Tutorial On
Spoofing Attack of Speaker Recognition

Prof. Haizhou Li, (haizhou.li@nus.edu.sg)
National University of Singapore, Singapore

Prof. Hemant A. Patil, (hemant_patil@daiict.ac.in)
DA-IICT, Gandhinagar, India

Ms. Madhu R. Kamble, (madhu_kamble@daiict.ac.in)
DA-IICT, Gandhinagar, India

Asia-Pacific Signal and Information Processing Association
Annual Summit and Conference 2017 (APSIPA ASC 2017)
Kuala Lumpur, Malaysia

Time Slot: 14.00-17.00 Date:12th Dec. 2017

Presenters

Prof. Haizhou Li

NUS, Singapore

Prof. Hemant A. Patil

DA-IICT, Gandhinagar, Gujarat

Madhu R. Kamble (Ph.D. Student)

DA-IICT, Gandhinagar, Gujarat

Agenda

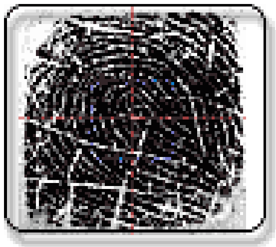
Part 1

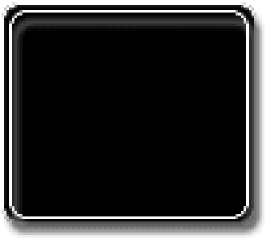
- **Introduction**
- **ASV System**
- Research Issues in ASV
- History of ASV Spoof
- Spoofing Attacks
- Speech Synthesis
- Voice Conversion

Part 2

- Mimics
- Twins
- Countermeasures
- Replay
- ASV Spoof 2015 Challenge
- ASV Spoof 2017 Challenge
- Future Research Directions

Voice Biometrics (ASV)

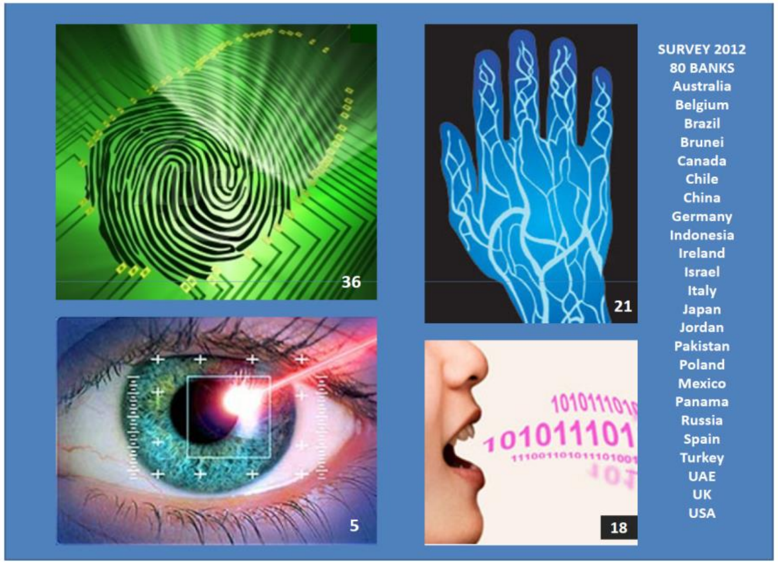




Citi is going to start using voice patterns to authenticate customers over the phone in Asia

Passwords are quickly becoming a thing of the past. And to paraphrase Martha Stewart, that's a good thing. Passwords are easy to guess (though somehow...

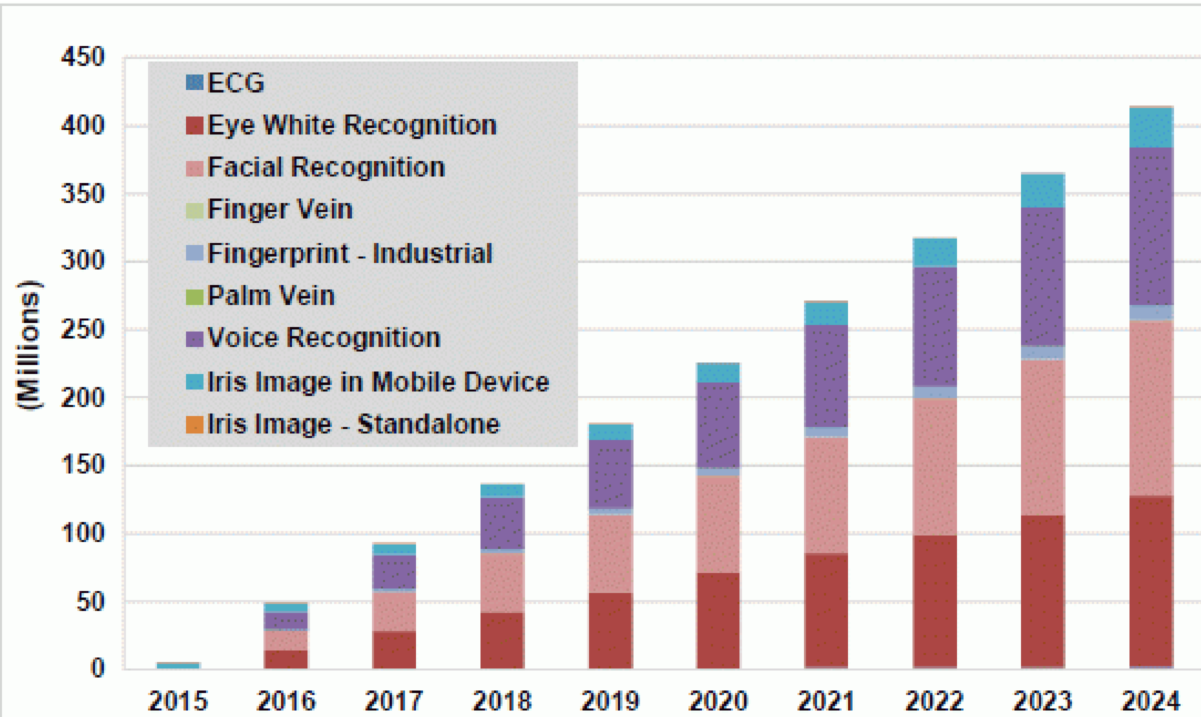
Future of Finance | Ian Kar | May 19, 2016



Various Biometric Spoofing

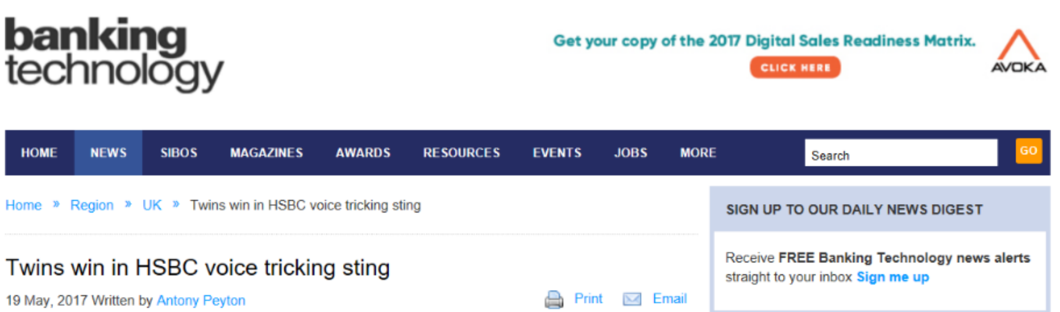
Spoofing

Voice Biometrics



Tractica Finance Biometrics Devices and Licenses by Modality, World Markets: 2015-2024
Kong Aik Lee, Bin Ma, and Haizhou Li, "Speaker Verification Makes Its Debut in Smartphone", IEEE SLTC Newsletter, 2013

Voice Biometrics


banking technology

Get your copy of the 2017 Digital Sales Readiness Matrix. CLICK HERE AVOKA

HOME NEWS SIBOS MAGAZINES AWARDS RESOURCES EVENTS JOBS MORE Search GO

Home » Region » UK » Twins win in HSBC voice tricking sting

Twins win in HSBC voice tricking sting

19 May, 2017 Written by Antony Peyton Print Email

SIGN UP TO OUR DAILY NEWS DIGEST

Receive **FREE Banking Technology news alerts** straight to your inbox **Sign me up**

- HSBC has been left red-faced after a BBC reporter and his non-identical twin tricked its voice ID authentication service.
- The *BBC* says its “Click” (a weekly TV show) reporter Dan Simmons created an HSBC account and signed up to the bank’s service. HSBC states that the system is secure because each person’s voice is “unique”.
- As *Banking Technology* reported last year, HSBC launched voice recognition and touch security services in the UK, available to 15 million banking customers. At that time, HSBC said the system “works by cross-checking against over 100 unique identifiers including both behavioural features such as speed, cadence and pronunciation, and physical aspects including the shape of larynx, vocal tract and nasal passages”.
- According to the *BBC*, the “bank let Dan Simmons’ non-identical twin, Joe, access the account via the telephone after he mimicked his brother’s voice.
- “Customers simply give their account details and date of birth and then say: ‘My voice is my password.’”
- Despite this biometric bamboozle, Joe Simmons couldn’t withdraw money, but he was able to access balances and recent transactions, and was offered the chance to transfer money between accounts.
- Joe Simmons says: “What’s really alarming is that the bank allowed me seven attempts to mimic my brothers’ voiceprint and get it wrong, before I got in at the eighth time of trying.”
- Separately, the *BBC* says a Click researcher “found HSBC Voice ID kept letting them try to access their account after they deliberately failed on 20 separate occasions spread over 12 minutes”.
- The *BBC* says Click’s successful thwarting of the system is believed to be “the first time the voice security measure has been breached”.
- HSBC declined to comment to the BBC on “how secure the system had been until now”.
- An HSBC spokesman says: “The security and safety of our customers’ accounts is of the utmost importance to us. Voice ID is a very secure method of authenticating customers.
- “Twins do have a similar voiceprint, but the introduction of this technology has seen a significant reduction in fraud, and has proven to be more secure than PINS, passwords and memorable phrases.”
- Not a great response is it? But very typical of the kind of bland statements that have taken hold in the UK. There is a problem and HSBC needs to get it fixed.
- The rest of the *BBC* report just contains security experts saying the same things – like “I’m shocked”. Whatever. No point in sharing such dull insight.
- You can see the full BBC Click investigation into biometric security in a special edition of the show on BBC News and on the iPlayer from 20 May.

Voice Biometrics

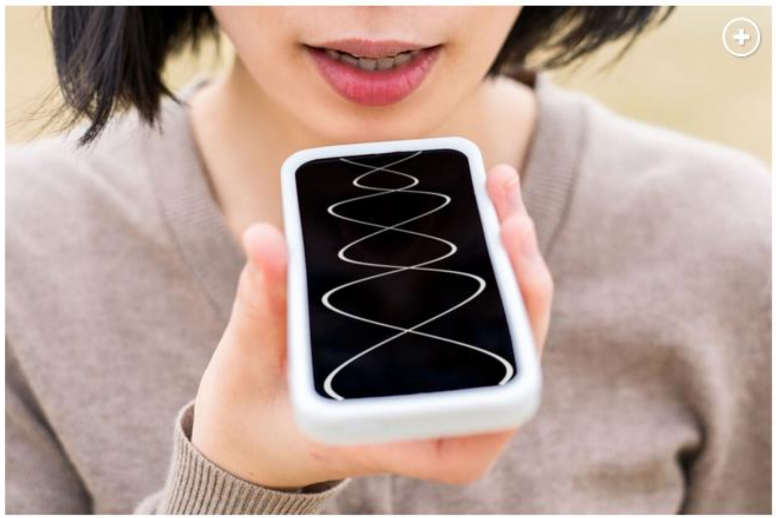
SECTIONS SEARCH **NEW YORK POST** TIPS

TECH

Terrifying AI learns to mimic your voice in under 60 seconds

By Mike Wehner, BGR

May 2, 2017 | 10:52am



Shutterstock

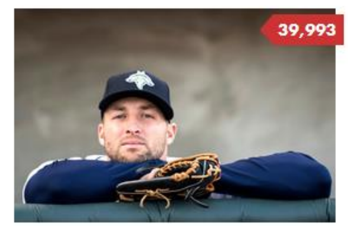
ORIGINALLY PUBLISHED BY:

BGR

When it comes to personal privacy and overall security, we often think of passwords, fingerprints, and even our own faces as being the keys that unlock our world, but what about your voice? If someone could perfectly mimic your voice, what kind of damage could they do? If they contacted people you know, could they lie their way into gaining private information about you?

ON NYPOST.COM

CNN boss in crosshairs if AT&T-Time Warner merger approved



Tim Tebow homers on first day after Mets promotion

The epidemic that's ruining youth sports

Voice Biometrics

Nuance deploys AI biometric security tools

19 May 2017 15:51 GMT

16 Jump to comments

Biometrics firm Nuance, which has focused on voice recognition, has announced a new multi-modal suite of biometric security solutions, driven by artificial intelligence (AI).

The new suit features facial and behavioural biometrics, as well as voice, with the company saying that these combine to provide advanced protection against fraud

Nuance has said that deep neural networks (DNN) are being used in the news solution along side advanced algorithms to detect "synthetic speech attacks".

"By combining a range of physical, behavioural, and digital characteristics to provide secure authentication and more accurately detect fraud across multiple channels - from the phone to the Web, mobile apps and more - Nuance's new Security Suite allows enterprises to attack fraud head-on, while at the same time offering an improved customer experience", wrote the firm.

In particular, the firm notes improved synthetic speech detection

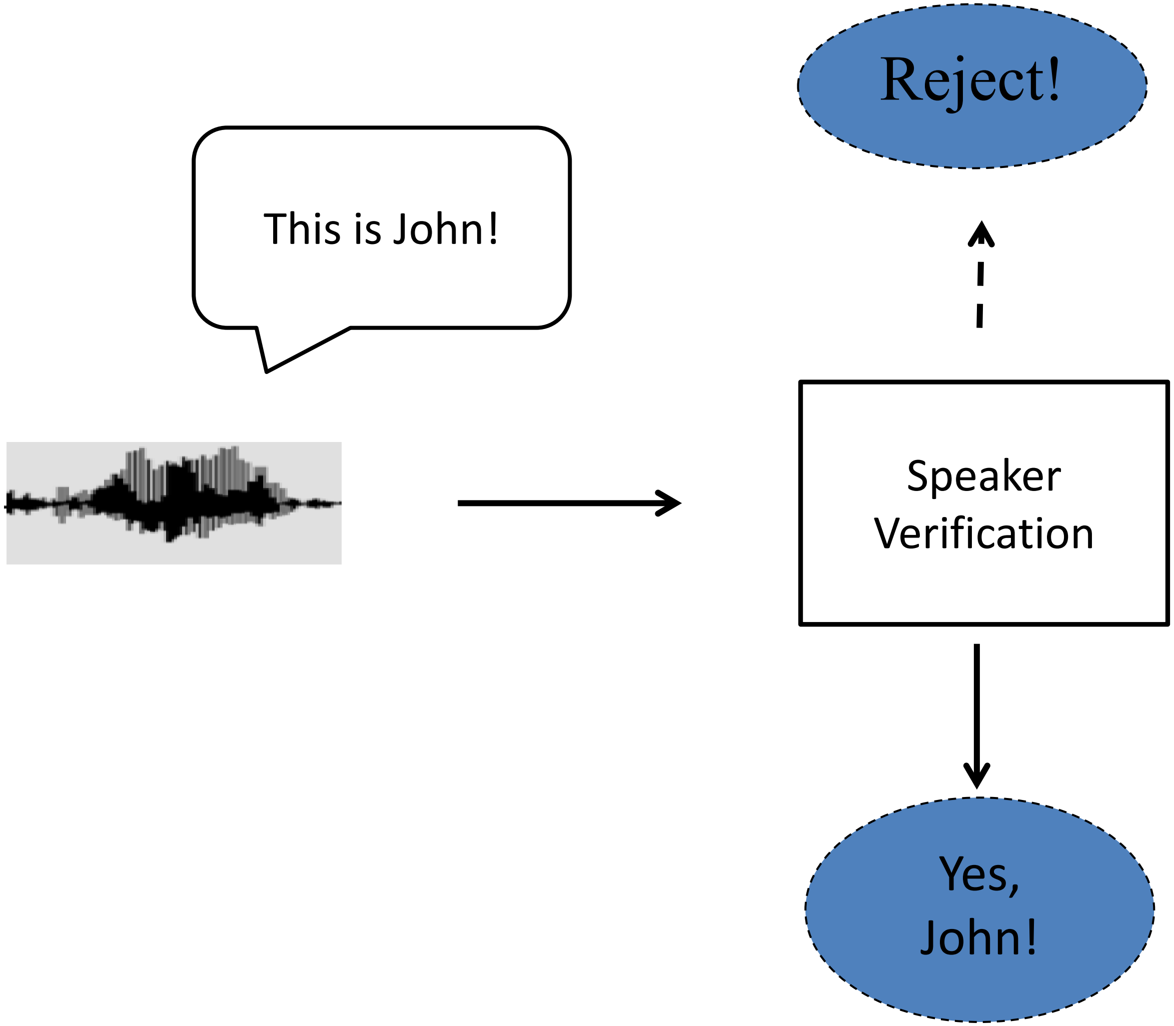
Other site news

IARPA awards $12.5 million contract to SRI

IriTech to showcase mobile iris-barcode solution

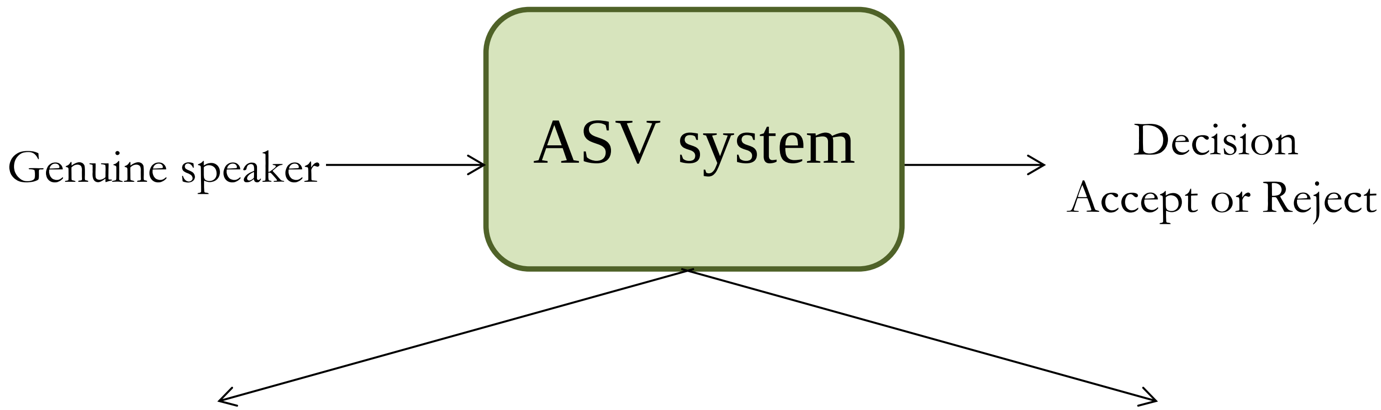
NEC tests face recognition with CBP at Dulles International Airport

Automatic Speaker Verification (ASV)



ASV System

Automatic speaker verification (ASV) system ***accepts*** or ***rejects*** a claimed speakers identity based on a speech sample.



Text-dependent

- Fixed or prompted phrases
- Higher accuracy
- Suited for authentication scenarios

Text-independent

- Any arbitrary utterances
- Call-center applications
- Surveillance scenarios

Block Diagram of ASV System

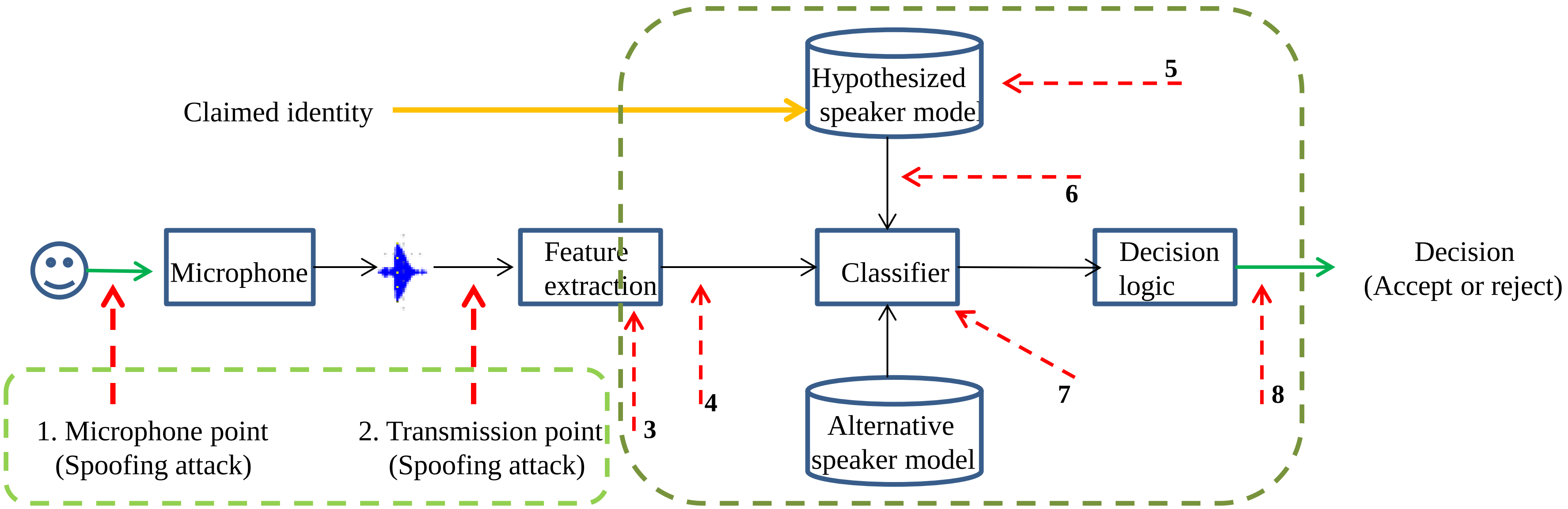
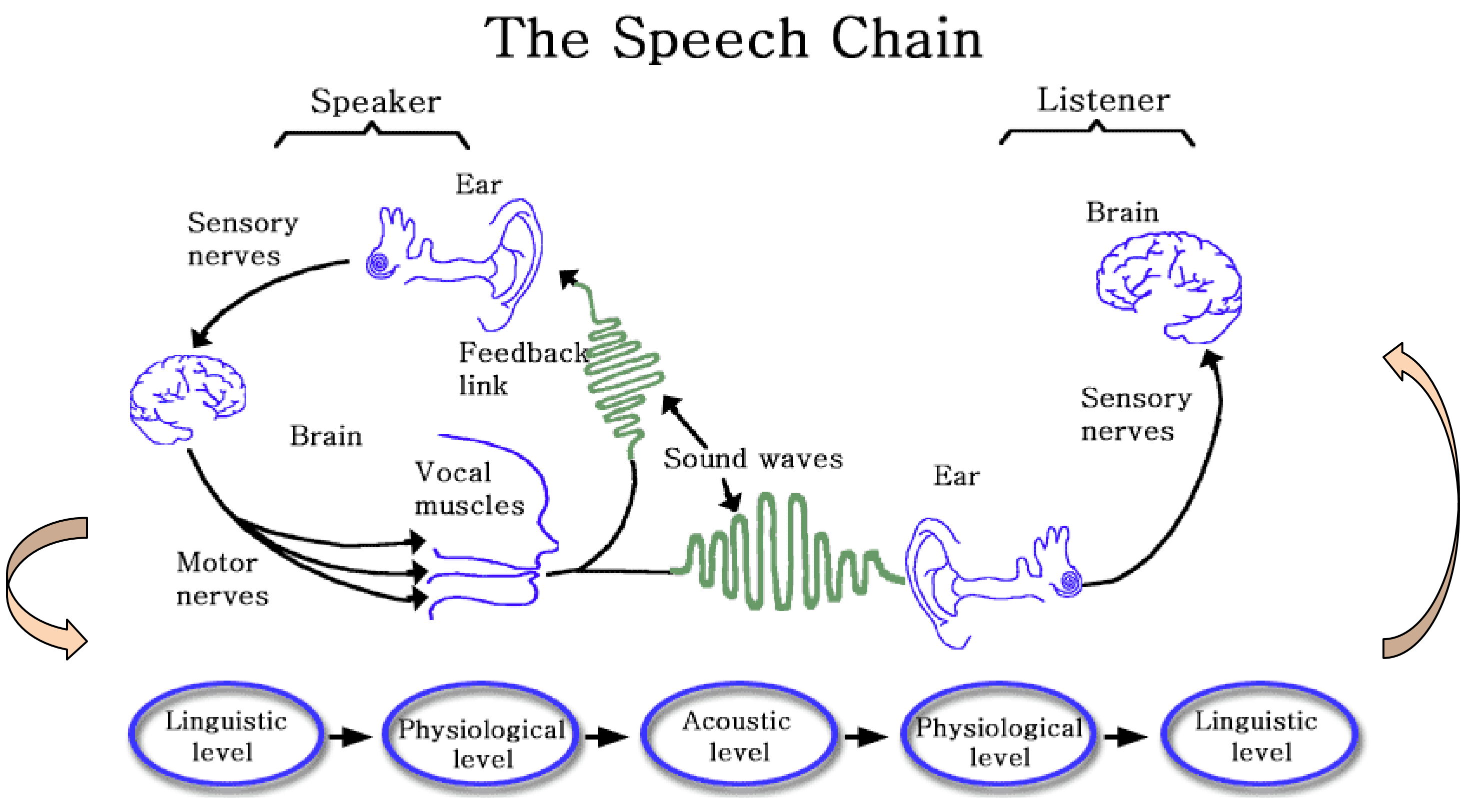


Figure 1: Brief illustration of an ASV system and eight possible attacks. After [1].

- **Direct Attacks**: Attacks applied at the microphone-level as well as the transmission-level – points 1 and 2.
- **Indirect Attacks**: Attacks within the ASV system itself – points 3 to 8.

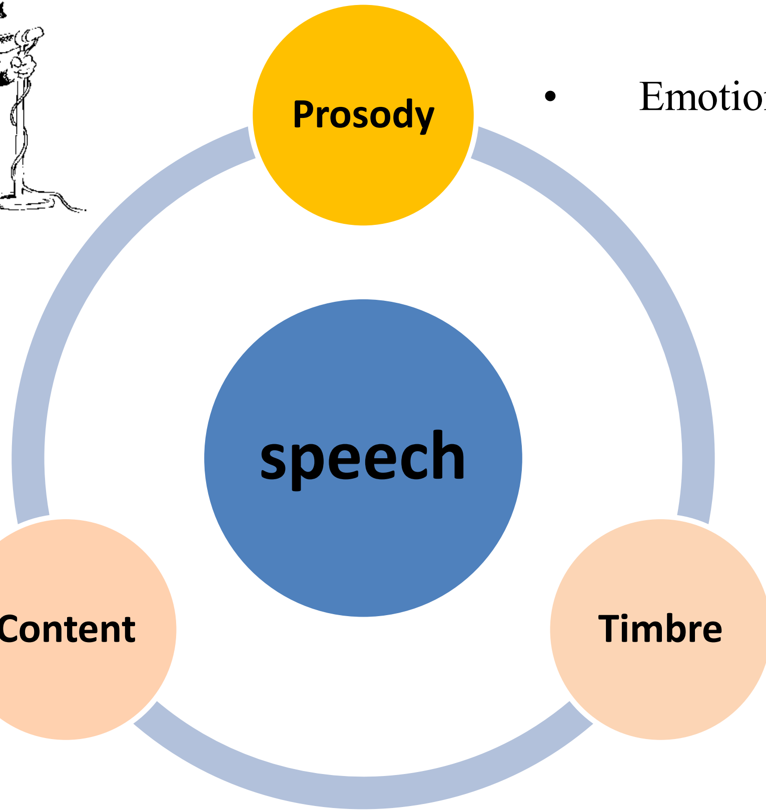
[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Comm.*, vol. 66, pp. 130-153, 2015.

Speech Chain

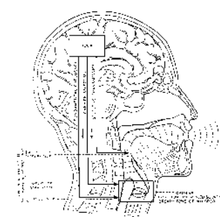


There are 3 subfields of Phonetics, i.e., Articulatory Phonetics, Acoustic Phonetics, and Auditory Phonetics. *Denes & Pinson (1993)*

Elements of Speech Signal



- Emotion to express



- What you want to say ?
- Who you are ?

