DNN Filter Bank Cepstral Coefficients for Spoofing Detection

Yu, Hong; Tan, Zheng-Hua; Zhang, Yiming; Ma, Zhanyu; Guo, Jun

Published in:
IEEE Access

DOI (link to publication from Publisher):
10.1109/ACCESS.2017.2687041

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 6, 2017, accepted March 20, 2017, date of publication March 24, 2017, date of current version April 24, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2687041

DNN Filter Bank Cepstral Coefficients for Spoofing Detection

HONG YU[1], ZHENG-HUA TAN[2], (Senior Member, IEEE), YIMING ZHANG[3], ZHANYU MA[1], (Senior Member, IEEE), AND JUN GUO[1]

[1]Pattern Recognition and Intelligent System Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark
[3]International School, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Z. Ma (mazhanyu@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61402047, in part by the Beijing Nova Program under Grant Z171100001117049, in part by the Beijing National Science Foundation under Grant 4162044, in part by the Scientific Research Foundation for Returned Scholars, Ministry of Education of China, in part by the Chinese 111 program of Advanced Intelligence, Network Service under Grant B08004, and in part by the OCTAVE Project, funded by the Research European Agency of the European Commission, in its framework programme Horizon 2020 under Grant 647850.

ABSTRACT With the development of speech synthesis techniques, automatic speaker verification systems face the serious challenge of spoofing attack. In order to improve the reliability of speaker verification systems, we develop a new filter bank-based cepstral feature, deep neural network (DNN) filter bank cepstral coefficients, to distinguish between natural and spoofed speech. The DNN filter bank is automatically generated by training a filter bank neural network (FBNN) using natural and synthetic speech. By adding restrictions on the training rules, the learned weight matrix of FBNN is band limited and sorted by frequency, similar to the normal filter bank. Unlike the manually designed filter bank, the learned filter bank has different filter shapes in different channels, which can capture the differences between natural and synthetic speech more effectively. The experimental results on the ASVspoof 2015 database show that the Gaussian mixture model maximum-likelihood classifier trained by the new feature performs better than the state-of-the-art linear frequency triangle filter bank cepstral coefficients-based classifier, especially on detecting unknown attacks.

INDEX TERMS Speaker verification, spoofing detection, DNN filter bank cepstral coefficients, filter bank neural network.

I. INTRODUCTION

As a low-cost and flexible biometric solution to person authentication, automatic speaker verification (ASV) has been used in many telephone or network access control systems, such as telephone banking [1]. Recently, with the improvement of automatic speech generation methods, speech produced by voice conversion (VC) [2], [3] and speech synthesis (SS) [4], [5] techniques has been used to attack ASV systems. Over the past few years, much research has been devoted to protect ASV systems against spoofing attack [6]–[8].

There are two general strategies to protect ASV systems. One is to develop a more robust ASV system which can resist the spoofing attack. Unfortunately, research has shown that all the existing ASV systems are vulnerable to spoofing attacks [9]–[13]. Verification and anti-spoofing tasks can not be done well in only one system at the same time.

The other more popular strategy is to build a separated spoofing detection system which only focuses on distinguishing between natural and synthetic speech [14]. Because of the advantage of being easily incorporated into existing ASV systems, spoofing detection has become an important research topic in anti-spoofing [6], [8], [10], [12], [15].

Many different acoustic features have been proposed to improve the performance of Gaussian mixture model maximum-likelihood (GMM-ML) based spoofing detection systems. In [8], relative phase shift (RPS) and Mel-frequency cepstral coefficients (MFCC) were used to detect SS attacks. A fusion system combining MFCC and group delay cepstral coefficients (GDCC) was applied to resist VC spoofing in [1]. Paper [16] compared the spoofing detection performance of 11 different features on the ASVspoof 2015 database [17]. Among others, dynamic linear frequency triangle filter bank cepstral coefficients (TFCC) feature performed best on

the evaluation set and the average equal error rate was lower than 1%.

Different from the aforementioned systems, some more general systems using machine learning methods were developed to model the difference between natural and synthetic speech more effectively. In [18]–[21], spoofing detection systems based on deep neural networks (DNNs) were proposed and tested, where a DNN was used as a classifier or feature extractor. Unfortunately, experimental results showed that, compared with the acoustic feature based GMM-ML systems, these DNN systems performed slightly better on detecting the trained/known spoofing methods, but much worse on detecting unknown attacks.

In the previous studies, when a DNN was used as a feature extractor, the output of the middle hidden layer was used as DNN features to directly train some other types of models, e.g., Gaussian mixture model (GMM) or support vector machine (SVM) [13], [19], [22]–[24].

If we use the short-term power spectrum as the input of a DNN and set the activation function of first hidden layer as ''linear'', the learned weight matrix between the input layer and the first hidden layer can be considered as a special type of learned filter bank. The number of this hidden layer nodes corresponds to the number of filter bank channels and each column of the weigh matrix can be treated as the frequency response of each filter. Unlike the conventional manually designed filter banks, the filters of the learned filter bank have different shapes in different channels, which can capture the discriminative characteristic between natural and synthetic speech more effectively. The DNN feature generated from the first hidden layer can be treated as a kind of filter bank feature.

Some filter bank learning methods such as LDA (Linear discriminant analysis) filter learning [25] and log Mel-scale filters learning [26] have been introduced in the literatures. These methods did not restrict the shapes of learned filters and the learned filter bank features were used on the speech recognition task.

In this paper, we introduce a new filter bank neural network (FBNN) by introducing some restriction on the training rules, the learned filters are non-negative, band-limited, ordered by frequencies and have restricted shapes. The DNN feature generated by the first hidden layer of FBNN has the similar physical meaning of the conventional filter bank feature and after cepstral analysis we obtain a new type of feature, namely, deep neural network filter bank cepstral coefficients (DNN-FBCC). Experimental results show that the GMM-ML classifier based on DNN-FBCC feature outperforms the TFCC feature and DNN feature on the ASVspoof 2015 data base [16].

