Chinese Journal of Electronics
Vol.27, No.2, Mar. 2018

Speech Magnitude Spectrum Reconstruction from MFCCs Using Deep Neural Network*

JIANG Wenbin[1], LIU Peilin[1] and WEN Fei[1,2]

(1. *Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*)

(2. *Air Control and Navigation Institution, Air Force Engineering University, Xi'an 710000, China*)

Abstract — This work proposes a Deep neural network (DNN) based method for reconstructing speech magnitude spectrum from Mel-frequency cepstral coefficients (MFCCs). We train a DNN using MFCC vectors as input and the corresponding speech magnitude spectrum as desired output. Exploiting the strong inference power of DNN, the proposed method has the capability to accurately estimate the speech magnitude spectrum even from truncated MFCC vectors. Experiments on TIMIT corpus demonstrate that the proposed method achieves significantly better performance compared with traditional methods.

Key words — Deep neural network (DNN), Mel-frequency cepstral coefficients (MFCCs), Spectrum reconstruction, Speech reconstruction.

I. Introduction

Mel-frequency cepstral coefficients (MFCCs) are widely used in speech signal processing, such as speech recognition[1,2] and speaker identification[3,4]. In some practical systems, such as Distributed speech recognition (DSR)[5,6], reconstructing the original speech signal from a stream of MFCC vectors is required. In the reconstruction process, inverting the MFCCs to speech spectrum is generally challenging, because a Mel-filter bank is applied to the original spectrum in generating the MFCCs, which makes the inversion problem under-determined.

In Refs.[7–9], the effect of the Mel-filter bank is equalized by subtracting the cepstral representation of the Mel-filter bank from the original speech MFCC vector, and then a high resolution Inverse discrete cosine transform (IDCT) is utilized to obtain an interpolated magnitude spectrum. This method can reasonably reconstruct the envelope of the original magnitude spectrum, but the spectral valleys may be distorted seriously. More recently, in Ref.[10], a Moore-Penrose pseudo-inverse of the Mel-scale weight function is used to get a least-squares solution. This method is efficient for removing the effect of the Mel-filter bank, but the reconstructed magnitude spectrum may be negative.

The emerging Deep neural network (DNN)[11,12] has recently achieved a great success in speech recognition[1], speech enhancement[13], *etc.* Theoretically, a DNN can be seen as a complicated non-linear function that maps the input data to the desired output data[14]. The inversion of MFCCs to magnitude spectrum consists of IDCT, Exponent (EXP), and the removal of Mel-filtering operations. That procedure can be functionally regarded as a complicated non-linear function, which motivates us to learn a DNN for the MFCCs inversion task.

In this paper, we propose a DNN-based method to inverse MFCC vectors to speech magnitude spectrum. The DNN is fed with MFCC vectors and optimized by minimizing the Minimum mean squared error (MMSE) between the DNNs output and desired speech magnitude spectrum. The main advantage of the DNN-based method over traditional methods is that, it learns a non-linear function to effectively remove the effect of the Mel-filter bank and has a strong inference power to accurately estimate the magnitude spectrum. Experimental results show that the proposed DNN-based method significantly outperforms traditional reconstruction methods.

II. Mel-Frequency Cepstrum Coefficients

MFCCs are defined as special cepstrum that a set of filters are applied to the power spectrum prior to the log and Discrete cosine transform (DCT) operations. These filters are designed based on human perception of pitch and are most commonly implemented in the form of a bank of triangular filters in Mel-scale[15]. The MFCCs,

*Manuscript Received Apr. 28, 2016; Accepted Jan. 6, 2017. This work is supported by the National Natural Science Foundation of China (No.61401501).

© 2018 Chinese Institute of Electronics. DOI:10.1049/cje.2017.09.018

$\boldsymbol{c}$, of the t-th frame speech, $\boldsymbol{s}_t = [s_t(0), s_t(1), \cdots, s_t(L-1)]^{\mathrm{T}}$($L$ is the frame length), is computed as (the subscript t is dropped to simplify notation)

$$\boldsymbol{c} = \mathrm{DCT}\{\log(\boldsymbol{y})\} = \mathrm{DCT}\{\log(\boldsymbol{\Phi x})\} \tag{1}$$

where $\boldsymbol{y} = \boldsymbol{\Phi x}$ is the Mel-filtered spectrum, $\boldsymbol{x} = |\mathrm{DFT}(\boldsymbol{s})|^2$ is the power spectrum of $\boldsymbol{s}$, and $\boldsymbol{\Phi} \in \mathbb{R}^{J\times L}$ (J is the number of Mel-filters) is a Mel-filter matrix in which each row represents a triangular window. For MFCCs, the recovery of $\boldsymbol{x} \in \mathbb{R}^{L\times 1}$ from $\boldsymbol{y} = \boldsymbol{\Phi x} \in \mathbb{R}^{J\times 1}$ is generally under-determined, since $J < L$.

