}$, and $\boldsymbol{H}^{\mathrm{H}}\ [\boldsymbol{H}\ \boldsymbol{H}^{\mathrm{H}}+\boldsymbol{I}_N]^{-1}$, respectively. The covariance matrix of $\boldsymbol{v}$ is expressed as $\boldsymbol{\Omega} = \mathbb{E}\{\boldsymbol{v}\boldsymbol{v}^{\mathrm{H}}\} = N_0\ \boldsymbol{W}^{\mathrm{H}}\ \boldsymbol{W}$. Let us consider ML detector of the filter output $\tilde{\boldsymbol{x}}$, instead of $\boldsymbol{y}$ in (1), which is represented as

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{\chi}\in\mathcal{X}^{M\times 1}} p(\tilde{\boldsymbol{x}}\,|\,\boldsymbol{\chi}), \tag{5}$$

where we have

$$\begin{aligned} p(\tilde{\boldsymbol{x}}\,|\,\boldsymbol{\chi}) &= \frac{1}{\pi^N \det[\boldsymbol{\Omega}]}\exp\left[-(\tilde{\boldsymbol{x}} - \boldsymbol{G}\boldsymbol{\chi})^{\mathrm{H}}\,\boldsymbol{\Omega}^{-1}(\tilde{\boldsymbol{x}} - \boldsymbol{G}\boldsymbol{\chi})\right] \\ &\propto \exp\left[2\Re\{\tilde{\boldsymbol{x}}^{\mathrm{H}}\boldsymbol{\Omega}^{-1}\boldsymbol{G}\boldsymbol{\chi}\} - \boldsymbol{\chi}^{\mathrm{H}}\boldsymbol{G}^{\mathrm{H}}\boldsymbol{\Omega}^{-1}\boldsymbol{G}\boldsymbol{\chi}\right] \\ &\triangleq \exp\left[2\Re\{\tilde{\boldsymbol{x}}^{\mathrm{H}}\boldsymbol{\Psi}\boldsymbol{\chi}\} - \boldsymbol{\chi}^{\mathrm{H}}\boldsymbol{\Gamma}\boldsymbol{\chi}\right]. \end{aligned} \tag{6}$$

Assuming $\boldsymbol{\Gamma}$ is a diagonal matrix, $\boldsymbol{\chi}^{\mathrm{H}}\ \boldsymbol{\Gamma}\boldsymbol{\chi}$ is a constant value while phase shift keying (PSK) signaling, resulting in

$$p(\tilde{\boldsymbol{x}}\,|\,\boldsymbol{\chi}) \propto \exp\left[2\Re\{\tilde{\boldsymbol{x}}^{\mathrm{H}}\boldsymbol{\Psi}\boldsymbol{\chi}\}\right]. \tag{7}$$

Furthermore, if the matrix $\boldsymbol{\Psi}$ can be approximated by a diagonal matrix with entries ψ_m, we have

$$p(\tilde{\boldsymbol{x}}\,|\,\boldsymbol{\chi}) \propto \prod_{m=1}^{M}\exp\left[2\Re\{\psi_m\tilde{x}_m^*\chi_m\}\right]. \tag{8}$$

In this case, the search problem (5) of ML is approximately simplified as

$$\hat{x}_m = \arg\max_{\chi_q \in \mathcal{X}} \Re\left\{\psi_m \tilde{x}_m^* \chi_q\right\}. \quad (9)$$

(9) implies the fact that the search space of Q^M in (5) shrinks to MQ. In this case, we have to mind the approximation of diagonalization of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$.

In large MIMO channels, we can enjoy the channel hardening effects. A breakthrough of large-scaled MUD is the large system limit, where the input and output dimensions M and N increase to infinity while the compression rate N/M is kept constant [2]. The channel hardening effects in the large system limit is obtained as

$$\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} \approx N\boldsymbol{I}_M, \quad \boldsymbol{H}\boldsymbol{H}^{\mathrm{H}} \approx M\boldsymbol{I}_N. \quad (10)$$

The resultant $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$ for each spatial filter approach diagonal matrices along with the increase of the dimensions of the large MIMO channel.

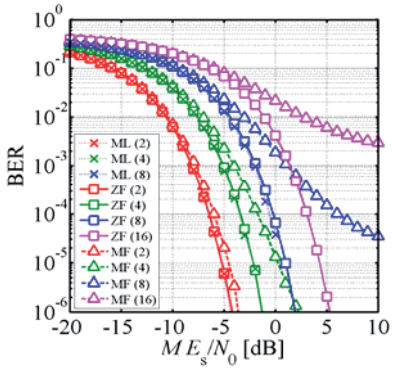


Fig. 1 BER performance in Large MIMO channels.

Bit error rate (BER) performances of ML, ZF, and MF are characterized in Fig. 1. The modulation scheme is Gray coded Quadrature PSK (QPSK). The number of the RX antennas N is 128. On the other hand, the number of the TX antennas M are 2, 4, 8, 16. The channel coefficients in $\boldsymbol{H}$ obey Rayleigh distribution of $\mathcal{CN}$ (0, 1). Due to the huge computational complexity, MLD of M = 16 is not shown in the figure. Comparing ZF with ML, the BER performances are almost the same. The fact implies that the approximation of (9) is tight. However, the performance of MF is getting worse with the increase of the number of TX antennas M due to insufficient channel hardening effects, although MF has an advantage in terms of low complexity in parallel computations without the matrix inversion.

Stochastic Signal Processing

To reduce complexity without sacrificing detection capability even in insufficient channel hardening effects, iterative detection schemes with soft interference cancellation have been investigated [3]. Iterative detection can be roughly classified into two types: iterative detection and decoding (IDD), and self-iterative detection (SID). In the development of IDD, turbo-principle was the most important rule, in which bit-wise log likelihood ratios (LLRs) are exchanged as prior beliefs between the symbol detector and channel decoder. The turbo-principle suggests that uncorrelated Gaussian signals should be approximately obtainable as prior beliefs in the iterative regime. Nevertheless, IDD is not preferable in practical large-scale systems because of the processing delay and large power consumption arising from its iterative decoding process.

A breakthrough in the development of large-scale MUD was the adoption of SID based on belief propagation (BP), which can be carried out without the use of a channel decoder in each iteration process [4] [5]. SID takes advantage of the channel hardening effect mentioned before. In practical MUD scenarios, however, a sufficient number of RX antennas may not be available owing to limits on the size, weight, cost, and/or power consumption of the receiver. Under such conditions, the convergence property of SID significantly deteriorates as a result of approximation errors. To mitigate the negative effects of approximation error, an adaptive scaled belief (ASB) [6] of MF output is proposed on the basis of Gaussian belief propagation (GaBP) [7].

Long Term Impacts

For the cost efficiency in 5G beyond considerations, the large MIMO is a promising technique to satisfy the requirements to cover enhanced mobile broadband (eMBB), massive MTC (mMTC), and ultra-reliable and low latency communications (URLLC). In the near future, the large MIMO signal detection will meet the stochastic parallel signal processing with the aid of machine learning and deep learning systems. However, wireless engineers must guarantee the ultra-reliability of the signal detection. To address this issue, the stochastic model of signals and involved approximations should be explored before leaving the matter to the machine learning engine. The author is confident that the fusion of belief propagation and deep learning with rigorous care will open new vistas in signal processing for wireless communications and networks.

References

[1] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," IEEE Commun. Surveys Tutorials, vol. 14, no. 4, pp.1941-1988, Fourthquarter, 2015.

[2] T. Tanaka, "A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors," IEEE Trans. Inf. Theory, vol. 48, no. 11, pp. 2888–2910, Nov. 2002.

[3] M. Tuchler and A. C. Singer, "Turbo equalization: An overview," IEEE Trans. Inf. Theory, vol. 57, no. 2, pp. 920–952, Feb. 2011.

[4] Y. Kabashima, "A CDMA multiuser detection algorithm on the basis of belief propagation," J.Phy. A: Math.Gen., vol. 36, no. 43, pp. 11111-11121, Oct. 2003.

[5] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," Proc. Nat. Acad. Sci., vol. 106, no. 45, pp. 18914–18919, Nov. 2009.

[6] T. Takahashi, S. Ibi, and S. Sampei, "Design of criterion for adaptively scaled belief in iterative large MIMO detection," IEICE Trans. on Commun., Feb. 2019 (to be published).

[7] P. Som, T. Datta, A. Chockalingam, and B. S. Rajan, "Improved large-MIMO detection based on damped belief propagation," in Proc. ITW 2010, Cairo, Egypt, Jan. 2010, pp. 1–5.

Algorithm/Architecture Co-design for Smart Systems in Cognitive Cloud and Reconfigurable Edge

Professor Chris G.G. Lee

PhD, SrMIEEE

Chair for APSIPA Signal Processing Systems Technical Committee

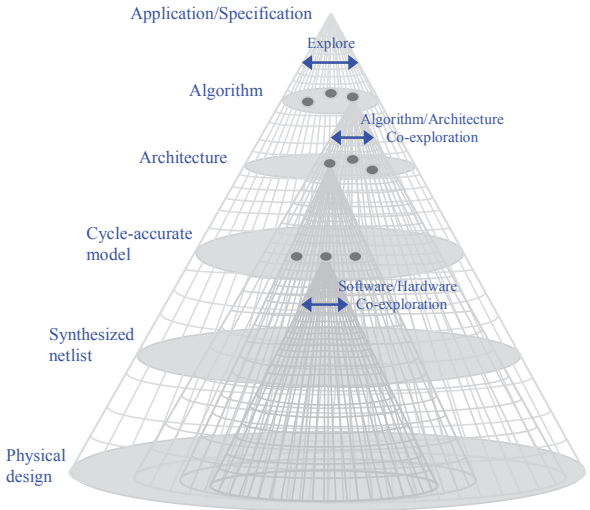
Professor of Department of Electrical Engineering, Director of Bioinfotronics Research Center, National Cheng Kung University

Chris Gwo Giun Lee is an investigator in the field of signal processing systems including multimedia and bioinformatics. His endeavors in system design, based on analytics of algorithm concurrently with architecture, has made possible computations on System-on-Chip cloud and edge platforms in resolving complex problems with both accuracy and efficiency. Having previously held leading and managerial positions in the industry such as System Architect in former Philips Semiconductor in Silicon Valley, Lee was recruited to NCKU in 2003 where he found and is currently directing the Bioinfotronics Research Center. Lee received his B.S. degree in electrical engineering from National Taiwan University and both his M.S. and Ph.D. degrees in electrical engineering from University of Massachusetts. He has contributed more than 130 original research and technical publications with the invention of 100+ patents worldwide. Lee serves as the AE for IEEE TSP and Journal of Signal Processing Systems. He was formerly the AE for IEEE TCSVT for which he received the Best Associate Editor's Award in 2011.

NIKLAUS EMIL WIRTH introduced the innovative idea that Programming = Algorithm + Data Structure. Inspired by this, we advance the concept to the next level by stating that Design = Algorithm + Architecture. With concurrent exploration of algorithm and architecture entitled Algorithm/Architecture Co-exploration (AAC), this methodology introduces a leading paradigm shift in advanced system design from System-on-a-Chip (SoC) to Cloud and Edge. As algorithms with high accuracy become exceedingly more complex and Edge/IoT generated data becomes increasingly bigger, flexible parallel or reconfigurable processing are crucial in the design of efficient signal processing systems requiring low power. Hence the analysis of algorithms and/or data for potential computing in parallel is crucial. With extension of AAC for SoC system designs to even more versatile platforms based on analytics architecture, system scope is readily extensible to cognitive cloud and reconfigurable edge computing, a cross-level-of abstraction topic which is introduced in this article.

Cross-Level-of-Abstraction in Smart Systems

In Design Space Exploration (DSE) introduced by Edward Lee of UC Berkley, profiling was commonly performed such that algorithms which are not capable of real time processing in software would be partitioned to be implemented in digital hardware accelerator. As shown in Figure 1, we have enlarged the design space to further include the algorithmic and architecture levels so that during exploration, in high accuracy and flexible cognitive cloud and reconfigurable edge computing platforms requiring yet higher speed performance and much lower power. Hence, in



Different instances or realizations

Levels	Symbols	Features	Time units
Algorithm		System functionality	Second
Architecture		System architecture	Number of cycles
IP (Macro)		IP functionality and micro-architecture	Number of cycles
Module		Arithmetic operation	Cycle
Gate		Logic operation	ns
Circuit		Voltage, current	ps
Device		Electron	ps

Feature in various level of abstraction

Figure 1. Cross Level of Abstraction Design Space Exploration

addition to the conventional Software/Hardware Co-design (SHC), the current smart system design methodology further extends to Algorithm/Architecture Co-design (AAC)!

As seen in Figure 2, enlarging the design space provides more degrees of freedom in the exploration of larger cloud and also edge computing systems possibly including the neuromorphic level in trading off higher efficiency and lower power consumption but of course with less flexibility and accuracy.

In common artificial intelligence, analytics algorithms are used to analyze speech, audio, image video data, etc. In current smart cross-level system design methodology

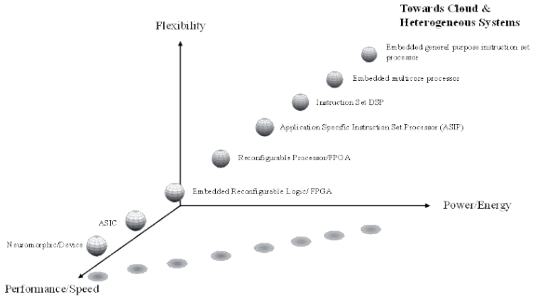


Figure 2. Spectrum of Platforms

different algorithmic realizations are analyzed to further increase efficiency and flexibility in constituting "analytics architecture".

Parallel/Reconfigurable Computing

In analytics architecture based smart cognitive cloud and reconfigurable edge computing, dataflow graph (DFG) models are used to represent different realizations or implementations of an algorithm also referred to as architecture instantiation. In DFG, a node or super node represents or models a processing element (PE) with varying data granularity ranging from an arithmetic unit, a CPU, a multicore processor, or a cluster of computing elements which may be centralized or distributed. The corresponding edges represent data communication between these processing elements. Having information on both algorithmic behavior and architectural information including both software and hardware, the DFG so introduced provides a mathematical representation which, as opposed to traditional linear difference equations, better models the underlying computational platform for systematic analysis thus providing flexible and efficient management of the computational and storage resources.

In this work, parallel and reconfigurable computing are formulated via DFG which are analogous to the analysis and synthesis equations of the well-known Fourier transform pair. In parallel computing, a connected component is eigen-decomposed to unconnected components for concurrent processing. For computation resource saving, commonalities in DFGs are analyzed for reuse when synthesizing or reconfiguring the platform.

Parallel computing requires data independency so that processing could be performed concurrently. As shown in the example in Figure 3. the DFG representing different possible realizations of an algorithm is analyzed using spectral graph theory [1][2][3]. Using this method, the Laplacian matrix of the DFG will then be constructed. After eigen-decomposition, the eigenvectors corresponding to the zero eigenvalues will represent unconnected components which could then be processed in parallel.

In Figure 3, the two eigenvectors with zero eigenvalues correspond to the two eigenvectors or unconnected components could be processed in parallel. It is also possible to compare the size of these unconnected components after eigen-analysis to study whether the parallelism is homogeneous or heterogeneous which provides important information which is critical in selecting or designing the instruction set architecture (ISA) for computing platform.

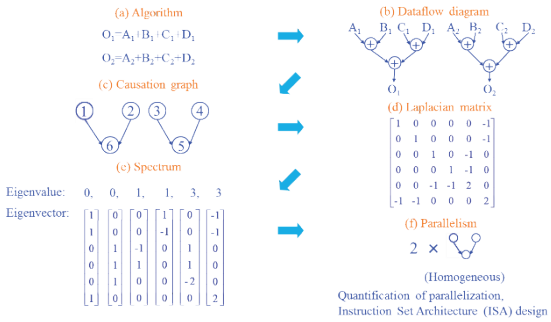


Figure 3. Degree of Parallelism: Eigen-analysis of DFG via Spectral Graph Theory

As shown in Figure 4 (a)(b), this outcome of the analytical software via the analytics or eigen-analysis of DFG for parallel computing serves as the retargetable compiler as shown in Figure 4 (c). As can be seen in Figure 4 (d), using the retargetable compiler the CPUs which are not required do not have to be turned on so as to save power. These then provides a cognitive cloud platform as depicted in Figure 5 with these analytic architecture [2][3] licensed by US based startup in Precision Medicine [4].

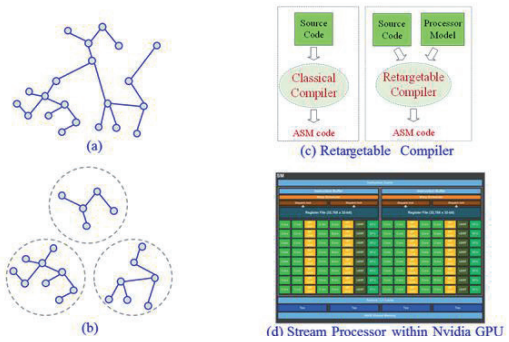


Figure 4. Mapping tasks onto parallel platforms

Reconfigurable computing extracts commonalities in the DFGs so that computation resources could be reused. As an example shown in Figure 6, when implementing different MPEG algorithms [5][6], we could lower the granularity and study the common parts of the DFG which could then be reused during the design. As seen in Figure 7 (a), when these lower granularity graphs are identified to be reusable for a specific task, they could be synthesized or reconfigured into a larger DFG as depicted in Figure 7 (b). Having this information from the DFG, microcode in microsequencer

were designed to control the reconfigurable parts of these DFGs.

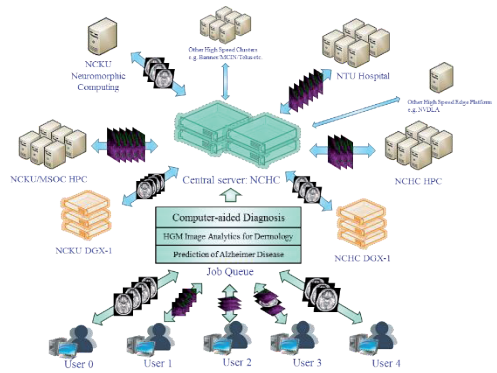


Figure 5. Cognitive parallel cloud computing platform

This methodology were used to design Edge platforms with higher efficiency, flexibility, lower data transfer rate and also lower power due to the reusable architecture.

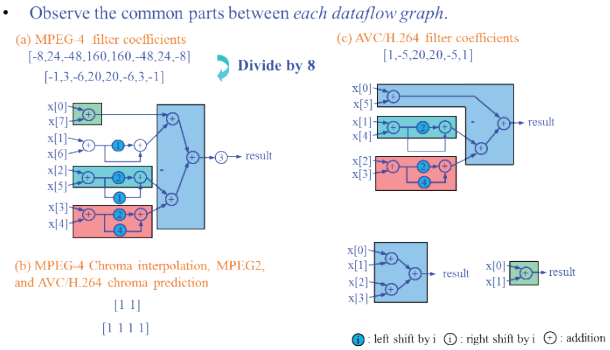


Figure 7. Commonality extraction using DFG in reconfigurable computing

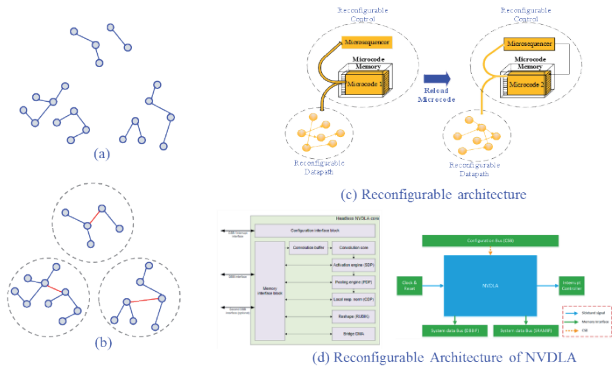


Figure 6. Mapping tasks on reconfigurable platforms

Long-term Impact

In the 1960's, Marshall McLuhan published the book entitled "The Extensions of Man", focusing primarily on Television, an electronic media as being the extension of human nervous system which from contemporary interpretation marks the previous stage of Big Data! Based upon mathematical fundamentals as foundations for complexity aware analytical signal and information processing algorithms especially in highly computational complex artificial intelligence (AI), intelligent, flexible, and efficient system architectures including both software and hardware are required be concurrently explored and designed, whereby the current innovative work envisions even further extension of human perceptual experiences and exchange of information, together with the expediting of field of signal information processing into yet another new era of Big AI Data, acquired from Internet-of-Things, to be computed upon cognitive cloud and reconfigurable edge.

References

[1] G. G. Lee, H.-Y. Lin, C.-F. Chen, T. Y. Huang, "Quantifying Intrinsic Parallelism Using Linear Algebra for Algorithm/Architecture Co-Exploration," IEEE Transactions on Parallel and Distributed Systems, vol. 23, iss. 5, pp. 944-957, May 2012.

[2] G. G. Lee, H.-Y. Lin, "Quantifying method for intrinsic data transfer rate of algorithms," USA, Patent No. US9092384 B2, Jul. 28, 2015.

[3] G. G. Lee, M. J. Wang, H.-Y. Lin, "Method and Algorithm Analyzer for Determining a Design Framework," USA, Patent No. US8621414 B2, Dec. 31, 2013.

[4] Boston, MA, June 1, 2015, GlobeNewswire

[5] G. G. Lee, C.-F. Chen, S. M. Xu, C. J. Hsiao, "High-Throughput Reconfigurable Variable Length Coding Decoder for MPEG-2 and AVC/H.264," Journal of Signal Processing Systems for Signal, Image, and Video Technology, vol. 82, iss. 1, pp 27-40, Jan. 2016

[6] G. G. Lee, W. C. Yang, H.-Y. Lin, M. S. Wu, "Space exploration method of reconfigurable motion compensation architecture," USA, Patent No. US8644391 B2, Feb. 4, 2014.

Photo Gallery: APSIPA ASC'2017 in Kuala Lumpur

Toward Future Research on Quality of Experience using Deep Convolutional Neural Networks

Professor Sanghoon Lee

PhD, SrMIEEE ***APSIPA IVM TC Chair***

Full Professor

Department of Electrical and Electronics Engineering

Yonsei University

Professor Sanghoon Lee is a Full Professor in Yonsei University, South Korea since he joined in 2003. He received his Ph.D. in E.E. from the University of Texas at Austin in 2000. From 1999 to 2002, he worked for Lucent Technologies. He was an Associate Editor of the IEEE Trans. Image Processing (2010-2014), an Editor of the Journal of Communications and Networks (JCN) (2009-2015), and a Guest Editor for IEEE Trans. Image Processing (2013) and Journal of Electronic Imaging (2015). He has been an Associate Editor of IEEE Signal Processing Letters (2014-) and a Chair of the IEEE P3333.1 Quality Assessment Working Group (2011-). He currently serves as a Chair of the APSIPA IVM Technical Committee (2018-). He has actively participated in international activities as a General Chair of the 2013 IEEE IVMSP Workshop, a Technical Program Co-chair of IEEE ICME 2018, APSIPA 2018, and the Exhibition Chair of ICASSP 2018. He received a 2015 Yonsei Academic Award, a 2012 Special Service Award from the IEEE Broadcast Technology Society, a 2013 Special Service Award from the IEEE Signal Processing Society, and a Best Student Paper Award of QoMEX (International Conference on Quality of Multimedia Experience) 2018, etc.

