A deep learning based framework for converting sign language to emotional speech

Nan Song* and Hongwu Yang*†‡ and Pengpeng Zhi*
* College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China
† Engineering Research Center of Gansu Province for Intelligent Information Technology and Application, Lanzhou 730070, China
‡ National and Provincial Joint Engineering Laboratory of Learning Analysis Technology in Online Education, Lanzhou 730070, China
E-mail: yanghw@nwnu.edu.cn

***Abstract*—This paper proposes a framework for converting sign language to emotional speech by deep learning. We firstly adopt a deep neural network (DNN) model to extract the features of sign language and facial expression. Then we train two support vector machines (SVM) to classify the sign language and facial expression for recognizing the text of sign language and emotional tags of facial expression. We also train a set of DNN-based emotional speech acoustic models by speaker adaptive training with an multi-speaker emotional speech corpus. Finally, we select the DNN-based emotional speech acoustic models with emotion tags to synthesize emotional speech from the text recognized from the sign language. Objective tests show that the recognition rate for static sign language is 90.7%. The recognition rate of facial expression achieves 94.6% on the extended Cohn-Kanade database (CK+) and 80.3% on the Japanese Female Facial Expression (JAFFE) database respectively. Subjective evaluation demonstrates that synthesized emotional speech can get 4.2 of the emotional mean opinion score. The pleasure-arousal-dominance (PAD) tree dimensional emotion model is employed to evaluate the PAD values for both facial expression and synthesized emotional speech. Results show that the PAD values of facial expression are close to the PAD values of synthesized emotional speech. This means that the synthesized emotional speech can express the emotions of facial expression.**

I. Introduction

Sign language is one of the most important communication methods among speech impaired and normal person. Since researches on sign language recognition have been widely concerned [1] , it have been a research hotspot in human-computer interaction. The wearing devices such as the data gloves are used for sign language recognition [2] in the very beginning. In recent years, machine learning methods such as Hidden Markov model (HMM) [3], Back Propagation (BP) neural network [4] and Support Vector Machine (SVM) [5] are introduced to sign language recognition for improving the recognition rate of sign language by computer vision-based approaches. At present, because deep learning becomes an important machine learning method, it also has been applied to sign language recognition [6] and has greatly improved the sign language recognition rate. Because emotion expression can make communication more accurate, facial expressions also play an important role in the communication between normal persons and speech impairments during their daily life. Therefore, facial expression recognition technologies also have a rapidly developing in the area of image processing. Several different methods such as SVM [7], Adaboost [8], Local Binary Pattern (LBP), Principal Components Analysis (PCA) [9] and Deep Learning [10] are applied to the facial expression recognition. At the same time, Deep Neural Network (DNN) - based speech synthesis methods are widely used in the field of speech synthesis [11], [12], by which the text information can be converted into speech. However, most of the existing researches mainly focus on sign language recognition, facial expression recognition and emotional speech synthesis individually. Some researches adopt information fusion methods to integrate facial expressions, body language and voice information to realize the emotion recognition under the multi-modal fusion [13]. In the study, the sign language recognition and voice information are fused to command the robot [14] and to achieve the wheelchair navigation control of robot [15]. These studies show that multi-modal information fusion has gradually become a trend. We have realized an emotional speech synthesis based on the HMM method. In order to improve the quality of emotional speech synthesis, we use the DNN-based method for emotional speech synthesis. We also realized a sign language to speech synthesis conversion [16], [17]. But the synthesized speech does not include changes in emotion so that ignore the speech expression of the deaf and mute people that make communication becomes easy for listeners to understand confusion.

To overcome the deficiency of current researches on sign language to emotional speech conversion, the paper proposed a DNN-based framework that combines sign language recognition, facial expression recognition and emotional speech synthesis together to realize the sign language to emotional speech conversion. We firstly use the static sign language recognition to obtain the text of sign language. At the same time, the expressed facial emotion information is obtained by the facial expression recognition. We also trained a set of DNN-based emotional speech acoustic models. The text and emotion information obtained from sign language recognition and facial expression recognition are finally converted into corresponding emotional speech with emotional speech synthesis.