In Eq.(1), the DCT and log operations are directly invertible, whereas the operation of applying the Mel-scale filter is not. A least-squares method has been used in Ref.[10] to estimate $\boldsymbol{x}$ from $\boldsymbol{y}$ as

$$\hat{\boldsymbol{x}} = \boldsymbol{\Phi}^{\dagger}\boldsymbol{y} \tag{2}$$

where

$$\boldsymbol{\Phi}^{\dagger} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}$$

is the Moore-Penrose pseudo inverse of $\boldsymbol{\Phi}$. In fact, Eq.(2) is the solution of the optimization problem which minimizes the Euclidean norm $\|\tilde{\boldsymbol{x}}\|_2$ subject to $\boldsymbol{y} = \boldsymbol{\Phi}\tilde{\boldsymbol{x}}$. The main problem of this method is that the estimated $\hat{\boldsymbol{x}}$ might be negative. Moreover, the minimization objective of this method is not as reasonable as the minimization objective $\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2$ for the inversion task.

Another efficient inversion method is to use equalization and interpolation operations[7,8]. In this procedure, firstly, a cepstral representation of the Mel-filters is calculated as

$$\boldsymbol{c_w} = \mathrm{DCT}\{\log(\boldsymbol{w})\} \tag{3}$$

where $\boldsymbol{w}$ is a vector consisted of the areas of the Mel-spaced triangular windows. Then, an equalized MFCC vector, $\boldsymbol{c}'$, is computed by

$$\boldsymbol{c}' = \boldsymbol{c} - \boldsymbol{c_w} \tag{4}$$

Finally, a high resolution (3933 dimensional) IDCT is applied to $\boldsymbol{c}'$ to get an interpolated magnitude spectrum. This method can reasonably reconstruct the envelope of the original magnitude spectrum, but it is still not sufficiently accurate in recovering the spectral valleys.

III. DNN Based Speech Magnitude Spectrum Reconstruction from MFCCs

In this section, we introduce the framework of the proposed DNN-based spectrum reconstruction method, and describe the training procedure of the DNN.

1. Framework of the proposed method

The framework of the proposed DNN-based spectrum reconstruction method is illustrated in Fig.1. $\boldsymbol{x}$ is the power spectrum of a speech frame, and $\boldsymbol{c}$ is the corresponding MFCC vector calculated by a series of MEL (*i.e.*, applying the Mel-filtering), log, and DCT operations. The corresponding inversion operations should include IDCT, EXP, and MEL^{-1} (*i.e.*, removing the effect of Mel-filtering). As discussed in Section II, the recovery of the original spectrum from the Mel-filtered spectrum is an under-determined problem. Thus, it is generally difficult to obtain the original spectrum $\boldsymbol{x}$. In an attempt to accurately reconstruct $\boldsymbol{x}$, we can learn a DNN that minimizes the square error between the desired spectrum $\boldsymbol{x}$ and the DNNs output $\tilde{\boldsymbol{x}}$.

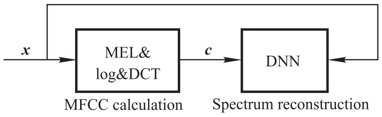


Fig. 1. Framework of the proposed DNN-based spectrum reconstruction

As shown in Fig.1, the input of the DNN is the MFCCs, $\boldsymbol{c}$, and the desired output is $\boldsymbol{x}$. In such a manner, we expect the DNN to learn a compositional functionality of the MEL^{-1}, IDCT and EXP functions, which well exploits specific features of speech signals.

2. Training procedure

We use Restricted Boltzmann machine (RBM)[16] to learn initial parameters of each layer of the DNN. Firstly, a Gaussian-Bernoulli RBM that has one visible layer of linear variables and one hidden layer of binary latent variables is trained by using patches of MFCC vectors as its training data. Then, a stack of Bernoulli-Bernoulli RBMs are trained by using the hidden activations of the previous RBM as its training data. Each of these RBMs is trained in an unsupervised greedy layer-wise fashion[14]. In the training, the weights and biases of each RBM are updated using the Contrastive divergence (CD) method. The learning procedure is illustrated in the left side of Fig.2.