The development of immersive display technology enables to represent the details of contents more naturally by providing a more realistic viewing environment while increasing immersion. In parallel, quality of experience (QoE) has been dealt with and discussed from both academy and industry to grade consumer products from the quality perspective. However, for quantification of QoE, it is very challengeable to analyze the human perception more accurately, even if it has been studied so many decades. Currently, there is no solid methodology to verify the human perception as a closed form objectively due to the limitation of human perception analysis. Recently, the deep convolutional neural network (CNN) has emerged as a core technology while breaking most performance records in the area of artificial intelligence via intensive training in accordance with the massive dataset. The main motivation of this article lies in finding of new insight on human perception analysis for QoE evaluation through visualization of intermediate node values. In this article, the emphasis is laid on how to utilize the CNN network as an analysis tool in addition to performance improvement. The new QoE predictor enables to find the human visual sensitivity for each image without using any prior information in comparison with conventional approaches. Toward the end, we provide a novel clue of how to obtain visual sensitivity for each pixel through visualizing the perceptual information, which is expected to be essentially applied for the QoE research in future including quality assessment (QA), visual discomfort prediction and virtual reality sickness assessment.

Human Visual Sensitivity for QoE

(a) (b) (c)

Fig. 1. Examples of distorted image and its visual sensitivity: (a) is a distorted image; (b) is an objective error map and (c) is a perceptual error map inferred by the HVS embedded deep model. Darker regions indicate more pixel-wise distorted pixels.

Just ahead of consumer electronics show (CES) 2018, Samsung Electronics debuted its new 146-inch modular television, 'The Wall', at a product unveiling showcase. As such, display technology has grown ever larger, more refined and more user immersive. Moreover, with the development of immersive displays including stereoscopic 3D (S3D), augmented reality and virtual reality, quality of experience (QoE) is important to demonstrate how much they provide high-quality user experience more objectively while ensuring users' viewing safety. In the academy research, QoE has been generally defined to cover various assessments including quality assessment (QA), contrast/sharpness assessment, visual presence assessment, visual discomfort prediction and virtual reality sickness assessment [1], [2], [3], [4], [5], [6], [7], [8], [9]. Since the ultimate observers of commercial contents are end-users, the primary goal of those metrics is to quantify the perceptual behavior of the human visual system (HVS) for each QoE task. Nevertheless, there is no solid work to quantify the QoE more generally due to the deep involvement of human perception, which has been actively discussed in the area of neuroscience, psychology, physiology and signal processing. One of the core HVS factors is a human visual sensitivity that explains which part of an image is more sensitive to the HVS, i.e., indicates the spatial strength of visual response to the visual information.

As aforementioned, the human visual sensitivity means the perception change of a user according to the characteristics

of given spatial signal [2], [3]. When the user views the visual contents, certain local spatial signals of the image are emphasized or masked according to the spatial characteristics of the image. Fig. 1 depicts an example of how much difference user perceives between a distorted image and the associated error signal. Fig. 1 (a) is a distorted image, (b) is the error map from the original where the darker region is more erroneous than the brighter one, and (c) is the visual sensitivity of being obtained from the deep-convolutional neural network (CNN) where the darker region is more sensitive to the HVS. In (b), it is shown that the rock region contains more errors than the sky region. However, in (a), it is apparent that the distortions of monotonous local areas such as sky look more errors than the rock regions from our observation. With a help of the visual sensitivity map in (c), it is obvious that the observed visual errors and sensitivity are highly correlated, which is quite different from what we observe from the objective errors in (b).

From the visual science perspective, this phenomenon is explained caused by the visual masking effect by human visual sensitivity, which has been analyzed by the change of QoE awareness according to various characteristics of contents. The visual masking effect means that the HVS exhibits different contrast sensitivity according to spatial pixel distribution. The contrast sensitivity function is a representative model for the description of this phenomenon [10]. Indeed, since the visual cortex is more complicatedly responsive when a human perceives the presence of texture, sub-band decomposition techniques such as Gabor filters or steerable pyramids have also been used for preprocessing to quantify the visual sensitivity numerically. Based on these findings, to accomplish the QoE task, researchers have attempted to embed human visual sensitivity in the metrics of performing the QA, visual discomfort prediction, etc.

For clarity, Fig. 2 depicts a simple comparison on how to apply the visual sensitivity for the accomplishment of image quality assessment (IQA) according to the methodology; conventional full-reference (FR)-IQA metric, deep no-reference (NR)-IQA model, HVS embedded deep IQA model and the future concept of a deep QoE model. Conventional IQA metrics extract feature maps that mimic the visual sensitivity of the HVS from content and predict quality through channel decomposition that reflects visual perception using Gabor filters or steerable pyramids as mentioned previously. Nonetheless, because the human visual behavior of content perception is very complicated to infer through those simple hand-crafted features, this generic process has followed performance limitations.

Recently, deep-learning technology, in particular, CNN has demonstrated significant performance improvement in the area of computer vision and signal processing even if the mechanism inside is unknown as a black-box method [2]. Beyond the classification framework, the black-box method has been successfully applied to regression problems in image/video QA and even visual discomfort prediction problems. Thanks to the benefit of voluminous data information, IQA metrics enable to demonstrate the state-of-the-art performance [2]. However, as described in the deep IQA flow of Fig. 2, it is not easy to analyze the physical meaning of the model by simply regressing subjective scores from the synthetic information of an image. Therefore, except the performance enhancement,

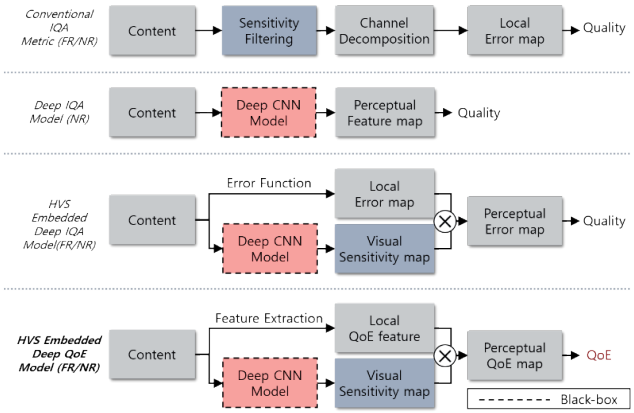


Fig. 2. Flowchart comparison of the conventional FR-IQA metric, deep NR-IQA model, HVS embedded deep IQA m odel and the concept of a future deep QoE model. FR and NR indicate full-reference and no-reference manner, respe ctively.

there has been a drawback that visual analysis looks impossible from the HVS perspective.

In order to overcome such a drawback of the CNN based model, we have investigated a new type of a CNN model of embedding human visual sensitivity, which enables to visualize the human perception while being accompanied by high predictive performance. The third row of Fig. 2 demonstrates a flow of the HVS embedded deep IQA model. The biggest difference from the existing CNN based model is that the visual sensitivity map is visualized as an intermediate result of the model and the visual weight of the given local error map can be analyzed. This model obtains the perceptual error map through an elementwise product of the weight and error maps, which are generated by inference based on characteristics of the input image. In addition, the local error map can be obtained in an FR manner using the reference image [1], or an NR manner inferring the error map from the distorted image itself [11]. Then, through a regression procedure of the perceptual error map to subjective quality scores made by users, the deep IQA model is trained.

Fig. 1 (c) shows the perceptual error map weighted by using the HVS embedded deep IQA where the dark areas represent more distorted pixels cognitively. In contrast to the objective error map in (b), since the area around the sky house is more monotonous, it can be seen that it has become relatively more emphasized in the perceptual error map. The main advantage of this approach is that it learns the visual sensitivity characteristics of the HVS without prior knowledge. The deep CNN can greatly improve the performance of the IQA model as well [13].

Long-term Impact

For the IQA work, the human visual sensitivity has been successfully visualized, which enables to provide deep insight on how the human perception is responsive to an input image. This approach is more advanced in the sense that the human perception is obtained without prior knowledge, which is different from conventional methods of obtaining handcraft features. So, how can this technology be applied to future QoE metrics? As we have introduced, display technology evolves to a larger, sharper, more immersive environment. The various QoE applica-

tions for this could fall in the area of S3D-visual discomfort prediction, sharpness and contrast assessment and virtual reality sickness assessment. Due to the intricately involved visual factors, currently, no solid numerical definition has not been published yet, but it is expected that new QoE metrics will be developed by modeling the HVS similarly to the mechanism used for IQA works. In general, the initial input of most content-oriented QoE is visual information, the visual sensitivity of visual content can be applied regardless of the types of tasks such as QA and visual discomfort prediction. At the fourth row of Fig. 2 shows an important clue on how to evolve the HVS embedded deep model to obtain the perceptual quality map of QoE. As shown in the figure, the visual sensitivity induced by the content itself can be mixed with its local QoE features (e.g. motion, depth, contrast, sharpness) and used as an element to predict QoE.

For example, in virtual reality sickness assessment work which is one of recent issue, visually induced motion sickness is caused by sensory mismatches between the motion perceived by the vestibular organ and the motion perceived by the HVS [9]. For this phenomenon, the distribution of the spatial texture has a great effect on the motion perception of the HVS. For example, in the monotonous background of Fig. 1 (a), the human is not aware of the motion that occurs in the content because there are relatively few temporal variations in which the motion is perceived. In contrast, for the rock and lighthouse, there are various spatial frequency components, so temporal variation is large, which makes the user more aware of motion. Thereby, it is expected that the motion component of an image and the weighting process of the visual sensitivity map extracted from the HVS embedded deep model can be effectively applied to calculate the visually perceived QoE.

Moreover, it is expected that for the S3D-visual discomfort prediction works [6], [7], such primitive approach of inferring visual sensitivity will lead to a higher performance improvement where the depth information induces the visual discomfort on the spatial domain. Furthermore, in other HVS based QoE studies such as 2D/3D visual presence assessment to quantify visual scene satisfaction of viewers [5], perceptual contrast/sharpness assessment model [4] and visual information measurement works [12], [8], the visual sensitivity also has utilized as a preprocessing method reflecting the users' visual perception. Based on these prospects, the HVS embedded deep model is expected to play a key role not only in improving the QoE field performance but also in exploring the visual perception.

Conclusion

In this article, we have introduced a new paradigm for the provision of how to resolve the human perception for future QoE tasks where the CNN model learns the human visual sensitivity. By using the visual sensitivity map learned by a deep model, we expect that each visual QoE feature map can be adopted directly to QoE metrics and it will lead to huge improvement on performance and perceptual analysis. In the constantly evolving immersive display technology, visual sensitivity studies are getting essential in the QoE field. In this regard, HVS embedded deep models will mark a new era in the QoE manner with the help of the infinite possibilities of artificial intelligence.

References

[1] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017.

[2] J. Kim and S. Lee, "Fully deep blind image quality predictor," IEEE Journal of selected topics in signal processing, vol. 11, no. 1, pp. 206– 220, 2017.

[3] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 130–141, 2017.

[4] H. Kim, S. Ahn, W. Kim, and S. Lee, "Visual preference assessment on ultra-high-definition images," IEEE Transactions on Broadcasting, vol. 62, no. 4, pp. 757–769, 2016.

[5] H. Oh and S. Lee, "Visual presence: Viewing geometry visual information of UHD S3D entertainment," IEEE Transactions on Image Processing, vol. 25, no. 7, pp. 3358–3371, 2016.

[6] J. Park, H. Oh, S. Lee, and A. C. Bovik, "3D visual discomfort predictor: Analysis of disparity and neural activity statistics," IEEE Transactions on Image Processing, vol. 24, no. 3, pp. 1101–1114, 2015.

[7] H. Oh, S. Lee, and A. C. Bovik, "Stereoscopic 3D visual discomfort prediction: A dynamic accommodation and vergence interaction model," IEEE Transactions on Image Processing, vol. 25, no. 2, pp. 615–629, 2016.

[8] K. Lee and S. Lee, "3D perception based quality pooling: Stereopsis, binocular rivalry, and binocular suppression," IEEE Journal of Selected Topics in Signal Processing, vol. 9, no. 3, pp. 533–545, 2015.

[9] J. Kim, W. Kim, S. Ahn, J. Kim, and S. Lee, "VR sickness predictor: Analysis of visual-vestibular conflict and VR contents," in International Conference on Quality of Multimedia Experience. QoMEX, 2018.

[10] S. J. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in Human Vision, Visual Processing, and Digital Display III, vol. 1666. International Society for Optics and Photonics, 1992, pp. 2–16.

[11] J. Kim and S. Lee, "Deep CNN-based blind image quality predictor," IEEE Transactions on Neural Networks and Learning Systems, 2018 (accepted).

[12] K. Lee and S. Lee, "A new framework for measuring 2D and 3D visual information in terms of entropy," IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 11, pp. 2015–2027, 2016.

Wireless Health Monitoring

In recent years, many countries have been facing the problem of a fast-aging population because of the increase of the life expectancy and the decrease of the birth rates. According to those facts, the cases where old people are staying alone are becoming very common. Remote monitoring in e-healthcare has become a hot issue in research fields and industries. To realize remote monitoring, the system should provide user's context including activities, locations, and environments' states, to doctors, nurses, and families. Currently, sensors have more attention for remote monitoring owing to low cost and high privacy protection, than video cameras. However, one main drawback of the sensors such as accelerometer, gyro and pressure, is that the user must hold or be close to the sensors. It is inconvenient and uncomfortable, and sometimes people forget to attach such sensors.
Instead of these cameras and wearable devices, we are developing sensing and monitoring systems that do not require images nor wearable devices, which we call "Wireless Health Monitoring". In this article we introduce some of wireless health monitoring systems and importance of them.

Smart Sensors that realize Wireless Health Monitoring

Array sensor is one the smart sensors that we developed. Fig. 1 shows a trial product of the array sensor. The array sensor exploits an antenna array on the receiver side and decomposes received signals into eigenvectors and eigenvalues by eigenvalue decomposition (EVD). Depending on applications, it uses these components and sometimes other features, such as received signal strength (RSS). When an event (e.g. falling down) occurs, the propagation environment changes, and thus the features such as eigenvector and eigenvalue related to the propagation also change. Based on the change of the features, we can detect an event without using camera. We can detect and classify simple activities, just based on the 1st eigenvalue spanning signal subspace. For classification of more complex activities, such as sitting in a bathtub and falling in a bathroom, the array sensor can classify by using machine learning algorithm based on those features obtained by the array sensor. Since the array sensor exploits not an exact direction of arrival (DOA) information but the change of radio propagation, it does not need a precisely-designed array antenna where its antenna positions are designed precisely and it needs calibration; just plural antennas are needed so that the array sensor can be realized at low cost and is easy to install. The array sensor can also realize passive localization using SVM based on those features in a fingerprinting manner.

The other smart sensor good for monitoring elderly people while keeping privacy is the activity recognition system using low-resolution thermopile sensor arrays.

Professor Tomoaki Ohtsuki

Ph.D., SrMIEEE, FIEICE

Past BoG Member and TC Chair of WMC APSIPA

Professor

Department of Information and Computer Science

Keio University

Tomoaki Otsuki (Ohtsuki) received the B.E., M.E., and Ph. D. degrees in Electrical Engineering from Keio University, Yokohama, Japan in 1990, 1992, and 1994, respectively. He is now a Professor at Keio University. He is engaged in research on wireless communications, optical communications, signal processing, and information theory. Dr. Ohtsuki is a recipient of the 1997 Inoue Research Award for Young Scientist, the 1997 Hiroshi Ando Memorial Young Engineering Award, Ericsson Young Scientist Award 2000, 2002 Funai Information and Science Award for Young Scientist, IEEE the 1st Asia-Pacific Young Researcher Award 2001, the 5th International Communication Foundation (ICF) Research Award, 2011 IEEE SPCE Outstanding Service Award, the 27th TELECOM System Technology Award, ETRI Journal's 2012 Best Reviewer Award, and 9th CHINACOM '14 Best Paper Award. He served a Chair of IEEE Communications Society, Signal Processing for Communications and Electronics Technical Committee. He served a technical editor of the IEEE Wireless Communications Magazine and an editor of Elsevier Physical Communications. He is now serving an Area Editor of the IEEE Transactions on Vehicular Technology and an editor of the IEEE Communications Surveys and Tutorials. He has served general-co chair, symposium co-chair and TPC co-chair of many conferences, including IEEE GLOBECOM and ICC2011. He gave tutorials and keynote speech at many international conferences including IEEE VTC, IEEE PIMRC, and so on. He was a Vice President of Communications Society of the IEICE and is President-Elect of Communications Society of the IEICE. He is a fellow of the IEICE and a senior member of the IEEE.

Fig. 1 Array sensor

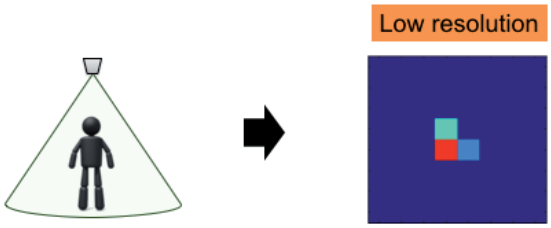


Fig. 2 Temperature distribution captured by low-resolution infrared array sensor placed on ceiling

The low-resolution thermopile array sensor has several infrared detectors (or pixels) inside. It can obtain the temperature distribution on a two-dimensional area. This kind of sensor is typically used for high performance home appliances (microwave oven and air conditioner), digital signage, automatic door, and so on. The activity recognition system using low-resolution thermopile sensor arrays installs the sensor on a ceiling and/or other places. Fig. 2 shows an example of temperature distribution obtained by the 6×6 low-resolution infrared array sensor place on ceiling. Different from images obtained by cameras, due to the low-resolution of sensor, we cannot distinguish individuals nor get much information about them, thus can keep privacy. Of course, it makes difficult to detect and classify activities, however, due to signal processing and machine learning algorithm, we can detect and classify even complex activities. The other advantage of the sensor is it can detect a person even in darkness by detecting infrared rays. The low-resolution infrared array sensor can also localize people easily.

Long-term Impact

One of the fundamental desires by people is to live healthy and peacefully. Healthcare is very important to realize it. Smart healthcare is a concept for an advanced healthcare. Smart healthcare is also expected in assisted living for elderly people. As mentioned above, the number of elderly people is rapidly increasing in many countries and that of elderly people living alone as well. Naturally, social costs for nursing care and medical expenses will rise. Wireless health monitoring is expected to support such a rapid aging society. One of the key factors for smart healthcare and wireless health monitoring is a privacy. In general, no one wants to be monitored with camera even by his/her child. Smart sensors that we introduced here will play an important role in our current and future aging society.

References

[1] S. Ikeda, H. Tsuji, and T. Ohtsuki, ``Indoor Event Detection with Signal Subspace Spanned by Eigenvector for Home or Office Security," Trans. of IEICE, E92-B no.7 pp. 2406-2412, July 2009

[2] T. Ohtsuki, ``(invited paper) Wireless Security and Monitoring System Using Array Antenna: Array Sensor," IEEE International Conference on Computing, Networking and Communications (ICNC2012), pp. 551-555, Maui, Hawaii, Jan. Feb. 2012

[3] J. Hong and T. Ohtsuki, ``State Classification with Array Sensor using Support Vector Machine for Wireless Monitoring Systems," Trans. of IEICE, Vol. E95-B, No.10, pp. 3088-3095, Oct. 2012

[4] Y. Inatomi, J. Hong, and T. Ohtsuki, ``Hidden Markov Model Based Localization Using Array Antenna," International Journal of Wireless Information Networks, Vol. 20, Issue 4, pp. 246-255, DOI: 10.1007/s10776-013-0211-y, June 2013

[5] J. Hong and T. Ohtsuki, ``Signal Eigenvector-based Device-Free Passive Localization using Array Sensor," IEEE Trans. on Vehicular Technology, vol. 64, no. 4, pp. 1354-1363, Apr. 2015

[6] Y. Agata, T. Ohtsuki, and K. Toyoda, ``Doppler Analysis Based Fall Detection Using Array Antenna," IEEE International Conference on Communications (ICC'2018), Kansas City, MO, USA, May 2018

[7] S. Mashiyama, J. Hong, and T. Ohtsuki, ``A Fall Detection System Using Low Resolution Infrared Array Sensor," IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC2014), Washington D.C., U.S.A., Sep. 2014

[8] S. Mashiyama, J. Hong, and T. Ohtsuki, ``Activity Recognition Using Low Resolution Infrared Array Sensor," IEEE International Conference on Communications (ICC'2015), pp. 495-500, London, UK, June 2015

[9] T. Ohtsuki, ``(Invited Paper) A Smart City Based on Ambient Intelligence," IEICE Transactions on Communications, Vol.E100-B, No.9, pp. 1547-1553, Sep. 2017

From Music Emotion Recognition to Music Video Generation

With the prevalence of mobile devices, video is widely used to record memorable moments such as weddings, graduations, and birthday parties. Popular websites such as YouTube have further boosted the phenomenon as broadcasting becomes easy. However, in music videos (MVs), movies, and television programs, music and video are often accompanied to complement each other to enhance emotional resonance and viewing experiences. Without soundtracks, most user-generated videos (UGVs) might look boring. Therefore, an automated process that can suggest a soundtrack to a UGV and make the UGV a music-compliant professional-like video is highly desirable. To this end, we have developed an emotion-based MV generation system that conducts soundtrack recommendation and video editing simultaneously.

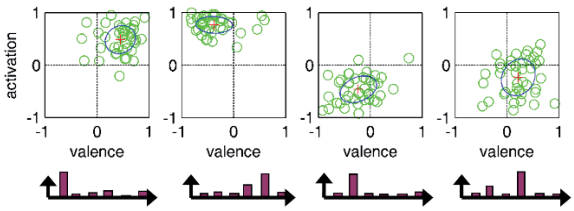


Figure 1: Subjects' annotations in the emotion space for four 30-second clips. The bottom figures show the acoustic feature representation for the corresponding clip in terms of the posteriori probabilities of an acoustic GMM.