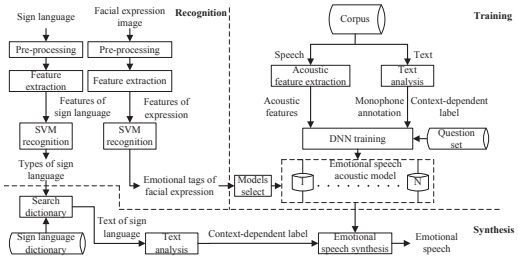

Fig. 1. The framework of sign language to emotional speech conversion for speech disorder.

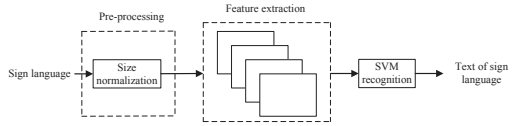

Fig. 2. Sign language recognition.

Fig. 3. Facial expression recognition.

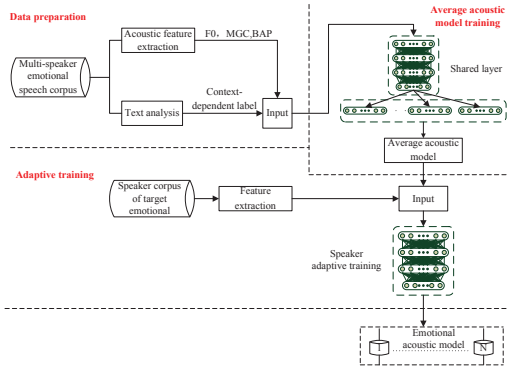

Fig. 4. Acoustic model training of emotional speech.

II. Framework of Sign Language to Emotional Speech Conversion

The framework of the proposed sign language to emotional speech conversion consists of three parts including sign language and facial expression recognition, acoustic models of emotional speech training and emotional speech synthesis, as shown in Fig. 1. In the recognition step, the categories of sign language are obtained by sign language recognition and the emotion tags are obtained by facial expression recognition. In the acoustic model training step, we use an emotional speech corpus to train a set of emotional acoustic models with a DNN-based emotional speech synthesis framework. In the speech synthesis step, the text is obtained from the recognized categories of sign language by searching a defined sign language-text dictionary. Then we generate the context-dependent label from text of sign language by a text analyzer. Meanwhile, the emotional speech acoustic models are selected by using the emotion tags that is recognized from facial expression. Finally the context-dependent labels are applied on the emotional acoustic models to generate acoustic parameters for synthesizing emotional speech.

A. Sign language Recognition

Sign language recognition includes pre-processing, feature extraction and SVM recognition. In the pre-processing stage, the size of the sign language image is normalized to 96×96. In the feature extraction stage, the pre-processed images are passed through the DNN model to obtain 128-dimensional sign language features on each input image. In the recognition phase, the SVM is used to classify the types of static sign language as shown in Figure 2.

B. Facial Expression Recognition

Facial expression recognition step includes pre-processing, feature extraction and SVM recognition as shown in Fig. 3. In the pre-processing stage, we process some unimportant background information in the original image that may affect the result of feature extraction. First of all, the original input image is detected by a detector with 68 facial landmark points and is adjusted to the edge of the landmark. Then the image is trimmed with the complete facial expression information. Some non-specific information is deleted after the image is cut off to obtain a 96×96 image as input of the neural network model. In the feature extraction stage, we use a DNN model for feature extraction to extract 128-dimensional features from each of the imported emoticons. After feature extraction, a trained SVM classifier is used to classify the facial expression with the extracted features to get emotion tags corresponding to facial expressions.

III. Training Acoustic Models of Emotional Speech

The paper trained a set of DNN-based emotional acoustic models with the multi-speaker emotional speech corpus as shown in Fig. 4. The framework consists of three parts including data preparation, average voice model (AVM) training and speaker adaptation.

A. Data preparation

In the data preparation phase, we use the WORLD vocoder [18] to extract the acoustic features from the multi-speaker emotional speech corpus. The acoustic features include the fundamental frequency (F0), the generalized Mel-generalized Cepstral (MGC), and the Band a periodical (BAP).

The paper adopt the initials and the finals as the unit of speech synthesis. A text analyzer is employed to obtain the initial, final, prosodic struct, word, and sentence information, which are used to form the context-dependent labels, through text normalization, grammar analysis, prosodic analysis, and phonological conversion by dictionary and grammar rule. The context-dependent label provides context information of the

speech synthesis units including the unit layer, syllable layer, word layer, prosodic word layer, phrase layer, and sentence layer [19].