II. FILTER BANK NEURAL NETWORKS

As a hot research area, deep neural networks have been successfully used in many speech processing tasks such as speech recognition [27]–[29], speaker verification [30], [31] and speech enhancement [12], [32], [33].

A trained DNN can be used for regression analysis, classification, or feature extraction. When a DNN is used as a feature extractor, due to lack of knowledge about the specific physical interpretation of the DNN feature, the learned feature can only be used to train some other models, directly. Further processing, such as cepstral analysis, can not be applied.

As one of the most classical features for speech processing, cepstral (Cep) features , e.g., MFCC and TFCC, have been widely used in most speech processing tasks.

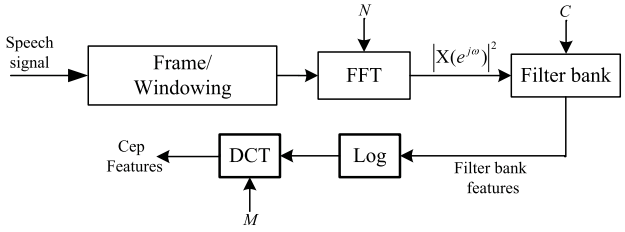

FIGURE 1. The processing flow of computing cepstral features, where *N*, *C*, and *M* stand for the FFT points, the number of filter bank channels, and the number of cepstral coefficients, respectively.

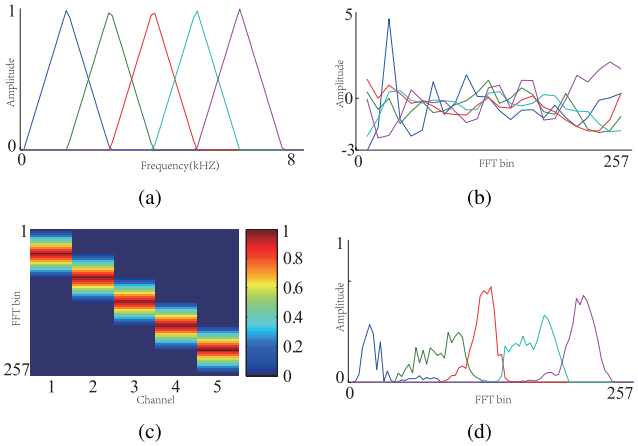

FIGURE 2. (a) A linear frequency triangular filter bank, (b) Learned filter bank without restriction, (c) Band-limiting mask matrix sampling from (a), (d) Learned filter bank with restriction.

Cep features can be created with the following procedure shown in Fig.1. Firstly, the speech signal is segmented into short time frames with overlapped windows. Secondly, the power spectrum $\left|\mathbf{X}\left(e^{jw}\right)\right|^2$ are generated by frame-wise N points fast Fourier transform (FFT). Thirdly, the power spectrum is integrated using overlapping band-limited filter bank with C channels, generating the filter bank features. Finally, after logarithmic compression and discrete cosine transform (DCT) on the filter bank feature, M coefficients are selected as the Cep feature.

As shown in Fig. 2(a), a representative of commonly filters bank used in Cep feature extraction are non-negative, band limited, sorted by frequency and have similar shapes in different channels. The similar shapes for all the channels are not suitable for the spoofing detection task because different

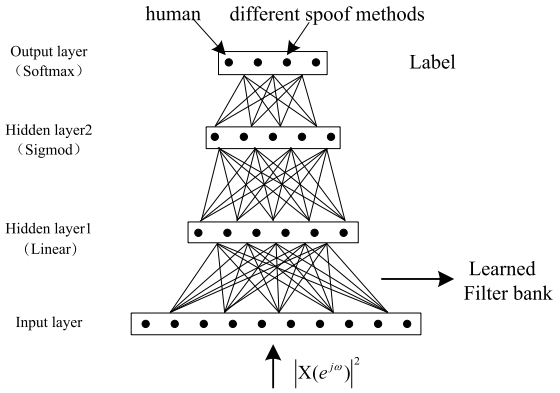

FIGURE 3. The structure of filter bank neural networks.

frequency bands may play different roles in spoofing attacks. This motivates us to use a DNN model to train a more flexible and effective filter bank.

As shown in Fig. 3 we build a FBNN which includes a linear hidden layer, a sigmoid hidden layer and a softmax output layer. The number of nodes in the output layer is N_{out}, where the first node stands for the human voice and the other nodes represent different spoofing attack methods. The same as computing Cep features, we also use the power spectrum as the input. Because the neural activation function of the first hidden layer is a linear function, the output of the first hidden layer can be defined as:

$$\mathbf{H}1 = \mathbf{F}\mathbf{W}_{\mathbf{fb}}, \tag{1}$$

where $\mathbf{F}$ is the input power spectrum feature with D dimension, $D = 0.5N + 1$. The weight matrix between the input layer and the first hidden layer is defined as a filter bank weight matrix $\mathbf{W}_{\mathbf{fb}}$ with dimensions $D \times C$. C is the number of nodes of the first hidden layer and also means the number of channels in the learned filter bank. Each column of $\mathbf{W}_{\mathbf{fb}}$ can be treated as a learned filter channel.

If we do not add any restrictions in the training processing, the learned filters will have the shapes as shown in Fig. 2.(b). Each channel can learn a different filter shape but the characteristics of a normal filter bank, such as non-negative, band-limit and ordered by frequency, can not be satisfied.