Music Emotion Recognition and its Application to Music Video Generation

We started this research with music emotion recognition. Fig. 1 shows human emotion annotations of music on the valance-arousal (VA) plane. Each figure is the annotations of different persons for one music clip. We can see that the annotations are subjective, but aggregation of annotations indeed exists. It seems that the dimensional emotion of a song can be described by a bivariate Gaussian distribution; therefore, we can predict the emotion of a song as a single Gaussian distribution. In our ACM Multimedia 2012 paper [1], we introduced a probabilistic framework, called *Acoustic Emotion Gaussians* (AEG), to jointly model the auditory features (represented by the posteriori probabilities of an acoustic GMM as shown in Fig. 1) and emotion annotations of a song. As shown in Fig. 2, AEG consists of two mixture models, namely an acoustic Gaussian mixture model (GMM) for auditory feature reference and a VA GMM. Given a music clip, we first compute the posteriori probabilities of the acoustic GMM

Professor Hsin-Min Wang

PhD

Secretary (2013-2014), Speech, Language and Audio (SLA) Technical Committee, APSIPA

Research Fellow

Institute of Information Science

Academia Sinica

Hsin-Min Wang received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1989 and 1995, respectively. In October 1995, he joined the Institute of Information Science, Academia Sinica, Taipei, Taiwan, where he is currently a Research Fellow. He also holds a joint appointment as a Professor in the Department of Computer Science and Information Engineering, National Cheng Kung University. He currently serves as an Editorial Board Member of IEEE/ACM Transactions on Audio, Speech and Language Processing and APSIPA Transactions on Signal and Information Processing. His major research interests include spoken language processing, natural language processing, multimedia information retrieval, machine learning, and pattern recognition. He was the President of the Association for Computational Linguistics and Chinese Language Processing (2013-2015). He received the Chinese Institute of Engineers Technical Paper Award in 1995 and the ACM Multimedia Grand Challenge First Prize in 2012. He was an APSIPA Distinguished Lecturer (2014–2015). He is a life member of APSIPA.

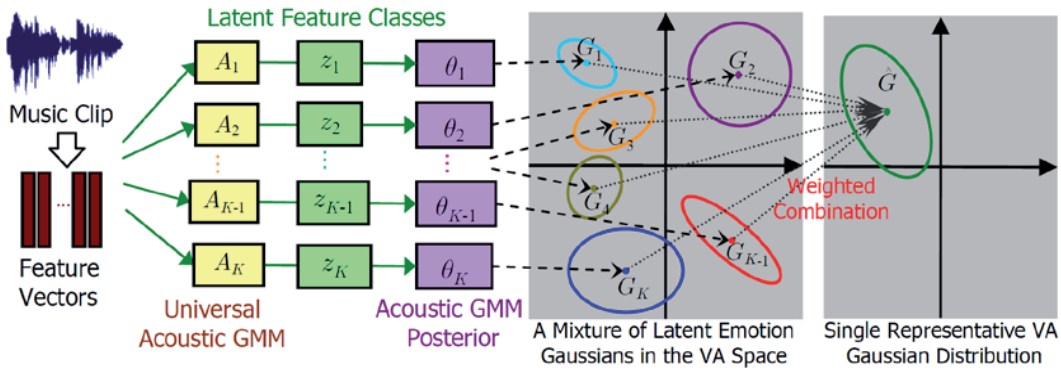


Figure 2: The generative process of the AEG model.

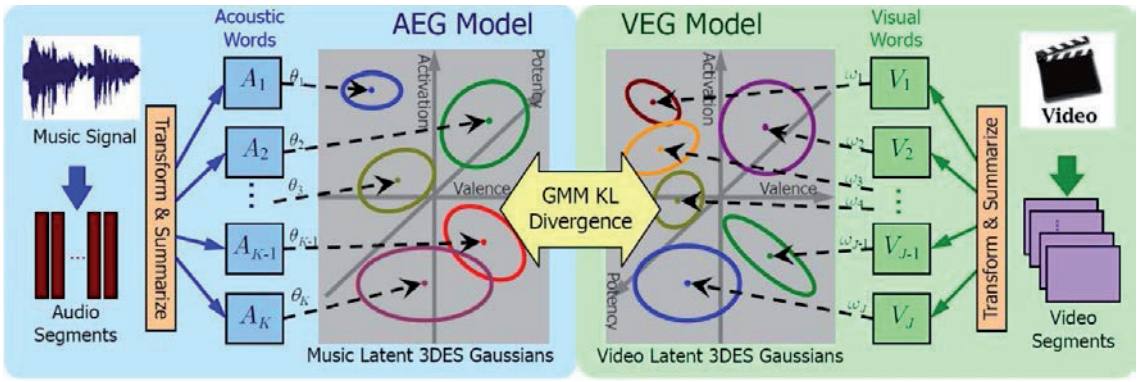


Figure 3. Illustration of the AVEG framework for music video generation.

components as the weights of the corresponding VA GMM components, and then find an approximate single VA Gaussian representation as the emotion annotation.

In ACM Multimedia 2012, Google proposed a multimedia grand challenge called automatic music video generation. The challenge involves two sub-tasks: the first is soundtrack recommendation; and the second is to recommend video to be played with a music clip. Since our music emotion annotation technique can be readily applied to video emotion recognition by modeling the relation between the visual features and the video emotion annotations, we can compute the distance between a music clip and a video clip based on their emotion Gaussian distributions. Therefore, we extended the AEG model to an *Acousticvisual Emotion Gaussians* (AVEG) model to fulfill the challenge [2]. The AVEG model was trained with a set of 65 emotion-annotated official MVs. This work finally won the Grand Challenge First Prize in ACM Multimedia 2012.

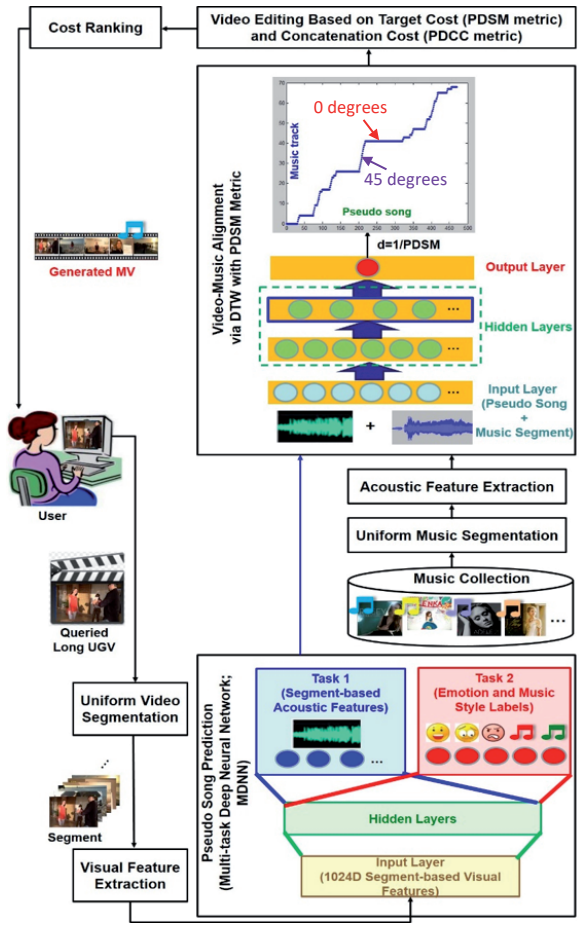


Figure 4: Music video generation based on semantic-oriented pseudo song prediction, matching, and video editing.

The music videos generated in this way are interesting but definitely not satisfactory. There are many issues remained to be solved. First, since the AVEG model matches music and video based on their average emotion annotations, it may recommend music with an emotional expression from happy to sad to a video with a reverse emotional expression. Second, even the music and video are with the same emotional category, the nonsynchronous emotional expression may still result in bad viewing experiences. Third, without object recognition, it can pair the image of a singing male and the vocal of a female, or pair the image of guitar playing and the piano accompaniment, and such mismatches also ruin the viewing experiences. Fourth, it only deals with soundtrack recommendation or video recommendation, and does not perform video editing.

Music Video Generation Based on Simultaneous Soundtrack Recommendation and Video Editing

Since 2012, we have been thinking about how to deal with the aforementioned issues. We mainly focus on the modeling of temporal emotion expression and the integration of soundtrack recommendation and video editing with the goal of suggesting a soundtrack to a long user generated video (UGV) and making the UGV a music-compliant professional-like video, and have proposed several less successful or incomplete methods [4-6]. Finally, in our ACM Multimedia 2017 paper [7], we proposed a semantic-oriented pseudo song prediction, matching, and video editing framework for MV generation. As shown in Fig. 4, given a UGV, it is first divided into a sequence of fixed-length short (e.g., 2 seconds) segments, and then a multi-task deep neural network (MDNN) is applied to predict the pseudo acoustic (music) features (or called the pseudo song) from the visual (video) features of each video segment. In this way, the distance between any pair of video and music segments of same length can be computed in the music feature space. Second, the sequence of pseudo acoustic (music) features of the UGV and the sequence of the acoustic (music) features of each music track in the music collection are temporarily aligned by the dynamic time warping (DTW) algorithm with a pseudo-song-based deep similarity matching (PDSM) metric. Third, for each music track, the video editing module selects and

concatenates the segments of the UGV based on the target and concatenation costs given by a pseudo-song-based deep concatenation cost (PDCC) metric according to the DTW-aligned result to generate a music-compliant professional-like video. Finally, all the generated MVs are ranked, and the best MV is recommended to the user.

It is worth noting that matching a music clip and a video clip in the music feature space could be more robust than matching them in the emotion space because the latter is susceptible to emotion recognition errors. In contrast, in the semantic-oriented pseudo song prediction method, the correlation among music, video, and semantic annotations such as emotion and music style is explored and modeled. In our experiments of music ranking conducted on 200 official MVs downloaded from YouTube, the new pseudo-song-based method could push ahead the ranking of the ground truth music track of a query video by 40 on average, compared to the above emotion-recognition-based method. From the time alignment results, we keep the video segments that correspond to the local paths at 45 degrees (cf. Fig. 4), because such locations indicate the synchronization of the video segment sequence and the music segment sequence. For those video segments aligned to a single music segment (i.e., the local paths at 0 degrees in Fig. 4), we select the video segment with the least target cost with respect to the music segment and concatenation costs with respect to the preceding and following video segments that have been selected. This idea is inspired by the target and concatenation costs widely used in unit selection-based text-to-speech (TTS) synthesis, whose goal is to synthesize a speech utterance that pronounces a given sentence with fluent quality. The MDNN for pseudo song prediction and the PDSM and PDCC metrics were trained by the same annotated official MV corpus mentioned above. The subjective 5-point mean opinion score (MOS) test on 5 long UGVs demonstrate that the generated music videos are generally satisfactory and can enhance human viewing and listening experiences.

We assume that the input UGV is longer than the candidate music tracks. If the UGN is shorter than the candidate music tracks, the system only performs soundtrack recommendation, i.e., it will skip the DTW process.

Long-term Impact

The current system could be very useful for end-users to accompany their home videos with suitable music, thereby making these videos more entertaining and attractive. However, the application is difficult to promote due to copyright protection. We may develop conditional music generation to replace soundtrack recommendation, i.e., to generate matching computer music based on the content of the video. Another advantage of this conditional music generation approach is that the video will be paired with a unique soundtrack. Since video editing is generally conducted by selecting and concatenating suitable video clips to best match the music track specified by a user, how to edit a video and generate matching music for it at the same time is a new research issue. Applying the technique to automatically generate an official MV given a music track is even more challenging. In addition to the problems of mismatch between the instrument being seen and the instrument being heard (or the gender being seen and the gender being heard) as mentioned above, the subtle out of synchronization between the lip motion and the singing voice is even harder to overcome. One possible solution is to generate matching computer animations according to the music content.

References

[1] J. C. Wang, Y. H. Yang, H. M. Wang, and S. K. Jeng, "The Acoustic Emotion Gaussians Model for Emotion-based Music Annotation and Retrieval," in Proc. of ACM MM 2012, pp. 89-98, October 2012.

[2] J. C. Wang, Y. H. Yang, I. H. Jhuo, Y. Y. Lin and H. M. Wang, "The Acousticvisual Emotion Gaussians Model for Automatic Generation of Music Video," in Proc. of ACM MM 2012, pp. 1379-1380, October 2012,

[3] J. C. Wang, Y. H. Yang, H. M. Wang, and S. K. Jeng, "Modeling the Affective Content of Music with a Gaussian Mixture Model," IEEE Trans. on Affective Computing, 6(1), pp. 56-68, March 2015.

[4] J. C. Lin, W. L. Wei, and H. M. Wang, "EMV-matchmaker: Emotional Temporal Course Modeling and Matching for Automatic Music Video Generation," in Proc. of ACM MM 2015, pp. 899-902, October 2015.

[5] J. C. Lin, W. L. Wei, and H. M. Wang, "DEMV-Matchmaker: Emotional Temporal Course Representation and Deep Similarity Matching for Automatic Music Video Generation," in Proc. ICASSP 2016, March 2016.

[6] J. C. Lin, W. L. Wei, and H. M. Wang, "Automatic Music Video Generation Based on Emotion-Oriented Pseudo Song Prediction and Matching," in Proc. of ACM MM 2016, pp. 372-376, October 2016.

[7] J. C. Lin, W. L. Wei, J. Yang, H. M. Wang, and H. Y. Liao, "Automatic Music Video Generation Based on Simultaneous Soundtrack Recommendation and Video Editing," in Proc. ACM MM 2017, pp. 519-527, October 2017.

Photo Gallery: APSIPA ASC'2013 in Kaohsung

VLSI Designs for Modern Block Ciphers in Constrained Environments

Professor M L Dennis Wong

BEng (Hons), PhD

TPC Co-Chair APSIPA ASC 2017

Deputy Provost

Heriot-Watt University Malaysia

Brief Biography:

M. L. D. Wong received his BEng(Hons) in Electronics and Communication Engineering and PhD in Signal Processing from the Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, UK. In 2004, Professor Wong joined the School of Engineering, Swinburne University of Technology Sarawak Campus as a Lecturer. Subsequently, he was appointed a Senior Lecturer in 2007 at the same institution. From 2011 to mid 2012, he was an Associate Professor at Xi'an Jiaotong Liverpool University, Suzhou, China. In June 2012 to Sept 2016, he was the Dean for Faculty of Engineering, Computing and Science at Swinburne Sarawak. Since October 2016, Professor Wong moved to Heriot-Watt University Malaysia to take up the role of Deputy Provost.

Professor Wong is also a Chartered Engineer registered with the Engineering Council UK; a Fellow of the IET; a Fellow and Chartered Professional Engineer with IEAust and a Senior Member of the IEEE. In 2014, he was electecd as the inaugural treasurer of the IEEE CIS Malaysia Chapter and subsequently the vice-chair in 2015. In 2016, he was elected as the inaugural chair for IEEE Sarawak Subsection.

In this age of ubiquitous computing, robust digital infrastructures enable users to generate over 2.5 quintillion bytes of data per day as of May 2018. The explosive increase in data generation creates new challenges and opportunities in various fields of study. In particular, ensuring information security for digital communication or transmission of these data has been an interesting challenge for several decades. With the advent of modern computing paradigm such as Internet of Things (IoT), we see an influx of highly constrained devices being interconnected to accomplish computational tasks. Said items include RFID devices, low-cost microcontrollers and sensor nodes in which hardware resources are extremely limited. To provide sufficient resistance to attacks from unauthorized adversaries, efficient hardware implementations of cryptographic primitives are necessary.

A secured cryptographic protocol incorporates several aspects ranging from secret key establishment to identity authentication. Block ciphers are a family of cryptographic primitives that are commonplace as components in such protocols. A block cipher is defined as a deterministic transformation that operates on a data block of fixed length in which the transformation is specified by a symmetric key used for both encryption and decryption. A secured block cipher ensures the encrypted data to be unintelligible without knowledge of the secret key. In this article, we explore the current state of hardware implementation of modern block ciphers and their optimizations.

Advanced Encryption Standard (AES)

AES is the most popular block cipher employed in digital cryptosystems. Originally known as the Rijndael cipher, the algorithm is standardized by the United States National Institute of Standards and Technology (NIST) in 2001. It operates on 128-bit data block and supports secret key lengths of 128, 192 or 256 bits. The design of the cipher is based on a substitution-permutation network (SPN) structure and features four transformations: non-linear substitution, bit permutation, linear transformation and key addition.

Among the four transformations in AES, the non-linear substitution (S-Box) is the most expensive in terms of hardware requirement. Naturally, hardware optimization of the cipher has been largely focused on said transformation. The input/output relationship of the AES S-Box can be referenced using the 8-to-8-bit truth table given in [1]. The mathematical equivalent to the AES S-Box transformation is the computation of multiplicative inverse over GF(2^8) followed by an affine transformation. The difficulty in deriving an economic circuit for this computation is mainly attributed to the complexity in evaluating the multiplicative inverse.

Composite Field Arithmetic (CFA)

To enable a more efficient construction for the AES S-Box, CFA can be adopted to map the computation to subfields of lower order. This approach has a significant benefit to the calculation of multiplicative inverse. In general, the process can be summarized in three steps:

1) Isomorphism function to map elements of the original field to a subfield.

2) Compute multiplicative inversion over the subfield.
3) Inverse isomorphism function to map the result to the original field.

In the case of AES, mapping of GF(2^8) to GF(((2^2)^2)^2) requires three stages of isomorphism and the associated field polynomials. These can be stated in general form as follow:

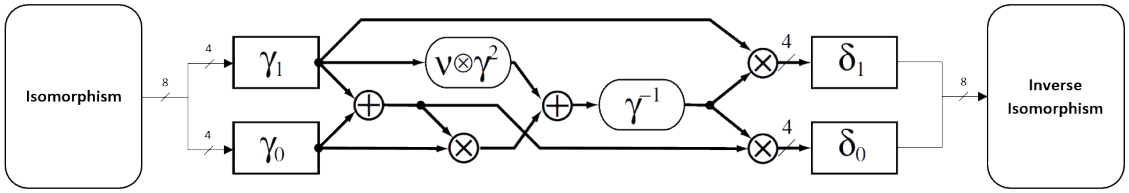


Figure 1: Example of a composite field architecture for AES S-Box.

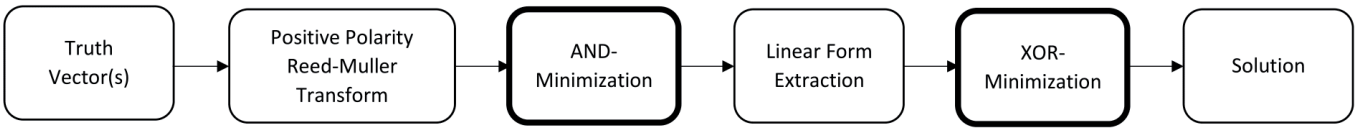


Figure 2: Two-step procedure in the proposed LMC algorithm.

1) GF(2^8)/GF(2^4): $r(y) = y^2 + \tau y + v$
2) GF(2^4)/GF(2^2): $s(z) = z^2 + Tz + N$
3) GF(2^2)/GF(2) : $t(w) = w^2 + w + 1$

In [2], we explored different coefficients $\{\tau, v, T, N\}$ in the field polynomial $r(y)$ and $s(z)$ to determine the composite field architecture with minimal arithmetic complexity. The optimized CFA architectures are also subjected to further architectural optimizations such as elimination of redundant common factor and merging of multipliers. Collectively, the above optimizations provided notable advantage for our proposed implementation in circuit area, critical path and throughput.

Low Multiplicative Complexity (LMC) Logic Design

To better minimize the circuit area of non-linear substitution circuits, we study a relatively new concept of LMC logic design. The premise of this approach is based on the Boyar-Peralta heuristic in [3] which suggests construction of a non-linear circuit with the minimal number of AND gates for low gate count implementation. The heuristic is suitable for complex functions which can be naturally decomposed into smaller components as it is not practical for large number of inputs. The original algorithm for LMC optimization is a two-step algorithm patented in [4]. Although effective in execution, the algorithm suffers from inconsistency due to reliance on randomness in its search process. In [5], we explored probabilistic improvement, new selection criterion and clearer parameters to improve the consistency and quality of results.

We also proposed a novel approach in [6] where we leverage the properties of polynomial decompositions to achieve LMC designs. A deterministic tree search algorithm is designed based on the decomposition procedure to search for solutions with optimal multiplicative complexity for a targeted function. Measures to include considerations for product sharing are described to maximize AND gate sharing between multiple functions of the same input set. The proposed algorithm is demonstrated to produce at least comparable (if not better) quality of results than the original algorithm on practical problems. Using the composite field architecture for the AES S-Box in [7] as basis, we applied the LMC algorithm to optimize the inversion circuit over GF(2^4) and observed significant improvement in circuit area over other CFA-optimized implementations.

Lightweight Block Ciphers

Although advanced optimization techniques have reduced the hardware footprint of the AES cipher significantly, the primitive is still not suitable for applications in heavily constrained environments. Instead, we see the introduction of new lightweight block ciphers to meet the stringent requirements. A non-exhaustive list of contemporary lightweight block ciphers can be reviewed in [8]. They are designed to have an advantage over AES mainly in circuit area and power/energy consumption.

Hardware Optimization of Lightweight Block Ciphers

Hardware optimizations on lightweight block ciphers are considered uncharted territories in comparison to the work done on AES. State-of-the-art implementations mostly rely on varying degree of architecture serialization to conserve hardware at the cost of several-fold increase in latency. However, this approach has severe implications in applications such as RFID where certain thresholds for response time must be met. To put it into perspective, a recent implementation of the PRESENT cipher in [9] reported a latency of 136 cycles in contrast to the recommended latency of 50 cycles for EPCGlobal Gen2 RFID in [10].

Despite various design differences, the core transformations involved in most lightweight block ciphers are relatively similar to each other. We explore hardware optimizations on common cryptographic transformations such as non-linear substitution, linear transformations, key scheduling mechanism and generation of round constants to enable hardware savings for round-based implementations of lightweight block ciphers. For example, non-linear substitution for lightweight block ciphers are mostly

computed using 4-bit S-Boxes which benefit directly from our LMC logic minimization technique as in the case of AES. On the other hand, linear transformations which involve finite field multiplications of different magnitude can benefit from circuit sharing. Through optimizations on the transformations, hardware savings can be achieved mostly through trade-offs with critical path instead of latency. This is much more desirable for the nature of applications for the lightweight block ciphers.

Long-term Impact

While the focus on constrained environments in this article may suggests the devaluation of the AES cipher, we emphasize that the role of the AES cipher is still prominent in the bigger picture. Devices on the higher end of the spectrum that can easily afford the hardware such as servers, desktops and smartphones are always recommended to implement the AES cipher for better security due to robust cryptanalysis on the cipher over the last two decades. Hence, optimization efforts on the cipher will continue to be valuable in the future.

On the other hand, lightweight block ciphers are relatively new. NIST has only recently reported on the plan to standardize lightweight primitives for constrained applications [11]. Hardware optimizations on these ciphers still have a lot of potential and will allow us to better understand the capabilities of each cipher. Implementation results will also help in the selection process for a standard when the time comes. The goal is to have an efficient cipher to be standardized alongside AES to fully cover applications between both ends of the device spectrum.

References

[1] *Advanced Encryption Standard (AES)*, National Institute of Standards and Technology Federal Inf. Process. Stds. (NIST FIPS) – 197, 2001.