Speaker Verification

+ Robust against channel effects and noise
- Difficult to extract
- A lot of training data needed
- Delayed decision making

High-level features

Phones, idiolect (personal lexicon), semantics, accent, pronunciation

Learned (behavioral)
Socio-economic status, education, place of birth, language background, personality type, parental influence

Prosodic & spectro-temporal features

Pitch, energy, duration, rhythm, temporal features

+ Easy to extract
+ Small amount of data necessary
+ Text- and language independence
+ Real-time recognition
- Affected by noise and mismatch

Short-term spectral and voice source features

Spectrum, glottal pulse features

Physiological (organic)
Size of the vocal folds, length and dimensions of the vocal tract

- Modeling the human voice production system
- Modeling the peripheral auditory system

Tomi Kinnunen and Haizhou Li, "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors", Speech Communication 52(1): 12--40, January 2010

Variants of Speaker Verification

- Mode of Text
- Text-Dependent
  - same text between enrolment and run-time test
- Text-Independent
  - different text between enrolment and run-time test

- Mode of Operation
- Speaker Identification
  - To identify the speaker from a population
- Speaker Verification
  - To verify if a claimed speaker identity is true

Tomi Kinnunen and Haizhou Li, “An Overview of Text-Independent Speaker Recognition: from Features to Supervectors”, Speech Communication 52(1): 12--40, January 2010

Text-independent Speaker Verification

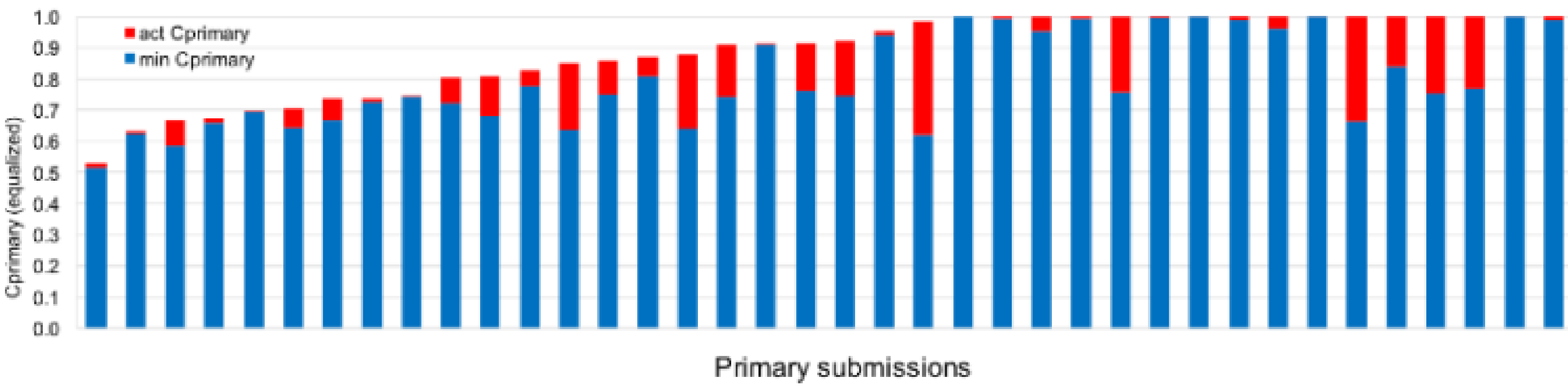


Figure 3: *Actual and minimum $C_{Primary}$ for SRE16 primary submissions.*

Sadjadi et al, The 2016 NIST Speaker Recognition Evaluation, INTERSPEECH 2017

Text-dependent Speaker Verification

ELSEVIER

Speech Communication

Volume 60, May 2014, Pages 56-77

Text-dependent speaker verification: Classifiers, databases and RSR2015

Anthony Larcher, Kong Aik Lee, Bin Ma, Haizhou Li

⊞ Show more

https://doi.org/10.1016/j.specom.2014.03.001

Get rights and content

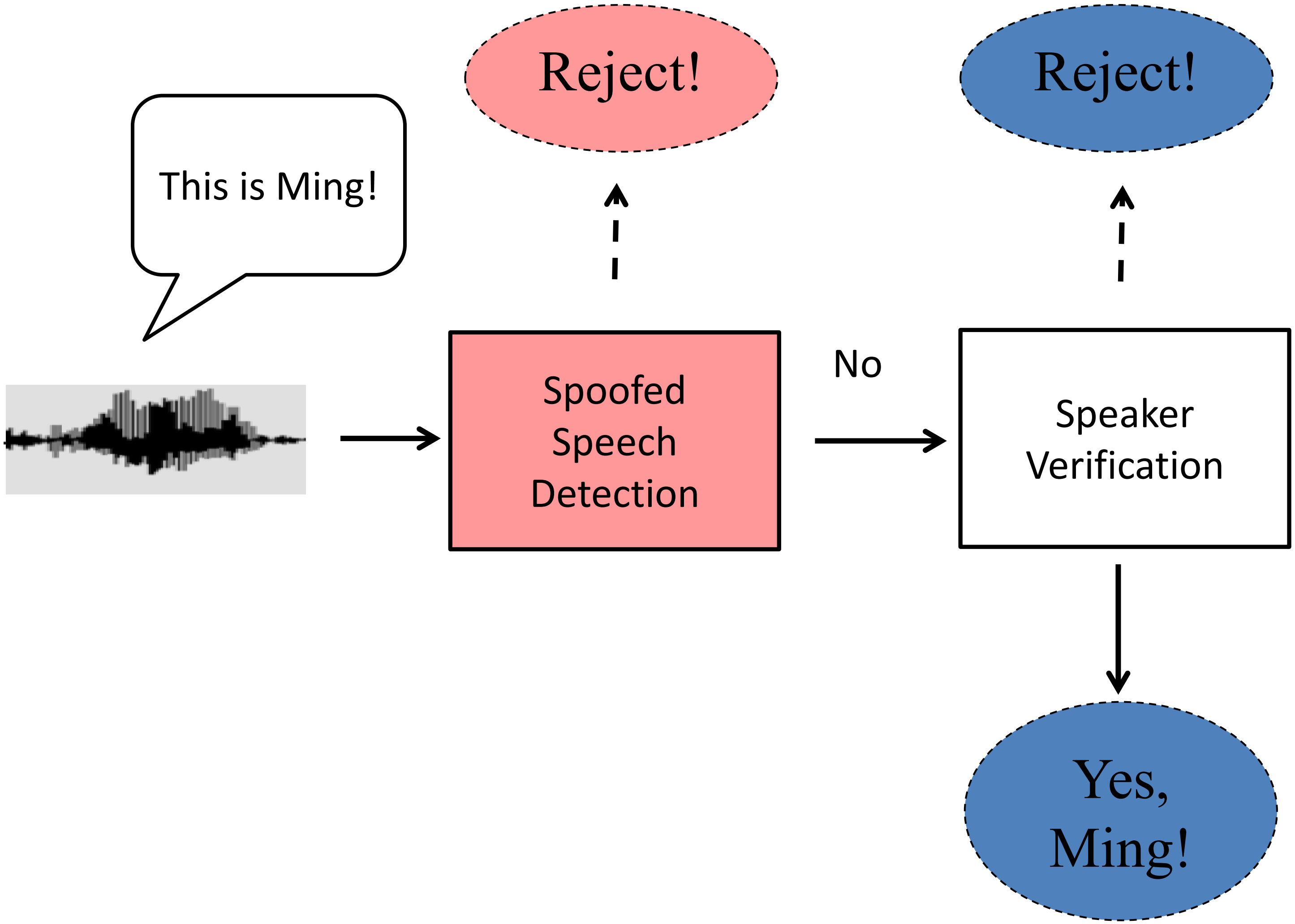
Spoofing: Speaker Verification

- Speaker Verification
  - transducer, channel
  - state of health, mood, aging
  - session variability
- Challenges and Opportunities
  - Systems assume natural speech inputs
  - More robust = more vulnerable
  - Machines and humans listen in different ways [1]
  - Better speech perceptual quality ≠ less artifacts [2]

[1] Duc Hoang Ha Nguyen, Xiong Xiao, Eng Siong Chng, Haizhou Li, "Feature Adaptation Using Linear Spectro-Temporal Transform for Robust Speech Recognition", IEEE/ACM Trans. Audio, Speech & Language Processing 24(6): 1006-1019 (2016).

[2] K.K.Paliwal, et al,"Comparative Evaluation of Speech Enhancement Methods for Robust Automatic Speech Recognition," Int. Conf.Sig. Proce.and Comm.Sys.,Gold Coast, Australia, ICSPCS, Dec. 2010.

Spoofing Attacks



Agenda

Part 1	Part 2

- Introduction
- ASV System
- **Research Issue in ASV**
- **History of ASV Spoof**
- **Spoofing Attacks**
- **Speech Synthesis**
- **Voice Conversion**

- **Mimics**
- **Twins**
- **Countermeasures**
- Replay
- ASV Spoof 2015 Challenge
- ASV Spoof 2017 Challenge
- Future Research Directions

How is Speech Produced?
- Physiological, Acoustic, Aeroacoustics

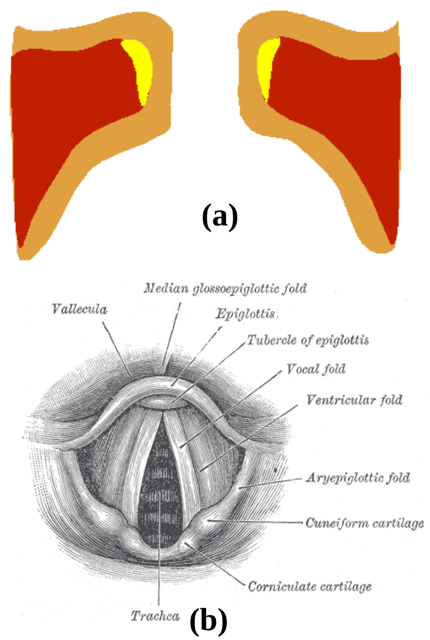


Figure 5. Simulation of vocal folds movement.

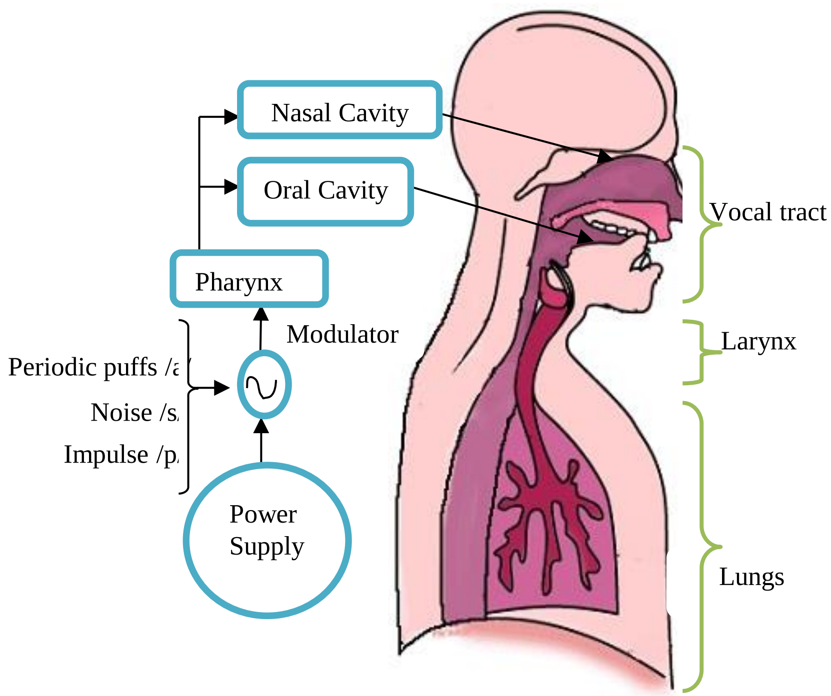


Figure 7. Human speech production system. .

M.D.Plumpe, T.F.Quatieri, and D.A.Reynolds, "Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification" 1999, IEEE

Jankowski, Charles Robert, Thomas F. Quatieri, and Douglas A. Reynolds. "Fine structure features for speaker identification." *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. Vol. 2. IEEE, 1996.

Speech Production (Contd.)

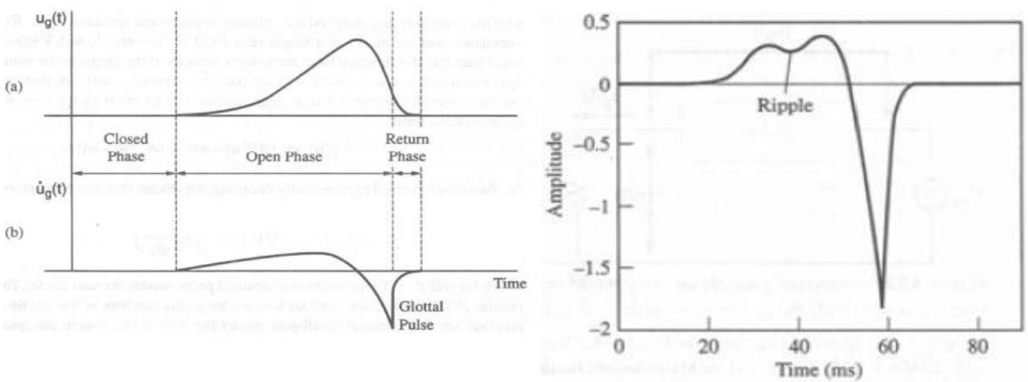


Figure 6: Glottal flow waveform and its derivative over one glottal cycle and ripples in the glottal derivative due to source/vocal tract interaction.

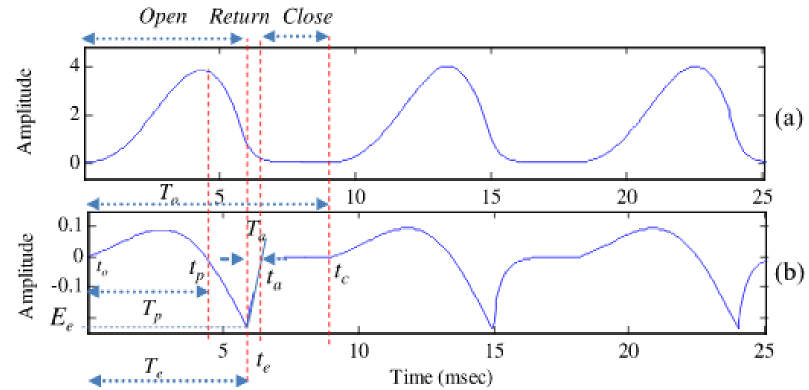


Figure 6.1: a) A schematic of g(t) and (b) the corresponding derivative of the g(t) along with various timing instants and the time periods used in the LF-model.

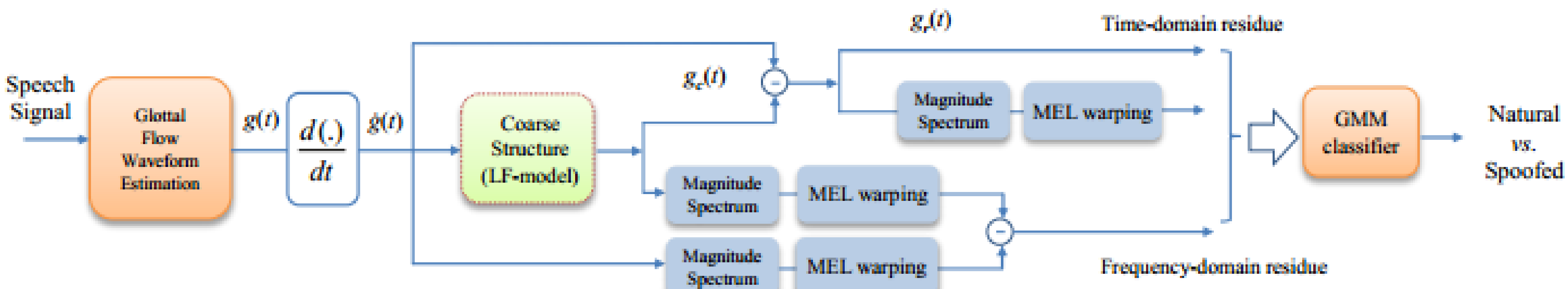


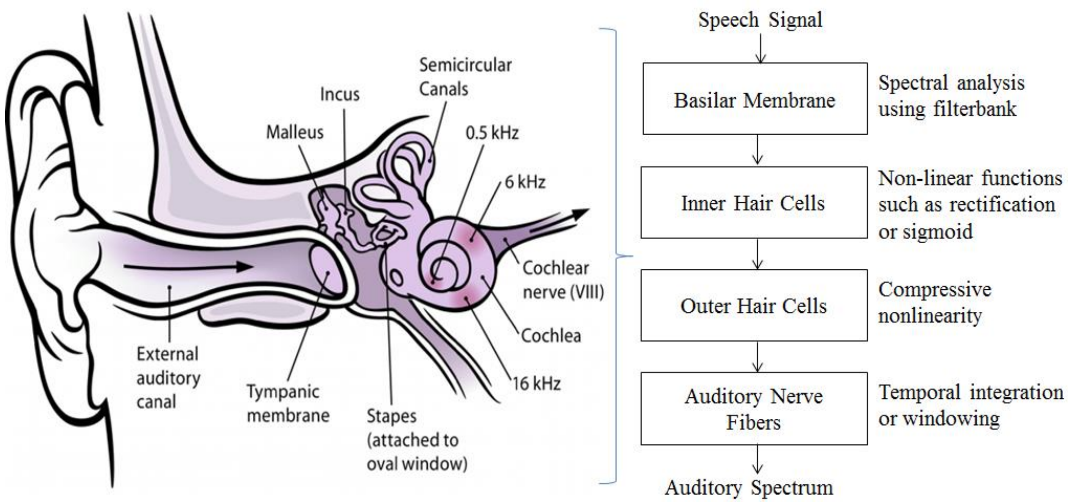
Figure 6.2: Schematic diagram of the S-F interaction feature extraction process (in time and frequency-domain) for the SSD task [1].

T. B. Patel and H. A. Patil, "Significance of source-filter interaction for classification of natural vs. spoofed speech," *IEEE Jour. on Selected Topics in Sig. Process. (JSTSP)*, vol. 11, no. 4, pp. 644 - 659, June 2017.

Hearing and Speech Perception

Threshold of hearing= $2 \times 10^{-5} N / m^2$

-> Process of detecting **energy** !



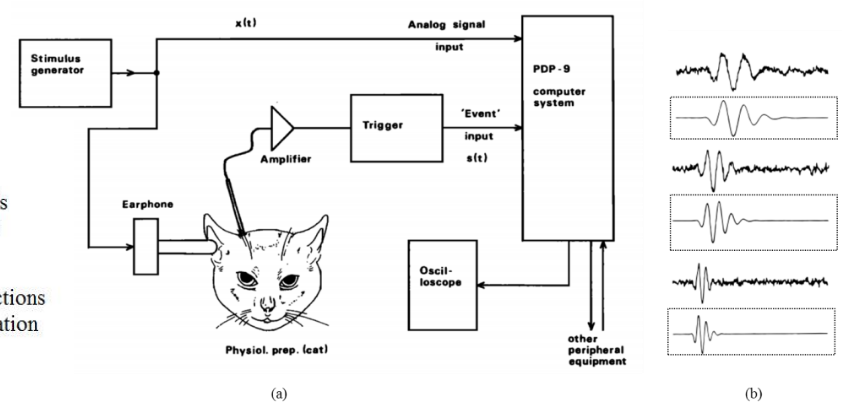


Figure 8.1. Physiological auditory filter estimation [2].

Figure 8. Early auditory processing and its corresponding mathematical representation [1].

Hearing lecture material from Prof. Laurence R. Harris. URL: http://www.yorku.ca/harris/ppt_files/
[1] Jan Schnupp, Israel Nelken and Andrew J. King, "Auditory Neuroscience – Making Sense of Sound", MIT Press 2012..
[2] L. H. Carney, T. C. Yin, "Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population", Journal of Neurophysiology, vol. 60, Pages 1653-1677, 1988

Speaker Biometrics

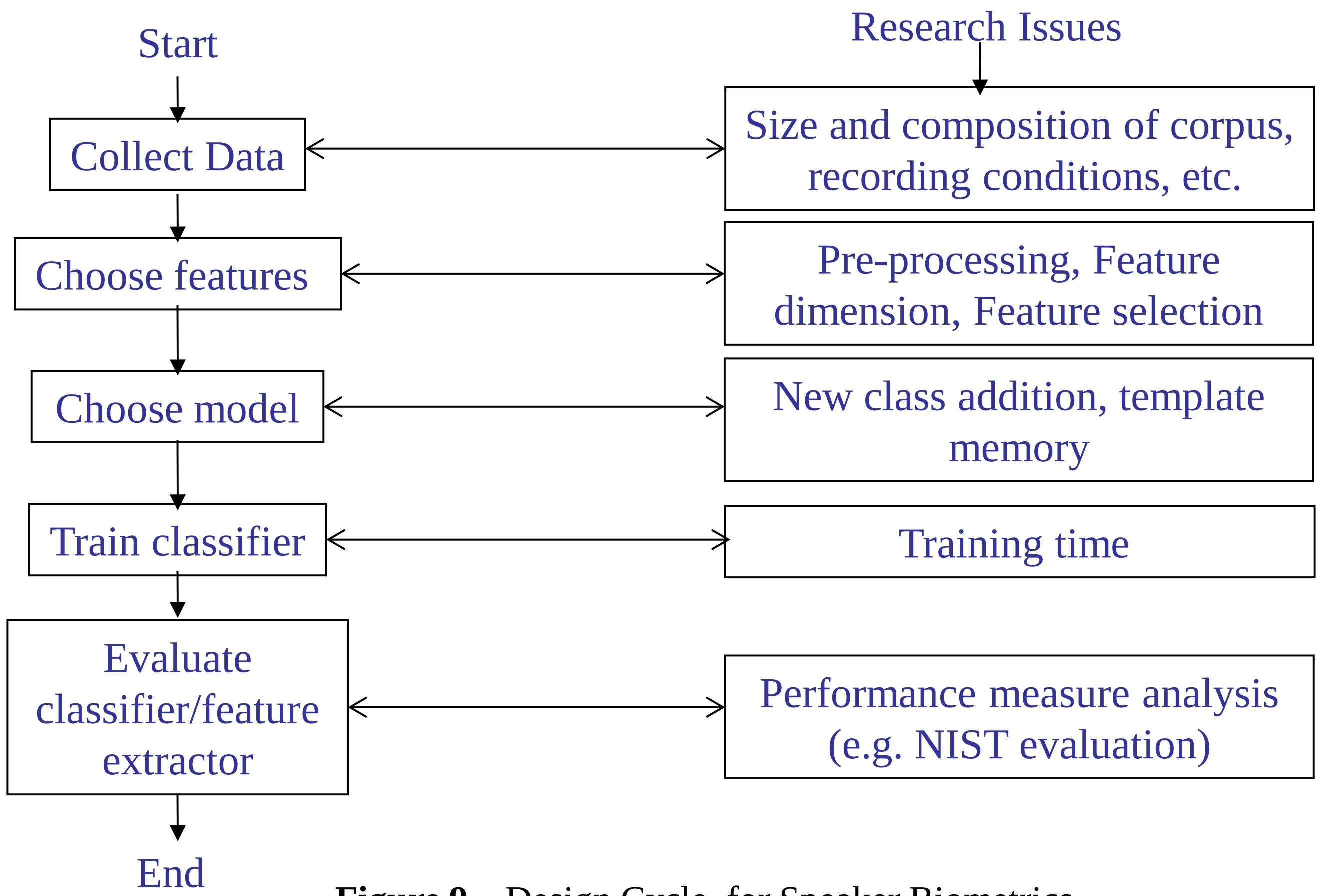


Figure 9. Design Cycle for Speaker Biometrics.

R. E. Duda, P.E. Hart and D. G. Stock, "Pattern Classification," Wiley, 2nd Ed., 2000.

Research Issues in Forensic Speaker Recognition

Research Issues in Forensic Speaker Recognition (comparison)

- Noise
  - Tape noise
  - Traffic or road noise
- Aural/Spectral Comparison *vs.* Automatic methods
- VoIP
  - codec
  - packet loss
  - packet reordering
  - network jitter
  - echo or foreign cross-talk
- Tape authentication
- Voice Disguise
  - Mimics
  - Whispered Voice
- Legal Issues
  - Fry Test
  - Daubert Test
- Stress
  - Anger
  - Fear
  - Anxiety
  - Lie detection

A. Neustein and Hemant A. Patil (Eds.), Forensic Speaker Recognition, Springer, Oct. 2011.

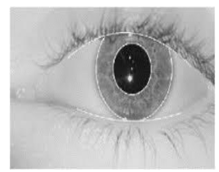
Categories of Biometric Identifications

Physiological

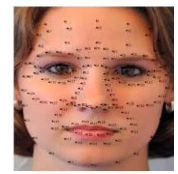
Fingerprint

Hand

Iris



Face



DNA

Behavioral

Voice

Keystroke

Signature

Issues in Voice Biometrics

Mimic Resistance

Physiological Characteristics

Behavioral Characteristics (skillfulness in mimicry)

Identification of Identical Twins or Triplets

Identification of Professional mimics

Similarity in spectral features

Similarity in prosodic features

- Physiological characteristics is challenging
- It has similar or identical vocal tract structure

Rosenberg, Aaron E. "Automatic speaker verification: A review." *Proceedings of the IEEE* 64.4 (1976): 475-487
Jain, Anil K., Salil Prabhakar, and Sharath Pankanti. "On the similarity of identical twin fingerprints." *Pattern Recognition* 35.11 (2002): 2653-2663.

Independent Problem: Spoof Detector

- Due to effect of spoofed speech on ASV systems, need of standalone detectors (natural *vs.* spoofed speech) arose.
- Spoofed speech → impersonated, replay, speech synthesis or voice converted.

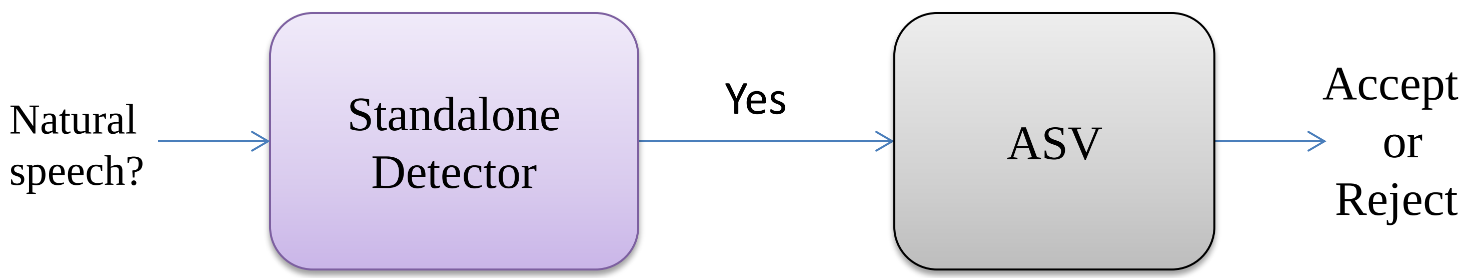


Figure: Automatic Speaker Verification (ASV) System.

- Recent trend is towards detecting synthetic and voice converted speech.

History of ASV Spoof

1999 — 2006 — 2013 — 2014 — 2015 — 2016 — 2017

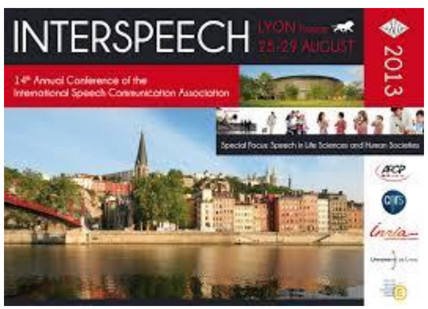
OCTAVE project starts

Small, purpose collected datasets

Adapted, standard datasets

IS 2013

Common, datasets, metrics, protocols



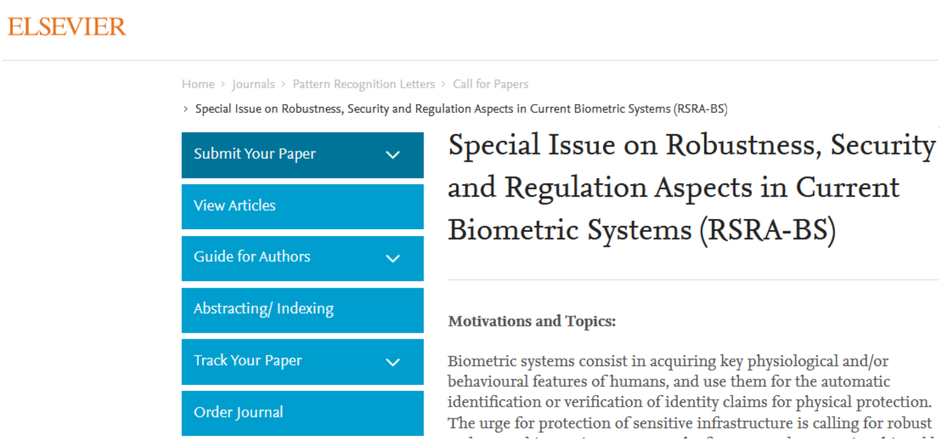
IS 2015

Common, datasets, synthetic speech

IS 2017

Common, datasets, replay, channel variation

Special Issues



Speech Synthesis (SS)

- Speech synthesis is the artificial production of **human speech.**
- Computer or instrument used is **Speech Synthesizer.**
- Text-To-Speech (TTS) synthesis is production of speech from normal language text.