In order to tackle this problem, we apply some restrictive conditions on $\mathbf{W}_{\mathbf{fb}}$ as

$$\mathbf{W}_{\mathbf{fb}} = \mathrm{NR}(\mathbf{W}) \odot \mathbf{M}_{\mathbf{bl}}, \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{D\times C}$, $\mathbf{M}_{\mathbf{bl}} \in \mathbb{R}^{D\times C}_{\geq 0}$ and $\odot$ means element wise multiplication.

$\mathrm{NR}(\cdot)$ is a non-negative restriction function which can make elements of $\mathbf{W}_{\mathbf{fb}}$ non-negative. Any monotone increasing function with non-negative output can be used. We select the sigmoid function:

$$\mathrm{NR}(x) = 1/(1 + exp(-x)). \tag{3}$$

$\mathbf{M}_{\mathbf{bl}}$ is a non-negative band-limiting shape restriction mask matrix which can restrict the filters of the learned filter bank to have limited band, regulation shape and ordered by frequency. $\mathbf{M}_{\mathbf{bl}}$ can be generated from any band-limited filter bank by frequency-domain sampling. Fig. 2.(c) shows a $\mathbf{M}_{\mathbf{bl}}$ sampling from a linear frequency triangular filter bank with five channels (Fig. 2.(a)).

W_{dc}, elements of $\mathbf{W}$, can be learned through stochastic gradient descent using equations (4) - (7):

$$W_{dc} = W_{dc} - \eta g_{new}, \tag{4}$$

$$g_{new} = (1 - m) \times g + m \times g_{old}, \tag{5}$$

$$g = \frac{\partial L}{\partial H1_c}\frac{\partial H1_c}{\partial \mathrm{W}_{dc}} = \frac{\partial L}{\partial H1_c}F_d M_{bl_{dc}}\frac{\partial \mathrm{NR}(W_{dc})}{\partial W_{dc}}, \tag{6}$$

$$\frac{\partial \mathrm{NR}(W_{dc})}{\partial W_{dc}} = \mathrm{NR}(W_{dc})[1 - \mathrm{NR}(W_{dc})], \tag{7}$$

where uppercase italic characters with subscripts mean elements of matrix and subscripts stand for indexes, $d \subseteq [1, D]$, $c \subseteq [1, C]$, η is the learning rate, m is the momentum, g is the gradient computed in backward pass, g_{old} is the gradient value in the previous mini-batch, and g_{new} is the new gradient for the current min-batch. L is the cost function and $\frac{\partial L}{\partial H1_c}$ can be computed by the standard back propagation equations for neural networks [34]. The learned filters with restrictions are illustrated in Fig. 2.(d), which are band limited, ordered by frequency and have different filter shapes in different channels.

Following the cepstral analysis steps we can generate a new kind of Cep features using the filter bank generated from FBNN, which is defined as deep neural networks filter bank cepstral coefficients (DNN-FBCC). The new feature can integrate the advantages of Cep feature and the discrimination ability of DNN model, which are specially suitable for the task of spoofing detection.

TABLE 1. Description of ASVspoof 2015 database.

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

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. DATABASE AND DATA PREPARATION

The performance of spoofing detection using the DNN-FBCC feature is evaluated on the ASVspoof 2015 database [17]. As shown in TABLE 1, the database includes three sub datasets without target speaker overlap: the training set, the development set and the evaluation set. We used the training set for FBNN and human/spoof classifier training. The development set and evaluation set were used for testing.

Training set and development set are attacked by the same five spoofing methods, where $S1$, $S2$ and $S5$ belong to VC method and $S3$, $S4$ belong to SS method. Regarding the evaluation set, besides the five known spoofing methods, there are another five unknown methods, where $S6$-$S9$ are VC methods and $S10$ is an SS method.

The speech signals were segmented into frames with 20*ms* length and 10*ms* step size. Pre-emphasis and a hamming window were applied on the frames before the spectrum computation. Paper [16] showed that all the frames of speech are useful for spoofing detection, so we did not apply any voice activity detection method.

B. FBNN TRAINING

The FBNN described in Section II was built and trained with computational network toolkit (CNTK) [35].

The output layer has five nodes, the first one is for human speech and the other four are for five known spoofing methods (S3 and S4 use the same label). The number of nodes in hidden layer H2 is set as 100, the cross entropy function was selected as the cost function L and the training epoch was chosen as 30. The mini-batch size was set as 128. $\mathbf{W}$ was initialized with uniform random numbers. η and m are set as 0.1 and 0 in the first epoch, 1 and 0.9 in the other epochs. Power spectrum of a frame with D dimension is used as input feature, the training label is the label for the utterance that the frame belongs to. The source code of FBNN is made publicly available.[1]

Some experimental results published in paper [36] and [16], show that the high frequency spectrum of speech is more effective for synthetic detection. In order to investigate the affect of different band-limiting and shape restrictions to the learned filter banks, we use four different manually designed filter banks to generate $\mathbf{M_{bl}}$: the linear frequency triangular filter bank (TFB), the linear frequency rectangular filter bank (RFB), the equivalent rectangular bandwidth (ERB) space Gammatone filter bank (GFB), and the inverted ERB space Gammatone filter bank (IGFB).

TFB and RFB equally distribute on the whole frequency region (Fig. 4(a), 4(c), 4(e) and 4(g)). GFB which has been successfully used in audio recognition [37]–[39], has denser spacing in the low-frequency region (Fig.4(i)) and IGFB gives higher emphasis to the higher frequency region(Fig.4(k)).

When using GFB and IGFB, the filter bank number C were set as 128, according to the suggestion of paper [39]. In order to compare with the results published in paper [16] and evaluate the effect of filter bank channel numbers on the learned filter banks, we set C as 20 and 128 when using TFB and RFB. When training 20-channel filter banks, the dimension of the input power spectrum is 257 (512 FFT bins). The spectrum dimension is 513 (1024 FFT bins) when training filter banks with 128 channels. Correspondingly, the number of nodes in the first hidden layer were also set as 20 and 128.