B. Training average voice model

During the training AVM, the paper trained a set of DNN models as the emotional AVM by using the linguistic features (binary and digital) as input and acoustic features as output. The linguistic features were obtained from context-dependent label of the text corpus. The acoustic features were extracted from speech corpus including MGC, BAP, F0 and voice/unvoiced (V/UV). During training, the DNN models share various hidden layers between different emotional speakers to model its language parameters. Duration models and acoustic models were trained by a stochastic gradient descent (SGD) [20] of back propagation (BP) algorithm. Finally, a set of speaker independent AVM were trained by the multi-speaker corpus.

The paper used a DNN structure including an input layer, a output layer and six hidden layer to train the AVM. The $tanh$ is used in the hidden layer and the linear activation function is used in the output layer. All speakers' training corpus share the hidden layer, so the hidden layer is a global linguistic feature mapping shared by all speakers. Each speaker has its own regression layer to model its own specific acoustic space. After multiple batches of SGD training, a set of optimal multi-speaker AVM model (average duration models and average acoustic feature models) is obtained.

During the training of the model, the DNN structure adopts a non-linear function to model the non-linear relationship between the linguistic features and acoustic features. Each hidden layer k uses the output $\mathrm{h}^k - 1$ of the previous layer to calculate the output vector h^k. $x = h^0$ is the input of the model. h^k can be defined as Equation 1,

$$\mathrm{h}^k = \tanh(b^k + w^k h^{k-1}) \tag{1}$$

Where, the b^k is offset vector and w^k is weight matrix. The mapping function uses $tanh$ in an node manner and can be replaced by other saturated non-linear functions such as Sigmod-shaped functions. The top level h^l and the supervised target output y are combined into a generally convex loss function $L(h^l, y)$.

The output layer of DNN uses the linear regression function as Equation 2,

$$\mathrm{h}_i^l = \frac{e^{b_i^l} + w_i^l h^{l-1}}{\Sigma_j e^{b_j^l + w_j^l h^{l-1}}} \tag{2}$$

Where w_i^l is the i line of w^l, h_i^l is positive, and $\Sigma_i h_i^l = 1$.

C. Speaker adaptation

In the speaker adaptation stage, a small corpus of the targeted emotions of the target speaker is used to extract acoustic features in the same way as AVM training, including F0, MGC, BAP and V/UV. Firstly, the speaker adaptation is performed by multi-speaker AVM model with the DNN models of the target emotional speaker to obtain a set of speaker-dependent adaptation models including duration models and acoustic models. The speaker-dependent model has the same DNN structure as the AVM, using six hidden layer structures, and the mapping function is the same as AVM. At the same time, The maximum likelihood parameter generation (MLPG) algorithm [21] is used to generate the speech parameters from emotional acoustic models.

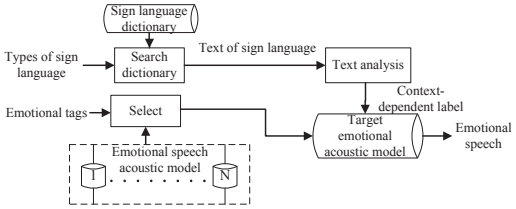

Fig. 5. Framework of sign language to emotional speech conversion.

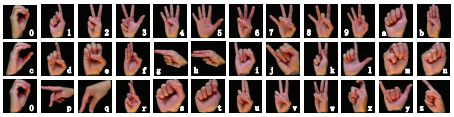

Fig. 6. Examples of 36 kinds of static sign language.

IV. Sign Language to Emotional Speech Conversion

Sign language to emotional speech conversion process is shown in Fig. 5. We have designed a sign language dictionary based on the meaning of the sign language types defined in "American Sign Language Dictionary" [22], which gives the semantic text corresponding to each sign language. In the process of conversion from sign language to emotional speech, a sign language category is obtained by sign language recognition. Then the sign language dictionary is searched to obtain the text of recognized sign language. Finally, text analysis is performed on the text of sign language to obtain the context-dependent label that is used to generate linguistic features as the input of the models. A set of speech assessment methods phonetic alphabet (SAMPA) is designed for labeling initials and finals. At the same time, the emotion tags are obtained by facial expression recognition to select the emotional acoustic model for synthesizing emotional speech.