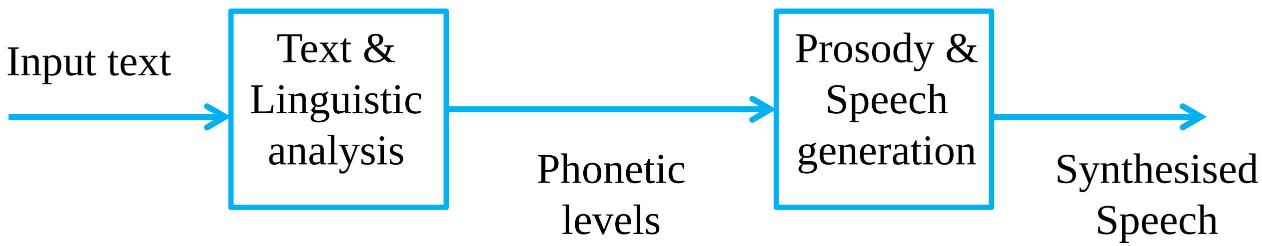
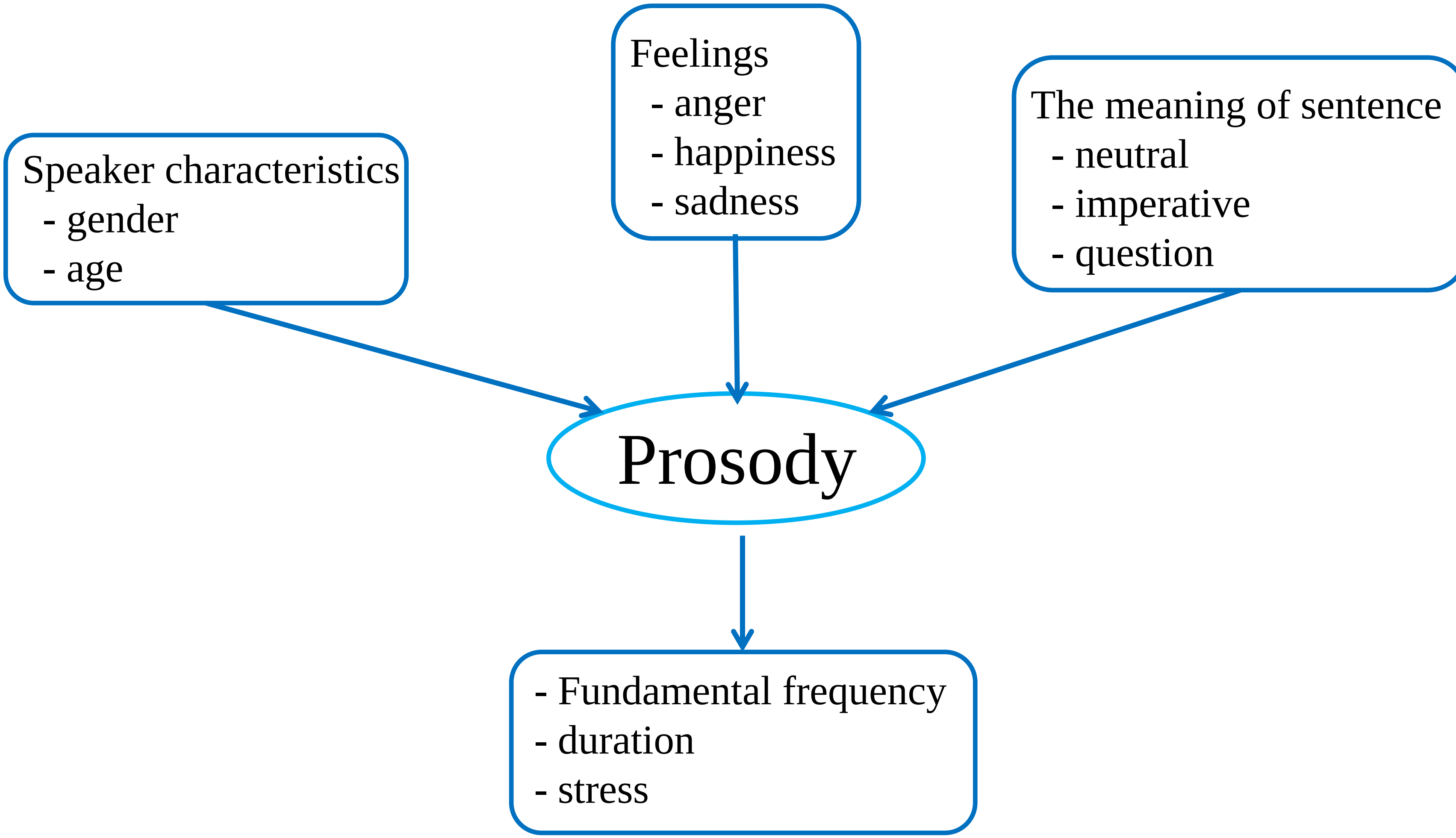


Figure 2: Simple TTS synthesis.

Figure 3: Stephen Hawkings with the TTS system [1].

[1] https://www.immortal.org/35333/stephen-hawkings-phd-thesis-now-available-free-online/

SS contd.


Speaker characteristics
- gender
- age
Feelings
- anger
- happiness
- sadness
The meaning of sentence
- neutral
- imperative
- question
Prosody
- Fundamental frequency
- duration
- stress

Application of SS

- General application
  - Reading and communication aid for visually challenged..
  - Deaf and vocally handicapped.

- Educational application
  - Spelling and language pronunciation
  - Telephone enquiry system.
  - Voice XML: Internet surfing using voice.

34

Voice Conversion (VC)

- Transform the speech of a (source) speaker so that it sound- like the speech of a different (target) speaker.

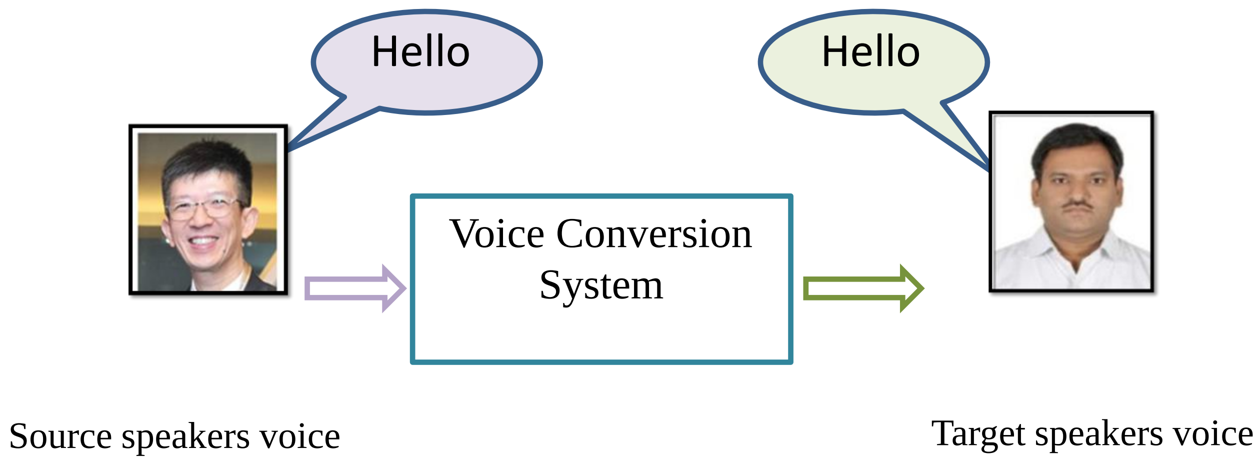


Figure 4: Schematic diagram of one-to-one voice conversion.

Application of VC

- Hiding identity of speaker

- Vocal pathology

- Voice restoration
- Speech-to-speech translation

- Dubbing of programs

Countermeasures

Spoofing technique	Accessibility (practicality)	Effectiveness (risk)		Countermeasure availability
		Text-independent	Text-dependent	
Impersonation	Low	Low	Low	Non-existent
Replay	High	High	Low to High	Low
Speech Synthesis	Medium to High	High	High	Medium
Voice Conversion	Medium to High	High	High	Medium

Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," Speech Communication, vol. 66, pp. 130–153, 2015

Mimic Resistance

- Referred to as human-mimicking, by altering their voices.
- Examples as : Twins, professional mimicry artists
- Challenging attack.
- No standard database available yet for both twins and mimics

D, Gomathi, Sathya Adithya Thati, Karthik Venkat Sridaran and Yegnanarayana B. “Analysis of Mimicry Speech.” INTERSPEECH (2012).

Mimic Resistance (contd.)

Mimic Resistance-> Physiological characteristics -> Identical twins

(a) Twins in childhood

(b) At the age of 28 years

Patil, H. A., & Basu, T. K. (2004, December). Detection of bilingual twins by Teager energy based features. In *Signal Processing and Communications, 2004. SPCOM'04. 2004 International Conference on* (pp. 32-36). IEEE.

Hemant A. Patil, *Speaker Recognition in Indian Languages: A Feature Based Approach*. Ph.D. Thesis, Department of Electrical Engineering, IIT Kharagpur, India, July 2005.

Mary, Leena, Anish Babu K. K, Aju Joseph and Gibin M. George. “Evaluation of mimicked speech using prosodic features.” 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (2013): 7189-7193.

Mimic Resistance (contd.)

Spectrographic Analysis:

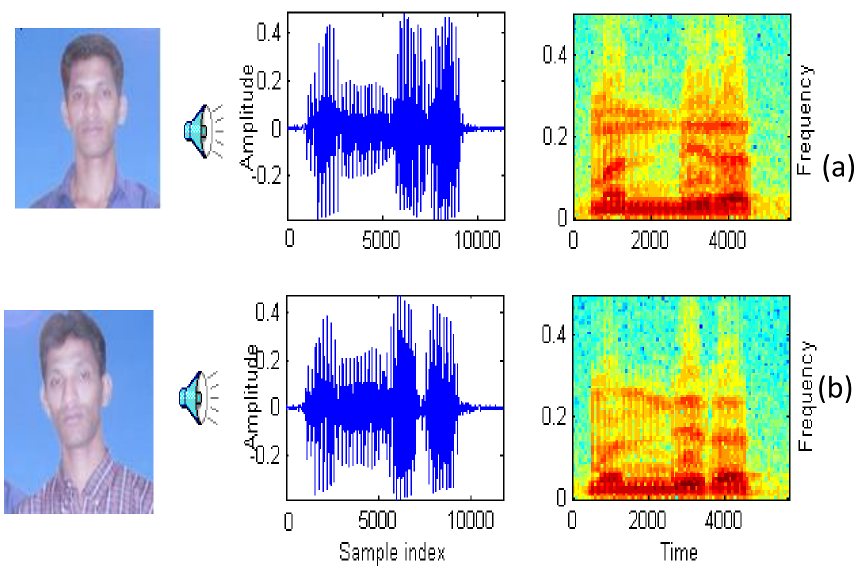


Fig ure 10. Speech signal and its spectrogram corresponding to the Marathi word , "Mandirat (in the temple)" spoken by identical twins: (a) Mr. Nilesh Mangaonkar, and (b) Mr. Shailesh Mangaonkar.

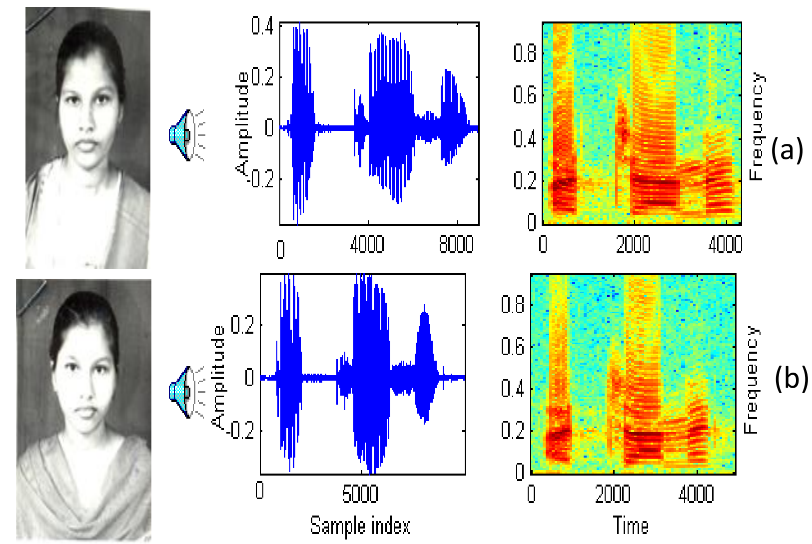


Figure 11. Speech signal and its spectrogram corresponding to the Hindi word, "Achanak (Suddenly)" spoken by identical twins: (a) Miss. Aarti Kalamkar, and (b) Miss. Jyoti Kalamkar.

Hemant A. Patil, *Speaker Recognition in Indian Languages: A Feature Based Approach*. Ph.D. Thesis, Department of Electrical Engineering, IIT Kharagpur, India, July 2005.

Results on Twins

Table 1. Success Rates (%) for 2^{nd} Order Polynomial Approximation with 60 s Training Speech

TEST (SEC)	VT-MFCC (DI=2)	T-MFCC (DI=1)	MFCC	LPCC	LPC
1	76.47 (88.23)	52.94 (70.58)	73.52 (88.23)	64.70 (79.41)	67.64 (79.41)
3	76.47 (88.23	70.58 (85.29)	76.47 (88.23)	67.64 (82.35)	67.64 (82.35)
5	79.41 (91.17)	70.58 (85.29)	79.41 (94.11)	64.70 (85.29)	67.64 (88.23)
7	85.29 (94.11)	73.52 (85.29)	79.41 (94.11)	70.58 (85.29)	70.58 (88.23)
10	85.29 (94.11)	76.47 (85.29)	76.47 (94.11)	82.35 (94.11)	70.58 (88.23)
12	82.35 (91.17)	79.41 (85.29)	79.41 (94.11)	79.41 (94.11)	73.52 (91.17)
15	82.35 (91.17)	73.52 (85.29)	79.41 (94.11)	79.41 (94.11)	73.52 (88.23)
Order 2	**81.09** (91.17)	71.00 (83.19)	77.73 (91.59)	72.68 (87.81)	70.16 (86.13)
Order *3*	84.87 (**95.37**)	86.13 (92.85)	**88.23** (**97.47**)	82.35 (93.27)	78.99 (93.69)

Hemant A. Patil and Keshab K. Parhi, "Variable length Teager energy based Mel cepstral features for identification of twins," *in* S. Chaoudhury *et. al.* (Eds.) *LNCS*, vol. 5909, pp. 525-530, 2009.

Fingerprint

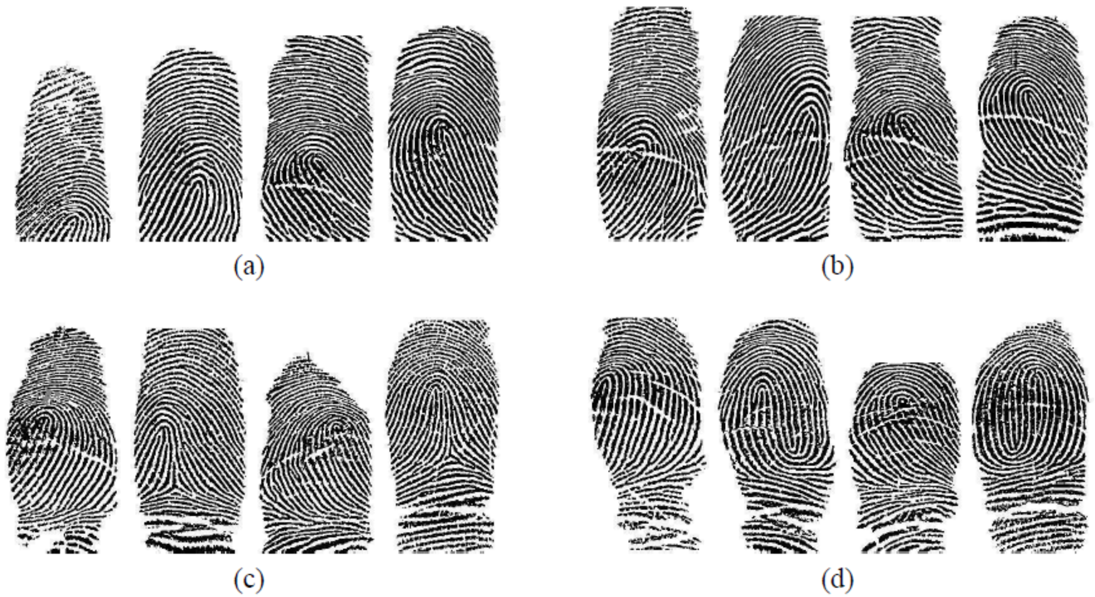
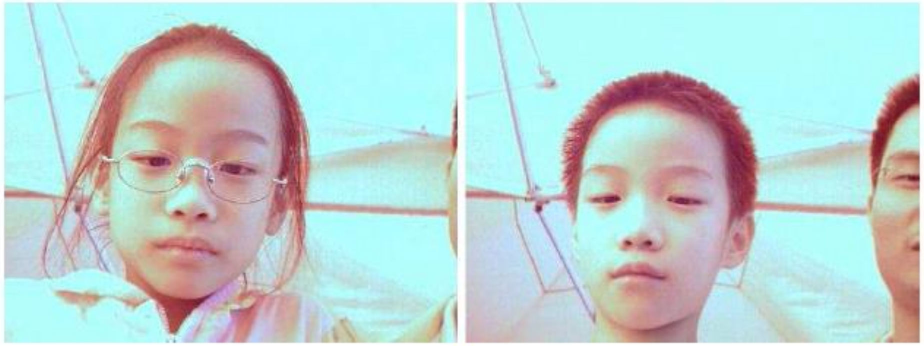


Figure 2.4: Fingerprint images of fingers 1, 2, 3, and 4 of the first twin (a), and the four images of the corresponding fingers of the second twin in an identical twin pair (b); similarly, (c) and (d) show fingerprint images of a non-identical twin pair. Note the similarity in ridge flow pattern between identical twins. All four corresponding fingers of identical twins in (a) and (b) have the same pattern type. But for non-identical twins in (c) and (d), only two corresponding fingers (no. 1 and 3) have the same pattern type.

Alessandra Aparecida Paulino , "CONTRIBUTIONS TO BIOMETRIC RECOGNITION: MATCHING IDENTICAL TWINS AND LATENT FINGERPRINTS," PhD. Thesis, Michigan State University, 2013.
Paone, Jeffrey R., et al. "Double trouble: Differentiating identical twins by face recognition." *IEEE Transactions on Information forensics and Security* 9.2 (2014): 285-295.

Twins

(a)



(b)

Figure 2.5: Face images of the first and second twin in (a) an identical twin pair, and (b) a non-identical twin pair.

Alessandra Aparecida Paulino , "CONTRIBUTIONS TO BIOMETRIC RECOGNITION: MATCHING IDENTICAL TWINS AND LATENT FINGERPRINTS," PhD. Thesis, Michigan State University, 2013.

Iris

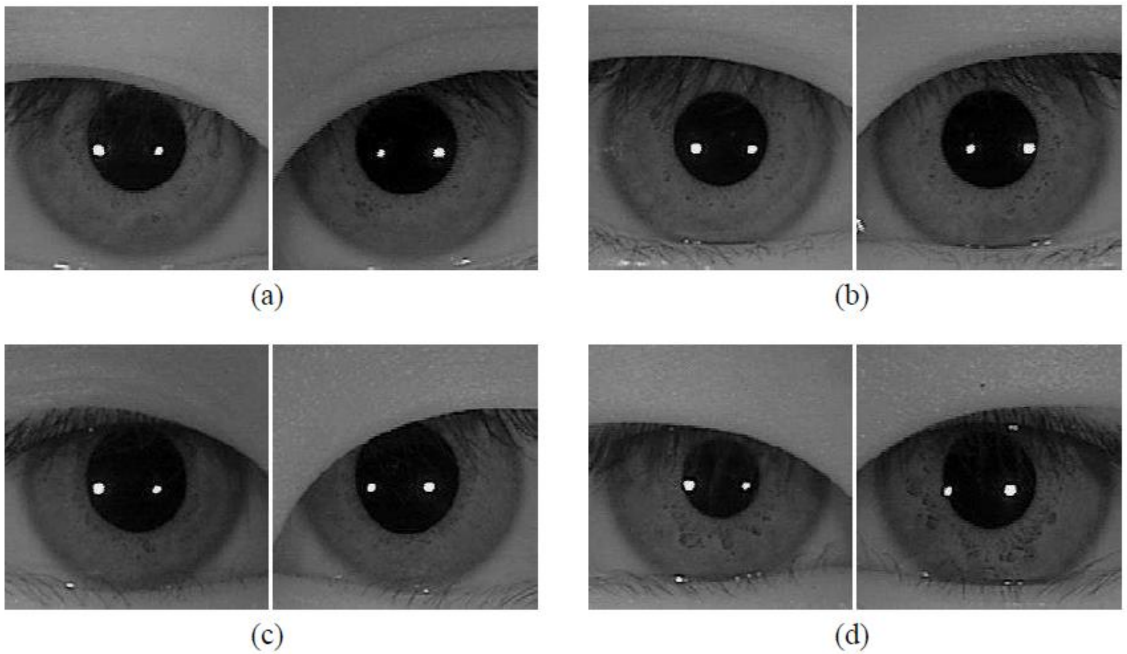


Figure 2.6: The left and right iris images of identical ((a) and (b)) and non-identical twin pairs ((c) and (d)).

Alessandra Aparecida Paulino , “CONTRIBUTIONS TO BIOMETRIC RECOGNITION: MATCHING IDENTICAL TWINS AND LATENT FINGERPRINTS,” PhD. Thesis, Michigan State University, 2013.

Literature on Twins

Table 1: Summary of studies on the biometrics of identical twins. Sets can include identical twin pairs as well as non-identical twin pairs

STUDY	YEAR	BIOMETRIC TRAIT	DATABASE SIZE
Daugman and Downing [62]	2001	Iris	1 set
Jain *et al.* [58]	2002	Fingerprints	94 sets
Kodate *et al.* [60]	2002	Face	10 sets
Han *et al.* [63]	2004	Fingerprints	66 sets
Patil and Basu [64]	2004	Voice	12 sets
Bronstein *et al.* [65]	2005	Face (3D)	1 set
Kong *et al.* [59]	2006	Palmprints	53 sets
Srihari *et al.* [66]	2008	Fingerprints	298 sets of twins
Ariyaeeinia *et al.* [61]	2008	Speech	49 sets
Sun *et al.* [67]	2010	Face, Fingerprints, Iris	51 sets
Hollingsworth *et al.* [68]	2011	Iris	76 sets
Phillips *et al.* [7]	2011	Face	126 sets
Pruitt *et al.* [69]	2011	Face	126 sets
Biswas *et al.* [70]	2011	Face	186 subjects[1]
Klare *et al.* [71]	2011	Face	126 sets
Vijayan *et al.* [72]	2011	Face (3D)	107 sets
Srinivas *et al.* [73]	2012	Face	126 sets
Tao *et al.* [74]	2012	Fingerprints	83 sets

Alessandra Aparecida Paulino , "CONTRIBUTIONS TO BIOMETRIC RECOGNITION: MATCHING IDENTICAL TWINS AND LATENT FINGERPRINTS," PhD. Thesis, Michigan State University, 2013.
Rosenberg, Aaron E. "Automatic speaker verification: A review." *Proceedings of the IEEE* Vol. 64. no.4 (1976): 475-487

Professional Mimics

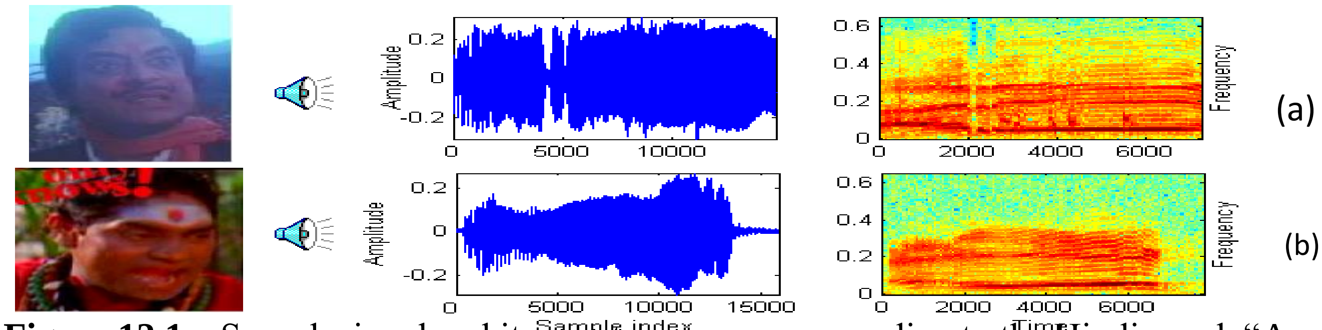


Figure 12.1. Speech signal and its spectrogram corresponding to the Hindi word, “Arrye” spoken by (a) target speaker viz. Mr. Jagdip and (b) professional mimic.

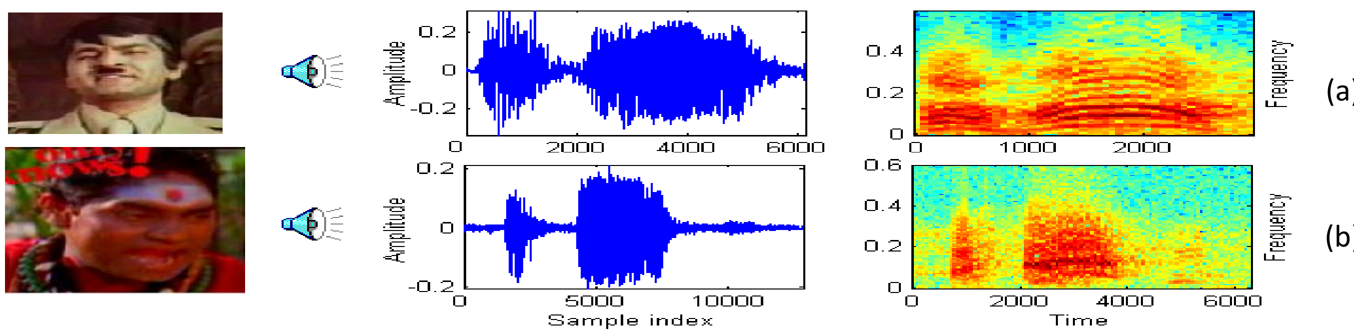


Figure 12.2. Speech signal and its spectrogram corresponding to the Hindi word, “Aahahha” spoken by (a) target speaker viz. Mr. Asrani and (b) professional mimic.

Patil, Hemant A., and Tapan Kumar Basu. "LP spectra vs. Mel spectra for identification of professional mimics in Indian languages." *International Journal of Speech Technology* 11, no. 1 (2008): 1-16.

Mimics (contd.)

Table 2: Results on Real Experiments.

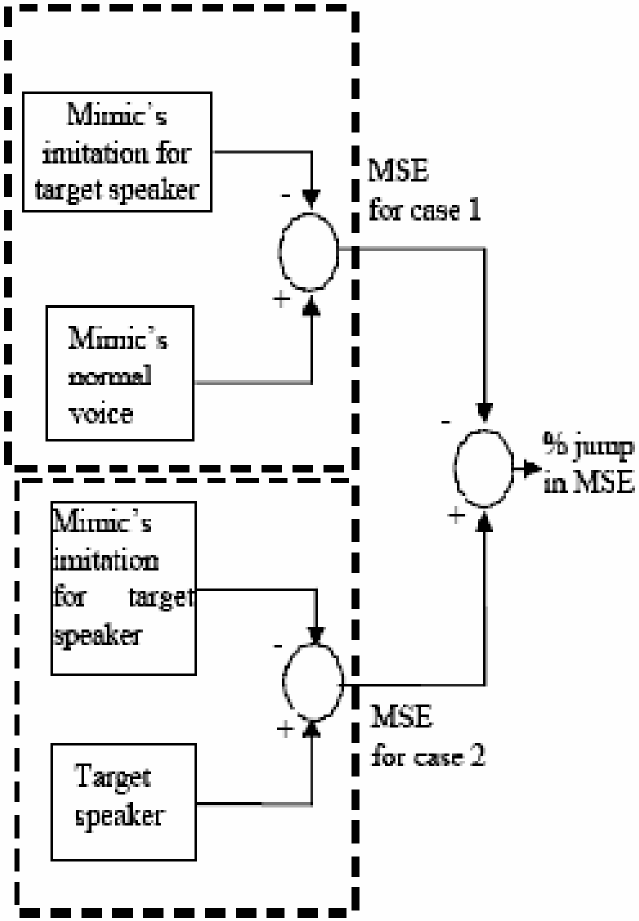
Average success rates (%) for real experiment with *2nd* order approximation (Hindi Mimic)		
TR / Feature	30s	60s
LPC	**98.09**	**99.04**
LPCC	**100**	**99.04**
MFCC	99.04	99.04
TMFCC	94.28	97.14

Table 3: Results on Fictitious Experiments

Average success rates (%) for *2nd* order approximation (Marathi Mimic)				
TR / Feature	30s	60s	90s	120s
LPC	**57.14**	**58.43**	**59.08**	**61.03**
LPCC	**62.98**	**64.28**	**66.23**	**65.58**
MFCC	50.64	49.34	49.34	50.66
TMFCC	27.26	26.61	27.26	27.91

Hemant A. Patil, P. K. Dutta and T. K. Basu, "Effectiveness of LP based features for identification of professional ,mimics in Indian languages", *in Int. Workshop on Multimodal User Authentication*, *MMUA'06*, Toulouse, France, May 11-12, 2006.