[2] M. M. Wong, M. L. D. Wong, A. K. Nandi, and I. Hijazin, "Construction of optimum composite field architecture for compact high-throughput aes s-boxes," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 6, pp. 1151-1155, Oct 2006.

[3] J. Boyar, P. Matthews, and R. Peralta, "Logic minimization techniques with applications to cryptology," *Journal of Cryptology*, vol. 26, no. 2, pp. 280-312, Apr 2013.

[4] R. Peralta and J. Boyar, "Method of optimizing combinational circuits," Nov. 20 2012, US Patent 8,316,338. [Online]. Available: https://www.google.com/patents/US8316338.

[5] J. J. Tay, M. L. D. Wong, M. M. Wong, C. Zhang, and I. Hijazin, "Low multiplicative complexity logic minimisation over the basis (AND, XOR, NOT)," *Electronics Letters*, vol. 52, no. 17, pp. 1438-1440, Aug. 2016.

[6] J. J. Tay, M. L. D. Wong, M. M. Wong, C. Zhang, and I. Hijazin, "A tree search algorithm for low multiplicative complexity logic design," *Future Generation Computer Systems*, vol. 83, pp. 132-143, 2018.

[7] D. Canright, "A very compact s-box for aes," in *Cryptographic Hardware and Embedded Systems – CHES 2005*, Springer Berlin Heidelberg, 2005, pp. 441-455.

[8] A. Biryukov and L. Perrin, Lightweight block ciphers. [Online]. Available: https://www.cryptolux.org/index.php/Lightweight_Block_Ciphers.

[9] C. A. Lara-Nino, A. Diaz-Perez, and M. Morales-Sandoval, "Lightweight hardware architectures for the present cipher in fpga," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 9, pp. 2544-2555, Sept 2017.

[10] M. J. O. Saarinen and D. Engels, "A do-it-all-cipher for rfid: Design requirements (extended abstract)," Cryptology ePrint Archive, Report 2012/317, 2012, http://eprint.iacr.org/2012/317.

[11] K. A. McKay, L. E. Bassham, M. S. Turan, and N. W. Mouha, "Report on lightweight cryptography," National Institute of Standards and Technology, Interagency/Internal Report (NISTIIR) – 8114, 2017.

Photo Gallery: APSIPA-ASC'2014 in Siem Reap

Affective Computing for Mental Health Care

Professor Chung-Hsien Wu

PhD

SLA Technical Committee Chair, APSIPA

Chair Professor

Department of Computer Science and Information Engineering

National Cheng Kung University, Taiwan

Chung-Hsien Wu *received the M.S. and Ph.D. degrees in electrical engineering from National Cheng Kung University (NCKU), Tainan, Taiwan, in 1987 and 1991, respectively. Since 1991, he has been with the Department of Computer Science and Information Engineering, NCKU. He became the Distinguished Professor and Chair Professor in 2004 and 2017, respectively. He also worked at Computer Science and Artificial Intelligence Laboratory of Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in summer 2003, as a Visiting Scientist. He was an Associate Editor of the IEEE Transactions on Audio, Speech and Language Processing (2010–2014) and the IEEE Transactions on Affective Computing (2010–2014). He is currently an Associate Editor of ACM Transactions on Asian and Low-Resource Language Information Processing*, and *APSIPA Transactions on Signal and Information Processing. He served as the Asia Pacific Signal and Information Processing Association (APSIPA) Distinguished Lecturer and Speech, Language and the Audio (SLA) Technical Committee Chair in 2013–2014. He was the President of the Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taiwan (2009-2011). He received the Outstanding Research Award of Ministry of Science and Technology, Taiwan, in 2010 and 2016. His research interests include deep learning, affective computing, speech recognition/synthesis, and spoken language processing.*

Affective Computing is a rapidly emerging field which has a goal to recognize the emotional state of a user for rational decision making, social interaction, perception, and memory [1]. There are numerous approaches to achieve the goal by using single/multiple modalities, including textual modality, audio-visual modalities and physiological modalities [2][3][4]. One of the most significant research domains of affective computing is directed towards the interrelation between emotions and human health, both mental and physical. In the past years, extensive research on affective computing has helped the medical community with technologies for better understanding of emotions, identifying their impact on health, and offering new techniques for diagnosis, therapy, and treatment of emotionally-influenced diseases [5]. With the growing and varied uses of human-computer interactions, people exhibit emotions that in certain contexts might influence their health. The availability and constant development of technologies that can facilitate the application of affective computing in the medical realm. Recently, the technology of affective computing is now poised on the threshold of usability for the process of monitoring, recognition, as well as expression of emotions for various medical purposes.

Audiovisual Emotion Recognition

Typically, a complete emotional expression is expressed by more than one utterance in natural conversation, and in more detail, each utterance may contain several temporal phases of emotional expression. Accordingly, a single HMM with left-to-right topology is unable to model the temporal course of emotional expression in natural conversation effectively [6]. Fig. 1 shows that when the emotional state (i.e., happiness) of Speaker 1 is evoked through conversation, Utterance 1 only covers the temporal phase of onset, while the apex and offset phases are covered in Utterance 2. For single HMM-based model training, the temporal information is lost by diverse training samples with complex temporal structures. Thus, for each emotional state, if all the temporal phases of the onset, apex, and offset are assumed to appear in one utterance, and a single HMM with left-to-right topology is used for emotional state modeling, the performance may be degraded for emotion recognition of multiple utterances with complex temporal structures in natural conversation. An effective bimodal fusion strategy in a real conversational environment is desirable to model the complex temporal structure.

To address this problem, a bimodal hidden Markov model (HMM)-based emotion recognition scheme, constructed in terms of sub-emotional states, which are defined to represent temporal phases of onset, apex, and offset, is adopted to model the temporal course of an emotional expression for audio and visual signal streams.

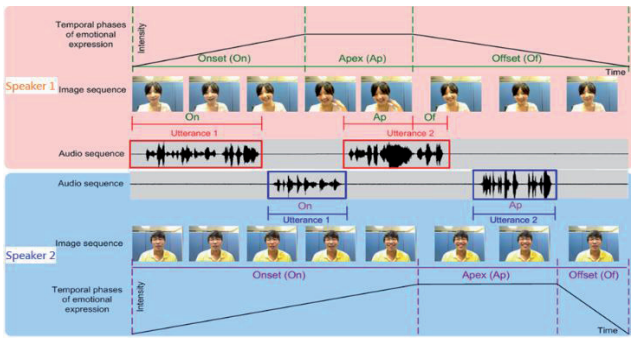


Fig. 1. An example of various temporal phases of happy emotional expression occurred to different utterances in a real conversational environment.

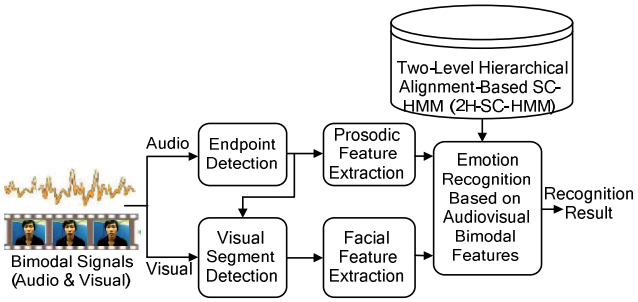


Fig. 2. Block diagram of the proposed audiovisual emotion recognition system.

A two-level hierarchical alignment mechanism is proposed to align the relationship within and between the temporal phases in the audio and visual HMM sequences at the model and state levels in a proposed semi-coupled hidden Markov model (SC-HMM). Furthermore, by integrating a sub-emotion language model, which considers the temporal transition between sub-emotional states, the proposed two-level hierarchical alignment-based SC-HMM (2H-SC-HMM) can provide a constraint on allowable temporal structures to determine an optimal emotional state. Fig. 2 shows the block diagram of the proposed bimodal emotion recognition mechanism. First, the input bimodal signals are divided into audio and visual parts. Because endpoint detection based on speech is more robust than that based on mouth movement, the start and end points of the speech segment were determined to obtain the time-aligned visual segment. The extracted speech and visual segments are then used to extract prosodic and facial features, respectively. Finally, using the extracted prosodic and facial features, the proposed 2H-SC-HMM is employed for emotion recognition. Experimental results show that the proposed approach can yield satisfactory results in both the posed MHMC and the naturalistic SEMAINE databases [7], and shows that modeling the complex temporal structure is useful to improve the emotion recognition performance, especially for the naturalistic database (i.e., natural conversation). The experimental results also confirm that the proposed 2H-SC-HMM can achieve an acceptable performance for the systems with sparse training data or noisy conditions.

Mood Disorder Detection

Mood disorder is a kind of mental illness which severely affects how you feel and think. According to the diagnostic and Diagnostic and Statistical Manual of Mental Disorders (DSM-V) published by American Psychiatric Association, mood disorder can be categorized into unipolar depression (UD) and bipolar disorder (BD) [8]. UD repeats the cycle of two states: euthymia and depression (low). Different from UD, BD experiences two opposite and extreme emotional states: mania (high) and depression (low) through euthymia, where euthymia is used to refer to the neutral mood. Fig. 3 shows the distinction between UD and BD with respect to three kinds of states. The patients with UD will only swing between euthymia and depression, while some patients with BD will swing between mania and depression through euthymia and some exhibit behaviors simultaneously associated with both manic and depressive episodes. Manic episodes and depressive episodes vary according to individual differences. In general, each period may take from a few days to a few weeks. Patients who suffer from BD will experience multiple periods of mood swing over a lifetime. There may be an interval of a few weeks, months or years between periods. In this interval, patients will get back to normal, namely "remission". In mania, patients will become excited, energetic, impulsive, and so on. In depression, patients will have feelings such as sadness, disinclination, and self-hatred. These two emotional states relapse regularly much more than normal people in intensity and duration.

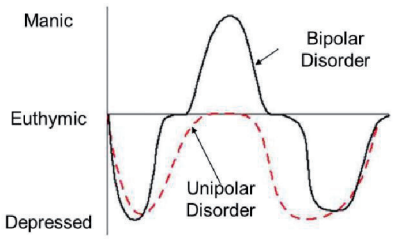


Fig. 3. Distinction between UD and BD with respect to three kinds of states.

As a mood is an emotional state, emotion expression or perception are relevant to mood disorder, and has been used for mood disorder detection. In previous studies, most of the approaches to mood disorder detection focused on long-term tracking [9]. To distinguish BD from UD, a mood database should be collected first for system training and evaluation. To the best of our knowledge, as there is no database for short-term detection for discriminating BD from UD currently, this work collected two databases containing a self-collected MHMC-EM emotional database and a clinician-collected CHI-MEI mood disorder database for model training and evaluation on mood disorder detection [10].

We cooperated with Chi-Mei Medical Center in Taiwan to collect a database containing the elicited facial expressions and speech responses of the patients with UD or BD. The serial number of the project approved by the Institutional Review Board (IRB) of Chi-Mei Medical Center is 10403-002. We used six emotional videos, including happiness, fear, surprise, anger, sadness and disgust to elicit expressions of the subjects. The subjects' facial expressions and the speech responses of the subjects in the following interviews with a clinician after watching each eliciting video were collected to construct the CHI-MEI mood speech database. Fig. 4 shows the mood data structure for each facial and speech response after watching the corresponding eliciting video.

Before data collection, each participant with BD or UD will be assessed by the doctor to check if his/her physical and mental state is stable before participating in the evaluation.

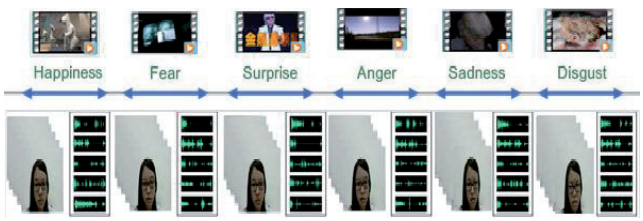


Fig. 4 The mood data structure for each facial and speech response after watching the corresponding eliciting video.

In this study, eliciting emotional videos are firstly used to elicit the patients' emotions. After watching each video clips, their facial expressions and speech responses are collected when they are interviewing with a clinician. In mood disorder detection, the facial action unit (AU) profiles and speech emotion profiles (EPs) are obtained respectively by using the support vector machines (SVMs) which are built via facial features and speech features adapted from two selected databases using a denoising autoencoder-based method. Finally, a Coupled Hidden Markov Model (CHMM)-based fusion method is proposed to characterize the temporal information. The CHMM is modified to fuse the AUs and the EPs with respect to six emotional videos.

As this work focuses on speech emotional expression and the changes of facial expressions responding to emotional stimuli, a database big enough with manual labels, consisting of labeled AUs and emotions, is important for training the speech emotion model and the AU model for emotion and AU profile generation. Because the collected CHI-MEI mood database is small and difficult for emotion labeling, in order to deal with the small data problem, we apply a domain adaptation method called Hierarchical Spectral Clustering-based denoising Autoencoder (HSC-DAE) to improve system performance. Inspired by the stacked denoising autoencoder domain adaptation, in this study, we first use a cluster-based linear transform method, Hierarchical Spectral Clustering, to adapt the source domain data to the target domain and generate the transformed data which are the source domain data with some "relevant noises". Then, we use the transformed data as the input to train a denoising autoencoder (DAE) to reconstruct the source domain data as the domain-adapted data for further process [10].

The system framework for mood disorder detection is shown in Fig. 5. In the framework, first, the facial features of the CK+ database and the speech features of the eNTERFACE database are extracted and adapted to that of the CHI-MEI database using the HSC-DAE. The adapted facial data and speech are then used for training the support vector machine (SVM)-based AU detector and the emotion detector, respectively. Based on the constructed SVM-based AU detector and the emotion detector, the AU profiles and the emotion profiles are generated for feature representation. As each AU profile only characterizes one facial image, the AU profile sequence corresponding to the entire image sequence for one question response is extracted for facial feature representation. Finally, the Coupled Hidden Markov Model (CHMM)-based multimodal fusion method integrating the AU and the emotion profile sequence which characterize the temporal context of facial expression and speech emotion is adopted for mood disorder detection. Experimental results show the promising advantage and efficacy of the CHMM-based fusion approach for mood disorder detection [11].

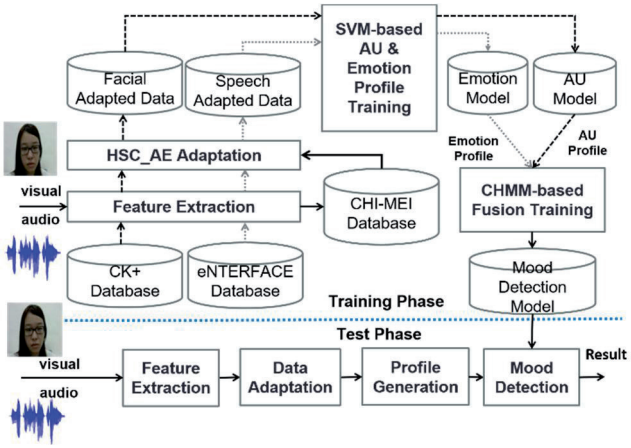


Fig. 5 The system framework for mood disorder detection

Long-term Impact

Previous research has shown that the medical community has started to realize that emotions play an important role in the preservation of human's mental health. Applying affective computing technology to medical practice is only in its beginning phase and many domains are yet to be explored. With the advancement in each of the sub-areas of affective computing based on text, speech, face and gesture expressions, and physiology, we can expect substantial increases in the interest for emotionally-intelligent applications in the mental health care domain. Currently, there are still several issues that need to be further explored in the future. First, most of the current recognition methods from all of the modalities are focused on acted versus naturally induced emotions. Besides, only a small set of basic emotions and the affective modalities (text, speech, face, gesture etc.) are still investigated separately which is different from how humans communicate emotions. Second, affective computing and medical informatics are currently two separate disciplines with only limited synergies realized until recently. It is desirable to come out a new integrative approach by sharing common analyses and domain knowledge to advance practical applications. Finally, exploring the emotional expression styles from different users is a viable direction for effective emotion recognition, which is not only related to the expression intensity, but also related to the expression manner and may be significantly associated to personality trait of the user.

References

[1] R. W. Picard, *Affective Computing*. MIT Press, 1997.

[2] C.-H. Wu and W.-B. Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels," *IEEE Trans. Affective Computing*, VOL. 2, NO. 1, January-March 2011, pp. 10~21.

[3] C.-H. Wu, J.-C. Lin, W.-L. Wei, "A Survey on Audiovisual Emotion Recognition: Databases, Features, and Data Fusion Strategies," *APSIPA Transactions on Signal and Information Processing*, Vol. 3, e12, published online: 11 November 2014.

[4] J.-C. Lin, C.-H. Wu and W.-L. Wei, "Error Weighted Semi-Coupled Hidden Markov Model for Audio-Visual Emotion Recognition," *IEEE Trans. Multimedia*, Vol. 14, No. 1, January 2012, pp.142~156.

[5] A. Luneski, E. Konstantinidis, and P. Bamidis, "Affective Medicine: a review of Affective Computing efforts in Medical Informatics. Methods of Information in Medicine," 49 (3), pp. 207-218, 2010.
[6] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Two-Level Hierarchical Alignment for Semi-Coupled HMM-Based Audiovisual Emotion Recognition with Temporal Course," *IEEE Trans. Multimedia*, VOL. 15, NO. 8, December 2013, pp.1880-1895.
[7] G. Mckeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpous of emotionally coloured character interactions," *Proc. IEEE Int'l Conf. on Multimedia and Expo*, pp. 1079–1084, 2010.
[8] A. P. Association, Diagnostic and statistical manual of mental disorders (fifth edn), American Psychiatric Association, 2011.
[9] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. La Torre, "Detecting depression from facial actions and vocal prosody," in Proc. IEEE International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1-7, 2009.
[10] K.-Y. Huang, C.-H. Wu, M.-H. Su, and Y.-T. Kuo, "Detecting Unipolar and Bipolar Depressive Disorders from Elicited Speech Responses Using Latent Affective Structure Model," in *IEEE Trans. Affective Computing*, DOI:10.1109/TAFFC.2018.2803178, 2018.
[11] T.-H. Yang, C.-H. Wu, K.-Y. Huang, M.-H.g Su, "Coupled HMM-based Multimodal Fusion for Mood Disorder Detection through Elicited Audio-Visual Signals," *Journal of Ambient Intelligence and Humanized Computing*, Special Issue on Media Computing and Applications for Immersive Communication, Vol. 8, No. 6, November 2017, pp.895–906.

Photo Gallery: APSIPA-ASC'2015 in Hong Kong

Image Quality Assessment for the Screen Content Images

In the era of multimedia communications, mobile and cloud computing, and the Internet of Things, the contents of digital images are no longer just limited to natural scenes. In fact, the contents of digital images nowadays can have a mixture of sources, such as natural scene, computer-generated graphics, texts, charts, maps, user's hand-writing and -drawing, and even some special symbols or patterns (e.g., logo, bar code, QR code) imposed and rendered by an electronic device or a photo editing software. Such kind of images is denoted as the *screen content images* (SCIs), and they are frequently encountered in various multimedia applications and services, such as online news and advertisement, online education, electronic brochures, remote computing, cloud gaming, to name a few. Refer to Fig. 1 for some SCI examples. It has been observed that the SCIs tend to have sharp edges, and high-contrast and vivid few colors in certain regions, which yield fairly different image characteristics from the natural images.

One critical issue associated with the SCIs is: how to conduct *image quality assessment* (IQA) for this kind of images? Since the human eyes are the final receiver of the images, IQA becomes an important issue in the field of image processing task with the goal of objectively evaluating the image quality in accordance with the *human visual system* (HVS). Besides being used for perceptual quality assessment, an IQA model for the SCIs can be also exploited as an effective *performance index* to guide the development of various SCI-based image processing algorithms (e.g., coding, interpolation, super-resolution, enhancement, and so on). Moreover, considering that the majority of existing IQA database and models are developed for the natural images, they cannot be directly exploited to conduct IQA for the SCIs. Therefore, IQA for the SCIs becomes an emerging and hot image technological research topic with both theoretical and practical values.

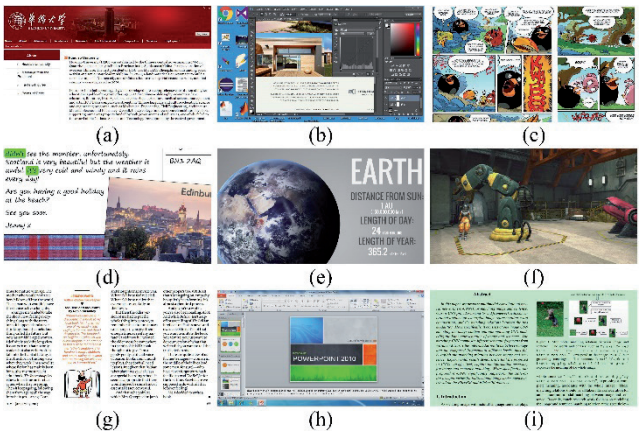


Fig. 1. Nine SCIs selected from our newly established SCI database (containing nearly 2,000 images) for demonstrations.

Professor Huanqiang Zeng

PhD, SrMIEEE ***TPC of APSIPA-ASC2017***

Professor

School of Information Science and Engineering

Huaqiao University, China

Huanqiang Zeng *received the B.S. and M.S. degrees from Huaqiao University, China and the Ph.D. degree from Nanyang Technological University, Singapore, all in electrical engineering. He was a Postdoctoral Fellow at the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong from 2012 to 2013, and a Research Associate at the Temasek Laboratories, Nanyang Technological University, Singapore in 2008. He is now a Professor at the School of Information Science and Engineering, Huaqiao University, Xiamen, China.*

His research interests are in the areas of image processing, video coding, and computer vision. He has published more than 80 papers in well-known international journals and conferences. He has been actively serving as the Associate Editor for IET Electronics Letters and International Journal of Image and Graphics, Guest Editor for multiple international journals (e.g., JVCIR), the General Co-Chair for IEEE ISPACS2017, the Technical Program Co-Chair for APSIPA ASC2017, the Area Chair for IEEE VCIP2015, the Technical Program Committee Member for multiple flagship international conferences. He received the Best Paper Award from CCSP2017. He is a senior member of IEEE, and a Member of International Steering Committee of International Symposium on Intelligent Signal Processing and Communication Systems.

IQA Database for the SCIs

To investigate IQA for the SCIs, one of our significant contributions is our newly established IQA database for the SCIs (denoted as SCID).