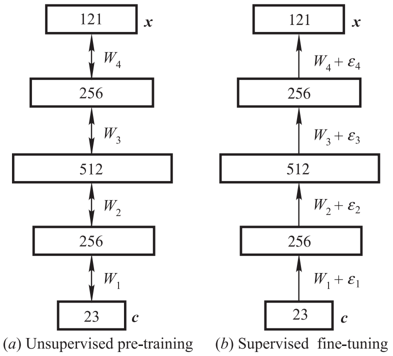


Fig. 2. (*a*) Illustration of unsupervised pre-training procedure. (*b*) Description of the fine-tuning procedure of DNN for spectrum reconstruction from MFCCs

Using these learned parameters as initialization, the DNN model is fine-tuned using the back-propagation algorithm[17], which is illustrated in the right side of Fig.2.

The parameters of this model are optimized by minimizing the average error over the training set

$$E=\frac{1}{N}\sum_{n=1}^{N}d(\boldsymbol{x}^{(n)},\tilde{\boldsymbol{x}}^{(n)})=\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(x_k^{(n)}-\tilde{x}_k^{(n)}\right)^2 \quad (5)$$

where N is the size of the training set, $\boldsymbol{x}^{(n)}$ and $\tilde{\boldsymbol{x}}^{(n)}$ denote the desired spectrum (*i.e.*, the training data) and the DNNs output of the n-th training sample, respectively, $d(\cdot)$ is a loss function that measured by square error.

The fine-tuning of the DNN model is a supervised training procedure that uses the MFCC vectors as input and the corresponding spectra as expected output. The adopted objective function is consistent with the ultimate minimization objective, $\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_2$. This makes the DNN capable to learn the desired nonlinear mapping function. Furthermore, using spectrum with full frequency bins as training data, the DNN can capture the context information between the frequency bins. In contrast, the methods in Refs.[7,8,10] do not exploit such features of speech signals.

IV. Evaluation

In this section, firstly, we describe the setup of the experiments. Then, the evaluation results of spectrum reconstruction are presented and discussed in detail. Finally, we give the results of computational complexity.

1. Setup

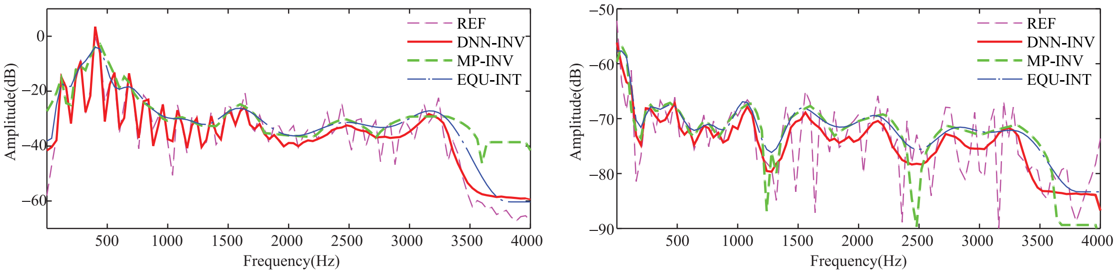
In this paper, we use TIMIT corpus[16] for the evaluation. There are totally 6300 utterances (about 5.4 hours) in this corpus. We randomly pick 5300 utterances as training set, 500 utterances as validation set, and the remainder 500 utterances as test set. The datasets for DNN training are prepared as follows. Firstly, all these speech signals are down-sampled to 8kHz and framed by a 25-ms window with 10-ms shift. Hence the length of each frame is 200 samples. Subsequently, a Discrete Fourier transform (DFT) is used to compute power spectra of these framed signals. Then, a series of Mel-filtering, log, and DCT operations are applied to calculate MFCC vectors. Finally, these MFCC vectors and power spectra are normalized to have zero mean and unit variance, respectively. The normalized MFCC vectors along with the corresponding power spectra constitute a pair of datasets for the DNN training. In the recovery stage, the outputs of the DNN are de-normalized to get the final power spectra.