Analysis of results through MSE.



$$MSE(n) = \frac{1}{N_D}\sum_{i=1}^{N_D}\left|\mathbf{x}_{tr_i}^{n} - \mathbf{x}_{te_i}^{n}\right|^2$$

where

$MSE(n)$ =Mean Square Error for nth frame.

$\mathbf{x}_{tr_i}^{n}$ =ith feature value in nth training feature vector for normal voice of the professional mimic (case 1) or normal voice of the target speaker (case 2).

$\mathbf{x}_{te_i}^{n}$ = ith feature value in nth testing feature vector for normal voice of mimic's imitations for the target speaker.

N_D =dimension of the feature vector.

Figure 13. Schematic for calculation of % jump in MSE.

Hemant A. Patil, P. K. Dutta and T. K. Basu, "Effectiveness of LP based features for identification of professional ,mimics in Indian languages", *in Int. Workshop on Multimodal User Authentication, MMUA'06*, Toulouse, France, May 11-12, 2006.

Mimic ID (contd.)

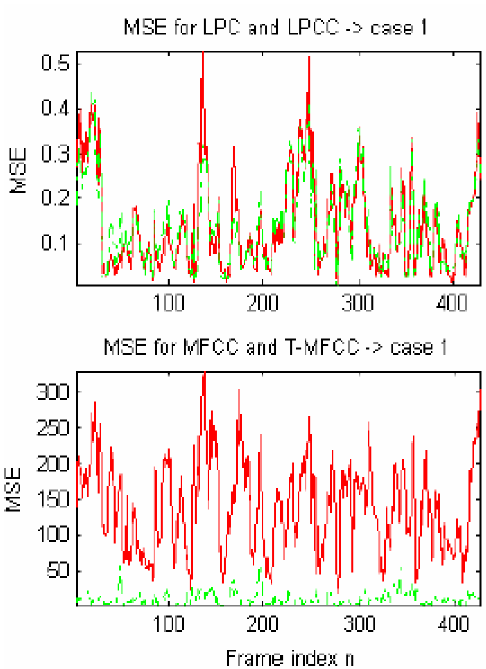


Figure 14. MSE for case 1

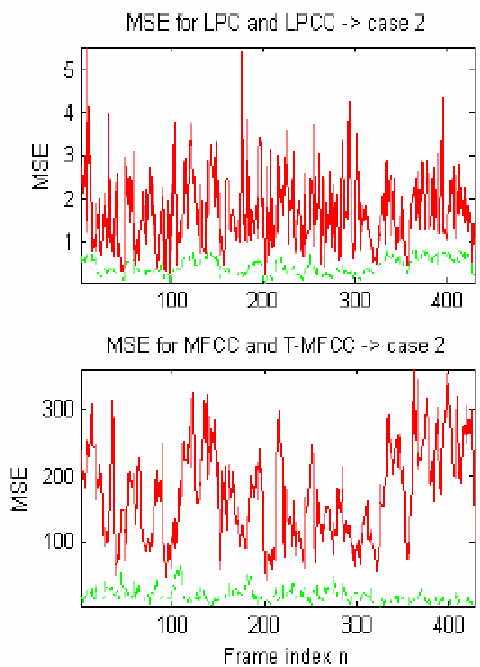


Figure 15: MSE for case 2

TABLE VIII
ANALYSIS OF RESULTS SHOWN IN TABLES 1-2 THEROGH OVERALL (OVER 429 FRAMES) MSE

FS / Av. MSE	LPC	LPCC	MFCC	T-MFCC
Case1	0.1433	0.1405	140.05	11.03
Case 2	1.7161	0.4211	172.05	19.28
% jump	**91.65**	**66.62**	18.16	42.79

Hemant A. Patil, P. K. Dutta and T. K. Basu, “Effectiveness of LP based features for identification of professional ,mimics in Indian languages”, *in Int. Workshop on Multimodal User Authentication, MMUA’06,* Toulouse, France, May 11-12, 2006.

MFCC

- Mel-Frequency Cepstral Coefficients (MFCC)

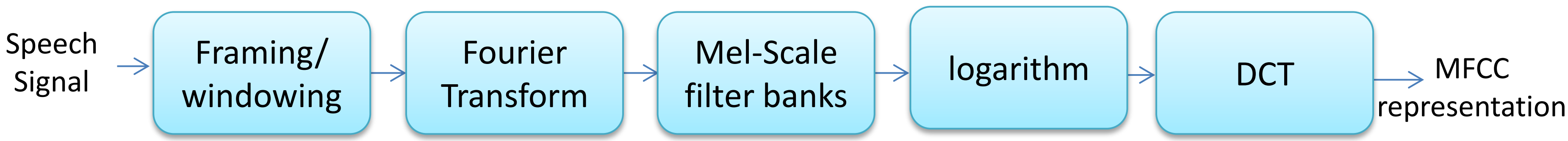


Figure 16: Schematic diagram of the MFCC feature extraction process After [1].

  - State-of-the-art features for speech processing applications.
  - *10-30* ms window
  - 28 (may vary) triangular filter banks
  - *12* static coefficients, *12* delta and *12* delta-delta

[1] S.B. Davis, and P. Mermelstein (1980), "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), pp. 357–366.

Cochlear Filter Cepstral Coefficients (CFCC)

- The CFCC feature extraction requires the following
  - Auditory Transform (AT) of speech
  - Motion of the Basilar Membrane (BM)
  - Nerve-spike density estimation
  - Loudness functions

[1] Q. Li, "An auditory-based transform for audio signal processing," in *IEEE Workshop on Applications of Sign. Process. to Audio and Acous* , New Paltz, NY, 2009.
[2] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 19, no. 6, pp. 1791-1801, 2011.

Cochlear Filters Response

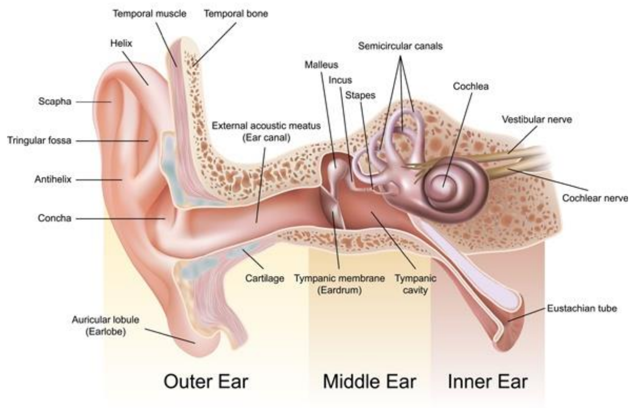


Figure 17: Anatomy of the ear [1].

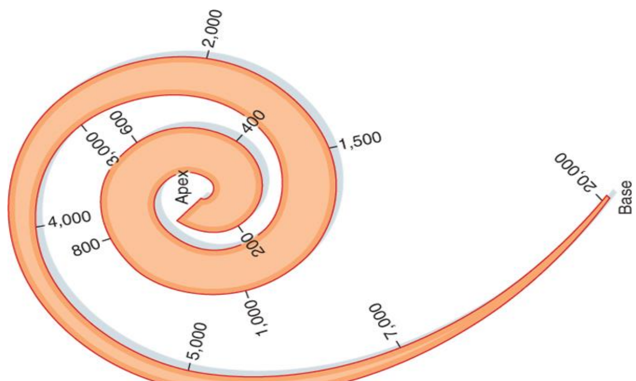


Figure 18: Cochlea's range of sensitivity of frequencies. (20 Hz – 20 kHz) [2].

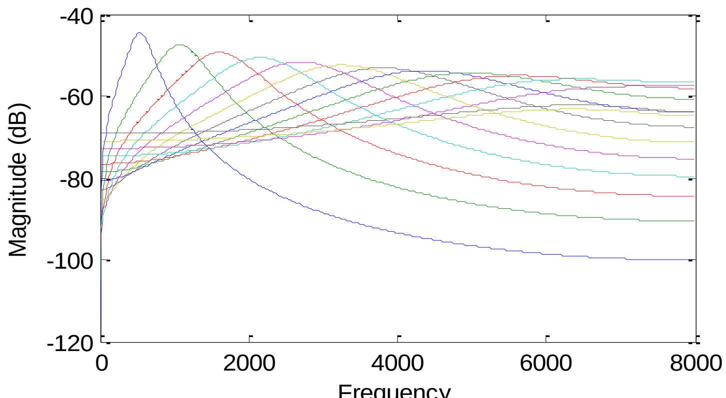


Figure 19: Responses of *28* cochlear filters on a linear scale with α=3 and $\beta=0.35$.

[1] [Available Online]: http://www.audiologyspecialists.com/anatomy-of-the-ear/
[2] [Available Online]: https://introtohearingscience.wordpress.com/.

CFCC (contd.)

- Auditory Transform (AT)
  - Speech signal $s(t)$ and cochlear filter impulse response $\psi(t)$.
  - The auditory transform of speech is given by [1]-[2]:

$$W(a,b) = s(t) * \psi_{a,b}(t),$$

where

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right),$$

$$= \frac{1}{\sqrt{a}}\left(\frac{t-b}{a}\right)^{\alpha} \exp\left[-2\pi f_L \beta\left(\frac{t-b}{a}\right)\right] \times \cos\left[2\pi f_L\left(\frac{t-b}{a}\right)+\theta\right] u(t-b).$$

→ factor a is the scale or dilation parameter

→ factor b is the time shift or translation parameter

→ parameters α and β determine the *shape* and *width* of the cochlear filter .

[1] Q. Li, "An auditory-based transform for audio signal processing," in *IEEE Workshop on Applications of Sign. Process. to Audio and Acous.*, New Paltz, NY, 2009.

[2] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 19, no. 6, pp. 1791-1801, 2011.

CFCC (contd.)

- Motion of the Basilar Membrane (BM) $h(a,b) = (W(a,b))^2; \quad \forall W(a,b)$
- Nerve spike density estimation $S(i,j) = \frac{1}{d} \sum_{b=l}^{l+d-1} h(i,b), \quad l = 1, L, 2L, \ldots; \ \forall i, j$
  - where d is the window length, and L is the window shift duration.
- Loudness functions
  - Scales of loudness functions as cubic root nonlinearity or
  - Logarithmic

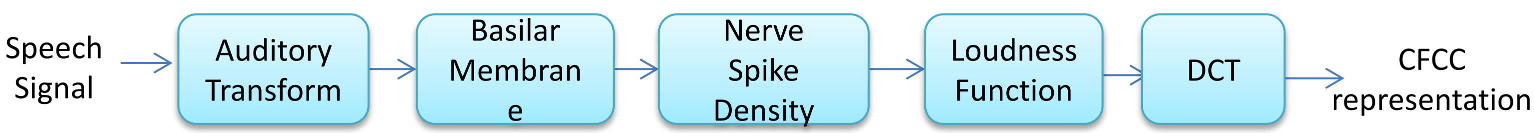


Figure 20: Schematic diagram of the auditory-based feature extraction algorithm named CFCC After [1].

[1] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 19, no. 6, pp. 1791-1801, 2011.

Proposed CFCC+IF features

ASV Spoof 2015 Challenge Winner System

- Instantaneous Frequency (IF)
  - Derivative of the unwrapped phase of the analytic signal derived from $s(t)$.
  - Apply IF to each subband signal framewise.

$$SIF(i,j) = \frac{1}{d}\sum_{b=l}^{l+d-1} IF(h(i,b)), \qquad l = 1, L, 2L, \ldots; \ \forall i, j$$

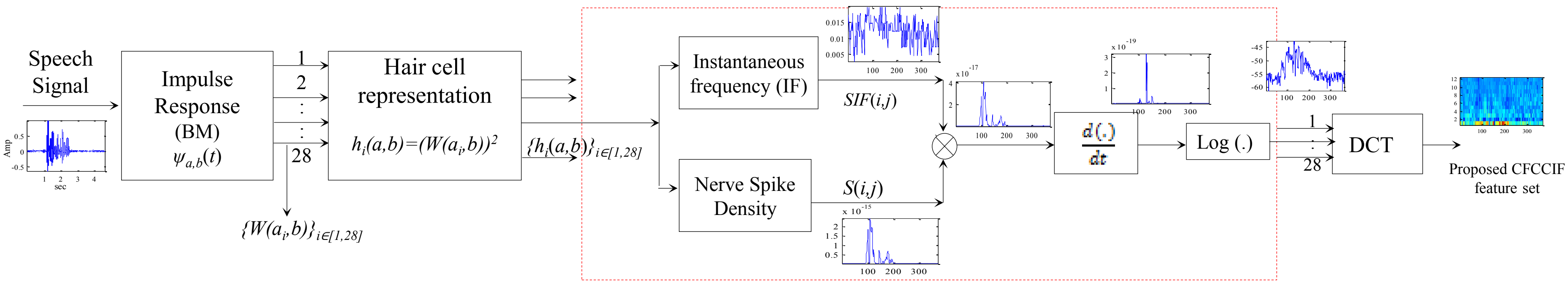


Figure 21: Block diagram for proposed CFCCIF feature extraction scheme After [1].

[1] T. B. Patel and H. A. Patil, "Significance of source-filter interaction for classification of natural vs. spoofed speech," IEEE Jour. on Selected Topics in Sig.Process. (JSTSP), vol. 11, no. 4, pp. 644 - 659, June 2017.
[2] Tanvina B. Patel and Hemant A. Patil, "Combining Evidences from Mel Cepstral, Cochlear Filter Cepstral and Instantaneous Frequency Features for Detection of Natural *vs.* Spoofed Speech," in *INTERSPEECH'15*, Dresden, Germany, September 6-10, 2015

Effect of CFCCIF Features

- Figure shows a speech signal (natural speech) the energy at outputs of the cochlear filterbanks.
  - CFCC alone
  - And by using IF features, i.e., CFCCIF

Observations

- CFCCIF enhances information representation.

(shown by dotted regions)

- Especially at higher frequencies

(which are known to be speaker-specific)

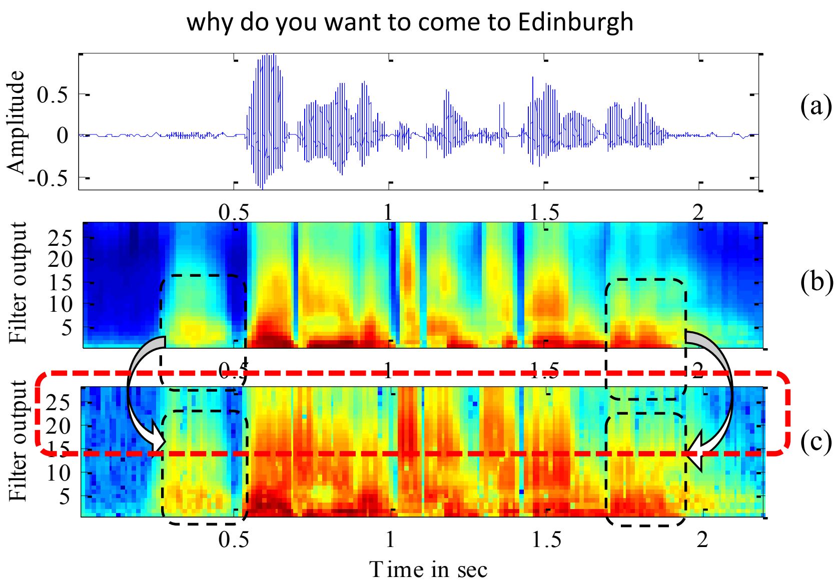


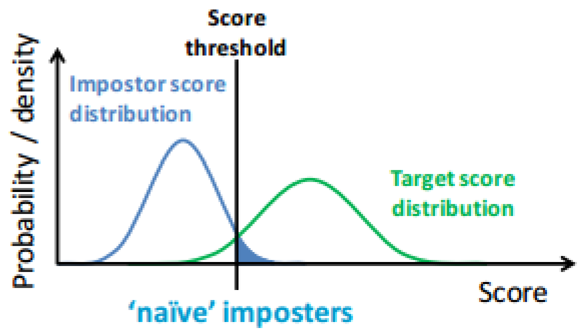
Figure 22: (a) Natural utterance (b) CFCC of *28* cochlear subband filters, and (c) CFCCIF of *28* cochlear subband filters [1].

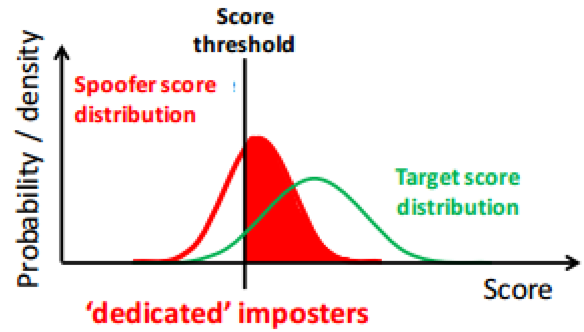
[1] Tanvina B. Patel and Hemant A. Patil, “Combining Evidences from Mel Cepstral, Cochlear Filter Cepstral and Instantaneous Frequency Features for Detection of Natural vs. Spoofed Speech,” in the *16th Annual Conference of International Speech Communication Association (ISCA), INTERSPEECH'15*, Dresden, Germany, September 6-10, 2015

Performance Measures

Table 4: Performance measures while spoofing ASV systems.

Trial	Decision	
	Accept	Reject
Target	Correct accept	False reject
Imposter	False alarm	Correct reject





A. Martin, G. Doddington, T. Kamm and M. Ordowski, "The DET curve in assessment of detection task performance," in *Proc. Eur. Conf. Speech Comm. Technol. (EUROSPEECH '97)*, Rhodes, Greece, pp. 1895-1898, 1997.

Adapted from: Spoofing and anti-spoofing a shared view of speaker verification, speech synthesis and voice conversio n APSIPA ASC tutoria 16th Dec. 2015

Performance Measures

- Equal Error Rate (EER)
  - Spoofed detected as natural (False Accept: FA)
  - Natural detected as spoofed (False Reject/Miss: FR)

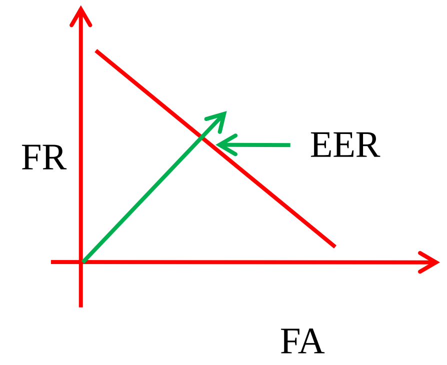


Table 5: Performance measures while spoofing ASV systems.

Actual\Detected	Natural	Spoofed
Natural	Correct	False Reject/ Miss Rate (FRR)
Spoofed	False Acceptance Rate (FAR)	Correct

- In spoofed attack: Minimize FAR -> avoids spoofed speech being detected as natural speech
- Detection Error Tradeoff (DET) Curve
  - % EER→ False acceptance rate = miss rate → FAR=FRR

A. Martin, G. Doddington, T. Kamm and M. Ordowski, "The DET curve in assessment of detection task performance," in *Proc. Eur. Conf. Speech Comm. Technol. (EUROSPEECH '97)*, Rhodes, Greece, pp. 1895-1898, 1997.

Results on Development Set

- Fusion of scores

$$LLk_{combine} = (1-\alpha_f)LLk_{MFCC} + \alpha_f LLk_{feature2}$$

Table 6: The score-level fusion % EER obtained on development set for D1, D2, and D3-dimensional feature vector [1].

Features with score-level fusion	**Dimension (*D*) of feature vector**	**EER (%) for varying values of α_f**										
		0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
MFCC+CFCC	D1: *12*-static	3.26	2.86	2.66	2.52	**2.43**	2.57	2.72	3.03	3.55	3.97	4.55
MFCC+(CFCCIF)		3.26	2.72	2.40	2.03	1.77	1.60	**1.52**	1.57	1.72	1.92	2.29
MFCC+CFCC	D2: *12*-static +*12* delta	2.17	1.83	1.54	1.40	**1.32**	1.32	1.46	1.63	1.89	2.23	2.60
MFCC+(CFCCIF)		2.17	1.83	1.46	1.23	1.03	0.97	**0.89**	0.89	0.97	1.14	1.40
MFCC+CFCC	D3: *12*-static +*12* delta + *12* (delta-delta)	1.60	1.32	1.14	0.97	**0.89**	0.89	0.92	1.00	1.17	1.34	1.54
MFCC+(CFCCIF)		1.60	1.37	1.14	1.00	0.86	**0.83**	**0.83**	0.92	1.03	1.17	1.52

[1] Tanvina B. Patel and Hemant A. Patil, "Combining Evidences from Mel Cepstral, Cochlear Filter Cepstral and Instantaneous Frequency Features for Detection of Natural *vs*. Spoofed Speech," in the *16th Annual Conference of International Speech Communication Association (ISCA), INTERSPEECH'15*, Dresden, Germany, September 6-10, 2015.

Results on Development Set

- Detection Error Tradeoff (DET) Curves
  - Lowest EER with MFCC+CFCC is $\alpha_f = 0.4$ and lowest EER with MFCC+CFCC is $\alpha_f = 0.6$
- CFCCIF has lower EER and better separation

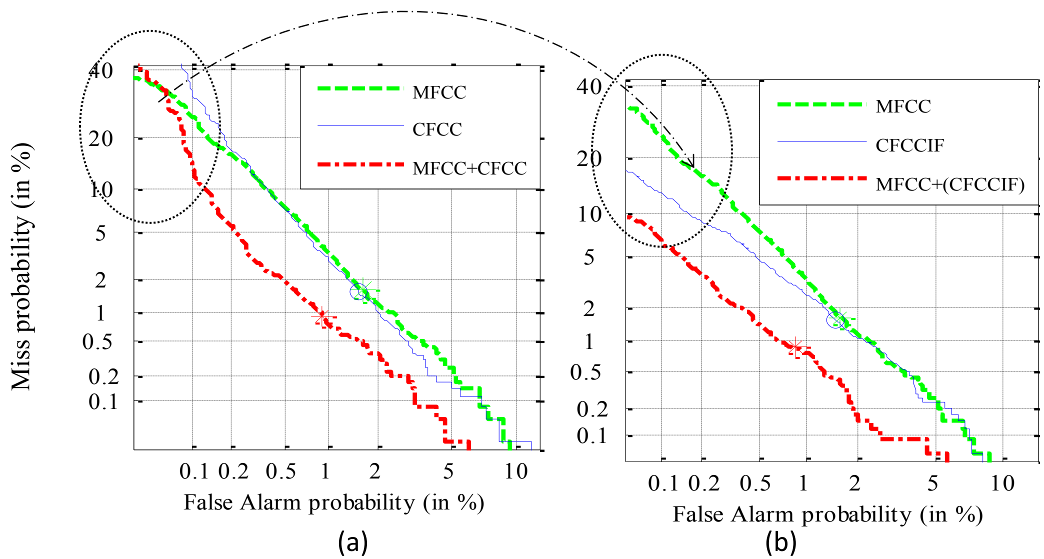
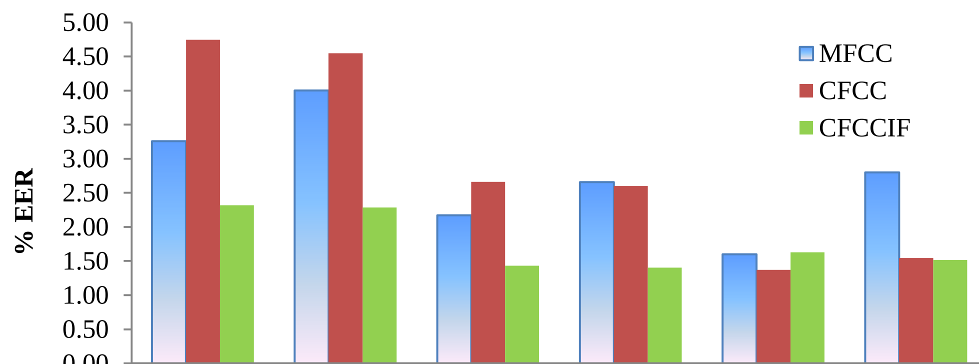


Figure 23: (a) DET curve for MFCC (--green), CFCC (blue), and their score-level fusion with α_f =0.4 (-.-red), (b) DET curve for MFCC (--green), CFCCIF (blue) and their score-level fusion with α_f =0.6 (-.-red) [1].

[1] Tanvina B. Patel and Hemant A. Patil, "Combining Evidences from Mel Cepstral, Cochlear Filter Cepstral and Instantaneous Frequency Features for Detection of Natural *vs*. Spoofed Speech," in the *16th Annual Conference of International Speech Communication Association (ISCA), INTERSPEECH'15*, Dresden, Germany, September 6-10, 2015.

Effect of Pre-emphasis

- *D1*- static features, *D2*- delta features and *D3*- delta-delta features
- The % EER of MFCC increases significantly without pre-emphasis.
- The % EER of CFCC and CFCCIF is almost with or without pre-emphasis.
- Proposed CFCCIF feature set gives less EER alone also.



	D1_P	D1_nP	D2_P	D2_nP	D3_P	D3_nP
MFCC	3.26	4.00	2.17	2.66	1.60	2.80
CFCC	4.75	4.55	2.66	2.60	1.37	1.54
CFCCIF	2.32	2.29	1.43	1.40	1.63	1.52

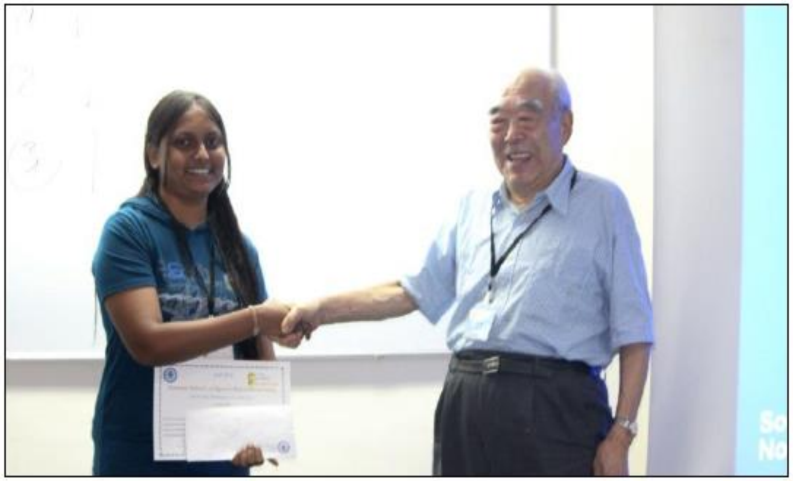
Figure: 24 Effect of pre-emphasis on % EER, using MFCC, CFCC and CFCCIF features (P=pre-emphasis and nP=no pre-emphasis on speech signal) [1].

[1] Tanvina B. Patel and Hemant A. Patil, "Combining Evidences from Mel Cepstral, Cochlear Filter Cepstral and Instantaneous Frequency Features for Detection of Natural vs. Spoofed Speech," in the *16th Annual Conference of International Speech Communication Association (ISCA), INTERSPEECH'15*, Dresden, Germany, September 6-10, 2015.

Results on the Evaluation Set

- **Attack-Independent:** Average % EER for all submissions

Sr. No.	Team	Known attacks	Unknown attacks	All attacks
1	A (DA-IICT)	0.408	2.013	1.211
2	B	0.008	3.922	1.965
3	C	0.058	4.998	2.528
4	D	0.003	5.231	2.617
5	E	0.041	5.347	2.694
6	F	0.358	6.078	3.218
7	G	0.405	6.247	3.326
8	H	0.67	6.041	3.355
9	I	0.005	7.447	3.726
10	J	0.025	8.168	4.097
11	K	0.21	8.883	4.547
12	L	0.412	13.026	6.719
13	M	8.528	20.253	14.391
14	N	7.874	21.262	14.568
15	O	17.723	19.929	18.826
16	P	21.206	21.831	21.518
Avg. of 16 submissions		3.337	9.294	6.3155



Dr. Tanvina B. Patel been awarded **ISCA supported** First Prize of Rs. 15,000 /- by Prof. Hiroya Fujisaki during 5 Minute Ph.D. Contest , S4P 2016, DA-IICT Gandhinagar.

[1] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, M. Sahidullah, A. Sizov, "ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge", *in INTERSPEECH* 2015, Dresden, Germany

Source-based Features for Spoofed Speech

- F_0 and *SoE* [1]
- Prediction [2]
- Fujisaki Model [3]

[1] Himanshu Bhavsar, Tanvina B. Patel and Hemant A. Patil, “Novel Nonlinear Prediction Based Features for Spoofed Speech Detection”, in INTERSPEECH 2016, San Francisco, 8-12 Sept. 2016.

[2] Tanvina B. Patel and Hemant A. Patil, "Effectiveness of Fundamental Frequency (F0) and Strength of Excitation (SoE) for Spoofed Speech Detection" in IEEE Int. Conf. Acoust., Speech and Signal Process., (ICASSP’16), Shanghai, China, pp. 5105-5109, 20-25th March, 2016.