V. Experimental Results

A. Sign Language Recognition

1) Sign language data: The sign languages used in the experiment contains a total of 2515 images of 36 static sign languages [23]. Before the experiment, the sign language images were pre-processed so that the size of each image was 96×96. An example of the pre-defined 36 static sign languages is shown in Figure 6.

2) Sign language recognition rate: We performed the SVM recognition experiments on the 36 static sign languages as shown in Figure 6. In the sign language recognition, five cross-validation experiments numbered from 1 to 5 respectively were performed on the test set to obtain the recognition rates. To compare our method with others, we also use the CNN

TABLE I
THE RECOGNITION RATE OF SIGN LANGUAGE RECOGNITION UNDER DIFFERENT TRAINING SETS.

Method	1	2	3	4	5	Average
CNN	89.5%	89.8%	90.6%	90.4%	89.3%	89.9%
DNN	91.4%	91.4%	89.7%	91.1%	89.9%	90.7%

Fig. 7. Examples of facial expression database.

TABLE II
FACIAL EXPRESSION RECOGNITION RATE UNDER DIFFERENT DATABASE

Emotion	CK+	JAFFE (Original)	JAFFE (Transform)
Angry	96.7%	86.6%	96.6%
Disgust	93.1%	82.7%	93.0%
Fear	97.6%	75.0%	96.8%
Happy	98.3%	80.6%	87.1%
Sad	92.9%	83.8%	96.8%
Surprise	89.0%	73.3%	98.3%
Average	94.6%	80.3%	94.8%

model and softmax classification to perform sign language recognition on same test sets. The sign language recognition rates are compared in Table I.

As can be seen from Table I, although the experimental sign language library contains multiple images of different human beings, the experimental method can still obtain a better recognition rate.

B. Facial expression recognition

1) Facial expression database data: In this paper, the extended Cohn-Kanade database (CK+) [24] and the Japanese Female Facial Expression (JAFFE) database [25] are used to train and test facial expression recognition. Each sequence image in the CK+ database begins with a neutral expression and ends with the emotional peak. The experimental database contains eight emotion categories. Contempt and neutral expression images are not used in the experiment. Images with obvious facial feature information are selected as a sample set. Six of the seven emoticons in the JAFFE database were tested without the use of neutral emoticons, each of which had an expression size of 256×256. Some examples of facial images are shown in Fig. 7.

2) Facial expression recognition rate: We carried out 5 cross-validation experiments on the CK+ database to obtain the corresponding recognition rates for the 6 facial expressions. We conducted three cross-validation experiments on the JAFFE database and obtained the corresponding recognition rates for the six facial expressions. The facial expression recognition results are shown in Table II.

It can be seen from Table II that the recognition rate on JAFFE in the original database is lower than that on CK+ database, mainly because the number of facial expression images in JAFFE original database is less than the number of facial expression images in CK+ database. In view of the above problems, the number of experimental images is increased by inverting the JAFFE database images in the experiment. The corresponding recognition rate of cross-validation experiments on six facial expressions after increasing 3 times of the database are shown in Table II.

Fig. 8. Examples of Convolution layer visualization.

3) Image visualization: In this paper, nn4.small2 neural network model [26] is used to extract the sign language image features and facial image features. Fig. 8 shows the output of a cropped sign language image and facial image after the first layer of the model. This figure shows the first 64 convolutions of all the filters of the input image. The network model definition is shown in Table III. The model including 8 Inception modules. Where "#3 $\times$ 3 reduce" and "#5 $\times$ 5 reduce" stands for the number of 1×1 filters in the reduction layer used before the 3×3 and 5×5 convolutions. there is a dimensionality reduction after the pooling it is denoted with p. The max pooling is denoted with m.

C. Emotional speech synthesis

1) Speech data: In our work, the emotional speech corpus contains 6 kinds of emotional speech recorded from 9 female speakers. Each speaker records 100 sentences for each emotion. The sample rate of the speech file is 16 kHz. 60-dimensional generalized Mel-Frequency Cepstral coefficients, 5-band non-periodic components and log fundamental frequency are extracted to compose a feature vectors. We use the WORLD vocoder to generate speech in steps of 5 milliseconds.