The settings for the DNN training are as follows: the DNN architecture is 23(or 13)-256-512-256-121. The datasets are divided into small "mini-batches" of 128 cases to speed-up the training. In the pre-training of each RBM, the momentum is set to 0.5, the learning rate is 0.001, and the number of epoch is 20. In the fine-tuning, the momentum is set to 0.9, and the initial learning rate is set to 0.1. The learning rate is gradually reduced by a factor of 0.9 when the decrease of the validation error between two consecutive epochs is less than 0.02%. The training process is stopped when the validation error decrease is less than 0.01%. We implement the DNN training algorithm in Python with the help of "Theano"[19], and carry out the training procedures on GTX Titan X GPU.

2. Qualitative evaluation of reconstruction performance

We compare the proposed DNN-based MFCCs inversion (DNN-INV) method with two popular methods, Moore-Penrose pseudo inverse (MP-INV)[10] and Equalization-interpolation (EQU-INT)[7,8]. Fig.3 shows the magnitude spectra reconstructed from a 23-D MFCC vector by DNN-INV, MP-INV, and EQU-INT for voiced frame and unvoiced frame. The original magnitude spectrum (REF) is also plotted for comparison. It can be clearly seen that the spectrum reconstructed by DNN-INV is more accurate than that reconstructed by MP-INV and EQU-INT. More specifically, for the voiced frame, both spectral peaks and spectral valleys of the original spectrum are well recovered by DNN-INV. In contrast, for MP-INV and EQU-INT, the resonant structure is nearly lost. For unvoiced frame, DNN-INV also shows better performance, but its advantage over MP-INV and EQU-INT is not so significant as that for voiced frame.

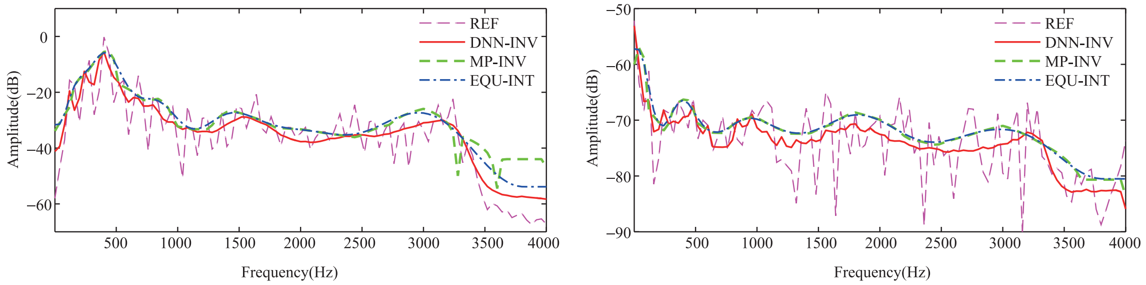
In practical systems (*e.g.*, Ref.[6]), the higher-order cepectra are usually truncated. Specifically, the 10 elements, C13 through C22, in the 23-D MFCC vector are discarded. In this case, we use the truncated 13-D MFCC vectors as input data to train the DNN, and the expected outputs are the original spectra. Using this strategy, the trained DNN can "predict" or "estimate" the spectra from



(*a*) Voiced frame

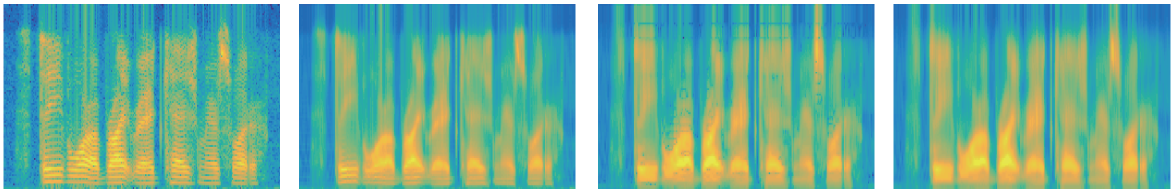
(*b*) Unvoiced frame

Fig. 3. Magnitude spectra reconstructed from a 23-D MFCC vector by DNN-INV, MP-INV, and EQU-INT



(*a*) Voiced frame (*b*) Unvoiced frame

Fig. 4. Magnitude spectra reconstructed from a truncated 13-D MFCC vector by DNN-INV, MP-INV, and EQU-INT



(*a*) Original (*b*) DNN-INV (*c*) MP-INV (*d*) EQU-INT

Fig. 5. Spectrograms obtained from (*a*) Original utterance, estimated from the truncated 13-D MFCC vectors; (*b*) DNN-INV; (*c*) MP-INV; (*d*) EQU-INT

the truncated 13-D MFCC vectors. In comparison, the MP-INV and EQU-INT methods have to apply a zero-padding strategy to get the 23-D MFCC vectors prior to the inversion process. However, the zero-padding strategy does not bring in any useful information.