[3] Tanvina B. Patel and Hemant A. Patil, "Analysis of Natural and Synthetic Speech using Fujisaki Model" in IEEE Int. Conf. Acoust., Speech and Signal Process., (ICASSP’16), Shanghai, China, pp. 5250-5254, 20-25th March, 2016.

Basis of using F_0 and SoE's

- For generating speech → Humans vary their vocal folds
- During fold movement → the F_0 contour and Strength of Excitation (SoE) varies
- F_0 and SoE from Glottal Flow Waveform (GFW) and speech signal are related

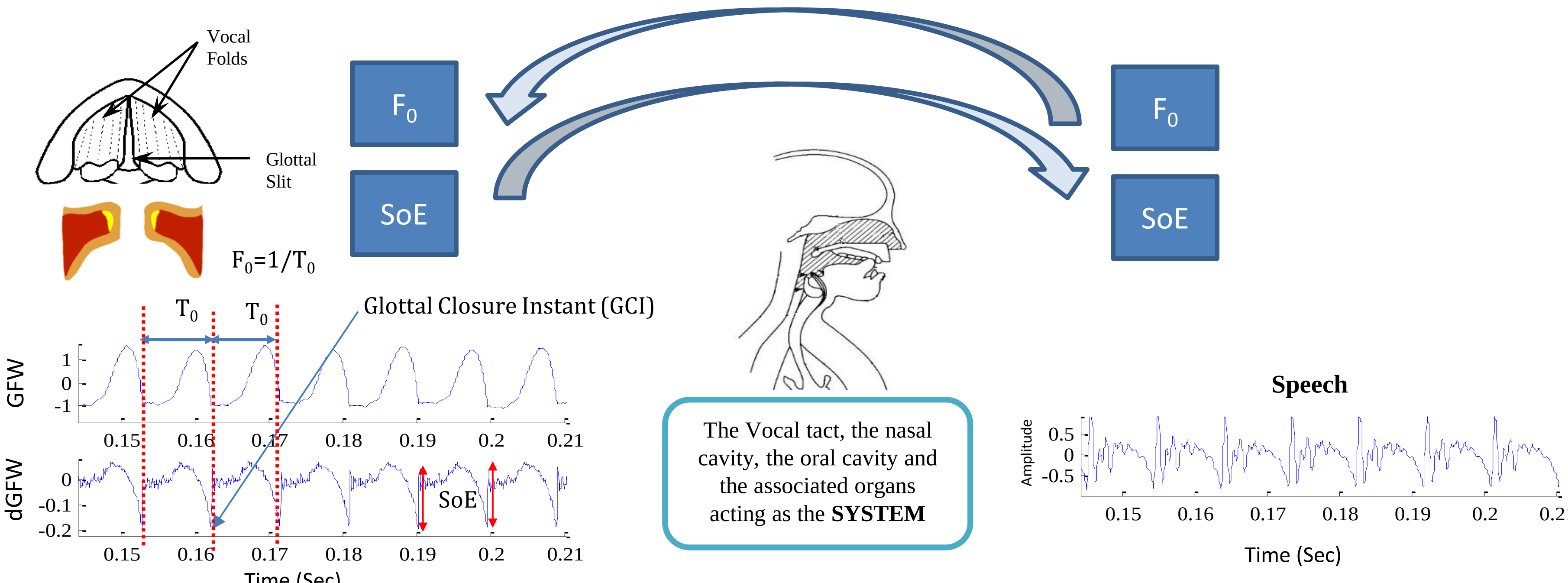


Figure 24: The source-system model of speech production.

Tanvina B. Patel and Hemant A. Patil, "Effectiveness of Fundamental Frequency (F0) and Strength of Excitation (SoE) for Spoofed Speech Detection" in IEEE Int. Conf. Acoust., Speech and Signal Process., (ICASSP'16), Shanghai, China, pp. 5105-5109, 2016.

Basis of using F_0 and *SoE's*

- Spoofed speech→ No actual vocal fold vibration
- F_0 and *SoE* from the estimated GFW and speech signal may or may not be related

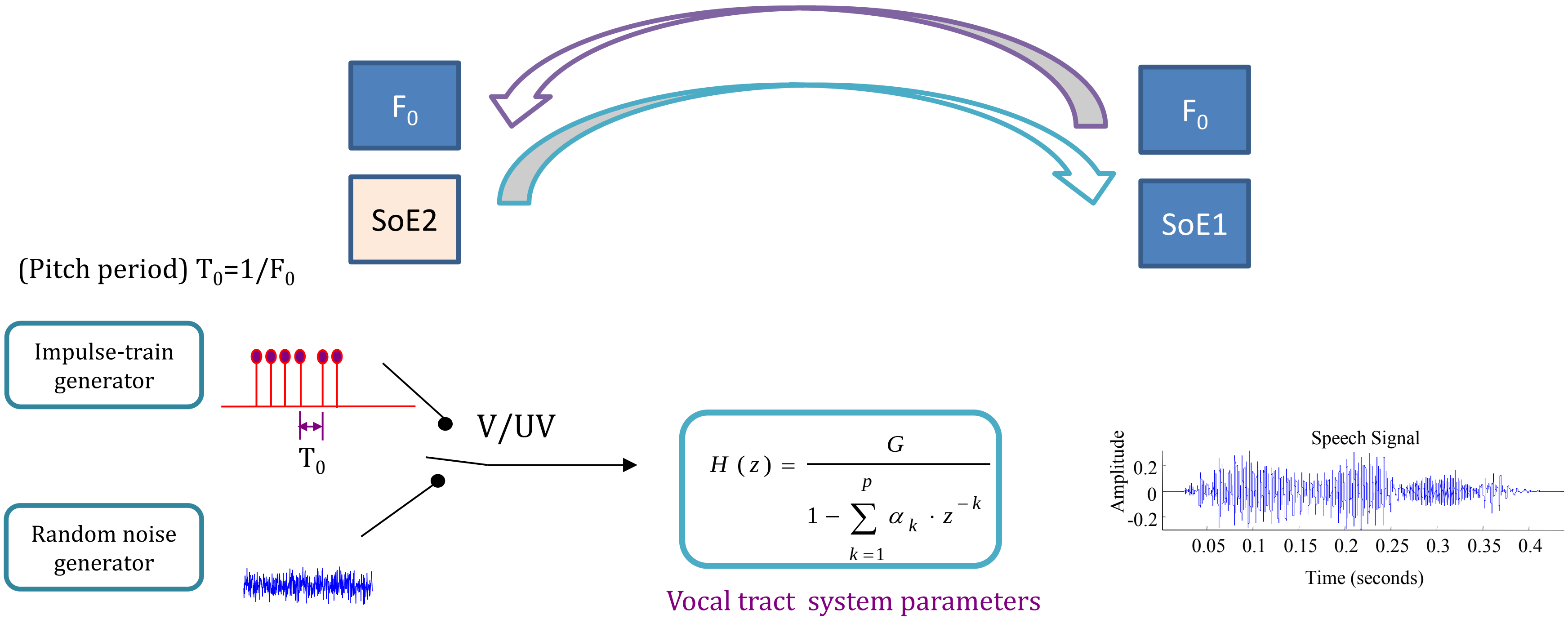


Figure 25: General source-system model of speech production [1].

[1] B. S. Atal, "Automatic recognition of speakers from their voices ," *Proc. of IEEE,* vol. 64, no. 4, pp. 460–475, 1976.
[2] Patel, Tanvina B., and Hemant A. Patil. "Effectiveness of fundamental frequency (F 0) and strength of excitation (SoE) for spoofed speech detection." *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016.

F_0 contour and *SoE1* from Speech

Zero frequency filtering (ZFF) method [1]

- Pass the signal through a resonator

$$H(z) = \frac{b_o}{(1-p_1 z^{-1})(1-p_2 z^{-1})}$$

- $w_r = 0 \rightarrow w_o = 0 \rightarrow$ and $p_2 = p_1^* = r$
- Remove trend from the filtered signal by subtracting the average over *10 ms*

$$y[n] = x[n] - \frac{1}{2N+1}\sum_{n=-N}^{N} x[n+m]$$

- GCI: Negative-to-Positive zero-crossing
- SoE: Slope at GCI

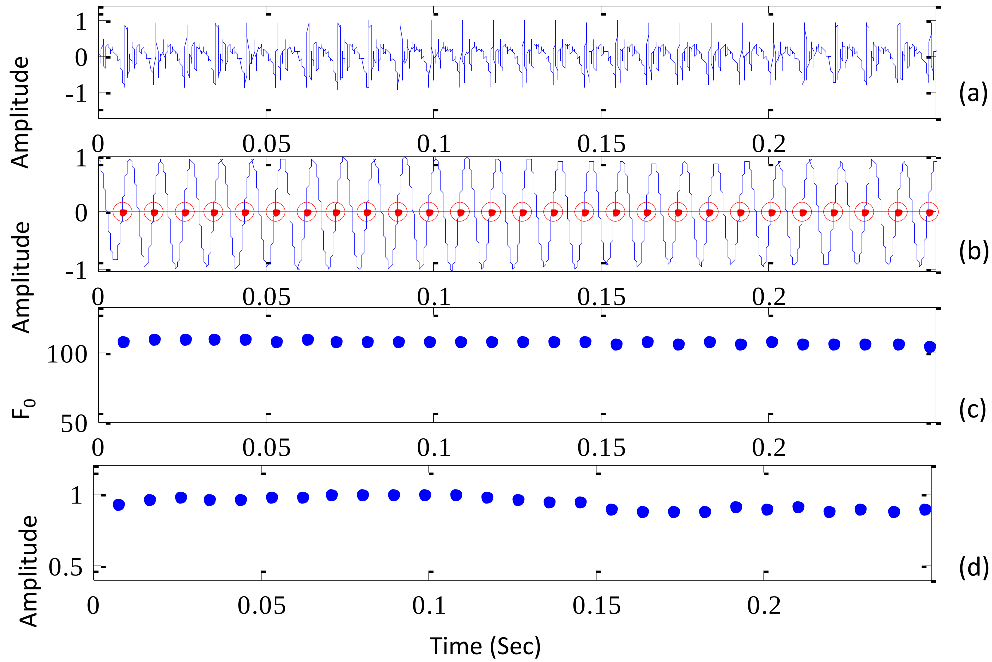


Figure 26: (a) speech segment (b) ZFF signal (c) F_0 contour from GCI locations (negative-to-positive zero-crossings) (d) SoE at GCI (slope at negative-to-positive zero-crossings).

[1] Sri Rama Murty, K. and Yegnanarayana, B., "Epoch extraction from speech signals," *IEEE Trans. on Speech and Audio Process.*, vol. 16, no. 8, pp. 1602-1613, November 2008.
[2] Patel, Tanvina B., and Hemant A. Patil. "Effectiveness of fundamental frequency (F 0) and strength of excitation (SoE) for spoofed speech detection." *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016.

Analysis on Natural *vs.* Spoofed Speech

There exist variations in excitation source

→

Dynamic features

ΔF_0

ΔSoE1

ΔSoE2

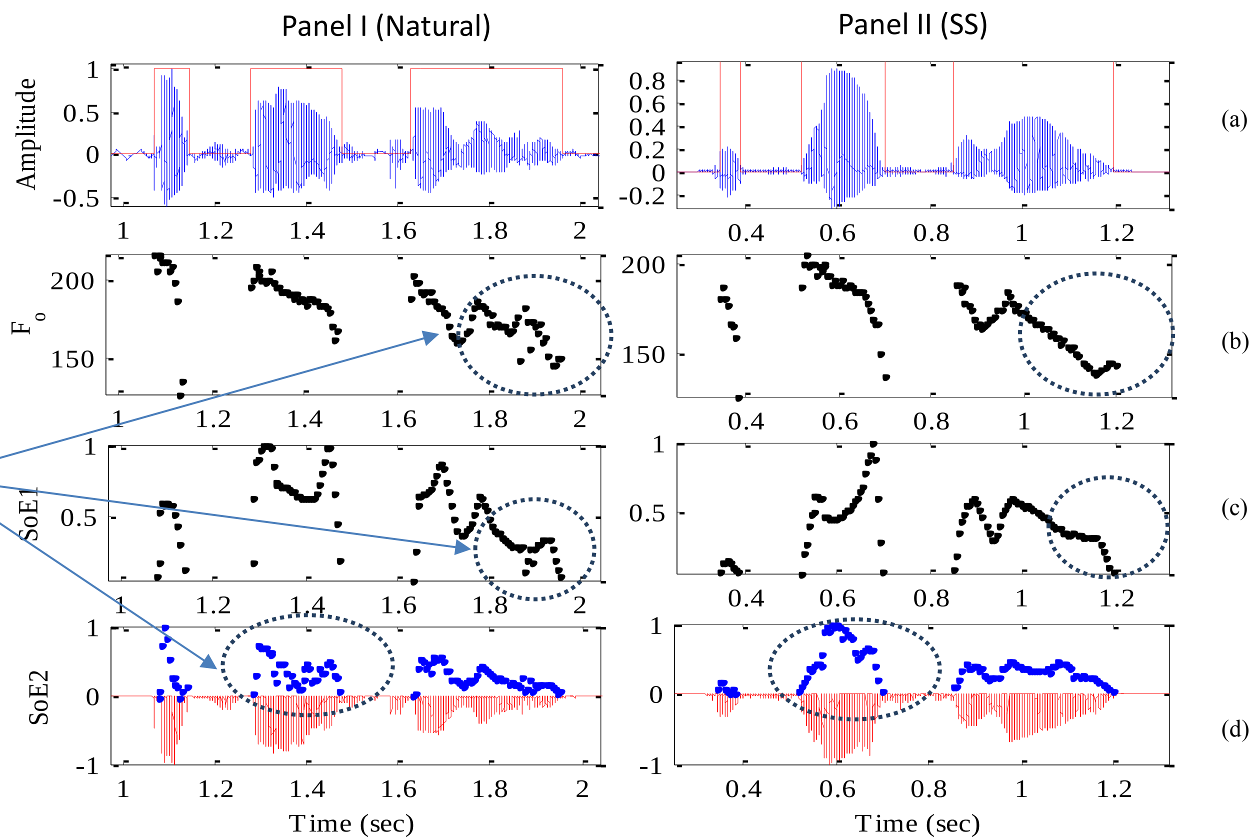


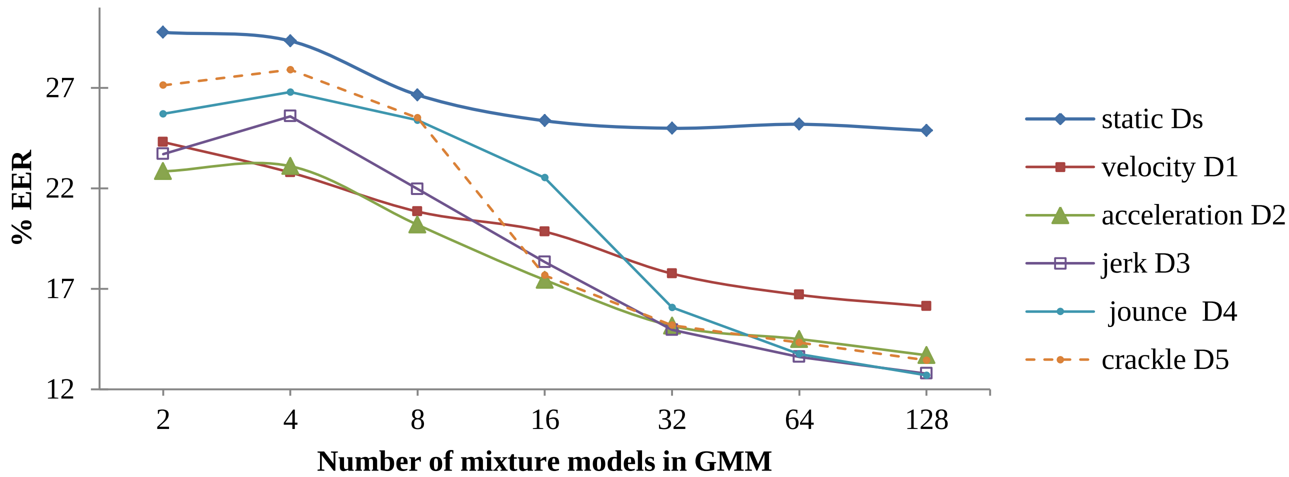
Figure 27: Panel I: Natural speech and Panel II: Spoofed speech (SS) (a) speech signal, (b) F_0 contour (c) normalized SoE1 at GCIs (d) the dGFW (estimated by IAIF) (red) and normalized SoE2 estimated from dGFW at GCI's. (dotted blue) [1]

[1] T. B. Patel and H. A. Patil, "Effectiveness of fundamental frequency (F_0) and strength of excitation (SoE) for spoofed speech detection," in *Proc.* IEEE *Int. Conf. on Acous. Speech and Sig. Process. (ICASSP)*, Shanghai, China, 2016, pp. 5105-5109.

[2] Patel, Tanvina B., and Hemant A. Patil. "Effectiveness of fundamental frequency (F 0) and strength of excitation (SoE) for spoofed speech detection." *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016.

Results on Development Set

Effect of source features and their dynamics [1]



% EER	Dimension
24.8%	3
16.1%	6
13.6%	9
12.7%	12
12.8%	15
13.4%	18

Figure 28 : The % EER obtained on the development set when the static and various dynamics, i.e., velocity, acceleration, jerk, jounce and crackle of static F_0, *SoE1* and *SoE2* are considered.

Observations

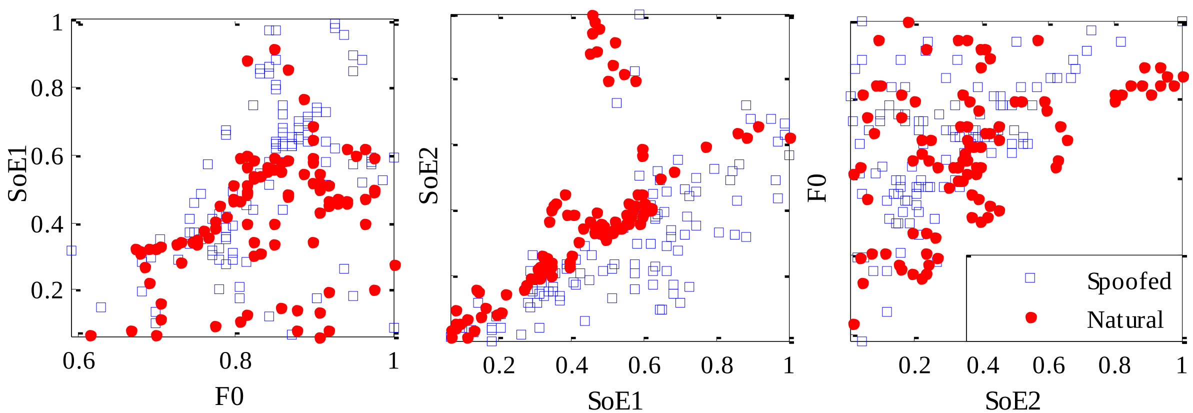
- % EER decreases significantly when dynamic information is added to static features.
- *D3* feature vector with *128* mixtures GMM is considered.

[1] T. B. Patel and H. A. Patil, "Effectiveness of fundamental frequency (F_0) and strength of excitation (SoE) for spoofed speech detection," in *Proc.* IEEE *Int. Conf. on Acous. Speech and Sig. Process. (ICASSP)*, Shanghai, China, 2016, pp. 5105-5109.

[2] Patel, Tanvina B., and Hemant A. Patil. "Effectiveness of fundamental frequency (F 0) and strength of excitation (SoE) for spoofed speech detection." *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016.

Correlation between Source Features

- The correlation coefficients between:
  - *F_0 vs. SoE1*, *SoE1 vs. SoE2* and *SoE2 vs. F_0* are:
  - *0.51*, *0.73* and *0.51* for natural speech and *0.34*, *0.645* and *0.45* for SS speech



(a) (b) (c)

Figure 29: Scatter plots (a) *F_0 vs. SoE1* (b) *SoE1 vs. SoE2* and (c) *SoE2 vs. F_0* for natural and spoofed (SS) speech.

Observation → correlations between features vary for natural and SS speech.

[1] T. B. Patel and H. A. Patil, "Effectiveness of fundamental frequency (F_0) and strength of excitation (SoE) for spoofed speech detection," in *Proc.* IEEE *Int. Conf. on Acous. Speech and Sig. Process. (ICASSP)*, Shanghai, China, 2016, pp. 5105-5109.

[2] Patel, Tanvina B., and Hemant A. Patil. "Effectiveness of fundamental frequency (F 0) and strength of excitation (SoE) for spoofed speech detection." *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016.

Constant Q Cepstral Coefficients

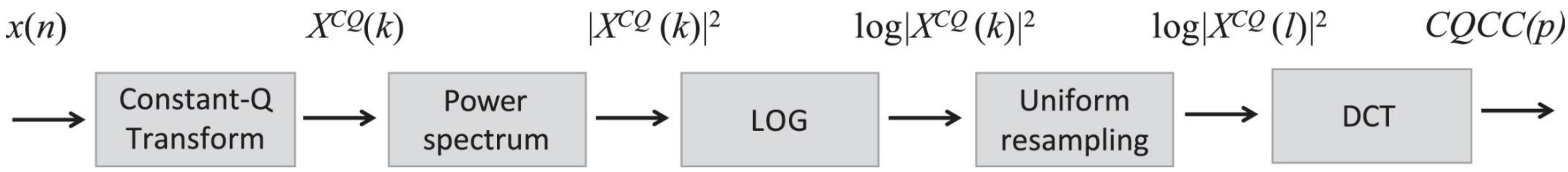


Figure 30: Block diagram of CQCC feature extraction

Constant Q cepstral coefficients (CQCCs) can then be extracted in a more-or-less conventional manner according to:

$$CQCC(p) = \sum_{l=1}^{L} \log \left| X^{CQ}(l) \right|^2 \cos \left[\frac{p \left(l - \frac{1}{2} \right) \pi}{L} \right] \quad (15)$$

where $p = 0, 1, ..., L-1$ and where l are the newly resampled frequency bins. The extraction of CQCCs is summarised in Figure 3. Our Matlab implementation of CQCC extraction can be downloaded from `http://audio.eurecom.fr/content/software`

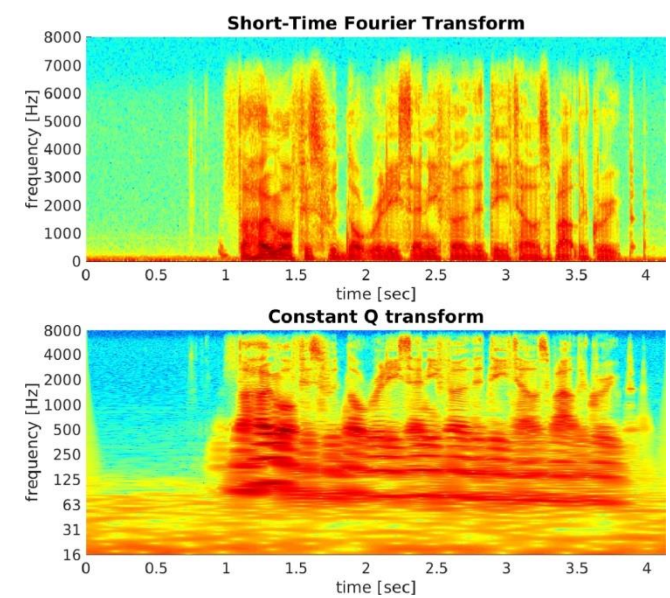


Figure 31: Spectrograms computed with the short-time Fourier Transform (top) and with the constant Q transform (bottom)

Todisco, M., Delgado, H., & Evans, N. (2016, June). A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients. In *Speaker Odyssey Workshop, Bilbao, Spain* (Vol. 25, pp. 249-252).

CQCC (contd.)

Table 5: *Performance in terms of average EER (%) for the best performing system, including individual results for each spoofing attack. Results for known and unknown attacks and the global average. Results for systems reviewed in Section 2 are included for comparison.*

	Known Attacks						Unknown Attacks						All
System	S1	S2	S3	S4	S5	Avg.	S6	S7	S8	S9	S10	Avg.	Avg.
CFCC-IF	0.101	0.863	0.000	0.000	1.075	0.408	0.846	0.242	0.142	0.346	8.490	2.013	1.211
i-vector	0.004	0.022	0.000	0.000	0.013	0.008	0.019	0.000	0.015	0.004	19.57	3.922	1.965
M&P feat.	0.000	0.000	0.000	0.000	0.010	0.002	0.010	0.000	0.000	0.000	26.10	5.222	2.612
LFCC-DA	0.027	0.408	0.000	0.000	0.114	0.110	0.149	0.011	0.074	0.027	8.185	1.670	0.890
CQCC-A	0.005	0.106	0.000	0.000	0.130	0.048	0.098	0.064	1.033	0.053	1.065	0.462	0.255

The **Best Paper award** sponsored by Agnitio went to **Massimiliano Todisco** for

Massimiliano Todisco, Héctor Delgado and Nicholas Evans.

A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients

Todisco, M., Delgado, H., & Evans, N. (2016, June). A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients. In *Speaker Odyssey Workshop, Bilbao, Spain* (Vol. 25, pp. 249-252).
http://www.odyssey2016.org/?p=Awards

Agenda

Part 1	Part 2

Part 1

- Introduction
- ASV System
- History of ASV Spoof
- Research Issues in ASV
- Spoofing Attacks
- Speech Synthesis
- Voice Conversion

Part 2

- Mimics
- Twins
- Countermeasures
- **Replay**
- **ASV Spoof 2015 Challenge**
- **ASV Spoof 2017 Challenge**
- **Future Research Directions**

Types of Spoofing Attacks

Spoofing Attacks on Voice Biometrics

IS 2015

IS 2017

Impersonation
- Physiological (Twins)
- Behavioral (Mimic)

Speech Synthesis
- Unit Selection(USS)
- HMM-based (HTS)

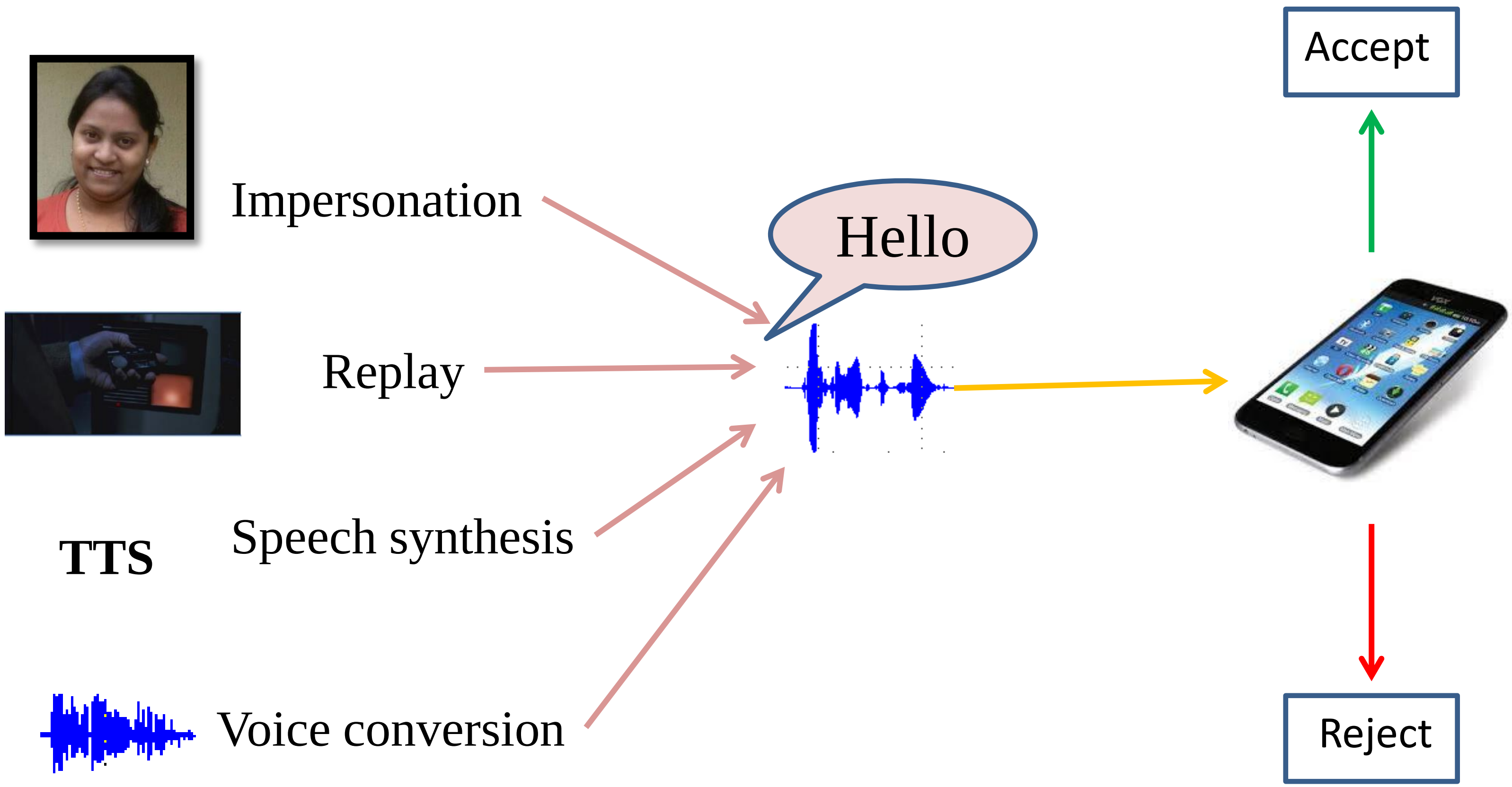
Voice Conversion
- Frame Selection
- Slope Shifting
- GMM
- KPLS
- Tensor

Replay
- Cut and Paste
- Tape recording
- Smart phone

Availability	**Low**	**High**	**High**	**Low/High**
Risk	Unknown	High	High	High

Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre and H. Li, “Spoofing and countermeasures for speaker verification: A survey,” *Speech Comm.*, vol. 66, pp. 130-153, 2015.