2) Emotion similarity evaluation: We use the emotional DMOS(EMOS) test to evaluate the emotional expression of synthesized emotional speech. 20 sentences of one emotion are synthesized for evaluation. The 10 subjects are played 20 original emotion speech files as a reference, and then we play 6 kinds of synthesized emotional speech files in sequence according to the emotion order. In the evaluation scoring process, it is conducted in accordance with the order in which the voices are played, and the subjects are asked to refer to the emotional expression experience in real life to score the

TABLE III
DEFINITION OF DEEP NEURAL NETWORK

type	Output size	#1 × 1	#3 × 3 reduce	#3 × 3	#5 × 5 reduce	#5 × 5	Pool proj
conv1 (7 × 7 × 3,2)	48 × 48 × 64						
max pool+norm	24 × 24 × 64						m 3 × 3,2
inception(2)	24 × 24 × 192		64	192			
norm+max pool	12 × 12 × 192						m 3 × 3,2
inception (3a)	12 × 12 × 256	64	96	128	16	32	m,32p
inception (3b)	12 × 12 × 320	64	96	128	32	64	ℓ_2,64p
inception (3c)	6 × 6 × 640		128	256,2	32	64,2	m 3 × 3,2
inception (4a)	6 × 6 × 640	256	96	192	32	64	ℓ_2,128p
inception (4e)	3 × 3 × 1024		160	256,2	64	128,2	m 3 × 3,2
inception (5a)	3 × 3 × 736	256	96	384			ℓ_2,96p
inception (5b)	3 × 3 × 736	256	96	384			m,96p
avg pool	736						
linear	128						
ℓ_2 normalization	128						

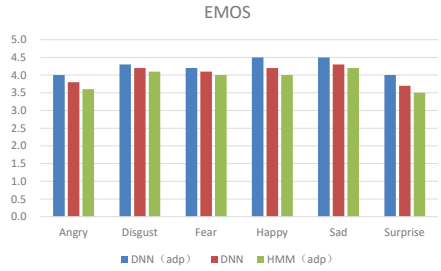

Fig. 9. The EMOS scores of synthesized emotional speech.

TABLE IV
THE RMSE OF DURATION AND FUNDAMENTAL FREQUENCY BETWEEN THE SYNTHESIZED EMOTIONAL SPEECH AND ORIGINAL EMOTIONAL SPEECH.

Emotional	F0/Hz			Dur/s		
	HMM (adp)	DNN	DNN (adp)	HMM (adp)	DNN	DNN (adp)
Angry	46.5	32.7	20.4	0.116	0.109	0.089
Disgust	41.7	34.5	28.7	0.141	0.135	0.092
Fear	48.5	38.7	22.7	0.131	0.128	0.086
Happy	52.1	45.1	21.8	0.104	0.103	0.102
Sad	38.5	35.4	22.1	0.230	0.228	0.202
Surprise	46.3	41.2	24.2	0.169	0.153	0.132

emotional similarity for each synthesized sentence according to the 5-point system. The results are shown in Fig. 9.

As can be seen from Fig. 9, the EMOS score of the emotional speech synthesized by the multi-emotional speaker adaptation method of DNN is higher than the EMOS score of the emotional speech synthesized by the DNN method and the HMM adaptation method. It also shows that the emotional speech synthesized by this method has better preference and natural emotion.

3) Objective evaluation: This paper calculates the Root Mean Square Error (RMSE) between the original speech and the synthesized speech in terms of duration and fundamental frequency. The results are shown in Table IV. From the Table IV, we can see that the RMSE of the DNN adaptation method is smaller than that of the DNN method and the HMM adaptation method. This means that the emotional speech synthesized by this method is closer to the original emotional speech. Therefore the synthesized emotional speech has better speech quality.

D. PAD Evaluation between Emotional Picture and Emotional Speech

In order to further evaluate the emotional expression similarity between the synthesized emotional speech and the original facial expression, we use the pleasure-arousal-dominance (PAD) three-dimensional emotional model to compare the difference of PAD values between the facial expression of pictures and emotional expression of the synthesized speech. In the paper, we use the abbreviated PAD emotion scale [27] to scale the PAD values of the facial expression of images and their corresponding synthesized emotional speech. First, all facial expression images were played at random to the subjects. The 10 subjects completed the PAD mood scale according to the emotional state of the images they felt when they observed the picture. Then the synthesized emotional speech is played randomly, which also requires the subjects to complete the PAD mood scale according to the emotional state that they feel when they listen to the emotional speech. Finally, the Euclidean distance of the PAD values between the facial expression picture and the emotional speech in the same emotional state is calculated. The results of the evaluation are shown in Table V. From Table V, we can see that the Euclidean distance of PAD value of facial expression and emotional speech is smaller in the same emotional state, indicating that the synthesized emotional speech can accurately reproduce the emotional state of facial expression. At the same time, The DNN speech PAD evaluation is better than the HMM speech evaluation. DNN emotional speech is closer to the emotion expressed by the expression image.