Fig. 4 shows the learned filter banks and their corresponding manually designed shape restriction filter banks. The trained filter banks include the DNN-triangle filter bank (DNN-TFB), the DNN-rectangle filter bank (DNN-RFB), the DNN-Gammatone filter bank (DNN-GFB) and the DNN-inverted Gammatone filter bank (DNN-IGFB). The flexible shapes that learned filters have in different frequency bands give more modeling power and this can potentially capture the difference between human and spoofed speech effectively. By observing learned filter banks, we can find that the learned filters have higher amplitudes in the high frequency region and lower amplitudes in the low frequency region (Fig. 4(f), 4(h), 4(l)), which highlights the more important of the high frequency region and inline with the finding in paper [36].

[1]The source codes and training config files of FBNN can be downloaded at http://kom.aau.dk/~zt/fbnn.zip

C. CLASSIFIER

In designing the classifier, we train two separated GMMs with 512 mixtures to model natural and spoofed speech, respectively. Log likelihood ratio is used as criterion of assessment, which is defined as:

$$\mathbf{ML}\,(X) = \frac{1}{T}\sum_{i=1}^{T}\left\{\log \mathrm{P}(\mathbf{X}_i|\lambda_{\text{human}}) - \log \mathrm{P}(\mathbf{X}_i|\lambda_{\text{spoof}})\right\}, \quad (8)$$

where X denotes feature vectors with T frames, λ_{human} and λ_{spoof} are the GMM parameters of human and spoof model, respectively.

TABLE 2. Description of manually designed Cep features and DNN-FBCC features used in the experiments.

	Feature Name	FFT (N)	Channel (C)	Coef. (M)	Filter bank
Manually designed Cep fearure	TFCC1	512	20	20	TFB
	TFCC2	1024	128	20	TFB
	RFCC1	512	20	20	RFB
	RFCC2	1024	128	20	RFB
	GFCC	1024	128	20	GFB
	IGFCC	1024	128	20	IGFB
DNN-FBCC	DNN-TFCC1	512	20	20	DNN-TFB
	DNN-TFCC2	1024	128	20	DNN-TFB
	DNN-RFCC1	512	20	20	DNN-RFB
	DNN-RFCC2	1024	128	20	DNN-RFB
	DNN-GFCC	1024	128	20	DNN-GFB
	DNN-IGFCC	1024	128	20	DNN-IGFB

D. COMPARISON WITH MANUALLY DESIGNED Cep FEATURES

We compare the spoofing detection performance between manually designed Cep features and DNN-FBCC features, as shown in Table 2. Manually designed Cep features include TFCC1/TFCC2 (linear frequency triangle filter bank cepstral coefficients), RFCC1/RFCC2 (linear frequency rectangle filter bank cepstral coefficients), GFCC (ERB space Gammatone filter bank cepstral coefficients) and IGFCC (inverted ERB space Gammatone filter bank cepstral coefficients), which are generated, respectively, by TFB, RFB, GFB, and IGFB as described in Section III-B. DNN-FBCC features include DNN-TFCC1/DNN-TFCC2, DNN-RFCC1/DNN-RFCC2, DNN-GFCC and DNN-IGFCC which are generated by learned filter banks DNN-TFB, DNN-RFB, DNN-GFB, and DNN-IGFB, respectively. Among the DNN-FBCC, TFCC1, DNN-TFCC1, RFCC1, DNN-RFCC1 are generated by 20-channel filter banks and other features are generated by 128-channel filter banks. The number of