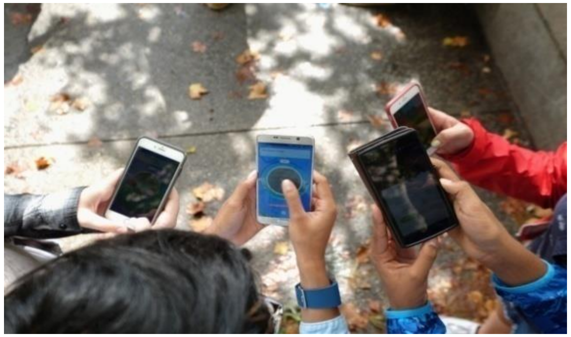
Spoofing Attacks



Replay

Using pre-recorded speech sample collected from a genuine target speaker is played back.

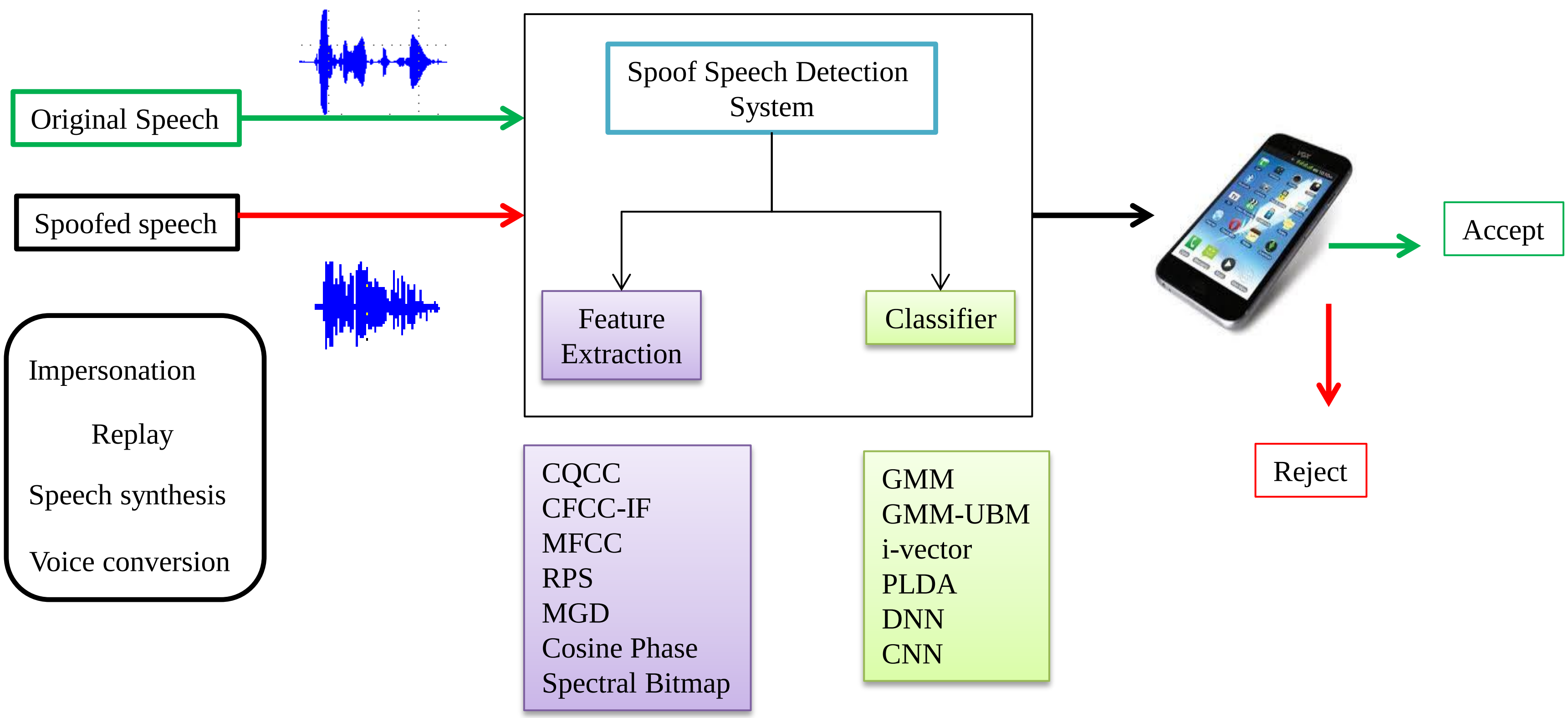
Harmful attack for text-dependent ASV system



Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Comm.*, vol. 66, pp. 130-153, 2015.

Spoof Speech Detection (SSD)

Due to effect of spoofed speech on ASV systems, need of standalone detectors (Natural *vs.* Spoofed speech) arose.



History of ASV Spoof

1999 — 2006 — 2013 — 2014 — 2015 — 2016 — 2017

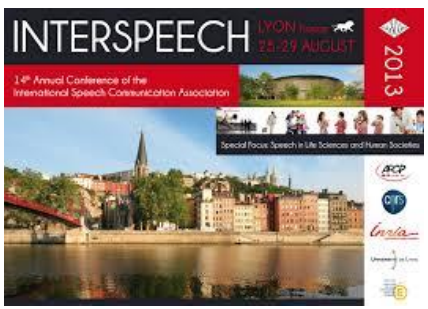
OCTAVE project starts

Small, purpose collected datasets

Adapted, standard datasets

IS 2013

Common, datasets, metrics, protocols



IS 2015

Common, datasets, synthetic speech

IS 2017

Common, datasets, replay, channel variation

Spoofing and Countermeasures for ASV 2013

- The INTERSPEECH 2013 special session in Spoofing and Countermeasures for ASV task.

- Motivation:

  – discussion and collaboration needed to organize the collection of standard datasets.

  – definition of metrics and evaluation protocols.

  – future research in spoofing and countermeasures for ASV.

[1] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," *in Proc. INTERSPEECH 2013*, Lyon France, 2013.

Key Differences

INTERSPEECH 2013	INTERSPEECH 2015	INTERSPEECH 2017
No standard Dataset	General data to all participants: Training, development (with key), Evaluation (without key)	General data to all participants: Training, development (with key), Evaluation (without key)
Spoofing and countermeasure for dedicated to ASV	No knowledge of ASV needed. Build detector for natural *vs.* spoof speech	No knowledge of ASV needed. Build detector for natural *vs.* spoof speech
Any spoof could be used	SS and VC spoof provided by organizers	Replay spoof provided by organizers
Performance measures evaluated independently	Uniformity in EER on the Evaluation set as evaluated by the organizers	Uniformity in EER on the Evaluation set as evaluated by the organizers
-	Text-independent	Text-dependent

ASV Spoof Challenge 2015

- Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASV spoof 2015 Challenge)

- Special session at INTERSPEECH 2015 → focus on spoofing detection.
- Develop → method/algorithm to discriminate human *vs*. spoofed speech (SS or VC)
- Database → generated from *10* (VC and SS) techniques.
- System expected to be reliable for both known and unknown attacks.
- No prior knowledge of ASV technology is needed.

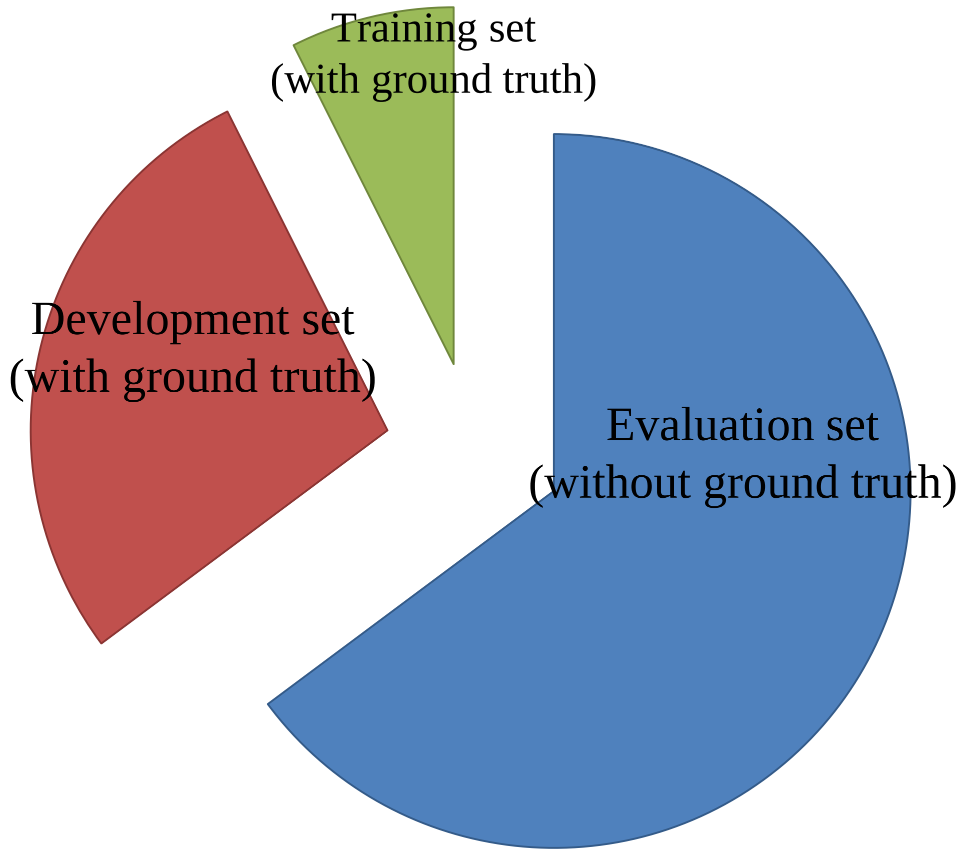
Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, M. Sahidullah, A. Sizov, "ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge", accepted in INTERSPEECH 2015, Dresden, Germany.

ASV Spoof 2015

Special session @INTERSPEECH 2015

Adapted from: Spoofing and anti-spoofing a shared view of speaker verification, speech synthesis and voice conversion APSIPA ASC tutorial 16th Dec. 2015

Database: Subsets



ASV Spoof 2015 Challenge Database

Table 7: Statistics of ASV Spoof 2015 Challenge datasets

Subset	Speakers		Utterances	
	Male	Female	Genuine	Spoofed
Training	10	15	3750	12625
Development	15	20	3497	49875
Evaluation	20	26	9404	184000

Training and development dataset: 5 spoofs (known)

Evaluation dataset : 10 spoofs (known and unknown)

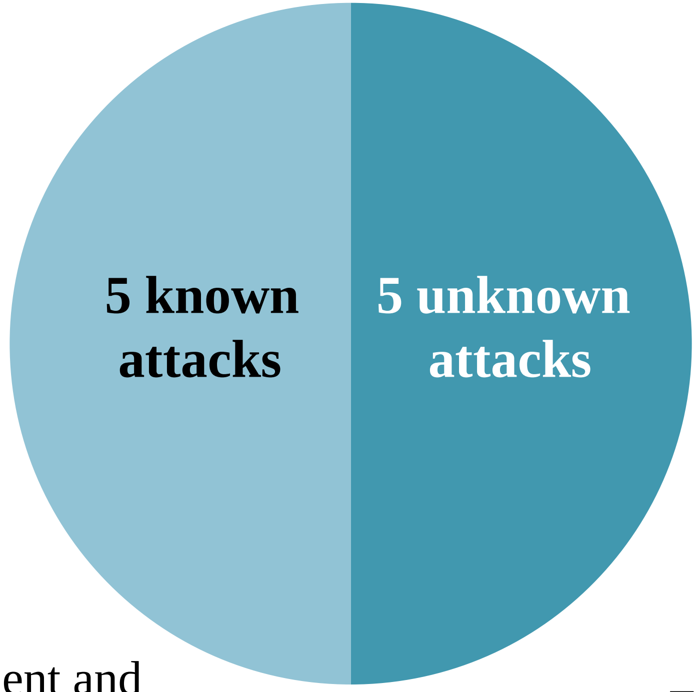
S3, S4, S10 : speech synthesis

S1, S2, S5, S6, S7, S8, S9 : voice conversion

Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," Speech Communication, vol. 66, pp. 130–153, 2015

Database: Spoofing algorithm

10 spoofing algorithms



Training, development and
Evaluation sets

Evaluation sets

Known and Unknown Attacks

- **S1-S5:** training, development and evaluation sets
  - S1: VC- Frame selection
  - S2: VC- Slope shifting
  - S3: TTS-HTS with 20 adaptation sentences
  - S4: TTS-HTS with 40 adaptation sentences
  - S5: VC- Festvox (http://festvox.org//)
- **S6-S10:** Only appear in evaluation sets
  - S6: VC- ML-GMM with GV enhancement
  - S7: VC- Similar to S6 but using LSP features
  - S8: VC- Tensor (eigenvoice)- based approach
  - S9: VC- Nonlinear regression (KPLS)
  - S10: TTS- MARY TTS unit selection

ASV Spoof 2015 Challenge Database

Table 8. Details of Spoofing Algorithm

Spoofing Algorithm	Type	Algorithm	Vocoder
Genuine	Natural	-	-
S1	VC	Frame Selection	STRAIGHT
S2	VC	Slope Shifting	STRAIGHT
S3	SS	HMM	STRAIGHT
S4	SS	HMM	STRAIGHT
S5	VC	GMM	MLSA
S6	VC	GMM	STRAIGHT
S7	VC	GMM	STRAIGHT
S8	VC	Tensor	STRAIGHT
S9	VC	KPLS	STRAIGHT
S10	SS	Unit Selection	-

Anti-spoofing Measures at the Challenge

❖ Countermeasures at the ASV spoof 2015 Challenge, INTERSPEECH 2015

Sr. No	Team	Features	Known attacks	Unknown attacks	All attacks
1	A (DA-IICT)	MFCC+CFCCIF	0.0408	2.013	1.211
2	B (STC)	MFCC, MFC, Cos-phase, MWPC	0.008.	3.922	1.965
3	C (SJTU)	RLMS, Spectrum, GD	0.058.	4.998	2.628
4	D (NTU)	LMS, RLMS, GD, MGD, IF, BPD, PSP	0.003	5.231	2.617
5	E (CRIM)	Cosine Normalized Phase, MGD, LP residual	0.041	5.347	2.694
6	F	Super vectors from MGD, Cos-phase, Fused with LB features	0.358	6.078	3.218
7	G	i-vector (MFCC, MFCC-PPP)	0.405	6.247	3.326
8	H	-	0.67	6.041	3.355
9	I	Iterative Phase Information	0.005	7.447	3.726
10	J	Fusion DNN (Spectrum + RPS)	0.025	8.168	4.097
11	K	Relative Phase Shift	0.21	8.883	4.547

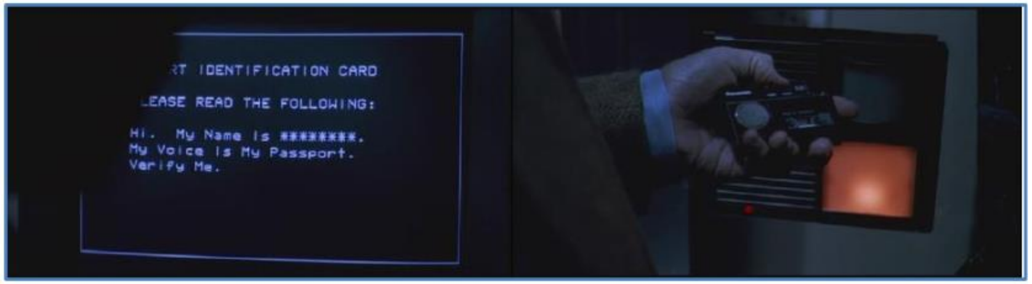
detection of natural vs. spoofed speech." Proc. Interspeech. 2015.

ASV Spoof 2017 Challenge

Statistics of ASV Spoof 2017 database.

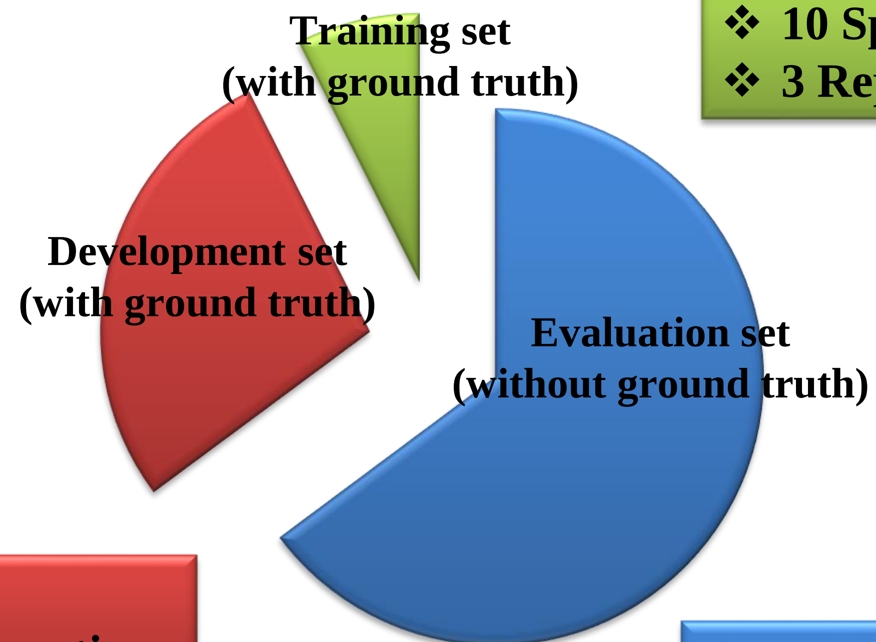
Table 9: Number of speakers and utterances in different datasets

Subset	Speakers	Utterances	
	Male	Genuine	Replay
Training	10	1508	1508
Development	8	760	950
Evaluation	24	1298	12922



T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Sahidullah, M. Todisco, and H. Delgado, "ASVspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2017.

Replay Database



- ❖ 10 Speakers
- ❖ 3 Replay configurations

- ❖ 8 Speakers
- ❖ 10 Replay configurations

- ❖ 24 Speakers
- ❖ 110 Replay configurations

Replay Configurations

Replay Configurations= Playback device + Environment +Recording device

Smartphone- smartphone

Headphone-PC mic

High-quality loudspeaker-
smartphone, anechoic room

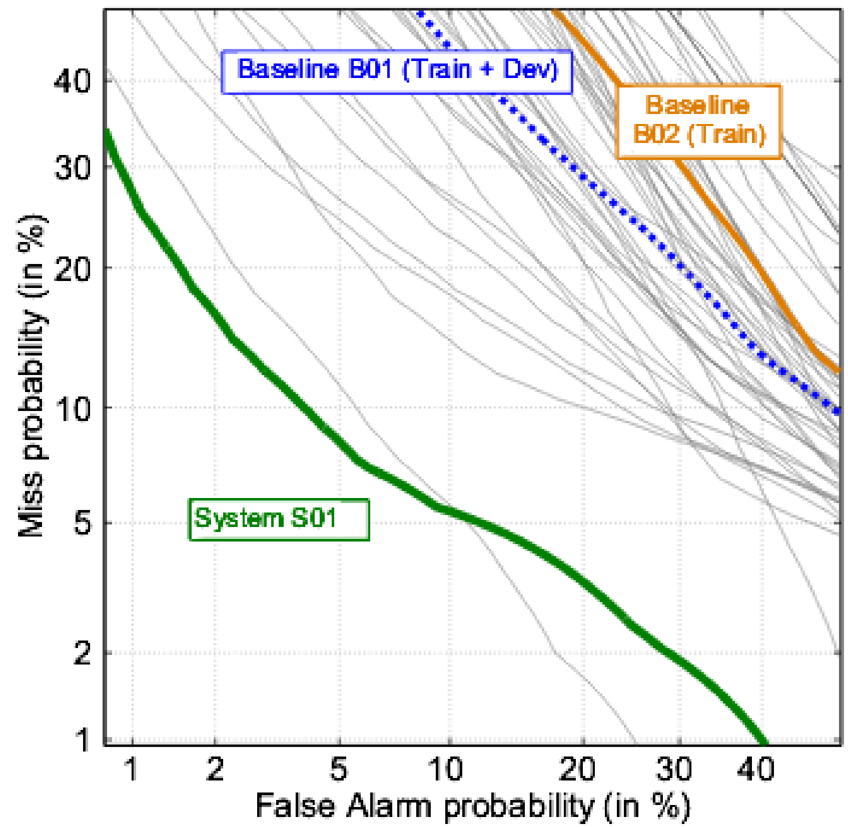
High-quality loudspeaker-
high-quality mic

Laptop line-out-
PC line-in using a cable

T. Kinnunen et al., "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 5395-5399.

ASV Spoof 2017 Challenge Results

ID	**EER**	*ID*	**EER**	*ID*	**EER**	*ID*	**EER**
S01	6.73	*S14*	22.04	*S26*	26.98	*S38*	31.59
S02	12.39	*S15*	22.23	*S27*	27.32	*S39*	31.76
S03	14.31	*S16*	22.41	*S28*	27.39	*S40*	32.59
S04	14.93	*S17*	23.11	*S29*	27.45	*S41*	34.78
S05	16.35	*S18*	23.19	*S30*	28.26	*S42*	35.57
S06	17.62	*S19*	23.53	*S31*	28.27	*S43*	36.05
S07	18.07	*S20*	23.85	*S32*	28.29	*S44*	37.2
S08	**18.33**	***B01***	**24.65**	*S33*	28.96	*S45*	38.15
S09	20.2	*S21*	24.66	*S34*	30.01	*S46*	38.51
S10	20.27	*S22*	25.1	***B02***	**30.17**	*S47*	39.06
S11	21.31	*S23*	25.19	*S35*	30.72	*S48*	45.82
S12	21.48	*S24*	26.21	*S36*	31.02	*D01*	7.39
S13	21.99	*S25*	26.51	*S37*	31.38	***Avg.***	**25.91**



S08: DA-IICT system B01: Baseline system (Pooled data) B02: Baseline system

Kinnunen, Tomi and Evans, Nicholas and Yamagishi, Junichi and Lee, Kong Aik and Sahidullah, Md and Todisco, Massimiliano and Delgado, H´ector, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection." submitted in INTERSPEECH, Stockhlom, Sweden, 2017.

Anti-spoofing Measures at the Challenge

- Countermeasures at the ASV spoof 2017 Challenge, INTERSPEECH 2017

Sr. No	Team	Features	Classifier	EER
1	S01	Power Spectrum, LPCC	CNN, GMM, TV, RNN	6.73
2	D01	MFCC, CQCC, WT	GMM, TV	7.39
3	S02	CQCC, MFCC, PLP	GMM-UBM, GSV-SVM, ivec-PLDA, GBDT, Random Forest	12.39
4	S03	MFCC, IMFCC, RFCC, LFCC, PLPCC, CQCC, SCMC, SSFC	GMM, FF-ANN	14.31
5	S04	RFCC, MFCC, IMFCC, LFCC, SSFC, SCMC	GMM	14.93
6	S05	Linear filterbank feature	GMM, CT-DNN with convolutional layer and time-delay layers	16.35
7	S06	CQCC, IMFCC, SCMC, Phrase one-hot encoding	GMM	17.62
8	S08 (DA-IICT)	IFCC, CFCCIF, Prosody	GMM	18.33
9	S10	CQCC	Residual Neural Network	20.27
10	S09	SFFCC	GMM	20.20
11	S11	CQCC	TV-PLDA	21.31
12	S12	CQCC	FF-DNN, BLSTM, GMM	21.48
13	S13	CQCC	GMM, ivector-SVM	21.99

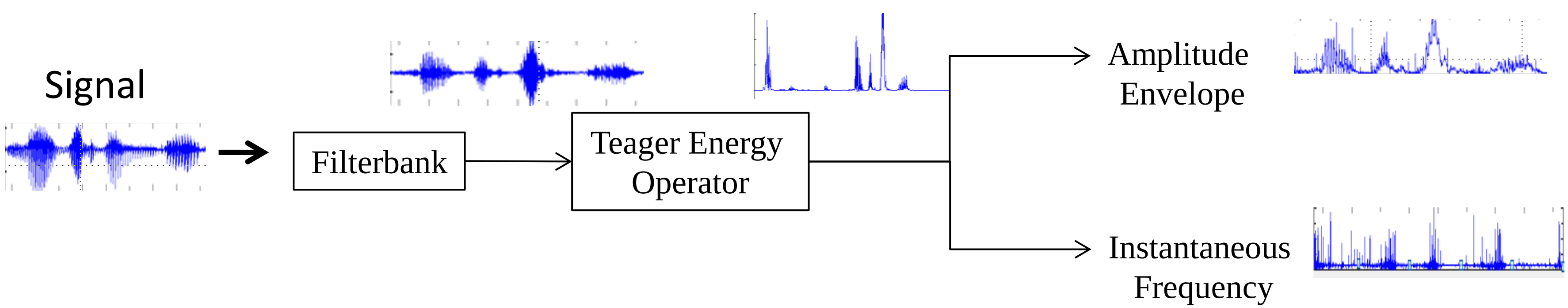
Teager Energy Operator (TEO)

We define TEO in Continuous Time domain as,

$$\psi\{x(t)\} = \dot{x}^2(t) - x(t)\ddot{x}(t)$$

$$x(t) = A\cos(\omega t)$$

$$\begin{aligned}\psi\{x(t)\} &= [-A\omega\sin(\omega t)^2 - A\cos(\omega t)(-\omega^2 A\cos(\omega t)] \\ &= A^2\omega^2(\sin^2(\omega t) + \cos^2(\omega t)) \\ &= A^2\omega^2\end{aligned}$$



Maragos, Petros and Kaiser, James F and Quatieri, Thomas F, "On separating amplitude from frequency modulations using energy operators," in IEEE ICASSP, vol. 2, San Francisco, California, USA, 1992, pp. 1–4.

Teager Energy Operator (TEO)

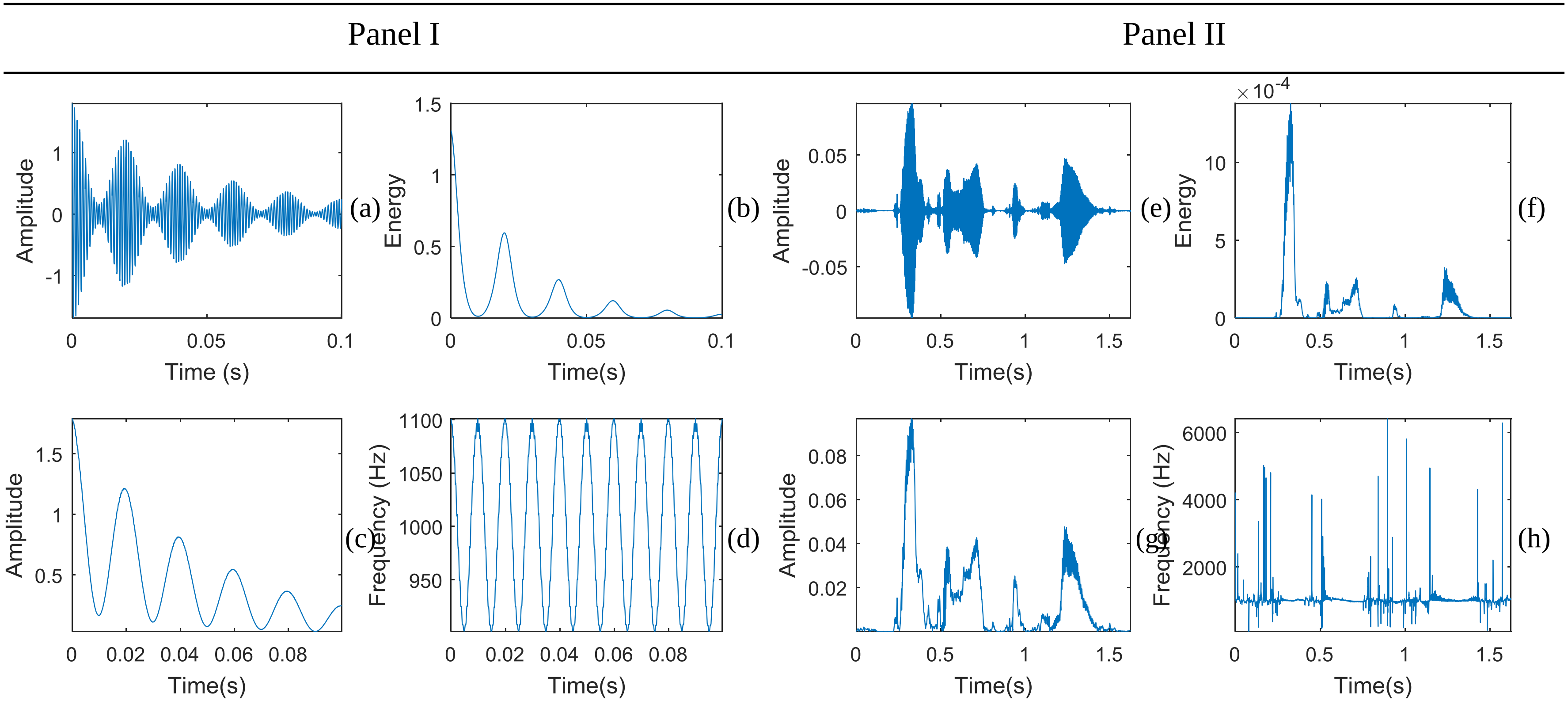


Figure 32: AM-FM estimation using the ESA on a synthetic (Panel I) and speech signal (Panel II) with "Johnson was pretty liar" utterance taken form ASV Spoof 2015 challenge database

(a) AM-FM signal of a = (0:998n(1 + 0:2cos((=80)n))) and x = a(cos(((=5)n) +sin((=40)n))),

(e) Filtered narrowband signal at fc = 1500 Hz,

(b-f) Teager energy,

(c-g) estimated amplitude envelope and

(d-h) estimated instantaneous frequency at fc = 1000 Hz for synthetic signal and 1500 Hz for speech signal.