VI. CONCLUSION

In this paper, we propose a framework of sign language to emotion speech conversion that integrate the facial expression recognition. Firstly, the recognized sign languages and facial

TABLE V
PAD EVALUATION FOR FACIAL EXPRESSION AND SYNTHESIZED EMOTIONAL SPEECH.

Emotional	Picture([-1,1])			HMM Speech([-1,1])			DNN Speech([-1,1])			HMM Euclidean distance	DNN Euclidean distance
	P	A	D	P	A	D	P	A	D		
Angry	-0.84	0.64	0.82	-0.83	0.60	0.84	-0.83	0.62	0.83	0.05	0.02
Disgust	-0.50	0.26	0.39	-0.46	0.19	0.39	-0.48	0.21	0.39	0.08	0.05
Fear	-0.39	0.62	-0.63	-0.35	0.67	-0.67	-0.37	0.65	-0.65	0.08	0.04
Happy	0.67	0.78	0.37	0.68	0.83	0.50	0.68	0.81	0.45	0.14	0.09
Sad	-0.24	-0.40	-0.75	-0.26	-0.35	-0.72	-0.26	-0.38	-0.73	0.06	0.03
Surprise	0.23	0.60	0.03	0.26	0.70	0.03	0.25	0.65	0.03	0.10	0.05

expression information are converted to the context-dependent label of the sign language and the corresponding emotional tags. At the same time, a DNN-based emotional speech synthesis is trained through an emotional corpus. Finally, emotional speech synthesis is done according to the emotional tags and the context-dependent labels of sign language text, so as to achieve the conversion from sign language and facial expression to emotional speech. The experimental results show that the average EMOS score of converted emotional speech is 4.2. At the same time, we use the PAD three-dimensional model to assess the expression of facial expression and synthesized emotional speech. The results show that the Euclidean distance of the PAD values are small between the facial expression of pictures and emotional expression of synthesized speech that demonstrate the converted emotional speech can express the emotional state of facial expression. Further work will use other deep learning method to optimize sign language recognition, facial expression recognition and emotional speech synthesis to improve the recognition rate and the quality of synthesized emotional speech.

ACKNOWLEDGMENT

The research leading to these results was partly funded by the National Natural Science Foundation of China (Grant No. 11664036, 61263036), High School Science and Technology Innovation Team Project of Gansu (2017C-03), Natural Science Foundation of Gansu (Grant No. 1506RJYA126), and Student Innovation Project of Northwest Normal University (CX2018Y162).

REFERENCES

[1] E. A. Kalsh, and N. S. Garewal, "Sign Language Recognition System," *International journal of computational engineering research,* vol. 3, no. 6, pp. 15–21, 2013.

[2] K. Assaleh, T. Shanableh, and M. Zourob, "Low complexity classification system for glove-based arabic sign language recognition," *International Conference on Neural Information Processing,* pp. 262–268, Springer, November 2012.

[3] V. Godoy, A. S. Britto, A. Koerich, J. Facon, and L. E. Oliveira, "An HMM-based gesture recognition method trained on few samples," *International Conference on Tools with Artificial Intelligence (ICTAI),* pp. 640–646, IEEE, November 2014.

[4] Z. Q. Yang, and G. Sun, "Gesture recognition based on quantum-behaved particle swarm optimization of back propagation neural network," *Journal of Computer Applications,* vol. 34, pp. 137–140, 2014.

[5] D. K. Ghosh, and S. Ari, "Static hand gesture recognition using mixture of features and SVM classifier," *International Conference on Communication Systems and Network Technologies (CSNT),* pp. 1094–1099, IEEE, April 2015.

[6] O. K. Oyedotun, and A. Khashman, "Deep learning in vision-based static hand gesture recognition," *Neural Computing and Applications,* vol. 28, no. 12, pp. 3941–3951, 2017.