Our developed SCID contains 40 *reference* SCIs and 1,800 *distorted* versions rendered from these reference SCIs. The 40 reference SCIs were thoughtfully identified from the Internet, and they cover a wide variety of image contents, including texts, graphics, symbols, patterns, and natural images. For demonstration, few reference SCIs from our database are shown in Fig. 1. Since various types of distortions could be inevitably introduced on SCIs during the acquisition, processing, compression, transmission, and display stages, our SCID database includes 9 types of distortions that are often encountered in practical applications. These 9 types of distortions include the Gaussian noise (GN), the Gaussian blur (GB), and the motion blur (MB), the contrast change (CC), color saturation change (CSC), color quantization with dithering (CQD), JPEG, JPEG2000 (J2K), and HEVC-SCC. For each distortion type, 5 levels of degradations (ranging from imperceptible level to highly annoying one) are generated and included in our database. As a result, we have produced 1,800 distorted SCIs for our database. After that, the subjective rating, *mean of score* (MOS) computation and its reliability analysis are performed by strictly following the standard ITU-R BT.500-13 to obtain a final reliable MOS for each distorted SCI. Finally, our SCID, consisting of the reference and the corresponding distorted SCIs, and the corresponding MOSs, is made publicly available in http://smartviplab.org/pubilcations/SCID.html. Our SCID database can be served as the 'ground truth' to quantitatively assess how accurate of the proposed IQA model compared with that of existing state-of-the-art models on the evaluation of SCIs.

IQA Model for the SCIs

To evaluate the perceptual quality of the SCIs, we mainly developed two IQA models for the SCIs, namely, the *edge similarity* (ESIM) and the *Gabor feature-based model* (GFM). In what follows, they will be succinctly discussed.

In the proposed ESIM, a parametric edge model is firstly used to extract two salient edge attributes, *edge contrast* and *edge width*, and this process will be applied to the distorted SCI and the reference SCI, respectively. This modeling process will be conducted at each pixel location, individually and independently. As a result, the extracted edge contrast and edge width information are expressed in terms of *maps*—i.e., the *edge contrast map* (ECM) and the *edge width map* (EWM), respectively. It is important to note that these maps have the same size as that of the input image. In addition, the *edge direction*, which is another salient edge feature, is considered and incorporated into our proposed IQA model. The *edge direction map* (EDM) will be generated directly from each SCI via our proposed edge direction computation method. In the second stage, the computed edge feature maps, one from the reference SCI and the other from the distorted SCI, will be compared to yield their edge similarity measurement. For example, the two ECMs, respectively obtained from the distorted SCI and the reference SCI, will be compared to arrive at the *edge contrast similarity* (ECS) map. Likewise, the *edge width similarity* (EWS) map and *edge direction similarity* (EDS) map will be generated based on the corresponding pair of EWMs and EDMs, respectively. The three generated similarity measurement maps will be combined to yield *one* measurement map, which is used as the input of the third, and the last, stage to compute the final ESIM score using our proposed edge-width-based pooling process.

The proposed GFM is motivated by the fact that an image representation yielded by a set of properly-chosen Gabor filters is highly consistent with the response or judgement as made by the HVS when the human eyes view the image. In our approach, therefore the Gabor features are first extracted from the *luminance* (i.e., the L component recorded in the LMN color space) of the reference and distorted SCI, separately. On this feature-extraction process, a specially-designed Gabor filtering (i.e., the imaginary part with odd symmetry) is conducted on the horizontal and the vertical directions, respectively. The obtained filtering results are combined to form the *Gabor feature map*. The *degree of similarity* measurement is then conducted on these maps for the luminance part and for the chrominance components independently, between the reference and distorted SCIs. Finally, the developed *Gabor-feature pooling* strategy is employed to combine these measurements and generate the final GFM score for the SCI under evaluation. Experimental simulation results have shown that the proposed ESIM and GFM not only yield higher consistency with the HVS perception on the evaluation of the SCIs but also require lower computational complexity, compared with that of the classical and several state-of-the-art IQA models.

Long-term Impact

As a newly emerged media, the SCI has attracted a lot of attentions from both academic and industrial communities. Consequently, a great progress has been made in the SCI-related research field, including the IQA for SCI as we discussed. However, it is worthwhile of mentioning that there are still many unsolved problems in this topic, for example, 1) the development of IQA for the SCI is highly related to its applications. New applications of the SCIs will definitely require new design of IQA model; 2) how to comprehensively evaluate and compare the performances of various IQA models across different IQA databases is still an essential and open topic in the field of IQA; 3) besides the IQA for the screen content image, the IQA for screen content video is also meaningful to be investigated. All these stress the need of new IQA databases and models in the future.

References

[1] Y. Fu, H. Q. Zeng, J. Q. Zhu, L. Ma, Z. K. Ni, and K.-K. Ma, "Screen Content Image Quality Assessment Using Multi-Scale Difference of Gaussian," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[2] Z. K. Ni, H. Q. Zeng, L. Ma, J. H. Hou, J. Chen, and K.-K. Ma, "A Gabor Feature-based Quality Assessment Model for the Screen Content Images," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4516-4528, September 2018.

[3] Z. K. Ni, L. Ma, H. Q. Zeng, J. Chen, C. H. Cai, and K.-K. Ma, "ESIM: Edge Similarity for Screen Content Image Quality Assessment" *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4818-4831, October 2017.

[4] Z. K. Ni, L. Ma, H. Q. Zeng, C. H. Cai, and K.-K. Ma, "Gradient Direction for Screen Content Image Quality Assessment" *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1394-1398, October 2016.

Disentangle Speech Information

Dong Wang

PhD

Vice Chair of SLA
Distinguish Lecture 2017
APSIPA

Associate Professor

Center for Speech and Language Technologies

Tsinghua University

Prof. Dong Wang received his Ph.D. degree (supported by a Marie Curie fellowship) from CSTR, University of Edinburgh, in 2010. He worked in Oracle China, IBM China, EURECOM France and Nuance US. He is now an Associate Professor with Tsinghua University, Beijing, China.

The research interest of Prof. Dong Wang includes speech processing, language processing and financial processing. He published more than 100 academic papers (including 3 best papers) and more than 20 patents. He is a senior member of IEEE, and has served in APSIPA as the vice chair of the SLA track since 2016. He also serves as the committee numbers of several international conferences, including NCMMSC, ISCSLP and O-COCOSDA.

A central difficulty of modern speech processing tasks is caused by the entanglement of information of different kinds at different levels, e.g., the intermixing of linguistic content, speaker traits and emotion. Most of the present speech processing algorithms focus on a particular information type and treats other as interference, which is clearly not the way that our human beings do in our brain. An ideal approach would process all the information simultaneously, and disentangle them in a harmonic way, i.e., letting different types of information help each other and support each other during the disentanglement. This leads to a new perspective for speech processing, which treats speech as a stream of information compound rather than raw signal. By this perspective, speech processing will be performed in an *information space*, and working in this space will transfer the conventional *speech signal processing tasks* to *speech information processing tasks*. This article will describe some of our recent work towards this direction, and envision the broad impact that the new idea may cause for the speech community.

Complexity caused by entanglement

Speech is a very special kind of signal, in the sense that it is the only complicated signal that humans produce intentionally and actively. Due to the special role of language in our daily life, speech signals involve very complex information, ranging from the low-level ones such as linguistic content, speaker traits, emotion, physical status, to the high-level ones often named as knowledge and philosophy. This multitude of information is embedded in the one-dimensional vibration and entangled with each other in an unknown manner, which makes speech processing tasks extremely difficult.

Human beings have dramatic capabilities to extract subtle information from complex speech signals, which is the foundation of our daily communication. One of the most eminent talents of our brain in the listening task is that it can process multiple types of information instantly and simultaneously. For instance, whenever we pick up the ringing phone and listen to the first word from the other side, we can quickly identify who is speaking, what he/she is saying, which language is used, and in what emotion the speaker is talking.

It has been a long-term dream to grant machines the same capability of extracting useful information from speech, motivated by the desire of establishing a friendly human-machine interface. After more than 50 years of research, tremendous progress has been achieved on some individual speech processing tasks, in particular automatic speech recognition (ASR) and speaker recognition (SRE). These achievements should be largely credited to the development of deep learning techniques and the accumulation of large data resources.

However, we are still far from human-like speech processing. A major shortage is that machines have not gained the capacity of disentangling information of different kinds that are intermixed in speech signals. Almost all the present algorithms are designed for processing single information, with little consideration of dependency among information of different types.

Information factorization

We call the speech information disentanglement process as information factorization. It is certainly a very difficult task, and what we have done is a small bit of the whole picture.

Our work started by stablishing a model that describes how information of different kinds is composed in speech. We then quickly found that many years ago (1996), Prof. Fujisaki had proposed an elegant information convolution framework that describes how speech is produced in the perspective of information, as shown in Fig. 1. In this framework, the speech signal is assumed to be a cascaded composition of various types of information, including linguistic content, para-linguistic alternation and non-linguistic modulation. Compared to the conventional source-filter model, the Fujisaki model is more information oriented: it does not care how the speech is produced

physically; instead, it is more about how the information is embedded in the signal. Following this model, information factorization can be achieved by a cascade deconvolution.

Although very elegant in concept, it is not easy to perform the deconvolution, as the exact way that different types of information are composed in speech is far from known. To solve this difficulty, many factorization methods hypothesize specific structures for the convolution process, usually Gaussian and linear. The famous GMM-UBM model in speaker recognition is a typical example. By this model, the acoustic space is firstly divided into multiple subregions that can be roughly regarded as phones, and then speaker traits are represented as the shift of the speaker GMM on each subregion. By this way, speech signals are factorized into two factors: linguistic content and speaker. This factorization is based on the Gaussian (each subregion) and linear (shift) assumption.

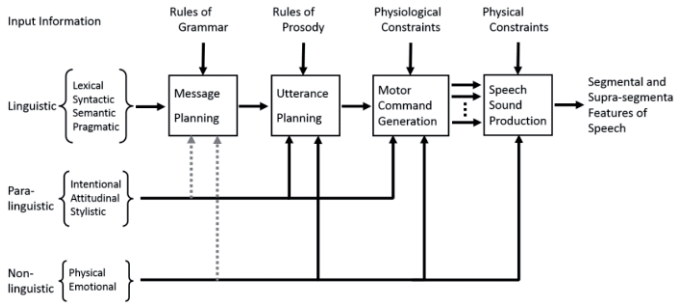


Fig.1 Fujisaki speech information convolution model [1].

The factorization methods with strong assumptions on convolutional structures are certainly not ideal, as the simplified structure cannot deal with the complex information composition processing in speech signals.

Experience from image processing

It is always a good idea to learn from other domains. Let's see an example of 'image arithmetic' reported by Radford and colleagues [2]. The authors found that a picture can be well represented by a set of latent factors extracted by a deep neural net, and in the latent space, pictures can be manipulated by very simple arithmetic operations. For example, it is possible to learn factors that represent the sun glass of a face image. Although we do not know which factor represents what, it is possible to compute the glass by subtracting an image without a glass from an image with a glass. The glass represented by the subtraction can be subsequently used to wear a glass for a new face, by a simple addition. This interesting example, as shown in Fig. 2, demonstrated that deep learning can learn high-level semantics, and in that semantic space the factorization will be very simple.

Fig.2 Information factorization in image processing [2].

Coming back to our speech information factorization task, we notice that if we can learn the high-level factors of speech by deep learning as well, the factorization will be straightforward. In other words, information factorization should be conducted with high-level factors rather than raw features (as the GMM-UBM model did).

Frame-level information factorization

Triggered by the idea of factorization on high-level factors, we can proceed with information factorization for speech. The simplest and ideal factorization is at the frame level, i.e., information can be disentangled within a single speech frame.

A key question here is: if all the interesting factors are short-time? We have known that the linguistic content is short-time: in speech recognition, phones can be identified with a single frame (0.02s) with certain accuracy. However, the property of speaker traits was under long-term controversy. Historically, features that were used for speaker recognition range from short-time LPC to long-term pitch variation, as well as discourse statistics such as word n-grams. A key empirical result we reported in 2017 is that the speaker can be identified by a speech segment as short as 0.3 seconds [3,4,5], as shown in Fig. 3. This demonstrated that speaker straits are largely short-time. This discovery paves the way to frame-based factorization.

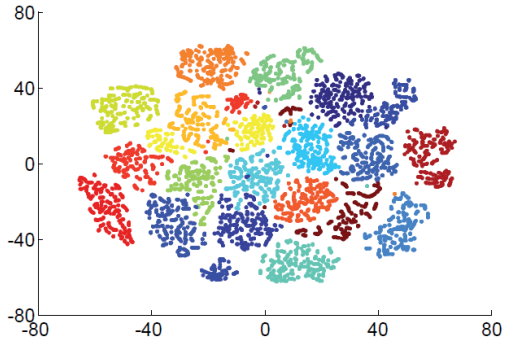


Fig.3 The frame-level speaker features extracted within 0.3 seconds, drawn by t-SNE. Each point is a sample, and each color is a speaker [3].

Following the experience learned from image processing, we need firstly extract high-level factors on which the linear factorization can be applied. This can be done by unsupervised learning. However it is not the best way, as our goal is to obtain task-oriented factors that can boost various speech processing tasks, such as ASR and/or SRE. We therefore designed a supervised factorization approach that uses task-related supervision to assist the factorization [6].

Fig.4 shows this architecture, which we call *cascaded deep factorization* (CDF). In this architecture, the information factor is represented by the output of the last hidden layer of a deep neural net that is designed for a particular task. The factor extracted from a particular task is then used as a conditional input of the network for the next task. This essentially forms a way that decomposes the information convolved in the speech signal sequentially. Compared to

Fig.1, it can be seen that CDF resembles the Fujisaki model, and can be regarded as its reverse process. Interestingly, we found that the factors extracted by the CDF approach can be used to recover the original speech with a rather high accuracy. This means that the factorization, with the help of deep learning, is not only possible, but also complete.

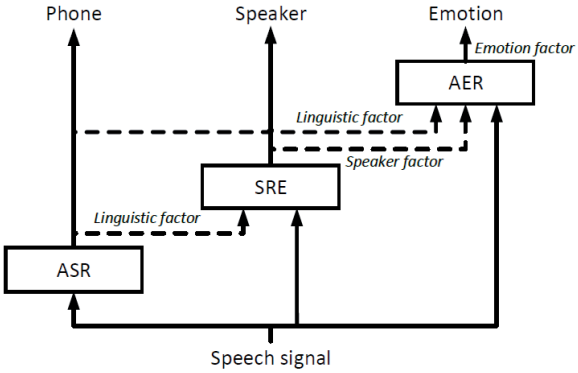


Fig.4 The cascade deep factorization (CDF) architecture. The linguistic factor, speaker factor and emotion factor are extracted sequentially, implementing the deconvolution processing for the Fujisaki model [6].

Sentence-level information factorization

We have demonstrated the possibility to factorize speech for short-time frames, but more accurate factorization must take into account the dynamic property of speech signals, i.e., the temporal dependency among frames. A possible way is to establish dynamic models for each task, and then combine them into a single process, where the outputs of all the tasks support each other. We call this architecture *collaborative learning*, as shown in Fig.5, where the output of each task is read off by the other task as a conditional variable for training and inferring at the next frame [7].

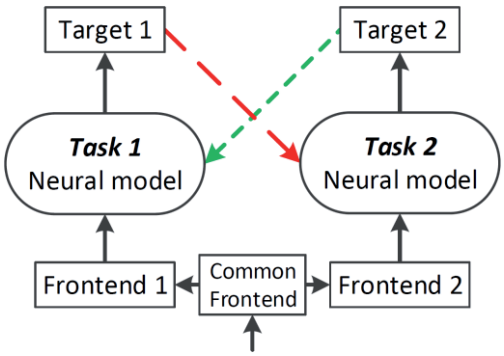


Fig.5 The collaborative learning architecture, where the outputs of the two tasks support each other for training/inference at the next frame [7].

Challenges for information factorization

The research on speech information factorization is just started, and lots of challenges remain unsolved. We just mention some most important issues.

- How to use unsupervised learning. There are lots of unlabeled speech data that allows machines learning the underlying factors of speech signals in an unsupervised way, e.g., by RBM or AE. A key shortcoming of the unsupervised factorization, as mentioned already, is that the learned factors do not clearly correspond to specific tasks, hence uneasy to explain and apply. Nevertheless, this factorization is undoubtedly useful. For example, supervised factorization can be conducted based on the factors learned by unsupervised learning, using a much simpler factorization model.

- How to deal with complicated dynamic dependency among factors. Although linguistic content and speaker traits have been demonstrated to be short-time, other factors are not. For example, emotion is widely regarded as a long-term property. We had found that the performance of emotion recognition was low even with the CDF approach, for which a reasonable hypothesis is that emotion is partly represented by long-term measurements, e.g., pitch. The collaborative training provides a way to incorporate temporal dependency between tasks, but it seems more research is required to deal with the complexity in temporal dynamics, a key nature of speech signals.

- How to extract independent factors. Although the CDF approach extract factors that are task-related, we have found that the speaker factor still involves some linguistic information, and vice versa. This information mixing means that the factorization is not perfect, and some exclusive-learning methods should be designed, e.g., adversarial training.

- How to deal with inclusive factors. Usually we assume the factors to extract are independent, but this is not the case in reality and some factors are dependent in nature. For example, the linguistic content and the language are two factors that are dependent. This dependence can be leveraged to improve performance on exacting either factor [8], but how to involve the dependency in information factorization is still an open question.

- Solid theoretical framework. Till now all the results we obtained are tentative and pragmatic. It is clear we need deep learning to discover high-level factors, and it might be also important to involve Bayesian framework to deal with the belief we have on the factors to extract (e.g., the dependency between linguistic content and language). How to combine these technical pieces actually has been beyond the scope of factorization, and is the key concern of the present machine learning research.

Long-term Impact

Speech information factorization may change our view for speech processing. Conventionally, speech processing is mostly signal processing, that designs various tools to discover and model the useful patterns for the target task. Deep learning essentially frees us from the low-level raw signals, and moves our attention to high-level information. This means that we may construct an information space and work more effectively over there, as shown in Fig.6.

The speech processing tasks, correspondingly, move from signal processing to information processing.

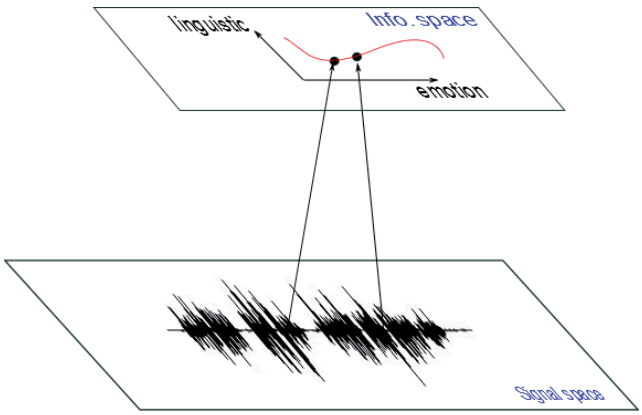


Fig.6 Mapping from signal space to information space.

In this information space, we can do lots of things than before. Firstly, the information space is a semantic space. Working in the information space is like working in the brain, rather than in the ear. The interference of noise and channel has been removed automatically, and rich information involved in speech signals can be extracted simultaneously. We can use these 'stereo information' to boost performance of individual tasks, even integrating information from other peripheral sensors, e.g., vision and touch. Most importantly, the information space is particularly good for representing insignificant factors. Taking emotion as an example: this factor is very subtle compared to linguistic content and speaker traits, which means that the variation in emotion is mostly masked by the change of other factors in the signal space. By moving to the information space, the most-varied factors become much more stable and so the emotion change can be readily inferred by the residual variation. This is the principle idea of CDF that conditions the inference for insignificant factors on prior-inferred significant factors.

Another key advantage of working in the information space is that we will have the capacity of manipulating speech signals 'freely'. As in the image example in Fig.2, the speech information factors are semantically meaningful, e.g., 'what is speaking', 'who is speaking' and 'in which emotion'. Since these factors have been identified, when a single factor is changed, the signal can be manipulated 'semantically'. This will bring us lots of freedom to handle speech, and lend itself to a broad range of applications. For example, this can be used to design low bandwidth speech codes by transferring factors that are important for the task only, e.g., linguistic content. Another application is for flexible speech synthesis. We can design the synthesis system with speech recording of a single speaker, or of a set of speakers, and change the voice by replacing the speaker factor to the factor of a targeting speaker. The same principle can be applied to speech conversion.

The problem we face today is how to jump from the signal space to the information space. Deep learning is a basic tool, and CDF and collaborative learning are just two tentative architectures. There must be many other ways that are much cleverer. Those others may or may not use deep learning, via either supervised or unsupervised approaches.

As a summary, we envision that many speech processing tasks today will probably be solved in twenty years, by mapping from the signal space to the information space, and applying information factorization therein. The gifted architecture of the Fujisaki model, which was hard to be materialized in the signal regime, will become a fundamental guidance for our work in the information space.

References

[1] Fujisaki H. Prosody, models, and spontaneous speech[M]//Computing prosody. Springer, New York, NY, 1997: 27-42.

[2] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.

[3] Li L, Chen Y, Shi Y, Tang Z, Wang D. Deep speaker feature learning for text-independent speaker verification. In INTERSPEECH 2017.

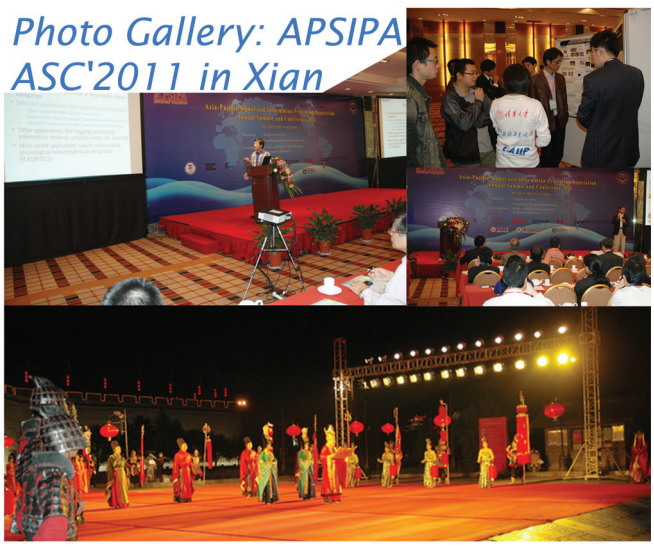
[4] Zhang M, Chen Y, Li L, Wang D. Speaker recognition with cough, laugh and "Wei"[J]. arXiv preprint arXiv:1706.07860, 2017.

[5] Zhang M, Kang X, Wang Y, Li L, Tang Z, Dai H, Wang D. Human and Machine Speaker Recognition Based on Short Trivial Events[J]. In ICASSP 2018.

[6] Li L, Wang D, Chen Y, Shi Y, Tang Z. Deep factorization for speech signal. In ICASSP 2018.

[7] Tang Z, Li L, Wang D, Vipperla R. Collaborative joint training with multitask recurrent model for speech and speaker recognition[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2017, 25(3): 493-504.

[8] Tang Z, Wang D, Chen Y, Li L, Abel A. Phonetic temporal neural model for language identification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(1): 134-144.



Do Androids Dream of Henri Poincaré with Hierarchical Optimization ?

Optimization using subliminal mechanism ?