Proposed ESA-IFCC Features

ESA-IFCC: Energy Separation Algorithm-Instantaneous Frequency Cosine Coefficients

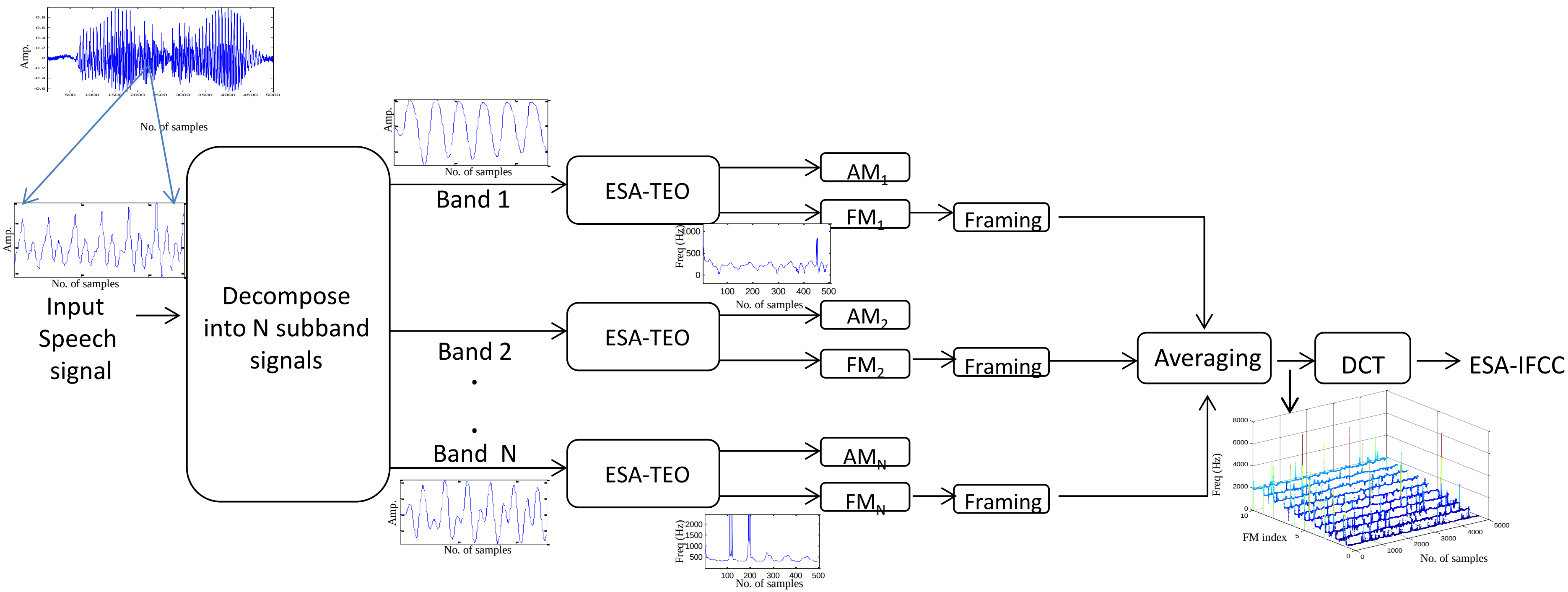


Figure 33: Block diagram of proposed feature

M. R. Kamble and H. A. Patil, "Novel energy separation based instantaneous frequency features for spoof speech detection" in European Signal Processing Conference (EUSIPCO), 2017

Variable length Energy Separation Algorithm (VESA)

In VESA, we modify original TEO to VTEO with change in equation given as:

TEO : $$\psi\{x(n)\} = x^2(n) - x(n+1)x(n-1)$$

VTEO : $$\psi_{DI}\{x(n)\} = x^2(n) - x(n+i)x(n-i)$$

i - indicates the dependency index (DI)

We used DESA-2 approach for VESA

$$AE = \frac{2\psi_{DI}\{x(n)\}}{\sqrt{\psi_{DI}\{x(n+1) - x(n+1)\}}} \qquad IF = \arcsin\left(\sqrt{\frac{\psi_{DI}\{x(n+1) - x(n-1)\}}{4\psi_{DI}\{x(n)\}}}\right)$$

H. A. Patil and K. K. Parhi, “Novel variable length Teager energy based features for person recognition from their hum,” in IEEE ICASSP, Dallas, Texas, USA, 2010, pp. 4526–4529.

H. A. Patil, M. R. Kamble, T. B. Patel, and M. H. Soni, "Novel variable length Teager energy separation based if features for replay detection," in INTERSPEECH, 2017.

Proposed VESA-IFCC Features

VESA-IFCC: Variable length Energy Separation Algorithm-Instantaneous Frequency Cosine Coefficients

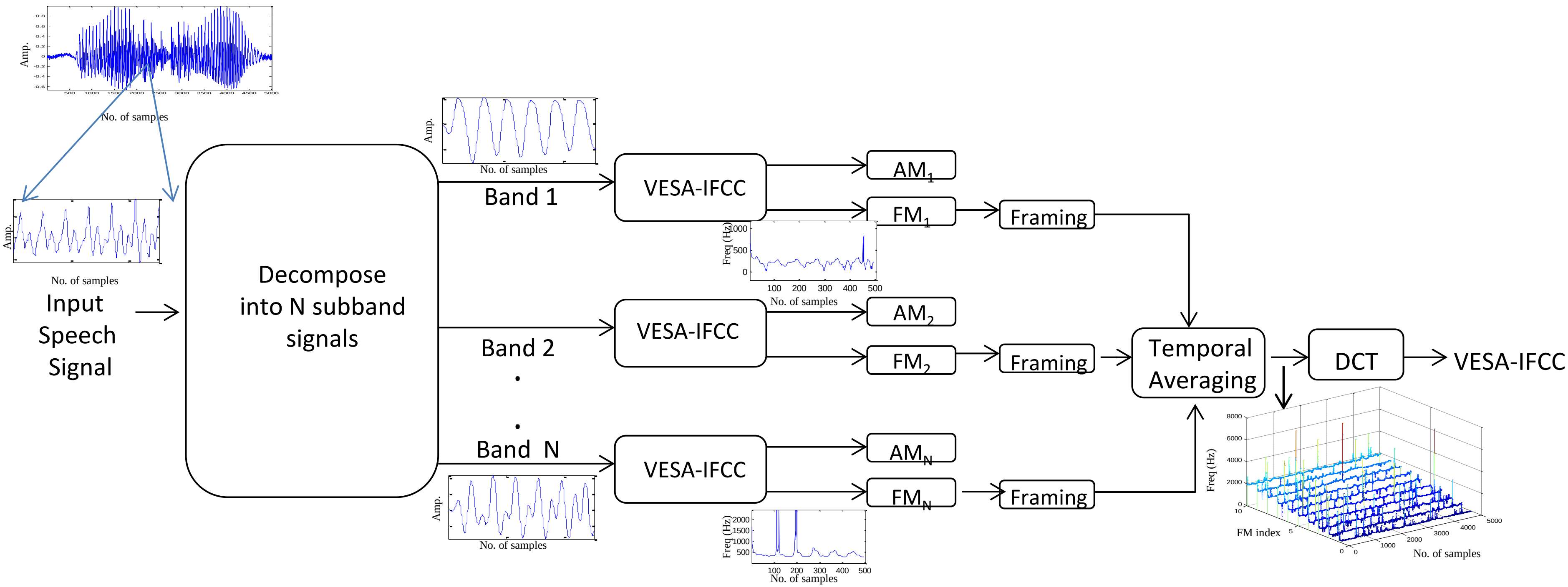


Figure 34: Schematic diagram to estimate proposed VTEO-based ESA-IFCC feature set.

H. A. Patil, M. R. Kamble, T. B. Patel, and M. H. Soni, "Novel variable length Teager energy separation based if features for replay detection," in INTERSPEECH, 2017.

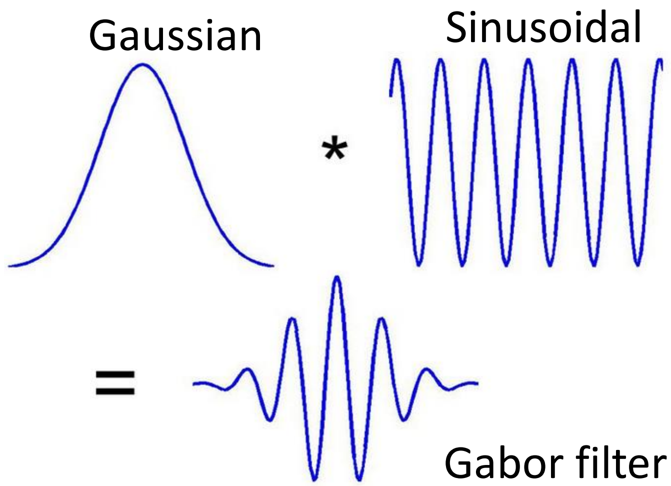
Gabor Filter

Gabor Filter: A Gabor filter is a combination of Gaussian filter and a sinusoidal term

Impulse response of Gabor filter

$$h(t) = \exp(-a^2 t^2)\cos(2\pi v t),$$

where *a* is the parameter for controlling the bandwidth
and *v* is the cutoff frequency

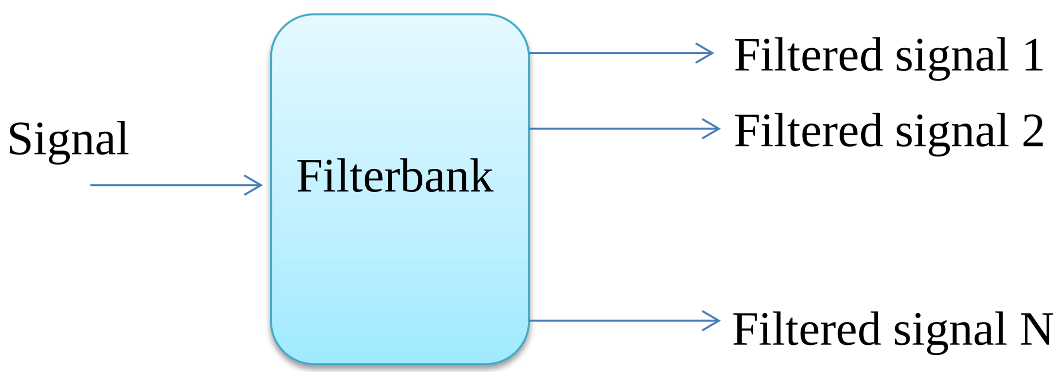


Gabor, D. (1946). Theory of communication. Journal of the Institute of Electrical Engineers, 93 , 429–457
Kleinschmidt, M., B. Meyer, and D. Gelbart. "Gabor feature extraction for automatic speech recognition”

Filterbank

Filterbank splits up signals into different frequency bands

> In signal processing, a filter bank is an array of band-pass filters that separates the input signal into multiple components, each one carrying a single frequency sub-band of the original signal.



Frequency Scales

ERB Scale

$$ERB = 6.23\,(v/1000)^2 + 93.39\,(v/1000) + 28.52$$

Mel Scale

$$Mel = 2595 \log_{10}(1 + v/700)$$

Linear Scale

$$Lin = v$$

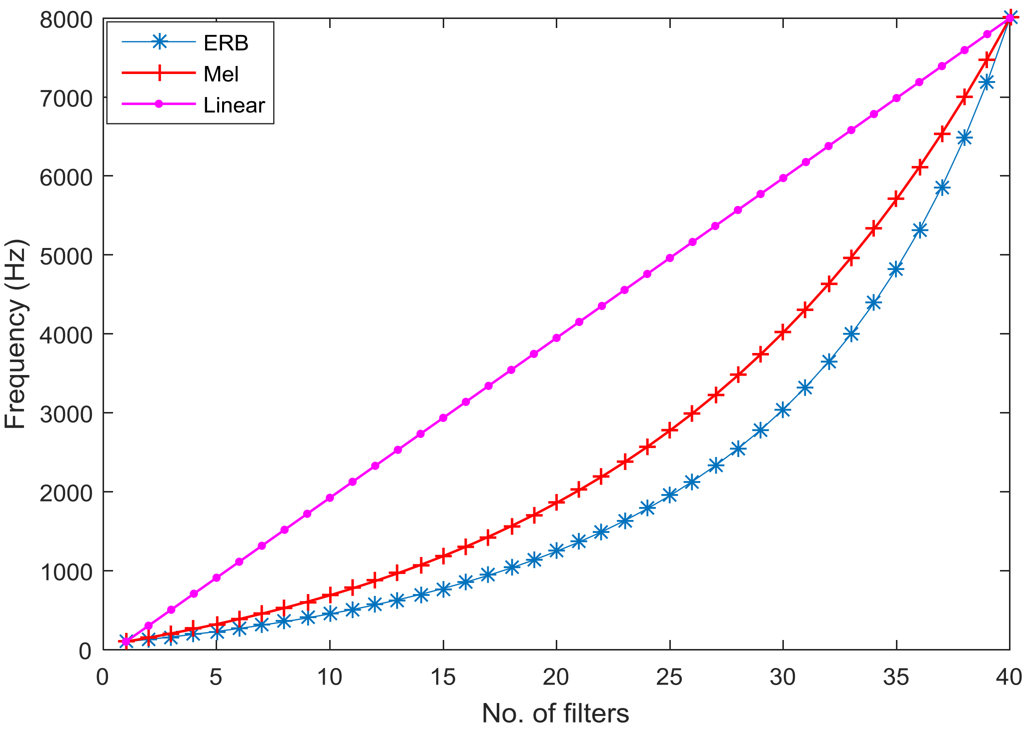


Figure 35: Frequency scales for ERB (blue), Mel (red) and linear (pink)

M. R. Kamble and H. A. Patil, Effectiveness of Mel scale-based ESA-IFCC features for classification of natural *vs*. spoofed speech," in Accepted in 7th International Conference on Pattern Recognition and Machine Intelligence (PReMI), (Kolkata, India), 2017.

Gabor Filterbank

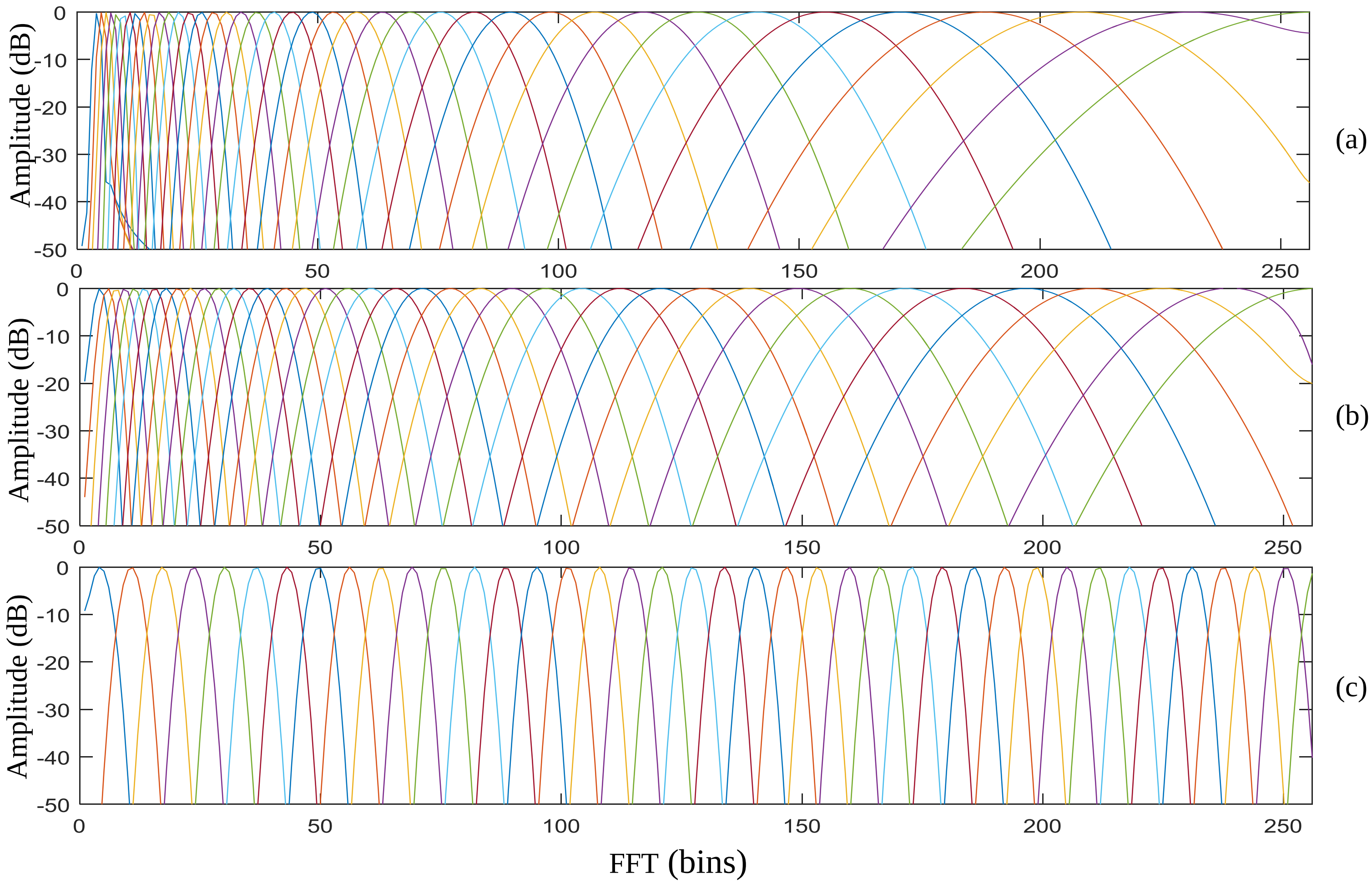


Figure 36: Frequency response of (a) ERB, (b) Mel and (c) linear frequency scales.

M. R. Kamble and H. A. Patil, Effectiveness of Mel scale-based ESA-IFCC features for classification of natural *vs*. spoofed speech," in Accepted in 7th International Conference on Pattern Recognition and Machine Intelligence (PReMI), (Kolkata, India), 2017.

Spectrographic Analysis with Gabor Filterbank

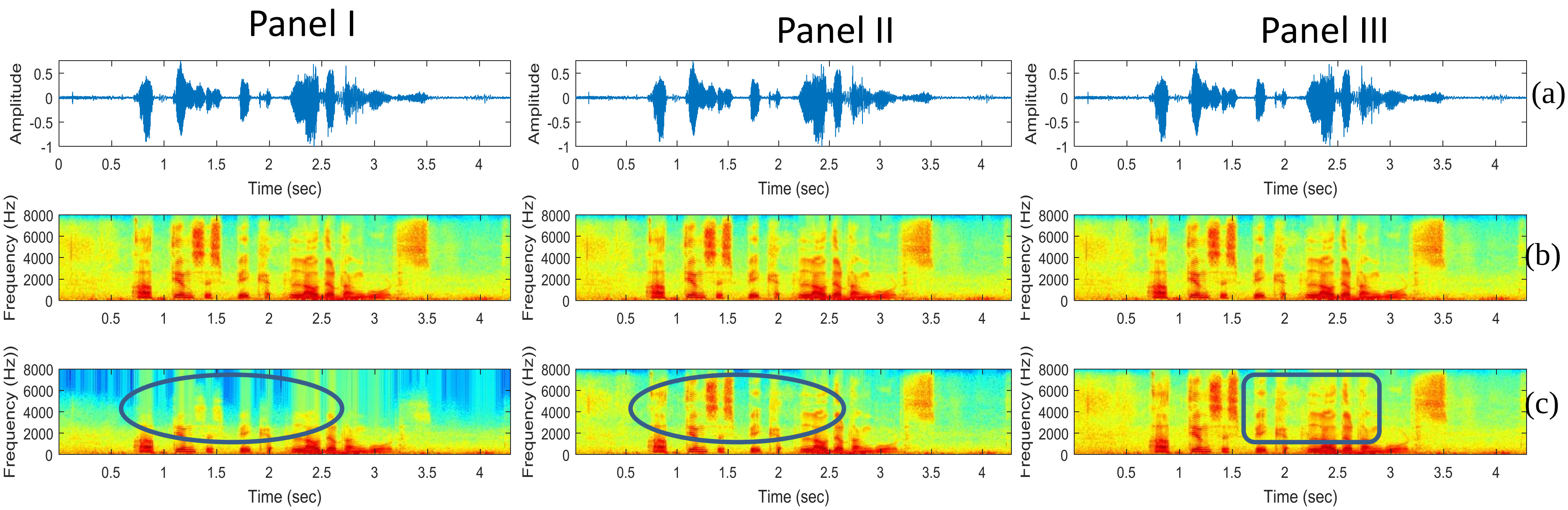


Figure 38: Spectrographic analysis (a) time-domain speech signal, (b) spectrogram and (c) energy density obtained after 40 subband Gabor filterbank of (Panel I) ERB (Panel II) Mel and (Panel III) Linear frequency scales

- Observations: Spectral energy obtained with linear frequency scales contains more speaker-specific information than ERB and Mel scale

Spectrographic Analysis for Replayed Speech

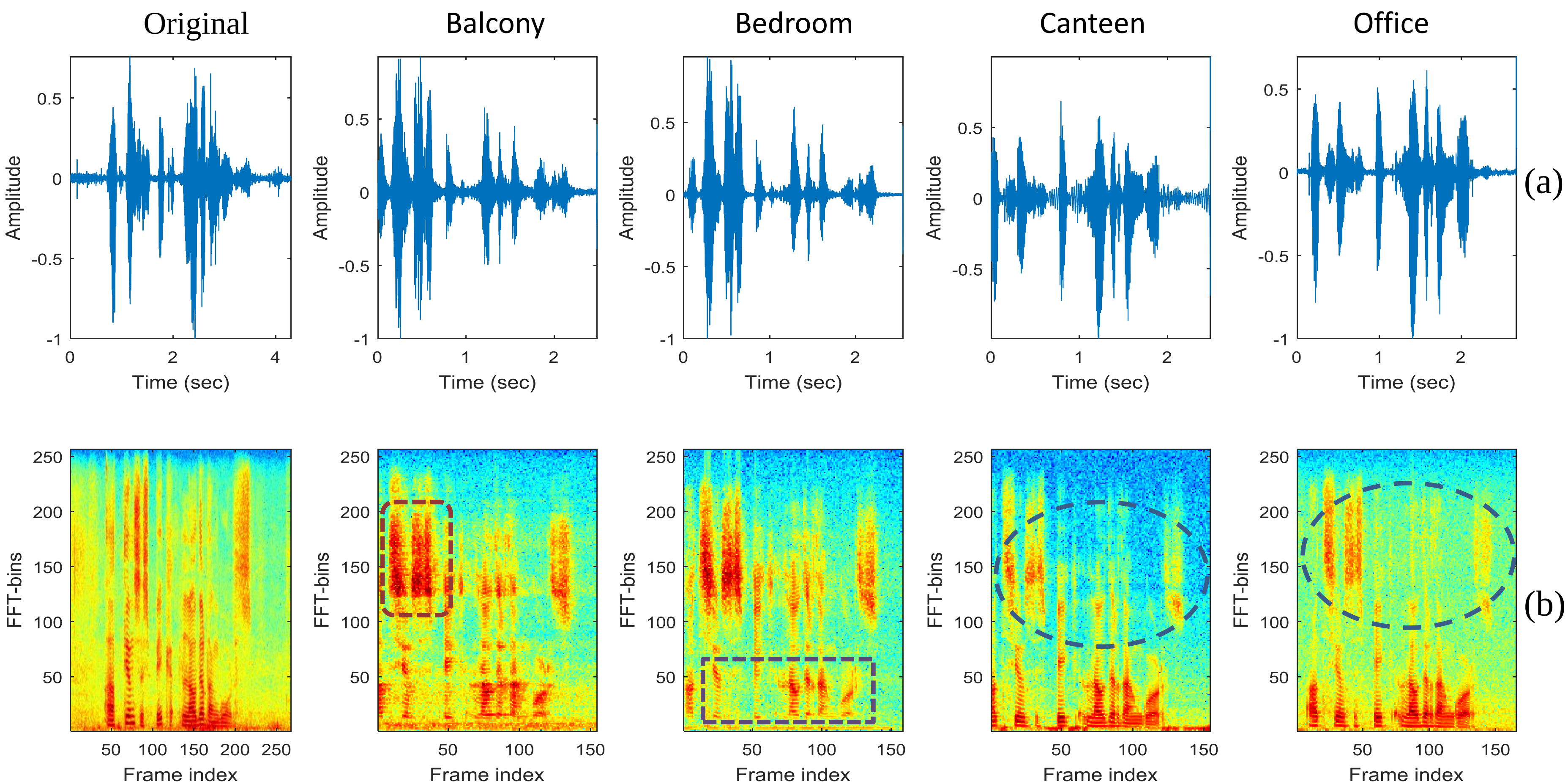


Figure 39. Spectrographic Analysis: (a) speech signal and (b) corresponding spectrogram

Experimental Setup ASV 2015

Table 11: Experimental setup used to extract the features on ASV 2015

Features	**GMM Models**	**Feature Dimension**	**Filterbank**	**No. of Filterbank**
MFCC	128	13	Butterworth	28
ESA-IFCC	128	13	Triangular	40

M. R. Kamble and H. A. Patil, Novel energy separation based instantaneous frequency features for spoof speech detection," in European Signal Processing Conference (EUSIPCO), (Kos Island, Greece), pp. 116-120, 2017.

Experimental Results ASV 2015

Table 12: Results on development set in % EER on ASV 2015

Features	Feature Dimension	EER		
		Static	Static+Δ	Static+Δ+ΔΔ
MFCC	39	6.98	6.75	6.14
A: ESA-IFCC	39	5.43	6.22	6.59
ESA-IFCC	120	6.38	7.47	7.18
MFCC+A	39	3.45	2.01	1.89

Table 13: Results on evaluation set in % EER on ASV 2015

Features	Known Attacks					Unknown Attacks					All Avg
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
MFCC	2.34	9.57	0.00	0.00	9.01	7.73	4.42	0.3	5.17	52.99	9.15
ESA-IFCC	2.53	5.29	0.00	0.00	13.54	12.09	3.61	4.52	4.13	36.14	8.18

M. R. Kamble and H. A. Patil, Novel energy separation based instantaneous frequency features for spoof speech detection," in European Signal Processing Conference (EUSIPCO), (Kos Island, Greece), pp. 116-120, 2017.

Results with Gabor Filterbank

Table 14: Details of feature extraction on ASV 2015

Features	MFCC	ESA-IFCC
No. of filters	40	40
Feature dimension	39 (13 S+D+DD)	39 (13 S+D+DD)
No. of mixtures in GMM	128	128
Frequency scale	Mel	ERB, Mel & Linear

Gaussian Mixture Models is used for binary classification

- No. of classes: 2
  - genuine class
  - spoof class

Log-Likelihood Ratio

LLR=log(LLK_Model1)-log(LLK_Model2),

M. R. Kamble and H. A. Patil, Effectiveness of Mel scale-based ESA-IFCC features for classification of natural *vs*. spoofed speech," in Accepted in 7th International Conference on Pattern Recognition and Machine Intelligence (PReMI), (Kolkata, India), 2017.