[7] C. C. Hsieh, M. H. Hsih, M. K. Jiang, Y. M. Cheng, and E. H. Liang, "Effective semantic features for facial expressions recognition using SVM," *Multimedia Tools and Applications,* vol. 75, no. 11, pp. 6663–6682, 2016.

[8] S. Prabhakar, J. Sharma, and V. Gupta, "Facial expression recognition in video using adaboost and SVM," *International Journal of Computer Applications,* vol. 3613, no. 1, pp. 672–675, 2014.

[9] M. Abdulrahman, T. R. Gwadabe, F. J. Abdu, and A. Eleyan, "Gabor wavelet transform based facial expression recognition using PCA and LBP," in *Signal Processing and Communications Applications Conference (SIU),* pp. 2265–2268, IEEE, April 2014.

[10] X. Zhao, X. Shi, and S. Zhang, "Facial expression recognition via deep learning," *IETE Technical Review,* vol. 32, no. 5, pp. 347–355, 2015.

[11] N. Hojo, Y. Ijima, and H. Mizuno, "DNN-Based Speech Synthesis Using Speaker Codes," *IEICE TRANSACTIONS on Information and Systems,* vol. 101, no. 2, pp. 462-472, 2018.

[12] B. Potard, P. Motlicek, and D. Imseng, "Preliminary work on speaker adaptation for dnn-based speech synthesis," *Idiap,* 2015.

[13] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, and K. Karpouzis, "Multimodal emotion recognition from expressive faces, body gestures and speech," in *IFIP International Conference on Artificial Intelligence Applications and Innovations,* pp. 375–388, Springer, September 2007.

[14] B. Burger, I. Ferrané, F. Lerasle, and G. Infantes, "Two-handed gesture recognition and fusion with speech to command a robot," *Autonomous Robots,* vol. 32, no. 2, pp. 129–147, 2012.

[15] D. A. Sinyukov, R. Li, N. W. Otero, R. Gao, and T. Padir, "Augmenting a voice and facial expression control of a robotic wheelchair with assistive navigation," *Systems, Man and Cybernetics (SMC),* pp. 1088–1094, IEEE, October 2014.

[16] H. Yang, X. An, D. Pei, and Y. Liu, "Towards realizing gesture-to-speech conversion with a HMM-based bilingual speech synthesis system," *International Conference on Orange Technologies (ICOT),* pp. 97–100, IEEE, September 2014.

[17] X. An, H. Yang, and Z. Gan, "Towards Realizing Sign Language-to-Speech Conversion by Combining Deep Learning and Statistical Parametric Speech Synthesis," *International Conference of Young Computer Scientists, Engineers and Educators (ICYCSEE),* pp. 678–690, Springer, August 2016.

[18] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *Ieice Transactions on Information and Systems,* vol. 99, no. 7, pp. 1877–1884, 2016.

[19] H. Yang, K. Oura, H. Wang, Z. Gan, and K. Tokuda, "Using speaker adaptive training to realize Mandarin-Tibetan cross-lingual speech synthesis," *Multimedia Tools and Applications,* vol. 74, no. 22, pp. 9927–9942, 2015.

[20] L. Deng, and D. Yu, "Deep learning: methods and applications," *Foundations and Trends®in Signal Processing,* vol. 7, no. 3, pp. 197–387, 2014.

[21] H. T. Hwang, Y. Tsao, H. M. Wang, Y. R. Wang, and S. H. Chen, "A probabilistic interpretation for artificial neural network-based voice conversion," *Signal and Information Processing Association Annual Summit and Conference (APSIPA),* pp. 552-558, IEEE, December 2015.

[22] E. Costello, *American sign language dictionary,* Random House Reference &, 2008.

[23] https://github.com/snrao310/ASL-Finger-Spelling-Recognition [Date of Access: October 17, 2017].
[24] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews "Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," *Computer Vision and Pattern Recognition Workshops (CVPRW),* pp. 94–101, IEEE, June 2010.
[25] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Automatic Face and Gesture Recognition,* pp. 200–205, IEEE, April 1998.
[26] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science,* 2016.
[27] L. Xiaoming, F. Xiaolan, and D. Guofeng, "Preliminary Application of the Abbreviated PAD Emotion Scale to Chinese Undergraduate," *Chinese Mental Health Journal,* vol. 22, no. 5, pp. 327–329, 2008.