Fig.4 shows the magnitude spectra reconstructed from a truncated 13-D MFCC vector by the compared methods. It again demonstrates that DNN-INV is more accurate than MP-INV and EQU-INT. Moreover, the spectrum recovered from the truncated 13-D MFCC vector by DNN-INV in Fig.4 is even closer to the original spectrum than the spectra recovered from the 23-D MFCC vector by MP-INV and EQU-INT in Fig.3. That is, compared with MP-INV and EQU-INT using the un-truncated MFCC vector, the DNN-INV can give a more accurate result even using the truncated MFCC vector.

Fig.5 compares the spectrograms estimated from the truncated 13-D MFCC vectors of a speech utterance. Compared with MP-INV and EQU-INT, DNN-INV obtains a more accurate spectrogram, especially for the harmonics. It should be noticed that the reconstructed spectrum of MP-INV contains negative values, which are set to absolute values in Fig.5. As a result, there are some abnormal pixels in the spectrogram reconstructed by MP-INV.

Similar results can be obtained from the un-truncated 23-D MFCC vectors, which also demonstrate the distinct advantage of DNN-INV. In the next evaluation, we examine the accuracy of the compared methods quantitatively.

3. Quantitative evaluation of reconstruction performance

We use Log-spectral distortion (LSD)[20] to measure the reconstruction error of magnitude spectrum. In addition, we use the well-known Least-squares estimate Inverse short-time Fourier transform magnitude (LSE-ISTFTM) algorithm[10,21] to estimate the time-domain speech signal from the reconstructed magnitude spectrum, in the absence of the phase spectrum. Perceptual evaluation of speech quality (PESQ), which has a high correlation with the subjective score[22], is used to evaluate the quality of the estimated speech signal.

Table 1 shows the average LSD of the reconstructed magnitude spectrum and the corresponding PESQ of the estimated speech signal. In general, the results from the un-truncated 23-D MFCC vectors are superior to that from the truncated 13-D MFCC vectors using one reconstruction method. An exception is the LSD of MP-INV. This is because the abnormal negative magnitude values make the LSD larger.

Table 1. Average LSD and PESQ results of the compared methods on the test set

Method	Criteria	MFCC vectors	
		C0–C23	C0–C12
DNN-INV	LSD	5.25	6.02
	PESQ	3.30	2.51
MP-INV	LSD	8.41	8.05
	PESQ	2.39	2.21
EQU-INT	LSD	7.08	7.43
	PESQ	2.45	2.20

It can be clearly seen from Table 1 that, DNN-INV has the lowest LSD, 5.26dB (about 37.5% lower than that of MP-INV and 25.7% lower than that of EQU-INT) for 23-D MFCC vectors, and 6.02dB (about 25.2% lower than that of MP-INV and 19.0% lower than that of EQU-INT) for truncated 13-D MFCC vectors. In terms of PESQ, DNN-INV obtains the highest scores, 3.30 (about 38.1%

higher than that of MP-INV and 34.7% higher than that of EQU-INT) for 23-D MFCC vectors, and 2.51 (about 13.6% higher than that of MP-INV and 14.1% higher than that of EQU-INT) for truncated 13-D MFCC vectors. Furthermore, DNN-INV shows a better performance even using truncated MFCC vectors when compared with MP-INV and EQU-INT using un-truncated MFCC vectors. The LSD of MP-INV and EQU-INT are 8.41dB and 7.08dB, respectively, for un-truncated MFCC vectors. In contrast, the LSD of DNN-INV is 6.02dB for the truncated MFCC vectors. Similar advantage of DNN-INV can also be seen in Table 1 in terms of PESQ.

Then, we evaluate the quality of the reconstructed speech with informal subjective preference evaluation. 100 speech utterances are random picked from the test set. 10 listeners (7 male and 3 female) are recruited for the evaluation. Each tester is asked to select the one he/she prefers for each utterance reconstructed by the compared methods. There are totally $100 \times 3 \times 10 = 3000$ selection results. The final statistic results are shown in Table 2. The preference ratio for DNN-INV, MP-INV and EQU-INT is 97.3%, 1.2%, and 1.5%, respectively. The subjective evaluation results in Table 2 accord well with the results in Table 1 under the objective metric.