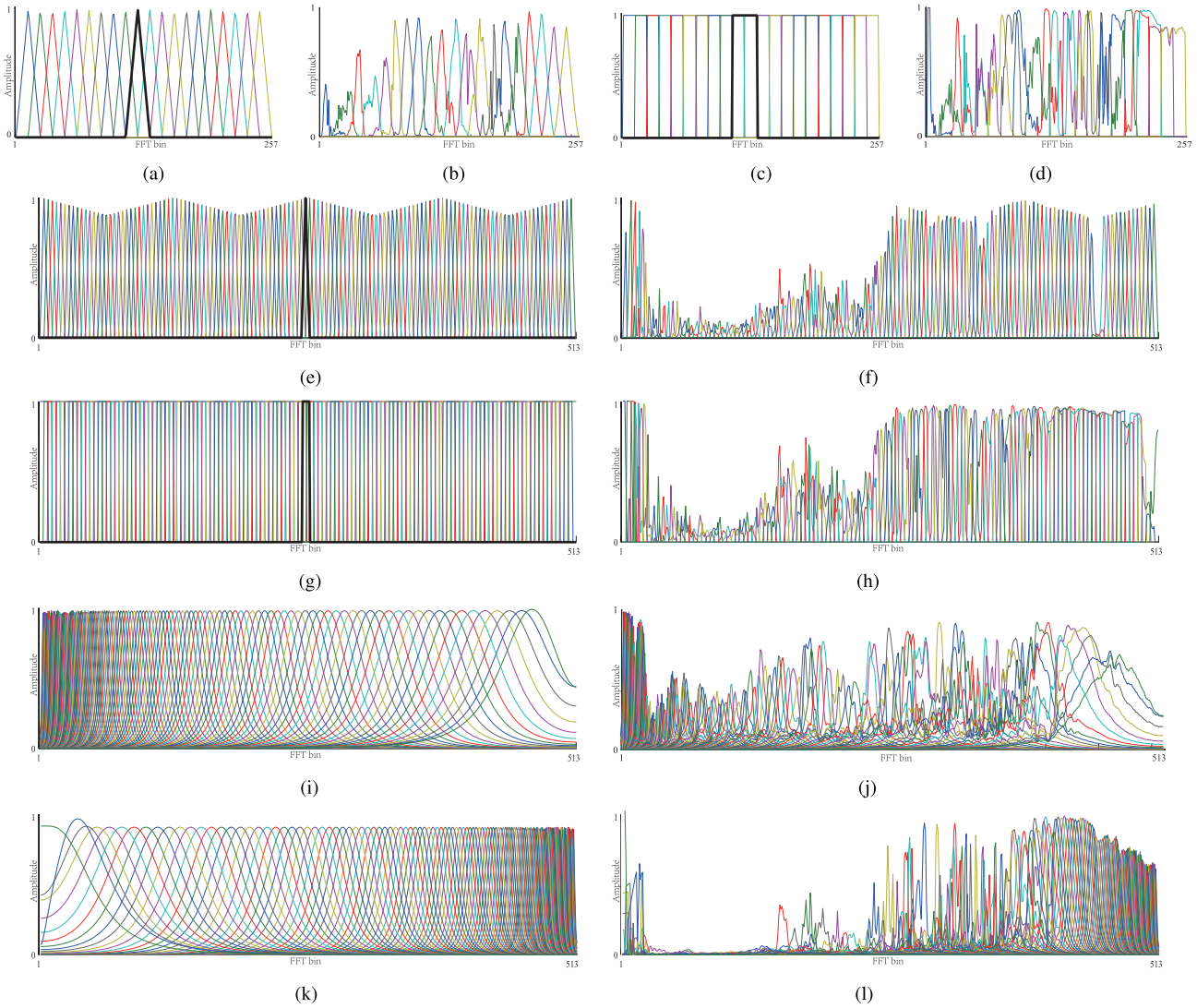

FIGURE 4. Filter banks used for shape restriction and corresponding learned filter banks. (a) TFB with 20 channels. (b) DNN-TFB with 20 channels. (c) RFB with 20 channels. (d) DNN-RFB with 20 channels. (e) TFB with 128 channels. (f) DNN-TFB with 128 channels. (g) RFB with 128 channels. (h) DNN-RFB with 128 channels. (i) GFB with 128 channels. (j) DNN-GFB with 128 channels. (k) IGFB with 128 channels. (l) DNN-IGFB with 128 channels.

coefficients M of all the features are set as 20 (including the 0'th coefficient).

Inspired by the work in [16], we use Δ and Δ^2 (first- and second-order frame-to-frame difference) coefficients to train the GMM-ML classifier. The equal error rate (EER) is used for measuring spoofing detection performance. The average EERs of different spoofing features on development and evaluation sets are shown in TABLE 3.

Among the manually designed Cep features, GFCC($\Delta\Delta^2$) generated by the filter bank with large spacing in the high-frequency region performs worst.

TFCC2($\Delta\Delta^2$), RFCC2($\Delta\Delta^2$) and IGFCC($\Delta\Delta^2$) perform better than TFCC1($\Delta\Delta^2$) and RFCC1($\Delta\Delta^2$) which means increasing the number of filter bank channels can extract more effective discriminative information for spoofing detection.

The six learned DNN-FBCC features outperform the corresponding manually designed Cep features. DNN-GFCC still works worst, which means the filter banks with wider bandwidth on the high frequency region are not suitable for the spoofing detection task.

DNN-RFCC1($\Delta\Delta^2$) generated by 20 channels DNN-RFB performs best on detecting known attacks, but works worse on unknown spoofing attacks. This indicates that the shape restrictions applied on FBNN affect the performance of spoofing detection. When a rectangle filter is selected (Fig. 4(c)), there are no special shape restrictions on the learned filters, and this makes the learned DNN-RFB overfit the trained/known attacks.

With the increase of filter bank channels and reduction of bandwidth of each filters, the shape restriction of RFB is further increased. As shown in Fig. 4(e)-4(h), shape

TABLE 3. Accuracies (Avg.EER in %) of different Cep features on the development and evaluation set.