Mathematical optimization has been a major driving force of modern AI technologies and data sciences (see, e.g., [1,2,3]). A question arises: **what is a groundbreaking optimization model that is expected to bring about dramatical evolution in next-generation AI ?** A valuable hint for this visionary question could be found only if we try to reveal the ingenious ways of thinking of exceptionally gifted humans, e.g., great mathematicians and grand masters of board games. In my student days long ago, I found, in *Mathematical Discovery* by Henri Poincaré (see, e.g., Chapter III of [4, pp.387-400]), the following impressive words: (i) *Everything happens as if the discoverer were a secondary examiner who had only to interrogate candidates declared eligible after passing a preliminary test* [4, p.391], (ii) *Of the very large number of combinations which the subliminal ego blindly forms, almost all are without interest and without utility. But, for that very reason, they are without action on the aesthetic sensibility; the conscious will never know them* [4, p.397]. Poincaré's words seem for me now to suggest that breakthrough ideas represented by outstanding mathematical discovery can be achieved through some mysterious process of double stage search where the first stage search is performed with the aid of a certain aesthetic sensibility in unconscious field. Similar words are also found in Yoshiharu Habu's explanation [5] on his aesthetic sensibility that unconsciously helps him select a breakthrough move in a crucial phase of shogi game.

Their explanations tempt me to formulate a hypothesis that their brains are exploiting simultaneously two different criteria, say Φ for the preliminary tests in their unconscious fields and ψ for the secondary tests in their conscious ones and to model their ingenious search as a certain computational process for the **hierarchical optimization**:

$$\left.\begin{array}{l}\text{minimize } \Psi(\mathbf{x}^\star) \\ \text{subject to } \mathbf{x}^\star \in \mathcal{S}_\star := \underset{\mathbf{x}\in\mathbb{R}^N}{\operatorname{argmin}}\ \Phi(\mathbf{x})\end{array}\right\} \quad (1)$$

rather than the traditional optimization model just for minimization of Φ, where the function Ψ is newly introduced for the second stage optimization. My amateur hypothesis does not contradict Sigmund Freud's psychoanalysis [6] saying that information stored in unconscious field of our brain has various level of difficulty for transforming it into available form in conscious field. Such a transform seems to correspond to the selection process explained by Poincaré and Habu in terms of aesthetic sensibility (Note: A similar consideration on the role of sub-conscious representation for creativity is found [7] but not in the context of optimization models). My naive imagination toward **optimization using subliminal mechanism** has also been motivated by studies on the neural associative memory [8] and the resting state fMRI [9].

Professor Isao Yamada

PhD, FIEEE

Technical co-chairs of APSIPA-ASC 2018

Professor

Department of Information and Communications Engineering,

Tokyo institute of Technology

Isao Yamada is a professor with the Department of Information and Communications Engineering, and the Director of the Global Scientific Information and Computing Center, Tokyo Institute of Technology. His current research interests are in mathematical signal processing, machine learning, nonlinear inverse problems, and optimization theory. He has been a Fellow of IEEE and IEICE since 2015. He received the MEXT Minister Award (Research Category) in 2016, the IEEE Signal Processing Magazine Best Paper Award in 2015, the IEICE Achievement Award in 2009, and the Docomo Mobile Science Award (Fundamental Science Division) in 2005.

What can we do for hierarchical optimization ?

The hierarchical optimization in (1) seems to be our ideal target but in reality the computation of its solution must be very challenging, as suggested by Tikhonov approximation theorem [10], even if Φ and Ψ are convex functions. To keep the currently achievable applicability by the state-of-the-art non-hierarchical convex optimization algorithms, we model the first stage cost function in (1) as

$$\Phi(\mathbf{x}) := f(\mathbf{x}) + \sum_{i=1}^{m} g_i(A_i\mathbf{x}), \quad (2)$$

where $f: \mathbb{R}^N \to (-\infty, \infty]$ and $g_i: \mathbb{R}^{N_i} \to (-\infty, \infty]$ $(i = 1,2,\dots,m)$ are convex but not necessarily differentiable everywhere, and $A_i \in \mathbb{R}^{N_i \times N}$ $(i = 1,2,\dots,m)$. Fortunately, a unified perspective from the view point of convex analysis and monotone operator theory (see, e.g., [11,12]) tells us that many convex optimization scenarios in data sciences, machine learning, and signal processing, appear as instances of the model (2) and that the so-called **proximity operators** of f and g_i $(i = 1,2,\dots,m)$ are available [11] as building blocks [13] of a computable

nonexpansive operator $T: \mathcal{H} \to \mathcal{H}$ and a bounded linear operator $\Xi: \mathcal{H} \to \mathbb{R}^N$ satisfying

$$\mathcal{S}_\star = \underset{\mathbf{x}\in\mathbb{R}^N}{\operatorname{argmin}}\ \Phi(\mathbf{x}) = \{\Xi(\mathbf{z}) \in \mathbb{R}^N \mid \mathbf{z} \in \operatorname{Fix}(T)\},$$

where $\mathcal{H}$ is a certain real Hilbert space, not necessarily $\mathcal{H} = \mathbb{R}^N$, and $\operatorname{Fix}(T) \coloneqq \{\boldsymbol{z} \in \mathcal{H} \mid T(\boldsymbol{z}) = \boldsymbol{z}\}$ is the set of all fixed points of T. Indeed, by plugging the nonexpansive operator T and the convex function $\Theta \coloneqq \Psi \circ \Xi$ into the **hybrid steepest descent method** [13,14]:

$$\mathbf{z}_{n+1} = T(\mathbf{z}_n) - \lambda_{n+1}\nabla\Theta(T(\mathbf{z}_n)) \qquad (3)$$

with a slowly decreasing sequence $(\lambda_n)_{n\geq 1} \subset [0, \infty)$, under reasonable conditions, we can generate a sequence $\Xi(\mathbf{z}_n)$ $(n = 0,1,2, \dots)$ which converges to a solution of (1).

An application to Cortes-Vapnik problem

To demonstrate the inherent applicability of the hierarchical convex optimization to machine learning problems, I conclude this article with a short introduction on our recent application [13] to a novel hierarchical convex relaxation of the Cortes-Vapnik problem [15, Sec.3]. This application has been made for sound extension of a central idea in the classical Support Vector Machine (SVM) [1] to be applicable to general training dataset:

$$\mathcal{D} := \{(\mathbf{x}_i, \mathfrak{L}(\mathbf{x}_i)) \in \mathbb{R}^p \times \{-1, 1\} \mid i = 1, 2, \dots, M\}$$
$$= \mathcal{D}_+ \cup \mathcal{D}_-(\mathcal{D}_\pm := \{\mathbf{x}_i \in \mathbb{R}^p \mid (\mathbf{x}_i, \pm 1) \in \mathcal{D}\}),$$

where $\mathfrak{L}(\mathbf{x}_i)$ is the binary label assigned to $\mathbf{x}_i$. Since the original Cortes-Vapnik problem was introduced as an NP-hard problem with a hierarchical structure, a simple convex relaxation (the soft margin SVM[1]):

$$\underset{(\mathbf{w},b)\in\mathbb{R}^p\times\mathbb{R}}{\text{minimize}}\ \Psi(\mathbf{w}, b) + C\Phi(\mathbf{w}, b), \qquad (4)$$

where $\Psi(\mathbf{w}, b) := \frac{1}{2}\|\mathbf{w}\|^2$,

$$\underbrace{\|\mathbf{w}\|\left[\sum_{\mathbf{z}^+\in\mathcal{D}_+} \mathrm{d}\left(\mathbf{z}^+, \Pi^{\geq+1}_{(\mathbf{w},b)}\right) + \sum_{\mathbf{z}^-\in\mathcal{D}_-} \mathrm{d}\left(\mathbf{z}^-, \Pi^{\leq-1}_{(\mathbf{w},b)}\right)\right]}_{=:\Phi(\mathbf{w},b)}, \text{ and}$$

$\mathrm{d}\left(\cdot, \Pi^{\geq+1}_{(\mathbf{w},b)}\right)$ and $\mathrm{d}\left(\cdot, \Pi^{\leq-1}_{(\mathbf{w},b)}\right)$ respectively stand for distances to closed half-spaces $\Pi^{\geq+1}_{(\mathbf{w},b)} := \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{w}^\top\mathbf{x} - b \geq +1\}$ and $\Pi^{\geq-1}_{(\mathbf{w},b)} := \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{w}^\top\mathbf{x} - b \leq -1\}$,

has been used extensively with a *tuning parameter* $C > 0$. However this naive relaxation induces a natural question: **Is the solution of (4) for general training data $\mathcal{D}$ really a mathematically sound extension of the classical SVM ?** This is because, by the complete loss of the hierarchical structure, the solution of (4) cannot reproduce, in general, the classical SVM that maximizes the margin among all error-free linear classifiers for linearly separable training dataset (Note: The above question is common even for the soft-margin SVM applied to the transformed data with nonlinear kernels [1]). Therefore, we newly formulate

$$\text{minimize } \Psi(\mathbf{w}^\star, b^\star) \text{ subject to}$$
$$(\mathbf{w}^\star, b^\star) \in \mathcal{S}_\star := \underset{(\mathbf{w},b)\in\mathbb{R}^p\times\mathbb{R}}{\operatorname{argmin}}\ \Phi(\mathbf{w}, b) \qquad (5)$$

as a much more faithful convex relaxation of the original Cortes-Vapnik problem than (4). Remark that the hierarchical convex relaxation (5) is well-defined even for linearly non-separable training dataset $\mathcal{D}$ and can reproduce perfectly the classical SVM for linearly separable dataset unlike (4). Fortunately, the problem (5) falls in the class of the hierarchical convex optimization problems of type (1) and (2), and therefore is solvable efficiently by combining the ideas in the hybrid steepest descent method and the art of proximal splitting [13].

[1] The function $\Phi(\mathbf{w}, \mathrm{b})$ is often expressed with a *hinge loss function* and serves as a convex relaxation of the number of misclassified training samples. The squared margin of the linear classifier with $(\mathbf{w}, \mathrm{b})$ is given by $2^{-1}\Psi^{-1}(\mathbf{w}, \mathrm{b})$.

References

[1] V. N. Vapnik, Statistical Learning Theory, Wiley, 1998.

[2] Y. A. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, 521, pp. 436-444, 2015.

[3] S. Theodoridis, K. Slavakis, and I. Yamada "Adaptive learning in a world of projections: a unifying framework for linear and nonlinear classification and regression tasks," IEEE Signal Process. Mag., 21, pp. 97-123, 2011.

[4] H. Poincaré, Science et Méthode, 1908 (translated by F. Maitland in 1952 as *Science and Method* which is collected in The Value of Science - Essential Writings of Henri Poincaré , pp. 357-572, Random House, 2001).

[5] Y. Habu and H. Shinohara, The present and future of AI, http://www.ntt.co.jp/activity/en/innovation/habu/, 2017.

[6] S. Freud, New Introductory Lectures on Psycho-Analysis (The complete psychological works of S. Freud, J. Strachey.ed.) W. W. Norton and Company, 1990.

[7] S. Shimojo, Saburiminaru inpakuto (in Japanese), Chikuma Shobo, Inc., 2008.

[8] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," Proc. Natl. Acad. Sci. U.S.A., 79, pp.2554-2558, 1982.

[9] M. E. Raichle, "The brain's dark energy," Scientific American, pp.44-49, March 2010.

[10] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," Soviet Math. Dokl., 4, pp.1035-1038, 1963.

[11] H. H. Bauschke and P. L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Space 2nd ed., Springer, 2017.

[12] P. L. Combettes and I. Yamada, "Compositions and convex combinations of averaged nonexpansive operators," J. Math. Anal. Appl., 425, pp. 55–70, 2015.

[13] I. Yamada and M. Yamagishi, "Hierarchical Convex Optimization by the hybrid steepest descent method with proximal splitting operators - Enhancements of SVM and Lasso," In: H. H. Bauschke, D. R. Luke, R. Burachik eds., Proceedings of Splitting Algorithms, Modern Operator Theory, and Applications 2017, 74 pp, Springer (in press).

[14] I. Yamada, "The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings," In: D. Butnariu, Y. Censor and S. Reich, eds., Inherently Parallel Algorithm for Feasibility and Optimization and Their Applications, Elsevier, pp.473-504, 2001.

[15] C. Cortes and V. N. Vapnik, "Support Vector Networks," Machine Learning, 20, pp.273-297, 1995.

APSIPA-ASC:

Asia Pacific signal and Information Processing Association - Annual Summit and Conference

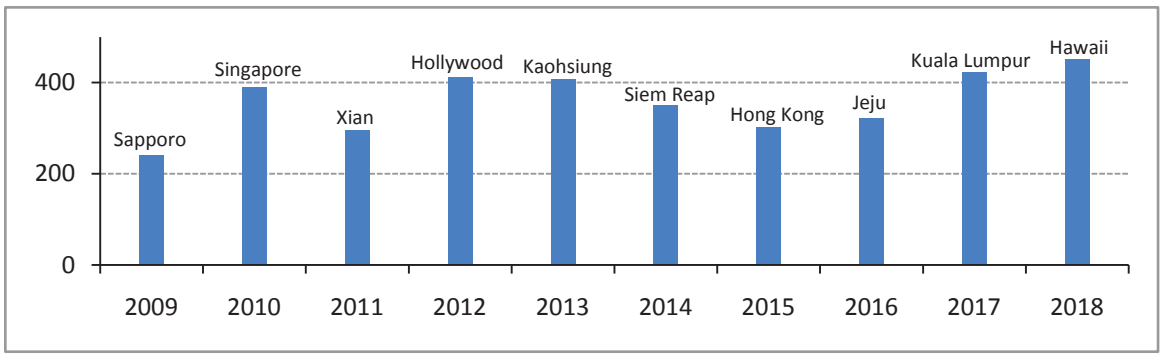
There are nine Asia Pacific Signal and Information Processing Association - Annual Summit and Conferences had been conducted successfully in various places, and the 10th one is going to be held in Hawaii, 12-15 November 2018, and there is an overwhelming support for over 450 attendees already. As shown below, we list some details of all APSIPA-ASCs, statistics on the number of attendees and the number of cities/regions involved in each year. It is seen that we have a very steady and healthy development of this summit and conference series.

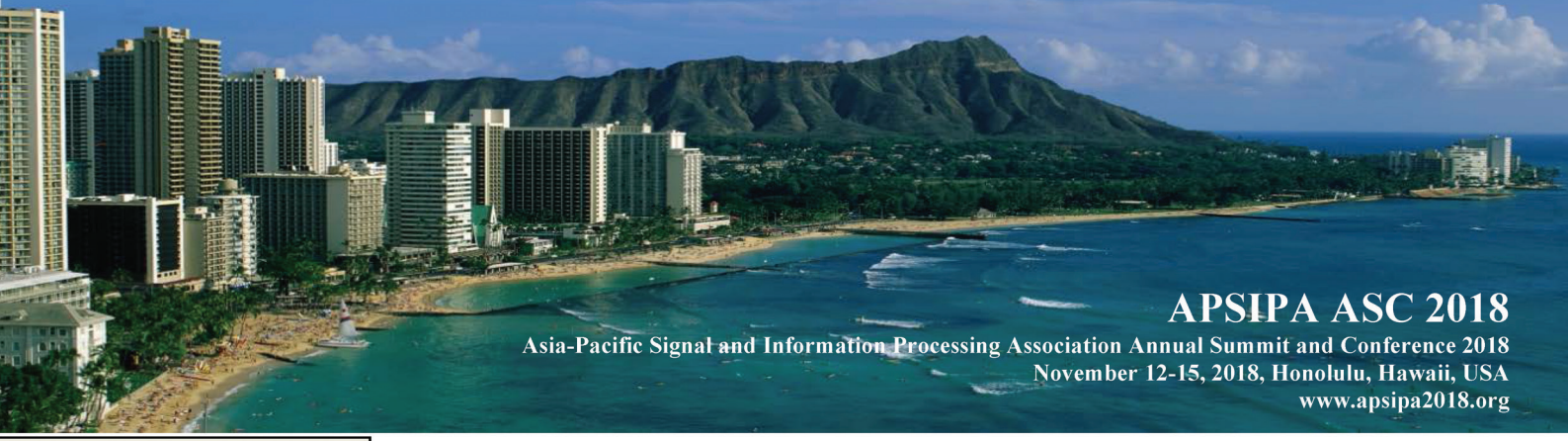
APSIPA-ASC 2018,	12-15 November 2018	Hawaii
APSIPA-ASC 2017,	12-15 December 2017	Kuala Lumpur
APSIPA-ASC 2016,	13-16 December 2016	Jeju
APSIPA-ASC 2015,	16-19 December 2015	Hong Kong
APSIPA-ASC 2014,	9-12 December 2014	Siem Reap
APSIPA-ASC 2013,	29 Oct-1 Nov 2013	Kaohsiung
APSIPA-ASC 2012,	3-6 December 2012	Hollywood
APSIPA-ASC 2011,	18-21 October 2011	Xi'an
APSIPA-ASC 2010,	14-17 December 2010	Singapore
APSIPA-ASC 2009,	4-7 October 2009	Sapporo

Attendance Record of APSIPA-ASC

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
	Sapporo	Singapore	Xian	Hollywood	Kaohsiung	Siem Reap	Hong Kong	Jeju	Kuala Lumpur	Hawaii
Attendees	240	390	294	411	408	350	302	322	422	450+
Countries/ Region (No.)	16	25	14	16	19	24	28	17	27	*

+ Initial figure only * Not available yet





APSIPA ASC 2018

Asia-Pacific Signal and Information Processing Association Annual Summit and Conference 2018
November 12-15, 2018, Honolulu, Hawaii, USA
www.apsipa2018.org

Organizing Committee

Honorary General Co-Chairs
Sadaoki Furui
K. J. Ray Liu

General Co-Chairs
Yih-Fang Huang
Anthony Kuh
Susanto Rahardja

Technical Program Co-Chairs
Hsueh-Ming Hang
Ming-Ting Sun
Z. Jane Wang
Isao Yamada
Woon-Seng Gan
Juin-In Guo
Kazunori Hayashi
Kazushi Ikeda
Sung Chan Jun
Sanghoon Lee
Shinsuke Ibi
Wen-Hsiao Peng

Finance Chair
Kenneth Lam

Plenary Co-Chairs
Helen Meng
Antonio Ortega

Forum Co-Chairs
Mingyi He

Special Session Co-Chairs
Toshihisa Tanaka
Yan Lindsay Sun
Zhi Tian
Chris Lee
Shoko Imaizumi
Supavadee Aramvith

Tutorial Co-Chairs
Homer Chen
Gene Cheung
H. Vicky Zhao

Publicity Co-Chairs
Chung-Nan Lee
Min Wu
Lei Xie

Local Arrangement Chair
June Zhang

Photo Gallery – Hawaii

The Organizers:

Thanks for their great work.

For more photos: http://www.apsipa.org/photo/apsipa_asc2018

THE 9TH APSIPA ASC 2017

DECEMBER 12-15, 2017, KUALA LUMPUR, MALAYSIA

ORGANIZING COMMITTEES

Advisory Committee Chairs
Sadaoki Furui (Toyota Technological Institute Chicago, USA)
Ray Liu (University of Maryland, USA)
CC-Jay Kuo (University of Southern California, USA)
Haizhou Li (National University of Singapore, Singapore)
Wan-Chi Siu (The Hong Kong Polytechnic University, Hong Kong)
Sanjit K Mitra (University of California, Santa Barbara, USA)

General Co- Chairs
Woon-Seng Gan (Nanyang Technological University, Singapore)
Kai-Kuang Ma (Nanyang Technological University, Singapore)
KokSheik Wong (University of Malaya, Malaysia)

TPC Co-Chairs
Gwo Giun Lee (National Cheng-Kung University, Taiwan)
Akira Hirabayashi (Ritsumeikan University, Japan)
Changchun Bao (Beijing University of Technology, China)
Toshihisa Tanaka (Tokyo University of Agriculture and Technology, Japan)
Jiwu Huang (Shenzhen University, China)
Kazunori Hayashi (Kyoto University, Japan)
Andy Khong (Nanyang Technological University, Singapore)
Huanqiang Zeng (Huaqiao University, China)
Dennis Wong (Heriot-Watt University Malaysia)

Finance Chair
Chee Seng Chan (University of Malaya, Malaysia)

Secretary / Web co-chairs
Vishnu Monn Baskaran (Multimedia University, Malaysia)
Yiqi Tew (Tunku Abdul Rahman University College, Malaysia)

Special-session Co-Chairs
Hitoshi Kiya (Tokyo Metropolitan University, Japan)
Yoshinobu Kajikawa (Kansai University, Japan)
Ma Bin (Institute of Infocomm Research, Singapore)
Lap Pui Chau (Nanyang Technological University, Singapore)
David Chuah (University of Malaya, Malaysia)

Forum Co-Chairs
Kenneth Lam (Hong Kong Polytechnic University, Hong Kong)
Waleed Abdullah (University of Auckland, New Zealand)
Shan Liu (MediaTek, USA)
John See (Multimedia University, Malaysia)

Tutorial Co-Chairs
Kong-Aik Lee (Institute of Infocomm Research, Singapore)
Yo-Sung Ho (Gwangju Institute of Science and Technology, S.Korea)
Vivian Yong (Petronas University of Technology, Malaysia)

Panel Session Co-Chairs
Thomas Zheng Fang (Tsinghua University, China)
Yoong Choon Chang (University of Tunku Abdul Rahman, Malaysia)

Publicity Co-Chairs
Sam Shi Chuang (Nanyang Technological University, Singapore)
Dong Wang (Tsinghua University, China)
Kosin Chamnongthai (King Mongkut's University of Technology Thonburi, Thailand)
Mohammad Faizal Ahmad Fauzi (Multimedia University, Malaysia)

Registration Co-Chairs
Chee Kau Lim (University of Malaya, Malaysia)
Sook Chin Yip (Multimedia University, Malaysia)

Publication Chair
Wai Lam Hoo (Tunku Abdul Rahman University College, Malaysia)

Local Arrangement Co-Chairs
Simying Ong (Taylor's University, Malaysia)
Chern Hong Lim (Tunku Abdul Rahman University College, Malaysia)

Sponsorship/ Exhibition Co- Chairs
Kok Soon Tey (University of Malaya, Malaysia)
Erma Rahayu (University of Malaya, Malaysia)

European / US liaison
Shujun Li (University of Surrey, UK)

Photo Gallery – Kuala Lumpur

For more photos: http://www.apsipa.org/photo/apsipa2017

The 8th APSIPA ASC 2016

December 13-16, 2016, Jeju, South Korea

Organizing Committee

Honorary Co-Chairs:
Sadaoki Furui (Tokyo Institute of Technology, Japan)
K.J. Ray Liu (University of Maryland, USA)
Wan-Chi Siu (Hong Kong Polytechnic University, Hong Kong)
Sang Uk Lee (Seoul National University, Korea)

General Co-Chairs:
Yo-Sung Ho (Gwangju Institute of Science and Technology, Korea)
C.-C. Jay Kuo (University of Southern California, USA)
Haizhou Li (Institute for Infocomm Research, A*STAR, Singapore)