Table 2. Subjective preference evaluation results of the compared methods on 100 speech utterances

Method	Perference ratio(%)
DNN-INV	97.3%
MP-INV	1.2%
EQU-INT	1.5%

4. Evaluation for unseen cases

We consider two unseen conditions, 1) training using clean speech, but test using noisy speech; 2) training using English corpus, but test using Chinese. In both conditions, the DNN is trained using the TIMIT corpus, and the training data is clean speech.

The evaluation results in the first condition using noisy speech (Gaussian noise) for test are shown in Table 3. The results show that DNN-INV distinctly outperforms MP-INV and EQU-INT at moderate to high SNRs, and its advantage is more significant at high SNRs. The evaluation results in the second condition using Chinese corpus for test are shown in Table 4. In this condition, DNN-INV also outperforms the signal processing based methods MP-INV and EQU-INT. The results imply that the DNN-based method has strong inference power in dealing with the unseen cases.

Table 3. Testing results of the compared methods on the test set corrupted by Gaussian noise at different SNRs

Method	0dB	5dB	10dB	20dB
DNN-INV	1.29	1.49	1.74	2.23
MP-INV	1.28	1.47	1.67	1.98
EQU-INT	1.28	1.49	1.71	2.06

Table 4. Testing results using Chinese corpus

Method	LSD	PESQ
DNN-INV	6.26	2.62
MP-INV	9.36	1.82
EQU-INT	7.68	2.14

5. Computational complexity

Finally, we analyze the computational complexity of the proposed method. For the proposed DNN-INV algorithm, the forward-propagation involves a number of matrix-vector multiplication and sigmoid operations. The main computational complexity in the DNN part is matrix-vector multiplication in each layer, and the overall computational load is $\sum_{l=1}^{L-1} O(n_l n_{l+1})$, where n_l is the size of lth layer, L is the total number of DNN's layers.

We have benchmarked DNN-INV, MP-INV and EQU-INT algorithms, which are all implemented in MATLAB, on an Intel i7@3.6-GHz PC. 1000 speech utterances (about 3000 seconds) are randomly picked from the TIMIT corpus for this evaluation, and the MFCC vectors are computed in advance. To inverse all these MFCC vectors, DNN-INV, MP-INV, and EQU-INT take 5.3, 0.5, and 3.1 seconds, respectively. The results indicate that the proposed DNN-based method can be used for real-time reconstruction in practical applications.

V. Conclusions

This paper proposed a novel DNN-based method for inverting MFCC vectors to speech magnitude spectrum. This method uses the MFCC vectors along with the corresponding spectra as a pair of datasets to train a DNN model. Evaluation results showed that the proposed method has the capability to obtain a more accurate spectrum, compared with two popular methods, MP-INV and EQU-INT. Moreover, due to the DNN's strong inference power, the proposed method shows a better performance even using truncated MFCC vectors, in comparison with MP-INV and EQU-INT using the un-truncated MFCC vectors. The proposed method can be applied to many MFCC-based speech reconstruction applications, such as server-side speech reconstruction of DSR[6] and low bit-rate speech coding[24,25].

References

[1] G. Hinton, L. Deng, D. Yu, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups", *IEEE Signal Processing Magazine*, Vol.29, No.6, pp.82–97, 2012.

[2] J. Xu, J. Pan and Y. Yan, "Agglutinative language speech recognition using automatic allophone deriving", *Chinese Journal of Electronics*, Vol.25, No.2, pp.328–333, 2016.

[3] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues", *IEEE Circuits & Systems Magazine*, Vol.11, No.2, pp.23–61, 2011.

[4] C. Liang, X. Zhang and Y. Yan, "Discriminative decision function based scoring method used in speaker verification", *Chinese Journal of Electronics*, Vol.21, No.4, pp.692–696, 2012.

[5] T. Ramabadran, A. Sorin, M. McLaughlin, *et al.*, "The ETSI extended distributed speech recognition (DSR) standards: Server-side speech reconstruction", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Quebec, Canada, pp.129–132, 2004.

[6] ETSI ES 202 212:2005, Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Extended Advanced Front-end Feature Extraction Algorithm; Compression Algorithms; Back-end Speech Reconstruction Algorithm.