Feature	Dev.						Eva.												
	known.						known					unknown					mean		
	S1	S2	S3	S4	S5	mean	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	known	unknown	all
TFCC1($\Delta\Delta^2$)	**0.00**	0.52	**0.00**	**0.00**	0.04	0.11	0.02	0.36	**0.00**	**0.00**	0.12	0.12	0.01	0.07	0.02	8.43	0.10	1.73	0.92
TFCC2($\Delta\Delta^2$)	0.03	0.57	**0.00**	**0.00**	0.04	0.13	0.01	0.46	**0.00**	**0.00**	**0.03**	0.04	**0.00**	0.18	0.02	6.98	0.10	1.44	0.77
RFCC1($\Delta\Delta^2$)	0.03	0.69	**0.00**	**0.00**	0.28	0.20	0.01	0.42	**0.00**	**0.00**	0.21	0.19	**0.00**	**0.00**	0.02	9.70	0.13	1.99	1.06
RFCC2($\Delta\Delta^2$)	0.01	0.59	**0.00**	**0.00**	**0.03**	0.13	**0.00**	0.47	**0.00**	**0.00**	0.04	0.04	**0.00**	0.07	0.01	7.69	0.10	1.56	0.83
GFCC($\Delta\Delta^2$)	0.08	2.98	**0.00**	**0.00**	0.86	0.78	0.01	1.91	**0.00**	**0.00**	0.50	0.39	0.09	0.04	0.12	25.5	0.48	5.22	2.85
IGFCC($\Delta\Delta^2$)	0.03	0.45	**0.00**	**0.00**	0.15	0.13	0.02	0.25	**0.00**	**0.00**	0.10	0.10	0.01	0.04	0.02	7.29	0.07	1.49	0.78
DNN-TFCC1($\Delta\Delta^2$)	0.01	0.68	**0.00**	**0.00**	0.13	0.16	0.02	0.57	**0.00**	**0.00**	0.14	0.19	0.02	0.12	0.04	7.26	0.15	1.53	0.84
DNN-TFCC2($\Delta\Delta^2$)	0.01	0.46	**0.00**	**0.00**	**0.03**	0.10	**0.00**	0.37	**0.00**	**0.00**	0.03	**0.03**	**0.00**	0.15	0.01	6.60	0.08	1.32	0.70
DNN-RFCC1($\Delta\Delta^2$)	0.03	**0.23**	**0.00**	**0.00**	0.18	**0.09**	**0.00**	**0.08**	**0.00**	**0.00**	0.11	0.08	**0.00**	0.01	0.01	15.0	**0.04**	3.01	1.52
DNN-RFCC2($\Delta\Delta^2$)	**0.00**	0.44	**0.00**	**0.00**	0.06	0.10	**0.00**	0.32	**0.00**	**0.00**	**0.03**	**0.03**	**0.00**	0.11	**0.00**	7.70	0.07	1.57	0.82
DNN-GFCC($\Delta\Delta^2$)	0.05	1.95	**0.00**	**0.00**	1.77	0.75	0.01	1.27	**0.00**	**0.00**	1.00	1.00	0.19	**0.00**	0.19	26.3	0.46	5.54	3.00
DNN-IGFCC($\Delta\Delta^2$)	0.03	0.30	**0.00**	**0.00**	0.26	0.12	**0.00**	0.19	**0.00**	**0.00**	0.13	0.01	**0.00**	**0.00**	**0.00**	**5.24**	0.06	**1.05**	**0.56**

restrictions of 128-channel TFB and RFB tend to be similar, which causes the learned filter banks, DNN-TFB and DNN-RFB, with 128 channels, also having the similar shapes. The spoofing detection performance of DNN-RFCC2($\Delta\Delta^2$) derived from 128-channel DNN-RFB is close to that of DNN-TFCC2($\Delta\Delta^2$) generated by 128-channel DNN-TFB. The over fitting problem of DNN-RFB is partially overcome by reducing the bandwidth of each filters.

When a Gammatone filter is chosen (IGFB, Fig.4(k)), the shape restriction can make the performance of DNN-IGFCC($\Delta\Delta^2$) better than the corresponding IGFCC($\Delta\Delta^2$) on both known and unknown attacks. In general, among all the investigated Cep features, DNN-IGFCC($\Delta\Delta^2$), generated by the learned filter bank which has denser spacing in the high frequency region and has the Gammatone shape restriction, performs best on ASVspoof 2015 data base and gets the best average accuracy, overall.

In summary, the learned filter banks produced by FBNN using suitable band limiting and shape restrictions can improve the spoofing detection accuracy over the existing manually designed filter banks by learning flexible and effective filters. DNN-FBCC, especially DNN-IGFCC($\Delta\Delta^2$), can largely in- crease the detection accuracy on unknown spoofing attacks.

E. COMPARISON WITH SOME OTHER DATA DRIVEN FEATURES

In this subsection we compare the performance of the DNN-IGFCC($\Delta\Delta^2$) feature with some other data driven features on spoofing detection tasks. All studied features are preprocessed with the same method described in Section III-A. The performance of studied features are evaluated using the GMM based spoofing detection model described in Section III-C. The neural networks used in this paper are all built and trained by CNTK with the same configuration used in FBNN training, in terms of training labels, loss function, learning rate, and learning epochs.

DNN-FBCC features are extracted by a learned filterbank with band-limiting and shape restrictions. In order to study the effect of these restrictions, we firstly investigate performance of the un-restricted filterbank (u-FB) feature extracted by learned filter banks without restrictions (Fig. 2(b)). We use power spectrum features with 513 dimensions (1024 FFT bins) as input and set the size of $\mathbf{W_{fb}}$ as 513×128. In the training process we ignore equation (2) and do not apply any restrictions on $\mathbf{W_{fb}}$.

Without non-negative restriction, cepstral analysis can not be applied on the learned u-FB feature. As DCT operation in cepstral analysis can be considered as a whitening method, in order to have fair quantitative comparison, we use principal component analysis (PCA) to whiten u-FB features and reduce the dimension from 128 to 20.

u-FB-PCA ($\Delta\Delta^2$) features with 40 dimensions are used for GMM model training. The performance of u-FB-PCA ($\Delta\Delta^2$) is shown in the third row of TABLE 4. It is very clearly shown that without restrictions the learned u-FB feature is not suitable for spoofing detection task.

TABLE 4. Accuracies (Avg.EER in %) of DNN-IGFCC($\Delta\Delta^2$) and some other data driven features on the evaluation set.

Feature(dim)	Known	Unknown	All
DNN-IGFCC($\Delta\Delta^2$)(40)	0.06	**1.05**	**0.56**
u-FB-PCA($\Delta\Delta^2$)(40)	23.35	25.26	24.30
LDA-FB(20)	23.02	40.71	31.87
MFCC-BN(60)	0.18	6.37	3.28
l-LMFB(20)	1.49	6.44	3.96
MFCC-BN($\Delta\Delta^2$)(120)	1.46	4.67	3.07
l-LMFB($\Delta\Delta^2$)(40)	0.18	3.2	1.69
DFB-BN(64)	14.26	25.22	19.73
DMCC-BN(64)	**0.03**	4.92	2.47
DLPCC-BN(64)	0.87	3.31	2.09
DPSCC-BN(64)	**0.03**	3.80	1.91
DPSCC-LSTM(64)	0.08	5.61	2.84

Then we compare the DNN-FBCC feature with three different data driven features widely used in speaker verification and speech recognition task.