Technical Program Co-Chairs:
Nam Ik Cho (Seoul National University, Korea)
Thomas Fang Zheng (Tsinghua University, China)
Hitoshi Kiya (Tokyo Metropolitan University, Japan)
Homer Chen (National Taiwan University, Taiwan)
Anthony Kuh (University of Hawaii at Manoa, USA)
Jiwu Huang (Shenzhen University, China)

Forum Session Co-Chairs:
Changick Kim (Korea Advanced Institute of Science and Technology, Korea)
Byeungwoo Jeon (Sungkyunkwan University, Korea)
Kwanghoon Sohn (Yonsei University, Korea)
Woon Seng Gan (Nanyang Technological University, Singapore)
Waleed Abdulla (University of Auckland, New Zealand)

Panel Session Co-Chairs:
Nam Soo Kim (Seoul National University, Korea)
Yong Man Ro (Korea Advanced Institute of Science and Technology, Korea)
Wonha Kim (Kyung Hee University, Korea)
Ming-Ting Sun (University of Washington, USA)
Kiyoharu Aizawa (University of Tokyo, Japan)

Special Session Co-Chairs:
Yung-Lyul Lee (Sejong University, Korea)
Jeongtae Kim (Ewha Womans University, Korea)
Hale Kim (Inha University, Korea)
Akihiko Sugiyama (NEC Corporation, Japan)
Mark Liao (Institute of Information Science, Academia Sinica, Taiwan)

Tutorial Co-Chairs:
Sang-Hoon Lee (Yonsei University, Korea)
Min-Cheol Hong (Soongsil University, Korea)
Kate Shim (Yonsei University, Korea)
Hsueh-Ming Hang (National Chiao-Tung University, Taiwan)

Publication Co-Chairs:
Chang-Su Kim (Korea University. Korea)
Dong-Gyu Sim (Kwangwoon University, Korea)
Chee Seng Chan (University of Malaya, Malaysia)
Tatsuya Kawahara (Kyoto University, Japan)

Publicity Co-Chairs:
Kyoungmu Lee (Seoul National University, Korea)
Joonki Paik (Chung-Ang University, Korea)
Ki Ryong Kwon (Pukyong National University, Korea)
Dongbo Min (Chungnam National University, Korea)
Yoshinobu Kajikawa (Kansai University, Japan)
Susanto Rahardja (Northwestern Polytechnical University, China)
Mrityunjoy Chakraborty (Indian Institute of Technology, India)
Kosin Chamnongthai (King Mongkut's University of Technology Thonburi, Thailand)
Thanh-Sach Le (Ho Chi Minh City University of Technology, Vietnam)

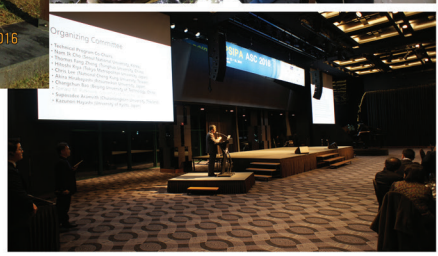
Web Co-Chairs:
Young-Woo Suh (Korean Broadcasting System, Korea)
Byung Tae Oh (Korea Aerospace University, Korea)
Jewon Kang (Ewha Womans University, Korea)

Finance Co-Chairs:
Sang-Keun Lee (Chung-Ang University, Korea)
Jong-Il Park (Hanyang University, Korea)
Kenneth Lam (Hong Kong Polytechnic University, Hong Kong)

Registration Co-Chairs:
Hong Kook Kim (Gwangju Institute of Science and Technology, Korea)
Jong Won Shin (Gwangju Institute of Science and Technology, Korea)

Local Arrangement Co-Chairs:
Jae Yun Lim (Jeju National University, Korea)
Kang-Sun Choi (KoreaTech, Korea)

Photo Gallery – Jeju



For more photos: http://www.apsipa.org/photo/apsipa2016

7th APSIPA ASC 2015

ASIA-PACIFIC SIGNAL AND INFORMATION PROCESSING ASSOCIATION
ANNUAL SUMMIT AND CONFERENCE 2015

DECEMBER 16-19, 2015
HONG KONG

apsipa2015.org

Honorary General Chair
Wan-Chi Siu, Hong Kong Polytechnic University

General Co-Chairs
Kenneth Lam, Hong Kong Polytechnic University
Helen Meng, Chinese University of Hong Kong
Oscar Au

Technical Program Co-Chairs
Changchun Bao, Beijing Univ. of Technology
Akira Hirabayashi, Ritsumeikan University
Jiwu Huang, Shenzhen University
Gwo Giun Lee, National Cheng Kung University
Daniel Lun, Hong Kong Polytechnic University
Tomoaki Ohtsuki, Keio University
Tomasz M. Rutkowski, University of Tsukuba
Sumei Sun, I2R, A*STAR

Finance Chair
Chris Chan, Hong Kong Polytechnic University

Secretary
Bonnie Law, Hong Kong Polytechnic University

Forum Co-Chairs
Wai-Kuen Cham, Chinese Univ. of Hong Kong
Homer Chen, National Taiwan University
King N. Ngan, Chinese University of Hong Kong
Ming-Ting Sun, University of Washington

Panel Session Co-Chairs
Shing-Chow Chan, University of Hong Kong
Dominic K.C. Ho, University of Missouri
Yo-Sung Ho, Gwangju Inst. of Science & Tech.
Yoshikazu Miyanaga, Hokkaido University

Special Session Co-Chairs
Mrityunjoy Chakraborty, India Inst. of Technology
Yui-Lam Chan, Hong Kong Polytechnic University
Lap-Pui Chau, Nanyang Technological University
Haojiang Deng, Chinese Academy of Sciences
Hsueh-Ming Hang, National Chiao-Tung University
Hitoshi Kiya, Tokyo Metropolitan University

Tutorial Co-Chairs
Waleed Abdulla, University of Auckland
Woon-Seng Gan, Nanyang Technological Univ.
Wing-Kuen Ling, Guangdong Univ. of Tech.
Lee Tan, Chinese University of Hong Kong

Registration Co-Chairs
Man-Wai Mak, Hong Kong Polytechnic University
Lai-Man Po, City University of Hong Kong

Publication Chair
Zheru Chi, Hong Kong Polytechnic University

Publicity Co-Chairs
Kiyoharu Aizawa, University of Tokyo
Yui-Lam Chan, Hong Kong Polytechnic Univ.
Hing Cheung So, City University of Hong Kong
Mark Liao, IIS, Academia Sinica
Thomas Fang Zheng, Tsinghua University

Local Arrangement Co-Chairs
Edward Cheung, Hong Kong Polytechnic Univ.
Frank Leung, Hong Kong Polytechnic University

Advisory Committee

Chairs:
Sadaoki Furui, Toyota Technological Institute at Chicago
Wen Gao, Peking University
C.-C. Jay Kuo, University of Southern California
Haizhou Li, Inst. for Infocomm Research, A*STAR
Ray Liu, University of Maryland

Members:
Thierry Blu, Chinese University of Hong Kong
Shih-Fu Chang, Columbia University
Liang-Gee Chen, National Taiwan University
Li Deng, Microsoft Research
Takeshi Ikenaga, Waseda University
Kebin Jia, Beijing Univ. of Technology
Anthony Kuh, University of Hawaii
Antonio Ortega, University of Southern California
Soo-Chang Pei, National Taiwan University
Susanto Rahardja, National Univ. of Singapore
Yodchanan Wongsawat, Mahidol University
Chung-Hsien Wu, National Cheng Kung Univ.

Photo Gallery – Hong Kong

For more photos: http://www.apsipa.org/photo/apsipa2015

978-988-14768-4-5©2018

Photo Gallery – Siem Reap

Organizer

Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI) Association of Thailand

Academic Sponsor

Asia-Pacific Signal and Information Processing Association (APSIPA)

Organizing Committees

Honorary Co-Chairs

Sadaoki Furui, Tokyo Institute of Technology, Japan
K. J. Ray Liu, University of Maryland, USA
Prayoot Akkaraekthalin, KMUTNB, Thailand

General Co-Chairs

Kosin Chamnongthai, KMUTT, Thailand
C.-C. Jay Kuo, University of Southern California, USA
Hitoshi Kiya, Tokyo Metropolitan University, Japan

Technical Program Co-Chairs

Pornchai Supnithi, KMITL, Thailand
Takao Onoye, Osaka University, Japan
Hsueh-Ming Hang, National Chiao Tung University, Taiwan
Anthony Kuh, University of Hawaii at Manoa, USA
Takeshi Ikenaga, Waseda University, Japan
Chung-Hsien Wu, National Cheng Kung University, Taiwan
Yodchanan Wongsawat, Mahidol University, Thailand
Oscar Au, HKUST, Hong Kong
Tomoaki Ohtsuki, Keio University, Japan

Forum Co-Chairs

Antonio Ortega, University of Southern California, USA
Waleed Abdulla, The University of Auckland, New Zealand
Homer Chen, National Taiwan University, Taiwan
Vorapoj Patanavijit, Assumption University, Thailand

Panel Session Co-Chairs

Mark Liao, IIS, Academia Sinica, Taiwan
Li Deng, Microsoft Research, USA
Jiwu Huang, Sun Yat-Sen University, China
Kazuya Takeda, Nagoya University, Japan

Special Session Co-Chairs

Minoru Okada, Nara Institute of Science and Technology, Japan
Gan Woon Seng, Nanyang Technological University, Singapore
Mrityunjoy Chakraborty, IIT Kharagpur, India
Supaporn Kiattisin, Mahidol University, Thailand
Supavadee Aramvith, Chulalongkorn University, Thailand

Tutorial Session Co-Chairs

Kenneth Lam, The Hong Kong Polytechnic University, Hong Kong
Toshihisa Tanaka, TUAT, Japan
Tatsuya Kawahara, Kyoto University, Japan
Sumei Sun, I²R, A*STAR, Singapore

Publicity Co-Chairs

Yoshio Itoh, Tottori University, Japan
Yo-Sung Ho, Gwangju Institute of Science and Technology, Korea
Thomas Fang Zheng, Tsinghua University, China
Chung-Nan Lee, National Sun Yat-sen University, Taiwan
Chalie Charoenlarpnopparut, Thammasat University, Thailand

Publication Co-Chairs

Yoshinobu Kajikawa, Kansai University, Japan
Nipon Theera-umpon, Chiangmai University, Thailand
Piya Warabuntaweesuk, Bangkok University, Thailand
Wisarn Patchoo, Bangkok University, Thailand

Financial Chairs

Rujipan Sampanna, Bangkok University, Thailand
Pairin Kaewkuay, ECTI, Thailand

Local Arrangement Chairs

Suttichai Premrudeeprechacharn, Chiangmai University, Thailand
Sermsak Uatrongjit, Chiangmai University, Thailand
Sathaporn Promwong, KMITL, Thailand

General Secretaries

Werapon Chiracharit, KMUTT, Thailand
Boonserm Kaewkamnerdpong, KMUTT, Thailand

For more photos: http://www.apsipa.org/photo/apsipa2014

Photo Gallery – Kaohsiung

Advisory Committee

***Chairs**

Sadaoki Furui, Tokyo Institute of Technology, Japan
Wan-Chi Siu, The Hong Kong Polytechnic University, Hong Kong
Biing-Hwang (Fred) Juang, Georgia Institute of Technology, USA
C.-C. Jay Kuo, University of Southern California, USA
Li Deng, Microsoft Research, USA

***Members**

Yoshikazu Miyanaga, Hokkaido University, Japan
H. Y. Mark Liao, Academia Sinica, Taiwan
H. M. Hang, National Chiao-Tung University, Taiwan
Antonio Ortega, University of Southern California, USA
Soo-Chang Pei, National Taiwan University, Taiwan
Liang-Gee Chen, National Taiwan University, Taiwan
Thomas Fang Zheng, Tsinghua University, Beijing , China
Shih-Fu Chang, Columbia University, USA
Tat-Seng Chua, National University of Singapore, Singapore

Organizing committee

***Honorary Co-Chairs**

Hung-Dun Yang, National Sun Yat-sen University, Taiwan
Samuel K.C. Chang, Chung Yuan Christian University, Taiwan
K. J. Ray Liu, University of Maryland, USA

***General Co-Chairs**

Chung-Nan Lee, National Sun Yat-sen University, Taiwan
Kiyoharu Aizawa, University of Tokyo, Japan
Chung-Hsien Wu, National Cheng Kung University, Taiwan

***Technical Program Co-Chairs**

Wei-Ying Ma, Microsoft Asia, China
Tatsuya Kawahara, Kyoto University, Japan
Hsin-Min Wang ,Academia Sinica, Taiwan
Mrityunjoy Chakraborty, Indian Institute of Technology, India
Mohan Kankanhalli, National University of Singapore, Singapore
Yo-Sung Ho, Gwangju Institute of Science and Technology, Korea
Jin-Jang Leou, National Chung-Cheng University, Taiwan
Takeshi Ikenaga, Waseda University, Japan
Hitoshi Kiya, Tokyo Metropolitan University, Japan
Yodchanan Wongsawat, Mahidol University, Thailand
Chia-Hung Yeh, National Sun Yat-sen University, Taiwan
Tomoaki Ohtuki, Keio University, Japan

***Forum Co-Chairs**

Yen-Kuang Chen, Intel, USA
Chia-Wen Lin, National Tsing-Hua University, Taiwan
Y. Tsao, Academia Sinica, Taiwan

***Panel Session Co-Chairs**

Kenneth Lam, The Hong Kong Polytechnic University, Hong Kong
Jen-Tzung Chien, Naitonal Chiao-Tung University, Taiwan

***Special Session Co-Chairs**

Wen-Nung Lie , National Chung-Cheng University, Taiwan
Jing-Liang Peng, Shandong University, China

***Tutorial Session Co-Chairs**

Woon Seng Gan, Nanyang Technological University, Singapore
Shu-Min. Li, National Sun Yat-sen University, Taiwan

***Publicity Chair**

Yuan-Hsiang Chang, Chung-Yuan Christian University, Taiwan

***Publication Co-Chairs**

Chiou-Ting Hsu, National Tsing-Hua University, Taiwan
Yi-Hsuan Yang, Academia Sinica, Taiwan

***Local Arrangement Co-Chairs**

Chia-Pin Chen, National Sun Yat-sen University, Taiwan
Chih-Wen Su, Chung-Yuan Christian University, Taiwan

***Sponsorship Co-Chairs**

Jia-Sheng Heh,, Chung-Yuan Christian University, Taiwan
Amy Lee, National Taiwan University, Taiwan
Wen-Huang Cheng, Academia Sinica, Taiwan

***Financial Chair**

Hsiao-Rong Tyan, Chung-Yuan Christian University

***Registration Co-Chairs**

Cheng-Wen Ko, National Sun Yat-sen University, Taiwan
Li-Wei Kang, National Yunlin Univ. of Sci. & Tech., Taiwan

For more photos: http://www.apsipa.org/photo/apsipa2013

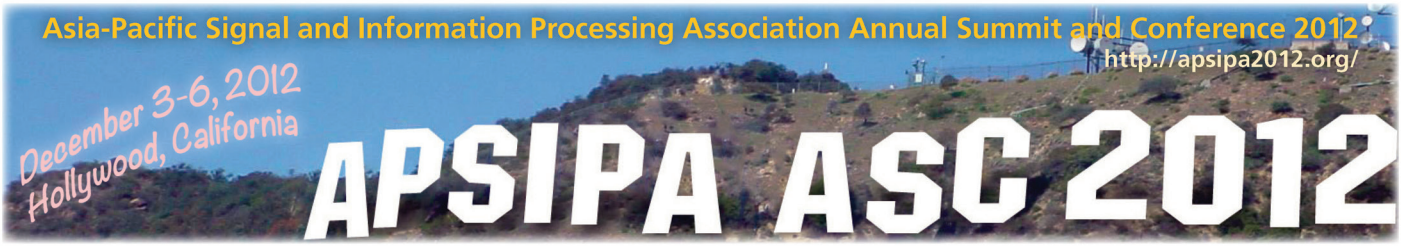


Photo Gallery – Hollywood

Organizing Committees

General Co-Chairs
C.-C. Jay Kuo, University of Southern California, USA
Shrikanth Narayanan, University of Southern California, USA
Antonio Ortega, University of Southern California, USA

Technical Program Co-Chairs
Richard Leahy, University of Southern California, USA (BioSiPS)
Toshihisa Tanaka, Tokyo University of Agriculture and Technology, Japan (BioSiPS)
Takao Nishitani, Tokyo Metropolitan University, Japan (SPS)
Tong Zhang, Rensselaer Polytechnic Institute, USA (SPS)
Yo-Sung Ho, Gwangju Institute of Science and Technology, Korea (IVM)
B. S. Manjunath, University of California, Santa Barbara, USA (IVM)
Tatsuya Kawahara, Kyoto University, Japan (SLA)
Bhaskar Rao, University of California, San Diego, USA (SLA)
Mrityunjoy Chakraborty, Indian Institute of Technology, India (SIPTM)
Anthony Kuh, University of Hawaii, USA (SIPTM)
Kwang-Cheng Chen, National Taiwan University, Taiwan (WCN)
Andreas Molisch, University of Southern California, USA (WCN)

Forum Co-Chairs
John Apostolopoulos, HP, USA
Ton Kalker, Huawei, USA
Chung-Sheng Li, IBM Research, USA

Panel Session Co-Chairs
Waleed Abdulla, The University of Auckland, New Zealand
Hitoshi Kiya, Tokyo Metropolitan University, Japan
Oscar Au, Hong Kong University of Science and Technology, Hong Kong

Special Session Co-Chairs
Krishna Nayak, University of Southern California, USA (BioSiPS)
Ioannis Katsavounidis, University of Thessaly, Greece (SPS)
Chih-Hung Kuo, National Cheng-Kung University, Taiwan (SPS)
Siwei Ma, Peking University, China (SPS)
Kyoung Mu Lee, Seoul National University, Korea (IVM)
Jingliang Peng, Shandong University, China (IVM)
Chia-Hung Yeh, National Sun-Yat-Sen University, Taiwan (IVM)
Haizhou Li, Institute for Infocomm Research, A*STAR, Singapore (SLA)
Helen Meng, Chinese University of Hong Kong, Hong Kong (SLA)
Y.-W. Peter Hong, National Tsing-Hua University, Taiwan (SIPTM)
Sau-Hsuan Wu, National Chiao-Tung University, Taiwan (SIPTM)
Jongwon Kim, Gwangju Institute of Science and Technology, Korea (WCN)
Wen-Kuang Kuo, National Cheng-Kung University, Taiwan (WCN)

Tutorial Session Co-Chairs
Jiwu Huang, Sun Yat-Sen University, China
Chang-Su Kim, Korea University, Korea
Qi Tian, University of Dallas at San Antonio, USA

Publicity Co-Chairs
Huan Chen, National Chung-Cheng University, Taiwan
Woon Seng Gan, Nanyang Technological University, Singapore
Ming-Sui Lee, National Taiwan University, Taiwan
Xiaokang Yang, Shanghai Jiao-Tung University, China

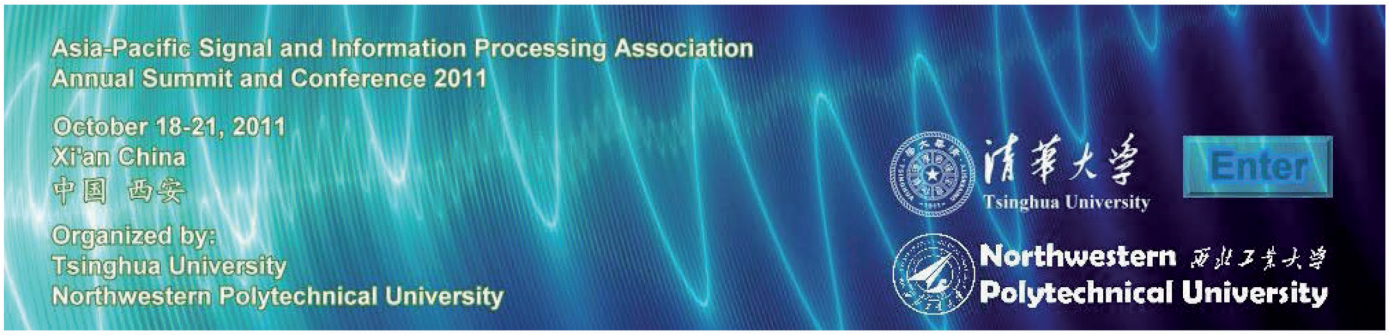
Publication Co-Chairs
Weisi Lin, Nanyang Technological University, Singapore
Jiaying Liu, Peking University, China
Hwangjun Song, Pohang Institute of Technology, Korea
Po-Chyi Su, National Central University, Taiwan

Local Arrangement Chair
Panayiotis G. Georgiou, University of Southern California, USA

Financial Chair
Gloria Halfacre, University of Southern California, USA

Registration Chair
Talyia Veal, University of Southern California, USA

For more photos: http://www.apsipa.org/photo/apsipa2012



Honorary Co-Chairs
Guangnan Ni, Chinese Information Processing Society of China
Biing-Hwang (Fred) Juang, Georigia Institute of Technology, USA
Xiaozhu Chen, Northwestern Polytechnical University, Xi'an

General Co-Chairs
Thomas Fang Zheng, Tsinghua University, Beijing
C. C. Jay Kuo, University of South California, USA
Yoshikazu Miyanaga, Hokkaido University, Japan

Technical Program Co-Chairs
Soo-Chang Pei, National Taiwan University, Taipei (BioSPS)
Liang-Gee Chen, National Taiwan University, Taipei (BioSPS)
Yo-Sung Ho, Gwangju Institute of Sci. & Technology, Korea (SPS)
Zhihua Wang, Tsinghua University, Beijing (SPS)
Akihiko (Ken) Sugiyama, NEC Corporation, Japan (IVM)
Yanning Zhang, Northwestern Polytech. University, Xi'an (IVM)
Li Deng, Microsoft, USA (SLA)
Jyh-Shing Roger Jang, National Tsing-Hua University, Hsinchu (SLA)
Jianguo Huang, Northwestern Polytech. University, Xi'an (SIPTM)
Chengqing Zong, Chinese Academy of Sciences, Beijing (SIPTM)
Zhi-Quan (Tom) Luo, University of Minnesota, USA (WCN)
Jing Wang, Tsinghua University, Beijing (WCN)

Forum Co-Chairs
Jhing-Fa Wang, National Cheng Kung University, Tainan
Hitoshi Kiya, Tokyo Metropolitan University, Japan
Tianling Ren, Tsinghua University, Beijing

Panel Session Co-Chairs
Lin-Shan Lee, National Taiwan University, Taipei
Jianwu Dang, JAIST, Japan/Tianjin University, Tianjin
Mingyi He, Northwestern Polytech. University, Xi'an
Kenneth Lam, The Hong Kong Polytechnic University, Hong Kong