[7] T. Ramabadran, J. Meunier, M. Jasiuk, *et al.*, "Enhancing distributed speech recognition with back-end speech reconstruction", *Proc. Europe Conference on Speech Communication and Technology*, Scandinavia, pp.1859–1862, 2001.

[8] B. Milner and X. Shao, "Speech reconstruction from Mel-frequency cepstral coefficients using a source-filter model", *Proc. Europe Conference on Speech Communication and Technology*, Denver, USA, pp.2421–2424, 2002.

[9] X. Milner and X. Shao, "Clean speech reconstruction from MFCC vectors and fundamental frequency using an integrated front-end", *Speech Communication*, Vol.48, No.6, pp.697–715, 2006.

[10] L. E. Boucheron, P. L. De Leon and S. Sandoval, "Low bit-rate speech coding through quantization of Mel-frequency cepstral coefficients", *IEEE Transaction on Audio, Speech, and Language Processing*, Vol.20, No.2, pp.610–619, 2012.

[11] G. Hinton, S. Osindero and Y. Teh, "A fast learning algorithm for deep belief nets", *Neural computation*, Vol.18, No.7, pp.1527–1554, 2006.

[12] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", *Science*, Vol.313, No.5786, pp.504–507, 2006.

[13] Y. Xu, J. Du, L. Dai, *et al.*, "A regression approach to speech enhancement based on deep neural networks", *IEEE Transaction on Audio, Speech, and Language Processing*, Vol.23, No.1, pp.7–19, 2015.

[14] Y. Bengio, "Learning deep architectures for AI", *Foundations and Trends in Machine Learning*, Vol.2, No.1, pp.1–127, 2009.

[15] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Transaction on Audio, Speech, and Language Processing*, Vol.28, No.4, pp.357–366, 1980.

[16] G. Hinton, "A practical guide to training restricted Boltzmann machines", *Momentum*, Vol.9, No.1, pp.3–17, 2010.

[17] D. Rumelhart, G. Hinton and R. Williams, "Learning representations by back-propagating errors", *Nature*, Vol.323, No.6088, pp.533–538, 1986.

[18] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database", *National Institute of Standards and Technology*, Gaithersburgh, Page 107, 1998.

[19] J. Bergstra, O. Breuleux, F. Bastien, *et al.*, "Theano: A CPU and GPU math compiler in python", *Proc. of the Python for Scientific Computing Conference*, Austin, TX, USA, pp.3–10, 2010.

[20] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC press, USA, 2013.

[21] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform", *IEEE Transaction on Audio, Speech, and Language Processing*, Vol.32, No.2, pp.236–243, 1984.

[22] ITU-T Recommendition P.862: 2001, Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs.

[23] D. Wang and X.W. Zhang, "Thchs-30: A free chinese speech corpus", *http://arxiv.org/abs/1512.01882*, 2015-10-7.

[24] W.B. Jiang, R.D. Ying and P.L. Liu, "Speech reconstruction for mfcc-based low bit-rate speech coding", *Porc. of IEEE International Conference on Multimedia and Expo Workshops*, Chengdu, China, pp.1–6, 2014.

[25] W.B. Jiang, P.L. Liu, and F. Wen, "An improved vector quantization method using deep neural network", *AEU – International Journal of Electronics and Communications*, Vol.72, No.1, pp.178–183, 2017.

JIANG Wenbin received the M.S. degree in electronic engineering from Hangzhou Dianzi University, China, in 2012. He is a Ph.D. candidate of Shanghai Jiao Tong University. His research interests include speech signal processing and machine learning. (Email: jwb361@sjtu.edu.cn)

LIU Peilin received the Ph.D. degree from the University of Tokyo majoring in electronic engineering in 1998 and worked there as a research fellow in 1999. He is a professor of Department of Electronic Engineering in Shanghai Jiao Tong University. His research interests include signal processing, low power computing architecture, application-oriented SoC design and verification. Prof. Liu is the chair of Shanghai Chapter of IEEE Circuit and System. (Email: liupeilin@sjtu.edu.cn)

WEN Fei received the B.S. degree from the University of Electronic Science and Technology of China (UESTC) in 2006, and the Ph.D. degree in communications and information engineering from UESTC in 2013. Since December 2012, he has been a lecturer at the Air Force Engineering University. Now he is a research fellow of Department of Electronic Engineering in Shanghai Jiao Tong University. His main research interests are statistical signal processing and machine learning. (Email: wenfei@sjtu.edu.cn)