Results on Development Set

- Performance measure on Equal Error Rate (EER)
- ESA-IFCC with linear scale has lower EER and better separation
- ESA-IFCC feature set with Mel and Linear scale has lower EER than MFCC alone for all dimensions

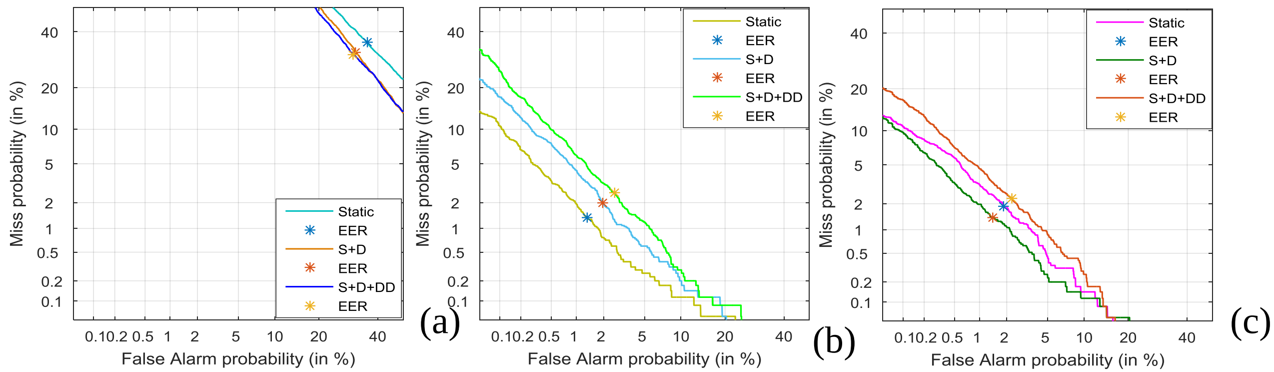


Figure 40: The DET curves for (a) ERB, (b)Mel and (c) Linear scale of ESA-IFCC feature set

Table 15: Results in EER on development set on ASV 2015

Frequency scales	Static	Static+D	Static+D+DD
MFCC	6.98	6.75	6.14
ESA-IFCC (ERB)	35.66	31.85	30.93
ESA-IFCC (Mel)	**1.32**	1.96	2.52
ESA-IFCC (Linear)	1.86	**1.39**	**2.23**

M. R. Kamble and H. A. Patil, Effectiveness of Mel scale-based ESA-IFCC features for classification of natural *vs*. spoofed speech," in Accepted in 7th International Conference on Pattern Recognition and Machine Intelligence (PReMI), (Kolkata, India), 2017.

Results on Development Set

- Score-level fusion

$$LLk_{combine} = (1-\alpha_f)LLk_{MFCC} + \alpha_f LLk_{feature2}$$

Table 16: Results of fused feature set in EER on development set

Frequency scales	Static	Static+D	Static+D+DD
MFCC+ERB	6.98	6.75	5.90
MFCC+Mel	1.01	1.35	1.50
MFCC+Linear	1.44	1.14	1.67

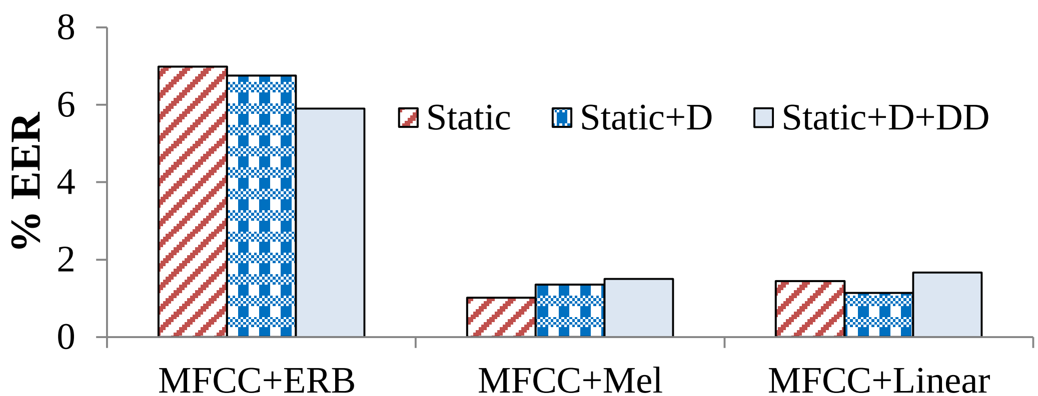


Figure 41: Bar graph result of score-level fusion of MFCC and proposed feature set

M. R. Kamble and H. A. Patil, Effectiveness of Mel scale-based ESA-IFCC features for classification of natural *vs*. spoofed speech," in Accepted in 7th International Conference on Pattern Recognition and Machine Intelligence (PReMI), (Kolkata, India), 2017.

Results on Evaluation Set

Table 17: Results in EER on evaluation set

Features	Known Attacks				Unknown Attacks						All Avg
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
MFCC	2.34	9.57	0.00	0.00	9.01	7.73	4.42	0.3	5.17	52.99	9.15
ESA-IFCC (Linear)	**0.85**	**1.85**	**0.00**	**0.00**	**3.03**	13.01	**1.63**	**0.23**	**1.89**	**33.37**	**5.58**

- Almost for all spoofing attacks ESA-IFCC features with linear scale performs better than MFCC
- Performance of S10 attack makes the overall EER lower than other attacks

M. R. Kamble and H. A. Patil, Effectiveness of Mel scale-based ESA-IFCC features for classification of natural *vs*. spoofed speech," in Accepted in 7th International Conference on Pattern Recognition and Machine Intelligence (PReMI), (Kolkata, India), 2017.

Selection of DI & Feature dimension

Table 18: Effect of DI in VESA-IFCC on the development set

DI	**1**	**2**	**3**	**4**	**5**
EER	7.65	6.61	6.65	6.63	6.99
DI	**6**	**7**	**8**	**9**	**10**
EER	7.41	8.57	6.46	4.61	7.17

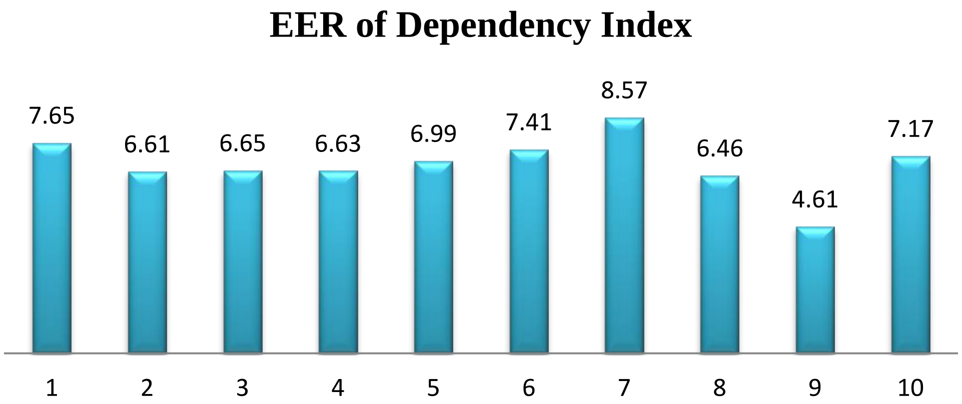
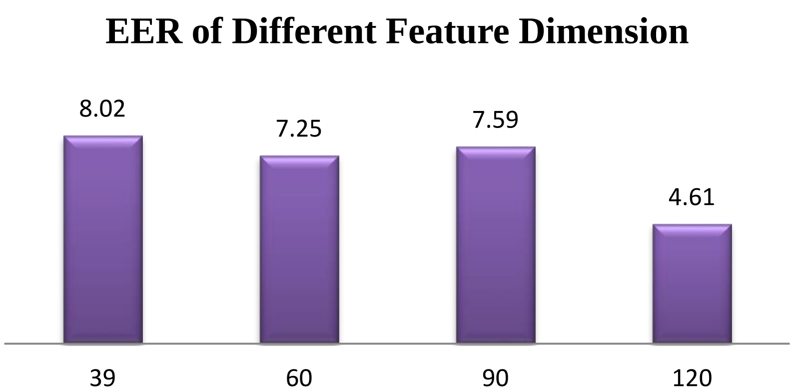


Table 19: Effect of Feature Dimension (FD) on the development set for D1=9

FD	**39**	**60**	**90**	**120**
EER	8.02	7.25	7.59	4.61

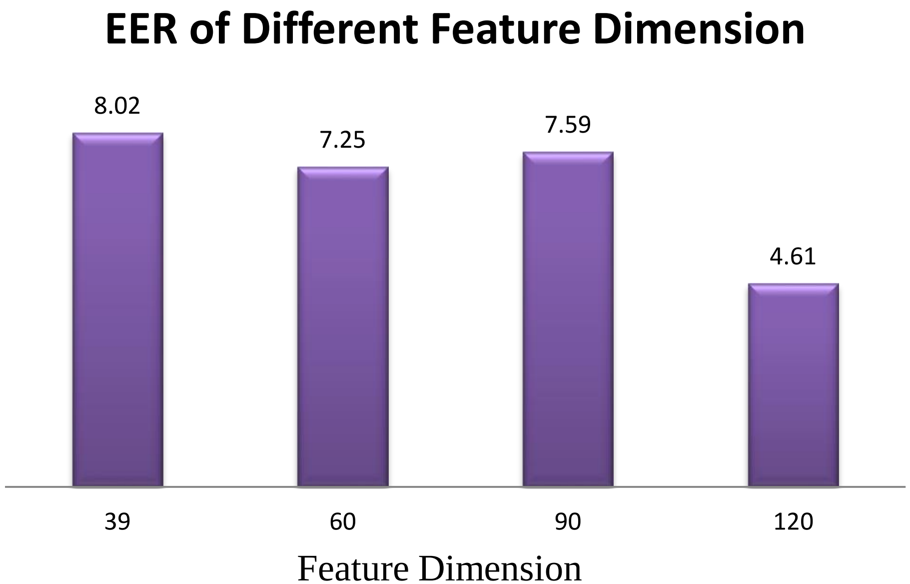


H. A. Patil, M. R. Kamble, T. B. Patel, and M. Soni, Novel variable length Teager energy separation based instantaneous frequency features for replay detection," in INTERSPEECH, Stockholm, Sweden, pp. 12-16, 2017

Selection of Feature dimension

Table 19: Effect of Feature Dimension (FD) on the development set for D1=9 with (static+delta+double delta)

FD	39	60	90	120
EER	8.02	7.25	7.59	4.61



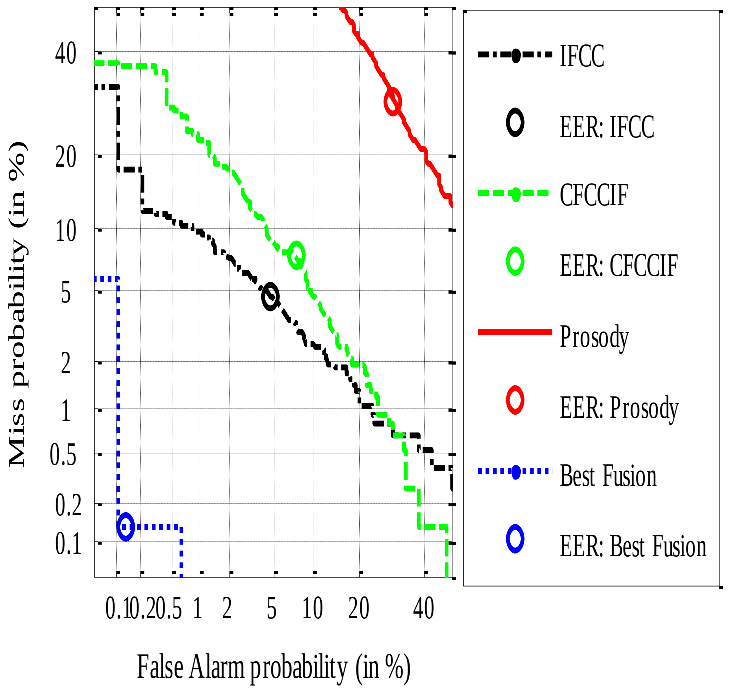
H. A. Patil, M. R. Kamble, T. B. Patel, and M. Soni, Novel variable length Teager energy separation based instantaneous frequency features for replay detection," in INTERSPEECH, Stockholm, Sweden, pp. 12-16, 2017

Post Evaluation Results

Table 20: Result in % EER on development and evaluation set with GMM classifier. * Primary Submission

Feature Set	Development	Evaluation
CQCC (Baseline)	11.06	30.17
A: CFCCIF	6.8	34.49
B: Prosody	29.40	31.40
C: VESA-IFCC	4.61	14.06
C+MFCC	1.47	17.93
C+CQCC	2.08	15.35
A+B+C	0.1263	**18.33***

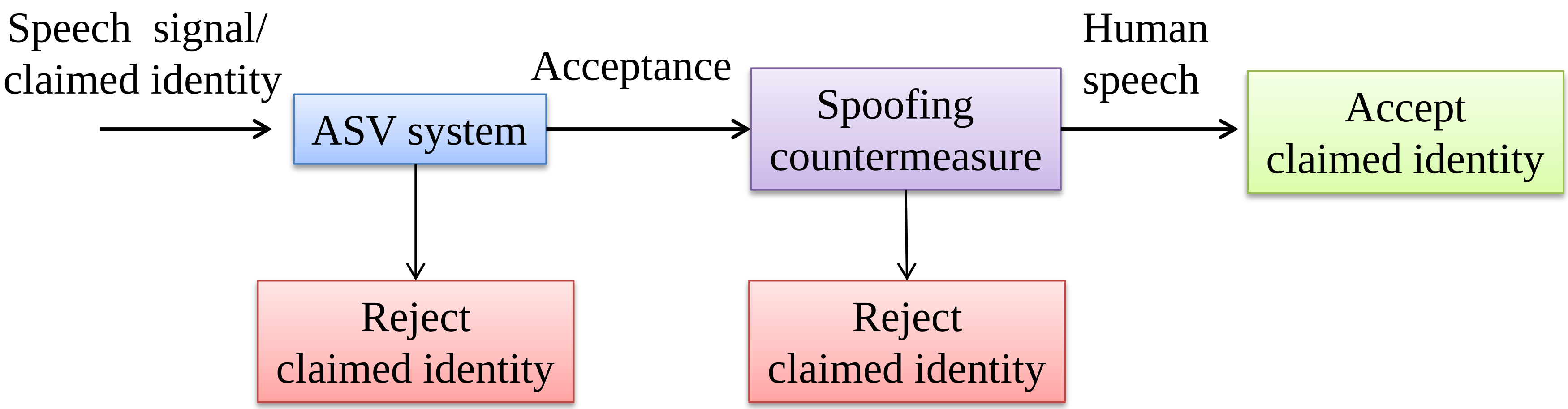
Features	EER
CQCC (Baseline)	24.65
VESA-IFCC	**15.50**
VESA-IFCC+CFCCIF+Prosody	23.68



The individual DET curves for IFCC, CFCCIF, prosody and the best fusion factor on the development set.

H. A. Patil, M. R. Kamble, T. B. Patel, and M. Soni, Novel variable length Teager energy separation based instantaneous frequency features for replay detection," in INTERSPEECH, Stockholm, Sweden, pp. 12-16, 2017

Spoofing ASV Systems with use of Countermeasures



Zhizheng Wu, et. al., "Anti-Spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance", *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 24, no. 4, pp 768-783, 2016.

Human *vs.* Machine

- Current spoof detectors almost contradict the human perception
- Spoofed speech accepted as genuine by humans is very well detected as spoof by detectors.

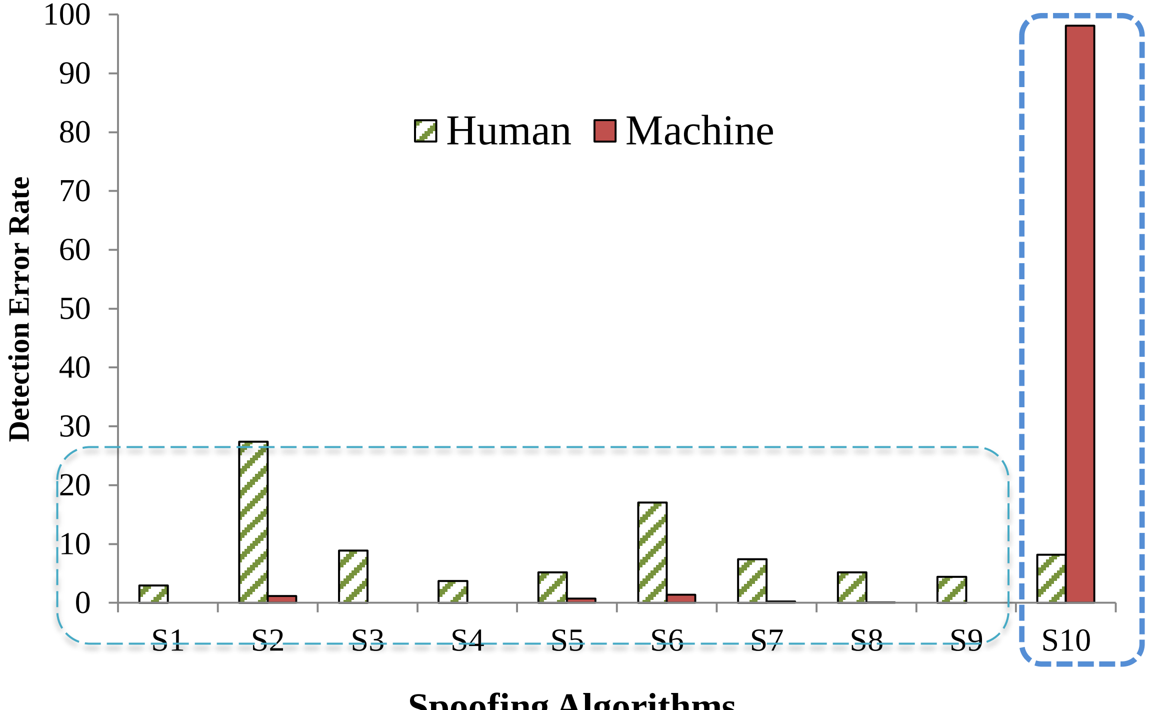


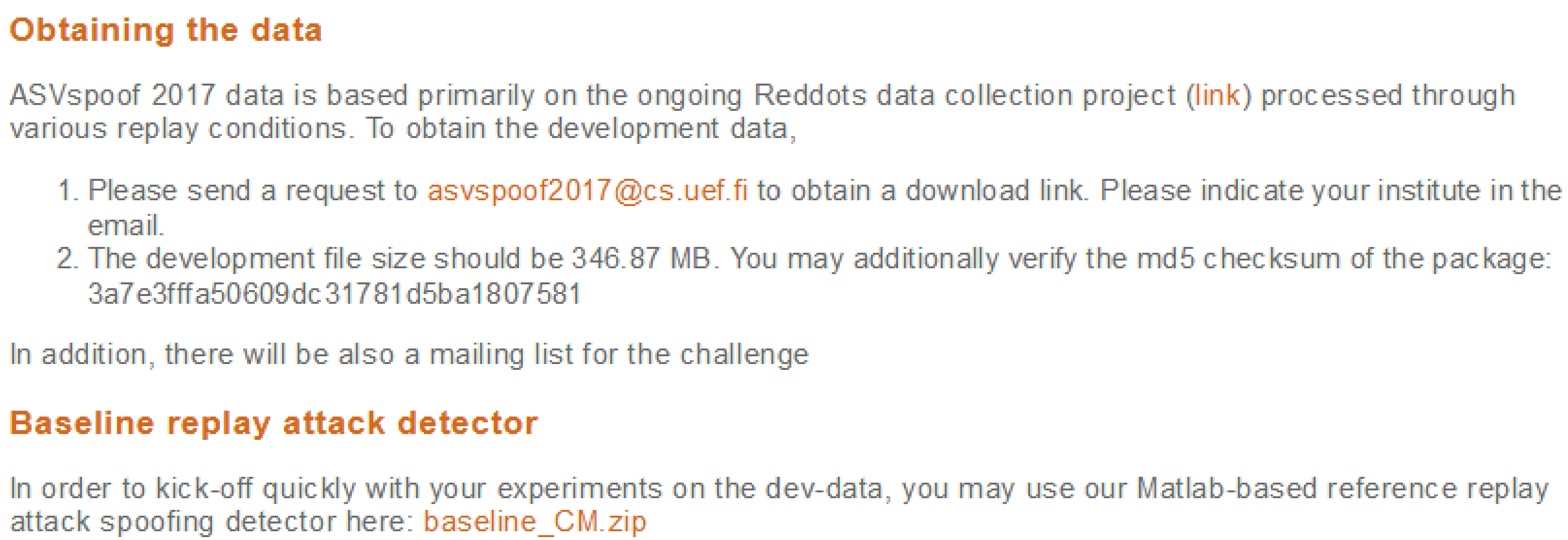
Figure 41: Human *vs.* Machine performance obtained via listening tests

[1] M. Wester, Z. Wu, and J. Yamagishi, "Human vs. machine spoofing detection on wideband and narrowband data," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2047-2051.

[2] Zhizheng Wu, et al., "Anti-Spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance", *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 24, no. 4, pp 768-783, 2016.

Baseline System for ASV Spoof 2017 Challenge

Download baseline CQCC-GMM system at URL: http://www.asvspoof.org/


Obtaining the data

ASVspoof 2017 data is based primarily on the ongoing Reddots data collection project (link) processed through various replay conditions. To obtain the development data,

1. Please send a request to asvspoof2017@cs.uef.fi to obtain a download link. Please indicate your institute in the email.
2. The development file size should be 346.87 MB. You may additionally verify the md5 checksum of the package: 3a7e3fffa50609dc31781d5ba1807581

In addition, there will be also a mailing list for the challenge

Baseline replay attack detector

In order to kick-off quickly with your experiments on the dev-data, you may use our Matlab-based reference replay attack spoofing detector here: baseline_CM.zip

Information of Challenge

Further information

Please refer to the earlier 2015 challenge edition here for general background. We will also keep adding other useful readings to this page.

The ASVspoof 2017 challenge overview paper to appear at INTERSPEECH 2017 is available:

> Tomi Kinnunen, Md Sahidullah, Hector Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, Kong Aik Lee, "**The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection**", manuscript, submitted to Interspeech 2017. [PDF]

T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. Gonzalez Hautamäki, D. Thomsen, A. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco, N. Evans, V. Hautamäki, K. A. Lee, "**RedDots Replayed: A New Replay Spoofing Attack Corpus for Text-Dependent Speaker Verification Research**", Proc. ICASSP 2017 [PDF]

Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, H. Delgado, "**ASVspoof: the Automatic Speaker Verification Spoofing and Countermeasures Challenge**", IEEE Journal on Selected Topics in Signal Processing (to appear, https://doi.org/10.1109/JSTSP.2017.2671435) [PDF]

M. Todisco, H. Delgado, N. Evans, "**Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification**", Computer Speech and Language (to appear, http://dx.doi.org/10.1016/j.csl.2017.01.001) [PDF]

Databases

ASVspoof 2017 data is based primarily on the ongoing Reddots data collection project

RedDots Project

https://sites.google.com/site/thereddotsproject/

ASV Spoof 2017

https://datashare.is.ed.ac.uk/handle/10283/2778

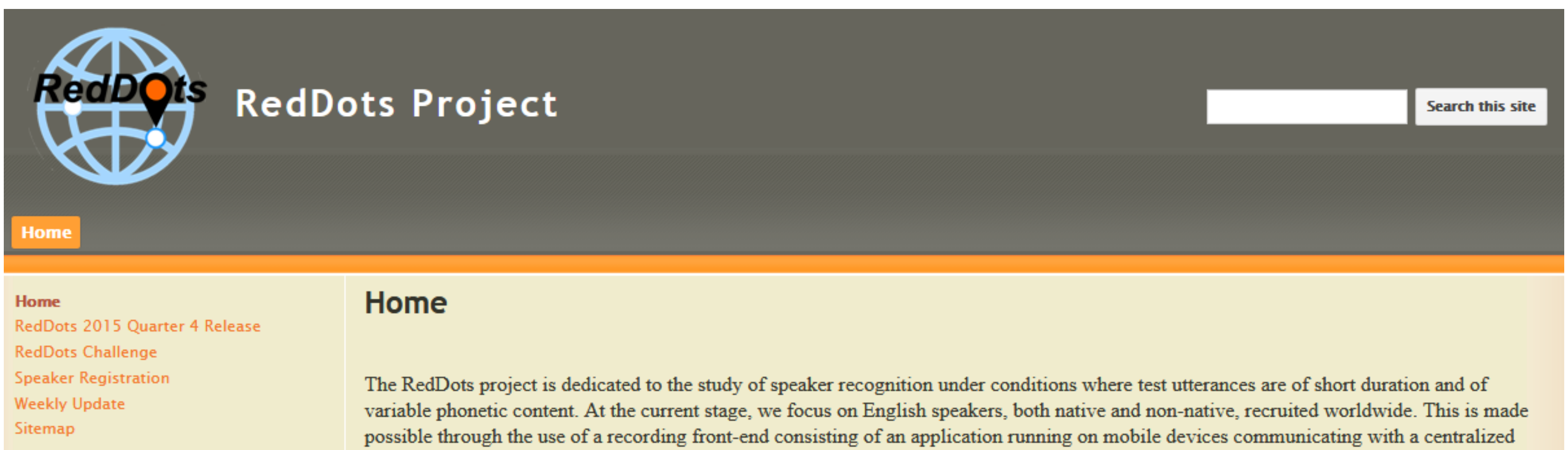
ASV Spoof 2015

https://datashare.is.ed.ac.uk/handle/10283/853

AV Spoof 2016: BTAS 2016

http://pythonhosted.org/bob.db.avspoof_btas2016/

ASV Spoof 2017



INTERSPEECH 2017

Download link for accepted papers

http://www.isca-speech.org/archive/Interspeech_2017/

Special Session: Interspeech 2017 Automatic Speaker Verification Spoofing and Countermeasures Challenge 1

The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection
Tomi Kinnunen, Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, Kong Aik Lee

Experimental Analysis of Features for Replay Attack Detection — Results on the ASVspoof 2017 Challenge
Roberto Font, Juan M. Espín, María José Cano

Novel Variable Length Teager Energy Separation Based Instantaneous Frequency Features for Replay Detection
Hemant A. Patil, Madhu R. Kamble, Tanvina B. Patel, Meet H. Soni

Countermeasures for Automatic Speaker Verification Replay Spoofing Attack : On Data Augmentation, Feature Representation, Classification and Fusion
Weicheng Cai, Danwei Cai, Wenbo Liu, Gang Li, Ming Li

Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features
Sarfaraz Jelil, Rohan Kumar Das, S.R. Mahadeva Prasanna, Rohit Sinha

Audio Replay Attack Detection Using High-Frequency Features
Marcin Witkowski, Stanisław Kacprzak, Piotr Żelasko, Konrad Kowalczyk, Jakub Gałka

Feature Selection Based on CQCCs for Automatic Speaker Verification Spoofing
Xianliang Wang, Yanhong Xiao, Xuan Zhu

INTERSPEECH 2017

Download link for accepted papers

http://www.isca-speech.org/archive/Interspeech_2017/

Special Session: Interspeech 2017 Automatic Speaker Verification Spoofing and Countermeasures Challenge 2

Audio Replay Attack Detection with Deep Learning Frameworks
Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, Vadim Shchemelinin

Ensemble Learning for Countermeasure of Audio Replay Spoofing Attack in ASVspoof2017
Zhe Ji, Zhi-Yi Li, Peng Li, Maobo An, Shengxiang Gao, Dan Wu, Faru Zhao

A Study on Replay Attack and Anti-Spoofing for Automatic Speaker Verification
Lantian Li, Yixiang Chen, Dong Wang, Thomas Fang Zheng

Replay Attack Detection Using DNN for Channel Discrimination
Parav Nagarsheth, Elie Khoury, Kailash Patil, Matt Garland

ResNet and Model Fusion for Automatic Spoofing Detection
Zhuxin Chen, Zhifeng Xie, Weibin Zhang, Xiangmin Xu

SFF Anti-Spoofer: IIIT-H Submission for Automatic Speaker Verification Spoofing and Countermeasures Challenge 2017
K.N.R.K. Raju Alluri, Sivanand Achanta, Sudarsana Reddy Kadiri, Suryakanth V. Gangashetty, Anil Kumar Vuppala

Summary and Conclusions

- ASV: Debut in smartphone
- NO standard databases for twins and mimics
- Same features do not perform uniformly on all the spoof attack
- Most of the participants in ASV Spoof 2017 Challenge achieved good results than the given baseline system (CQCC)
- Need of generalized countermeasure for all spoofing attacks
- There is still a long way to go towards a real generalized countermeasure

Future Research Directions

- Generalised countermeasures
- Speaker - dependent countermeasures
- Use of both direct and physical access
- Signal degradation conditions
- Combined spoofing attacks and fused countermeasures
- Noise and channel variability
- ASV Spoof 2019 ?

A (possible) special session at INTERSPEECH 2018

- http://vc-challenge.org/

Acknowledgements

- Authorities of DA-IICT, Gandhinagar, India and NUS, Singapore.
- Organizers of APSIPA ASC 2017
- Organizers of ASV Spoof 2015 and 2017 Challenge.
- Department of Electronics and Information Technology (DeitY), New Delhi, Govt. of India for their kind support to carry out research work.
- University Grants Commission (UGC) for providing Rajiv Gandhi National Fellowship (RGNF)
- All members of Speech Research Lab of DA-IICT, Gandhinagar.

Speech Research Group at DA-IICT


Thank you