Special Session Co-Chairs
Alex Kot, Nanyang Technological University, Singapore
Masato Akagi, JAIST, Japan
Weibin Zhu, Beijing Jiaotong University, Beijing

Tutorial Session Co-Chairs
Waleed Abdulla, University of Auckland, New Zealand
Jiwu Huang, Sun Yat-sen University, Guangzhou
Koh Soo Ngee, Nanyang Technological University, Singapore

Publicity Co-Chairs
Namsoo Kim, Seoul National University, Korea
Mrityunjoy Chakraborty, IIT Kharagpur, India
Qing Wang, Northwestern Polytechnical University, Xi'an

Publication Co-Chairs
Antonio Ortega, University of Southern California, USA
Eng Siong Chng, Nanyang Technological University, Singapore
Ying Li, Northwestern Polytechnical University, Xi'an

Local Arrangement Co-Chairs
Lei Xie, Northwestern Polytechnical University, Xi'an
Xiaojun Wu, Tsinghua University, Beijing
Jiangbin Zheng, Northwestern Polytechnical University, Xi'an

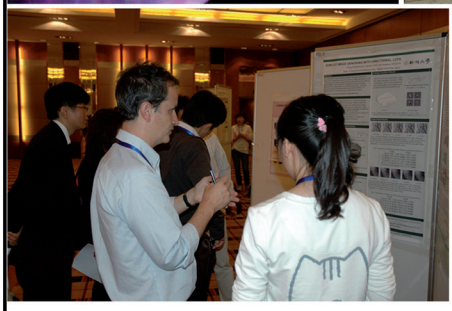
Sponsorship Co-Chairs
Guoqing Wang, Aviation Industry Corporation of China, Shanghai
Xi Xiao, Tsinghua University, Beijing
Dongmei Jiang, Northwestern Polytechnical University, Xi'an

Industrial & Government Advisor
Yong Qin, IBM Research - China, Beijing
Claus Bauer, Dolby Laboratories Intl. Services (Beijing), Beijing

Financial Co-chairs
Qiang Zhou, Tsinghua University, Beijing
Runping Xi, Northwestern Polytechnical University, Xi'an
Jinqiu Sun, Northwestern Polytechnical University, Xi'an

Registration Co-Chairs
Xinbo Zhao, Northwestern Polytechnical University, Xi'an
Zhonghua Fu, Northwestern Polytechnical University, Xi'an
Tao Yang, Northwestern Polytechnical University, Xi'an

Photo Gallery – Xi'an



For more photos: http://www.apsipa.org/photo/apsipa2011

APSIPA Annual Summit and Conference

APSIPA ASC 2010

December 14 -17, 2010
Biopolis, Singapore
http://apsipa2010.i2r.a-star.edu.sg

Honorary Co-Chairs
Kin Mun Lye
Institute for Infocomm Research, A*STAR, Singapore
Meng Hwa Er
Nanyang Technological University, Singapore
Sanjit Mitra
University of California Santa Barbara, USA

General Co-Chairs
Susanto Rahardja
Institute for Infocomm Research, A*STAR, Singapore
Benjamin Wah
University of Illinois at Urbana-Champaign, USA

Technical Program Co-Chairs
Haizhou Li
Institute for Infocomm Research, A*STAR, Singapore
Shri Narayanan
University of Southern California, USA
Nam Ling
Santa Clara University, USA
Eliathamby Ambikairajah
University of New South Wales, Australia
Zhi Ning Chen
Institute for Infocomm Research, A*STAR, Singapore

Forum Co-Chairs
Mark Liao
Institute of Information Science Academia Sinica, Taiwan
Lin-shan Lee
National Taiwan University, Taiwan

Panel Session Co-Chairs
C.C. Jay Kuo
University of South California, USA
Mohan Kankanhalli
National University of Singapore, Singapore

Special Session Co-Chairs
Alexander C. Loui
Eastman Kodak Company, USA
Susanne Boll
University of Oldenburg, Germany

Tutorial Session Co-Chairs
Pasi Franti
University of Joensuu, Finland
King Ngi Ngan
Chinese University of Hong Kong, Hong Kong
Daniel Racoceanu
National Center for Scientific Research, France

Publicity Co-Chairs
Hwee Hwa Pang
Singapore Management University, Singapore
Alireza Ahrary
Fukuoka Industry, Science and Technology Foundation, Japan
Rajalida Lipikorn
Chulalongkorn University, Thailand

Publication Co-Chairs
Roger Zimmermann
National University of Singapore, Singapore
Norihide Kitaoka
Nagoya University, Japan
Eng Siong Chng
Nanyang Technological University, Singapore

Sponsorship Chair
Jongwon Kim
Gwang-Ju Institute of Science & Technology, Korea

Local Arrangement Chair
Lekha Chaisorn
Institute for Infocomm Research, A*STAR, Singapore

Industrial & Government Advisor
Arthur Fong
Media Development Authority, Singapore

Photo Gallery – Singapore

For more photos: http://www.apsipa.org/photo/apsipa2010

2009 APSIPA Annual Summit and Conference
APSIPA ASC 2009

October 5 - 7, 2009
Sapporo Convention Center, Sapporo, Japan

Organizing Committee:

Honorary Chair
Sadaoki Furui
Tokyo Institute of Technology, Japan

General co-Chairs
Yoshikazu Miyanaga
Hokkaido University, Japan
K. J. Ray Liu
University of Maryland, USA

Technical Program co-Chairs
Hitoshi Kiya
Tokyo Metropolitan Univ., Japan
Tomoaki Ohtsuki
Keio University, Japan
Mark Liao
Academia Sinica, Taiwan
Takao Onoye
Osaka University, Japan

Forum co-Chairs
Wan-Chi Siu
Hong Kong Polytechnic Univ., Hong Kong
Xinhua Zhuang
University of Missouri , USA
Waleed Abdulla
University of Auckland , New Zealand

Panel Session co-Chairs
Kiyoharu Aizawa
Tokyo University, Japan
Min Wu
University of Maryland , USA
Ling Guan
Ryerson University , Canada

Special Session co-Chairs
Nobuo Hataoka
Tohoku Institute of Tech., Japan
Antonio Ortega
University of Southern California , USA
Kosin Chamnongthai
KMUTT , Thailand
Jiwu Huang
Sun Yat-sen University , China

Tutorial Session co-Chairs
Thomas Zheng
Tsinghua University, China
Mrityunjoy Chakraborty
IIT Kharagpur , India
Subhrakanti Dey
University of Melbourne , Australia
Helen Meng
Chinese Univ. of Hong Kong , Hong Kong

Publicity co-Chairs
Naohisa Ohta
Keio University, Japan
Susanto Rahardja
*I^2R A*STAR, Singapore*
Chong-Yung Chi
National Tsing-Hua University, Taiwan
Namsoo Kim
Seoul National University, Korea

Publication co-Chairs
Hideaki Sakai
Kyoto University, Japan
Ying-Chang Liang
*I2R, A*STAR, Singapore*

Fundraising Chair
Yoh'ichi Tohkura
National Institute of Informatics, Japan

Finance Chair
Akira Taguchi
Musashi Institute of Technology, Japan

Photo Gallery – Sapporo

For more photos: http://www.apsipa.org/photo/apsipa2009

Sample Documents of APSIPA Governance

APSIPA Bylaws

Article 1 - Name and Residence

The name of the association is "Asia-Pacific Signal and Information Processing Association (APSIPA)". The Association is registered in Hong Kong.

Article 2 - Mission

Mission - APSIPA is a non-profit organization with the following objectives:

- providing education, research and development exchange platforms for both academia and industry
- organizing common-interest activities for researchers and practitioners
- facilitating collaboration with region-specific focuses and promoting leadership for worldwide events
- disseminating research results and educational material via publications, presentations, and electronic media
- offering personal and professional career opportunities with development information and networking

Article 3 - Field of Interest

The field of interest of APSIPA concerns all aspects of signals and information including processing, recognition, classification, communications, networking, computing, system design, security, implementation, and technology with applications to scientific, engineering, health, and social areas.

Article 4 - Membership

4.1 Individual members are those who have paid individual membership dues. APSIPA supports three types of individual membership: full membership, student membership, and life membership.

- All members are entitled to equal rights and privileges, except that student members are not entitled to vote.
- Any member who fails to pay his/her membership dues will have their membership revoked accordingly.

4.2 APSIPA also supports a form of institutional membership termed a "Patron membership" if an institution, organization, company, or laboratory pays the Patron membership fee. A patron can participate in the Patron Forum and enjoy the networking and programs offered by APSIPA.

Article 5 - Officers and Board of Governors

5.1 The Board of Governors (BoG) consists of Officers and Members-at-Large. All members of BoG must be members of the Association. There are twelve Members-at-Large, and the Officers are:

- President
- President-Elect
- Past-President
- Vice President - Conferences
- Vice President - Industrial Relations and Development
- Vice President - Institutional Relations and Education Program
- Vice President - Member Relations and Development
- Vice President - Publications
- Vice President - Technical Activities

5.2 Election

5.2.1 Members-at-Large of the BoG are elected by direct vote of the voting Members of the Association. The officers are elected by the BoG.

- One-third of the total (12) Members-at-Large will be elected annually.
- The term of a Member-at-Large is three years.
- All officers have two-year terms.
- No more than three Members-at-Large shall come from the same country.

5.2.2 Eligibility for Re-election

- Vice Presidents may be elected to the same office for no more than two consecutive terms.
- Members-at-large may be elected for no more than two consecutive terms.

5.3 Officer Responsibilities

The President manages all aspects of the Association and represents the Association, whether internally or externally, to promote and protect the interests of the Association and its members.

- The President-Elect is the Treasurer of the Association. He/she is also in charge of long-term planning.
- The Past-President is in charge of awards, nominations and elections. He/she chairs the Award and Election Board.

5.3.1 Vice Presidents are accountable to the President for their areas of responsibility.

5.3.2 Vice President - Conferences is responsible for all aspects of technical conferences, workshops, and professional meetings, including conference publications, and co-sponsoring of existing meetings. He/she chairs the Conference Board.

5.3.3 Vice President - Industrial Relations and Development is responsible for developing industrial patrons, industry forum and networking, and outreach to industry. He/she chairs the Industry Board and Forum.

5.3.4 Vice President - Institutional Relations and Education Program is responsible for developing institutional patrons and building relations with existing agencies, societies, and associations within each affiliated country. He/she is also responsible for developing educational programs to meet the needs and demands of different regions and members. He/she chairs the Institution Board and Forum.

5.3.5 Vice President - Member Relations and Development is responsible for programs related to members, branches, membership development, and marketing, and fostering a strong international presence. He/she chairs the Branch and Membership Board.

5.3.6 Vice President - Publications is responsible for all activities related to print and electronic products, such as journals, magazines, and on-line offerings. He/she chairs the Publications Board.

5.3.7 Vice President - Technical Activities is responsible for overseeing the technical committees and their technical activities. He/she chairs the Technical Activities Board.

5.4 Board of Governors Meetings

5.4.1 The BoG must hold one formal meeting annually. Special BoG meetings may be held at the request of the President or four members of the BoG. A majority of the voting members of the BoG constitutes a quorum. When a quorum is present, a majority vote is necessary to pass motions.

5.4.2 Business may be conducted by means other than formally held meetings when the matter can be adequately handled via letter, electronic ballot, conference call, or electronic mail interchange, etc.

5.5 Executive Committee (EXCOM)
Between formal and special BoG meetings, business will be managed by the EXCOM consisting of all the officers. Actions of EXCOM must be ratified by the BoG in its next meeting.

5.6 Operations

5.6.1 Minutes of each BoG and EXCOM meeting will be distributed to the BoG within 30 days of the meeting.

5.6.2 Members of the BoG and EXCOM must receive notice of formal meetings no less than 30 days prior to the scheduled date.

5.6.3 If a quorum is not present at a duly called BoG or EXCOM meeting, actions may be formulated but are not effective until ratified by letter, electronic mail, or conference call. A majority vote of that specific body with the quorum is required for ratification. Approved decisions will be recorded in the minutes of that meeting.

5.6.4 Business at Association meetings shall be conducted according to Robert's Rules of Order (latest revision).

5.6.5 The BoG may relieve volunteers in appointed/assigned positions of their responsibilities.

Article 6 - Technical Committees

6.1 Technical Committees are established to promote and achieve the technical objectives of the Association. Technical Committees may be created, merged, or dissolved by resolution of the BoG.

6.2 The Chair of a new Technical Committee is appointed for two years by the Vice President - Technical Activities with the approval of the President. During this period a mentor is assigned to the committee by the Vice President - Technical Activities. Subsequently, the Chair will be elected by members of the Technical Committee.

6.3 Elections for Technical Committee Chairs are held every two years for a two-year term. A Chair cannot serve more than two consecutive terms of office.

6.4 General policies and procedures are provided to guide technical committees and may be modified for the individual technical committee with approval of the Vice President - Technical Activities. Policies must include officer positions and election procedures.

6.5 Each Technical Committee will have a technical scope that may be modified when appropriate, upon approval of the Vice President - Technical Activities and with consent of the BoG.

Article 7 - Budget and Finance

7.1 Each year the Treasurer is responsible for the development of the annual budget which must be approved by the BoG.

7.2 The Treasurer monitors revenues and expenses, providing periodical review of the Association finances and recommends adjustments needed to ensure financial stability. A complete financial report is presented by the Treasurer annually.

7.3 The Treasurer will follow orderly procedures for disbursement of funds, providing sufficient checks and balances and appropriate record keeping. A budgeted expenditure requires no further approval beyond approval of the Treasurer.

7.4 The Treasurer will cooperate with Association officials to carry out financial audits when requested. The results of these audits will be presented to the BoG.

Article 8 - Regional and Local Branches

Each country, region, or local area, can establish a branch, provided they have a minimum of 10 members with the approval of Vice President - Member Relations and Development. A branch shall elect a chair and promise to comply with APSIPA Policies and Rules.

Article 9 -Conferences and Workshops

9.1 To accomplish the mission and objectives of APSIPA, an annual flagship conference called APSIPA Summit shall be held during the fall, where the conference, workshops, institution forum, industry forum, BoG meeting, standing boards and committees meetings shall take place.

9.2 APSIPA shall also financially sponsor other conferences and workshops of interest to its members. All the financially sponsored conferences and workshops shall comply the policies and rules of APSIPA.

9.3 All conferences and workshops proposals shall be approved by Conference Board and overseen by Vice President - Conferences. To ensure sound financial stability, each conference/workshop shall budget a 10% profit margin. After each conference/workshop, financial audit shall take place. The organizers of the hosting local branches shall retain 25% final profit for branch activities and development, and the rest shall return to the Treasurer of APSIPA.

9.4 APSIPA shall also technically sponsor conferences and workshops, determined by quality, interest, and strategic alliance as perceived by its members. All such requests shall be reviewed and approved by Conference Board.

Article 10 - Adoption, Modification, and Amendment of Bylaws

The Bylaws shall be constantly modified and amended via the growth of the Association, in charge by the President. A Bylaws modification and/or amendment shall be thoroughly reviewed and approved with majority 2/3 votes by EXCOM, which shall bring for the approval of BoG in the annual BoG meeting with majority 2/3 votes.

APSIPA Sadaoki Furui Prize Paper Award Guideline

1. Name and number of the Award: APSIPA awards ONE APSIPA Sadaoki Furui Prize Paper Award selected each year from the APSIPA Transactions.
2. Award Committee:
   a) The Award Committee is co-chaired by the VP of Publications and the EiC of the APSIPA Transactions.
   b) The Award Committee includes THREE elected members, who are nominated by members of the BoG, Advisory Board and APSIPA Transactions Editorial Board. They are elected by the BoG. Each member has a term of two years, and can serve up to two consecutive terms.
   c) The Award Committee is in charge of collecting nominations and selecting an award candidate from nominations received.
3. Nomination and Awarding Process:
   a) The nominated paper should be published on the APSIPA Transactions within a FIVE-year window ending at the end of the previous calendar year.
   b) The Award Committee accepts nominations from the general public, including members and non-members of APSIPA. The deadline for nominations is August 1
   c) The Award Committee selects a candidate for recommendation to the BoG by September 1.
   d) After approval by the BoG, the award recipient is announced on October 1.
   e) The award is granted during the APSIPA ASC.
4. This Guideline is approved and revised if necessary by the BoG.

Friend Labs

Motivation and Background

One of the key missions of APSIPA is to provide education, research and development exchange platforms for both academia and industry. One way to accomplish this mission is to recruit academia and industrial labs to become APSIPA Friend Labs.

Benefits

All APSIPA Friend Labs will be listed in the APSIPA website. Each lab has one page to post lab information, photos and a link to the lab home page. The provided data in the on-line application form will be used to generate the friend lab page in the APSIPA website. We will encourage mutual visit and information sharing among APSIPA Friend Labs.

Application Criterion

An academia or industrial lab is qualified to become an APSIPA friend lab if it has at least 10 current or former lab members who are full or associate members of APSIPA. A person can become an APSIPA associate member by joining the APSIPA Group in Linkedin. A person can become an APSIPA full member by clicking the "Join Us" button in the up-right corner of the APSIPA homepage and following the given instructions.

Application Process

Each lab has to apply to become an APSIPA Friend Lab. Please fill out the on-line application form which is available from the Website of APSIPA. The approval process will take about one week.

APSIPA Friend Lab Promotion Committee

Chair:	Hitoshi Kiya, Tokyo Metropolitan University, Japan
Secretary:	Yoshinozu Kajikawa, Kansai University, Japan
Members:	Kosin Chamnongthai, King Mongkut's University of Technology Thonburi, Thailand Wen-Huang Cheng, Academia Sinica, Taiwan Woon Seng Gan, Nanyang Technological University of Singapore, Singapore Anthony Kuh, University of Hawaii, USA Sanghoon Lee, Yonsei University, Korea Jiaying Liu, Peking University, China Tomek Rutkowski, University of Tsukuba, Japan Osamu Takyu, Shinshu University, Japan Toshihisa Tanaka, Tokyo University of Agriculture and Technology, Japan KokSheik Wong, University of Malaya, Malaysia Chia-Hung Yeh, National Sun Yat-sen University, Taiwan

APSIPA Transactions on Signal and Information Processing (TSIP) Editorial Board Guidelines

1. Editor-in-Chief (EiC)
   a) Editor-in-Chief (EiC) is responsible for all editorial processes of TSIP, and works closely with the Publisher and the VP of Publications.
   b) EiC candidate is elected by EiC Election Committee, which is chaired by VP of Publications and consists of APSIPA President, President-Elect, current EiC and immediate past EiC. The EiC Election Committee accepts nominations from the members of APSIPA, and selects a candidate for recommendation to the BoG. After approval by the BoG, a new EiC is appointed.
   c) EiC has a term of TWO years, and can serve up to TWO consecutive terms.
   d) EiC must submit an annual report in the BoG meeting.
2. Editorial Board (EB)
   a) Editorial Board members have commitment to the editorial processes of TSIP; they are in charge of the paper review process and they should contribute to the growth of TSIP.
   b) New EB members are nominated by EiC and current EB members, and appointed by EiC.
   c) EB members have a term of TWO years, and can be re-appointed by EiC.
3. Advisory Board
   a) Advisory Board gives advices to EiC and EB.
   b) APSIPA President, President-Elect and immediate past EiC are ex-officio members of the Advisory Board, and other members are nominated and appointed by EiC.
4. Editorial Board Meeting
   a) Editorial Board meeting is to be held in APSIPA ASC annually and chaired by EiC.
   b) All EB members are invited as voting members, and all Advisory Board members are invited as non-voting members.
5. This Guideline has been approved and any further modification must be approved by the BoG.

APSIPA Newsletter

APSIPA is pleased to publish an online newsletter for the Signal and Information Processing research interest community. All are welcome to send us their contributions to publish it in APSIPA Newsletter. All contributions should not be more than double column one page with font size 10 including images. The contributions could be (but not limited) articles or notes on:

1. Review on a recently published book or monograph (preferably within 5 years)
2. Interesting technology or research line
3. Interesting scientific fact, progress, development, ... etc
4. Science and technology in history and pioneers
5. Announcements specially for interns, scholarships, PG studies, fellowships, and postdocs
6. Lectures in signal and information technology (description and link to the video of presentation)
7. Any other material you think suitable for inclusion in the newsletter

All submissions should be sent to Dr. Bonnie law (ennflaw@polyu.edu.hk) with a subject heading 'APSIPA Newsletter'. We reserve the right to carry on minor editing to all submissions to meet our editorial procedure. Copyright of material remains with the original author who grants us the right to make it available on APSIPA website.

2018 APSIPA Officers

President:	Wan-Chi Siu, The Hong Kong Polytechnic University, Hong Kong (2018)
Past Presidents:	Sadaoki Furui (2009-2012), C.C. Jay Kuo (2013-2014)
Immediate Past President:	Haizhou Li, National University of Singapore, Singapore (2018)
President-Elect:	Hitoshi Kiya, Tokyo Metropolitan University, Japan (2018)
VP - Conferences:	Thomas Fang Zheng, Tsinghua University, China (2019)
VP - Industrial Relations and Development:	Guan-Ming Su, Dolby Labs, Sunnyvale, CA, USA (2019)
VP - Institutional Relations and Education Program:	Woon-Seng Gan, Nanyang Technological University, Singapore (2018)
VP - Member Relations and Development:	Yoshinobu Kajikawa, Kansai University, Japan (2019)
VP - Publications:	Kenneth Lam, The Hong Kong Polytechnic University, Hong Kong (2019)
VP - Technical Activities:	Anthony Kuh, University of Hawaii at Manoa, USA (2019)

Members-at-Large :

Kosin Chamnongthai, King Mongkut's University of Technology Thonburi, Thailand (2019)
Homer Chen, National Taiwan University, Taiwan (2018)
Nam Ik Cho, Seoul National University, Korea (2019)
Hsueh-Ming Hang, National Chiao-Tung University, Taiwan (2018)
Yo-Sung Ho, Gwangju Institute of Science and Technology (GIST), Korea (2020)
Tatsuya Kawahara, Kyoto University (2020)
Bonnie Ngai-Fong Law, The Hong Kong Polytechnic University, Hong Kong (2020)
Kai-Kuang Ma, Nanyang Technological University, Singapore (2018)
Chung-Nan Lee, National Sun Yat-sen University, Taiwan (2019)
Kazuya Takeda, Nagoya University, Japan (2018)
Susanto Rahardja, Northwestern Polytechnical University, (2020)
Toshihisa Tanaka, Tokyo University of Agriculture and Technology (2019)

APSIPA:
A non-profit making organization with limited liability incorporated in Hong Kong under the Companies Ordinance, since 23 July 2009.
APSIPA Headquarters Address:
APSIPA, Asia Pacific Signal and Information Processing Association
Department of Electronic and Information Engineering
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
Website: http://www.apsipa.org/headquarters.htm

978-988-14768-4-5©2018

10th APSIPA Anniversary Magazine:

The Era of Signal and Information Technologies and their Long-Term Prospective