LDA filter bank feature (LDA-FB) [25] is generated by a 20 channels LDA filter bank which is learned by power spectrum feature with 257 dimensions.

MFCC bottle neck feature (MFCC-BN) [22] is produced by the middle hidden layer of a five-hidden-layer DNN, and the nodes number of hidden layers are set as 2048, 2048, 60, 2048 and 2048, respectively. The DNN is trained by a block of 11 frames of 60 MFCC (static+$\Delta\Delta^2$) features.

The log-normalized learned mel-scale filter bank feature (l-LMFB) is generated by a neural network introduced in [26] which also use the power spectrum with 257 dimension as

input and the log-normalized output of middle hidden layer is used as features. l-LMFB also belongs to learned filter bank features and we set the channel number of learned mel-scale filter bank as 20.

Static and dynamic ($\Delta\Delta^2$) features are used for spoofing detection model training, respectively. The experimental results in TABLE 4 show that among these three kinds of features, the simple data driven filter bank feature LDA-FB is not suitable for the spoofing detection task. While, MFCC-BN and l-LMFB generated from complex neural networks work much better. Especially, l-LMFB which is also extracted by a filter-bank learning method perform best. However, as there are no shape and amplitude restrictions applied on the learned filter banks, l-LMFB performs worse than DNN-IGFCC, especially on unknown spoofing detection.

We also compare DNN-IGFCC($\Delta\Delta^2$) with some DNN based bottle neck (BN) features used for spoofing detection tasks. The published results show that dynamic features are more useful for spoofing detection. Following the suggestion in paper [21], we use four dynamic features to generate DNN-BN feature, including dynamic mel-scale filter bank (DFB) feature, dynamic Mel-frequency cepstral coefficients (DMCC), dynamic product spectrum-based cepstral coefficients (DPSCC) and dynamic linear predication cepstral coefficients (DLPCC). DFB, DMCC and DPSCC are extracted by a mel-scale filter bank with 20 channels and the coefficient numbers of these four features are set as 20.

The feature extraction DNN has five sigmoid hidden layers with node numbers being 1000, 1000, 1000, 1000 and 64, respectively. The output of the fifth layer is used as the DNN-BN feature. The input layer consists of a block of 15 successive dynamic ($\Delta\Delta^2$) features, so the dimension of the input layer is $40 \times 15 = 600$. The softmax output layer also have five nodes, which is the same as the setting of FBNN. All the learned features are also whitened by the PCA method. From the experimental results in TABLE 4, we can observe that the DPSCC-BN feature, which includes both amplitude-frequency and phase information, gives the best performance [40]. It works a little better than DNN-IGFCC($\Delta\Delta^2$) on known spoofing attacks but perform worse on unknown attacks because of the over-fitting problem.

We also use the same DPSCC features to train a long short term memory (LSTM) networks based BN feature extractor which includes two LSTM layers with 1000 nodes and a full connection sigmoid hidden layer with 64 nodes. The PCA whitened DPSCC-LSTM-BN feature with dimension 64 still perform worse than the DNN-IGFCC($\Delta\Delta^2$) feature.

Generally speaking, DNN-FBCC features, especially DNN-IGFCC($\Delta\Delta^2$), which are generated by learned restrictive filter banks, perform better on the spoofing detection tasks than the other data driven features.

IV. CONCLUSIONS AND FURTHER WORKS

In this paper, we introduced a filter bank neural network with two hidden layers for spoofing detection. During training, a non-negative restriction function and a band-limiting mask matrix were applied on the weight matrix between the input layer and the first hidden layer. These restrictions made the learned weight matrix non-negative, band-limited, shape restriction and ordered by frequency. The weight matrix can be used as a filter bank for cepstral analysis. Experimental results show that cepstral coefficients (Cep) features produced by the learned filter banks were able to distinguish the natural and synthetic speech more precisely and robustly than the manually designed Cep features and general DNN features. Recently, some new speech synthesis technologies based on neural networks has been published [41], it encourage us to develop more robust feature to defence the spoofing attacks.

REFERENCES

[1] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 7234–7238.

[2] Z. Wu, E. S. Chng, and H. Li, "Conditional restricted Boltzmann machine for voice conversion," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP)*, Jul. 2013, pp. 104–108.

[3] E. Helander, H. Silen, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 3, pp. 806–817, Mar. 2012.

[4] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1. May 1996, pp. 373–376.

[5] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 7962–7966.

[6] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and antispoofing in the *i*-vector space," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 821–832, Apr. 2015.

[7] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2016, pp. 2119–2123.

[8] J. Sanchez, I. Saratxaga, I. Hernáez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 810–820, Apr. 2015.

[9] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4401–4404.

[10] P. L. D. Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 8, pp. 2280–2290, Oct. 2012.

[11] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification—A study of technical impostor techniques," in *Proc. Eurospeech*, vol. 99. 1999, pp. 1211–1214.

[12] Z. Yang, J. Lei, K. Fan, and Y. Lai, "Keyword extraction by entropy difference between the intrinsic and extrinsic mode," *Phys. A, Statist. Mech. Appl.*, vol. 392, no. 19, pp. 4523–4531, 2013.

[13] J. Lei *et al.*, "A universal framework for salient object detection," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1783–1795, Sep. 2016.

[14] M. Sahidullah *et al.*, "Integrated spoofing countermeasures and automatic speaker verification: An evaluation on ASVspoof 2015," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 2016, pp. 1700–1704.

[15] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 2012, pp. 1700–1703.

[16] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 2015, pp. 2087–2091.

[17] Z. Wu *et al.*, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," *Training*, vol. 10, no. 15, p. 3750, 2015.

[18] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 2015, pp. 1–5.

[19] J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 2015, pp. 1–5.

[20] N. Chen, Y. Qian, H. Dinkel, B. Chen, and K. Yu, "Robust deep feature for spoofing detection-the SJTU system for ASVspoof 2015 challenge," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 2015, pp. 2097–2101.

[21] M. J. Alam, P. Kenny, V. Gupta, and T. Stafylakis, "Spoofing detection on the ASVspoof2015 challenge corpus employing deep neural networks," in *Proc. Speaker Lang. Recognit. Workshop Odyssey*, 2016, pp. 270–276.

[22] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, vol. 237. 2011, p. 240.

[23] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4153–4156.

[24] Z. Ou *et al.*, "Utilize signal traces from others? A crowdsourcing perspective of energy saving in cellular data communication," *IEEE Trans. Mobile Comput.*, vol. 14, no. 1, pp. 194–207, Jan. 2015.

[25] L. Burget and H. Heřmanský, "Data driven design of filter bank for speech recognition," in *Proc. Int. Conf. Text, Speech Dialogue*, 2001, pp. 299–304.

[26] T. N. Sainath, B. Kingsbury, A.-R. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2013, pp. 297–302.

[27] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[28] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[29] J. Lei, C. Zhang, Y. Fang, Z. Gu, N. Ling, and C. Hou, "Depth sensation enhancement for multiple virtual view rendering," *IEEE Trans. Multimedia*, vol. 17, no. 4, pp. 457–469, Apr. 2015.

[30] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4052–4056.

[31] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1096–1104.

[32] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[33] S. Gholami-Boroujeny, A. Fallatah, B. P. Heffernan, and H. R. Dajani, "Neural network-based adaptive noise cancellation for enhancement of speech auditory brainstem responses," *Signal, Image Video Process.*, vol. 10, no. 2, pp. 389–395, 2016.

[34] D. Williams and G. Hinton, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–538, 1986.

[35] D. Yu *et al.*, "An introduction to computational networks and the computational network toolkit (Invited Talk)," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 2014.

[36] H. Yu, A. Sarkar, D. A. L. Thomsen, Z.-H. Tan, Z. Ma, and J. Guo, "Effect of multi-condition training and speech enhancement methods on spoofing detection," in *Proc. IEEE 1st Int. Workshop Sens., Process. Learn. Intell. Mach. (SPLINE)*, Jul. 2016, pp. 1–5.

[37] A. Adiga, M. Magimai, and C. S. Seelamantula, "Gammatone wavelet cepstral coefficients for robust speech recognition," in *Proc. IEEE Region 10 Conf. (31194) TENCON*, Oct. 2013, pp. 1–4.

[38] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1684–1689, Dec. 2012.

[39] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2009, pp. 4625–4628.

[40] D. Zhu and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 1. May 2004, p. I-125.

[41] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," *CoRR*, Sep. 2016.

HONG YU received the master's degree in signal and information processing from Shandong University, Jinan, China, in 2006. He is currently pursuing the Ph.D. degree with the Beijing University of Posts and Telecommunications, Beijing, China. From 2006 to 2013, he was as a Lecturer with Lu Dong University, Shandong, China. He has been a Visiting Ph.D. Student with Aalborg University, Aalborg, Denmark, since 2015. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in image processing, speech processing, data mining, biomedical signal processing, and bioinformatics.

ZHENG-HUA TAN received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1999. He was a Visiting Scientist with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA, an Associate Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, and a Post-Doctoral Fellow with the Department of Computer Science, Korea Advanced Institute of Science and Technology, Daejeon, South Korea. In 2001, he joined the Department of Electronic Systems, Aalborg University, Aalborg, Denmark, where he is currently an Associate Professor. He has authored extensively in these areas in refereed journals and conference proceedings. His research interests include speech and speaker recognition, noise-robust speech processing, multimedia signal and information processing, human-robot interaction, and machine learning. He has served as an Editorial Board Member/Associate Editor of the *Computer Speech and Language* Elsevier, the *Digital Signal Processing* Elsevier, and the *Computers and Electrical Engineering* Elsevier. He was a Lead Guest Editor of the *IEEE Journal of Selected Topics in Signal Processing*. He has served as a Program Co-Chair, Area and Session Chair, Tutorial Speaker, and Committee Member of many major international conferences.

YIMING ZHANG is currently pursuing an undergraduate degree with the Beijing University of Posts and Telecommunications, Beijing, China, and with the Queen Mary University of London, U.K. His research includes pattern recognition and data mining.

ZHANYU MA (SM'17) received the Ph.D. degree in electrical engineering from the KTH Royal Institute of Technology, Sweden, in 2011. From 2012 to 2013, he was a Post-Doctoral Research Fellow with the School of Electrical Engineering, KTH Royal Institute of Technology. He has been an Associate Professor with the Beijing University of Posts and Telecommunications, Beijing, China, since 2014. He has been also an Adjunct Associate Professor with Aalborg University, Aalborg, Denmark, since 2015. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in multimedia signal processing, data mining, biomedical signal processing, and bioinformatics.

JUN GUO received the B.E. and M.E. degrees from the Beijing University of Posts and Telecommunications (BUPT), China, in 1982 and 1985, respectively, and the Ph.D. degree from the Tohuku Gakuin University, Japan, in 1993. He is currently a Professor and a Vice-President with BUPT. He has authored over 200 papers on the journals and conferences, including Science, Nature Scientific Reports, the IEEE Transactions on PAMI, the *Pattern Recognition*, AAAI, CVPR, ICCV, and SIGIR. His research interests include pattern recognition theory and application, information retrieval, content-based information security, and bioinformatics.