INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Unsupervised Representation Learning Using Convolutional Restricted Boltzmann Machine for Spoof Speech Detection

Hardik B. Sailor, Madhu R. Kamble, Hemant A. Patil

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India

{sailor_hardik, madhu_kamble, hemant_patil}@daiict.ac.in

Abstract

Speech Synthesis (SS) and Voice Conversion (VC) presents a genuine risk of attacks for Automatic Speaker Verification (ASV) technology. In this paper, we use our recently proposed unsupervised filterbank learning technique using Convolutional Restricted Boltzmann Machine (ConvRBM) as a front-end feature representation. ConvRBM is trained on training subset of ASV spoof 2015 challenge database. Analyzing the filterbank trained on this dataset shows that ConvRBM learned more low-frequency subband filters compared to training on natural speech database such as TIMIT. The spoofing detection experiments were performed using Gaussian Mixture Models (GMM) as a back-end classifier. ConvRBM-based cepstral coefficients (ConvRBM-CC) perform better than hand crafted Mel Frequency Cepstral Coefficients (MFCC). On the evaluation set, ConvRBM-CC features give an absolute reduction of 4.76 % in Equal Error Rate (EER) compared to MFCC features. Specifically, ConvRBM-CC features significantly perform better in both known attacks (1.93 %) and unknown attacks (5.87 %) compared to MFCC features.

Index Terms: Automatic Speaker Verification, spoofing, countermeasures, ConvRBM, filterbank.

1. Introduction

Automatic Speaker Verification (ASV) or voice biometrics is the task of verifying the claimed identity of a person from his or her voice with the help of machines [1]. However, the practical ASV systems are vulnerable to the biometric attacks. The major forms of the attack known today includes voice conversion (VC) [2], speech synthesis (SS) [3], replay [4], and impersonation [5], which are known to degrade the performance of ASV systems [1]. The general countermeasure approach is one of the solutions to focus on feature representation and statistical pattern recognition techniques. In particular, feature representation forms key task for spoof speech detection (SSD). The aim is to distinguish between genuine and impostor speech by capturing the key discriminative features between two speech signals. This might suggest that the design of spoofing countermeasures should better focus on feature representation, rather than on the advanced or complex classifiers [6, 7].

The details of various approaches used for the SSD task both for the ASVspoof 2015 challenge and post evaluation results are given in [8]. Most of the ASV systems used Fourier transform (FT) magnitude-based and phase-based features from the speech signal. The detailed analysis of various features for ASVspoof 2015 challenge database is presented in [7]. The representation of a speech signal based on human speech perception is of significant interest in developing features for speech processing applications. Classical auditory models were developed to mimic the human auditory system in the 1980s. These auditory models are based on mathematical modeling of auditory processing or psychophysical and physiological experiments. Mel Frequency Cepstral Coefficients (MFCC) [9] are one of the state-of-the-art auditory-based features for spoof detection. The ASVspoof 2015 challenge winner system uses auditory-inspired features Cochlear Filter Cepstral Coeffcients-Instantaneous Frequency (CFCCIF) [10, 11]. Recently, auditory-inspired Constant-Q Cepstral Coefficients (CQCC) also perform better compared to MFCC features specifically in detecting spoof by unit selection speech synthesis system [12], [13]. Such handcrafted features rely on simplified auditory models [14], [15].

Representation learning has gained a significant interest for feature learning in various signal processing areas including speech processing [16]. For the SSD task, various approaches were proposed using representation learning (i.e., deep learning) techniques. In [17], deep features were learned using deep neural network (DNN) from the Mel filterbank. Deep features were learned using deep learning architectures, namely, DNN and Recurrent Neural Networks (RNN) and applied on various classifiers such as linear discriminant analysis (LDA) [18]. DNN-based bottleneck features were also used with Gaussian Mixture Models (GMM) classifier in [19]. An end-to-end learning from the raw speech signals for the SSD task is also recently proposed [20] where it was evaluated for BTAS2016 challenge database [21]. These studies show that instead of conventional handcrafted features, the features learned using the machine learning techniques perform better for the SSD task. Recently, we proposed a Convolutional Restricted Boltzmann Machine (ConvRBM) for unsupervised filterbank learning directly from the raw speech signals. The review of different methods for unsupervised filterbank learning is given in [22]. ConvRBM filterbank was shown to perform better than MFCC and Mel filterbank features for speech recognition task [22], [23].

In this paper, we propose to exploit our approach of the unsupervised filterbank learning using ConvRBM for the SSD task. The filterbank learned using ConvRBM was used to extract the features from the genuine and spoofed speech signals. Compared to our earlier works [22, 23], here we have used an Adam optimization [24] in ConvRBM training. The experiments on ASV 2015 database shows that ConvRBM-based features perform better than MFCC features.

2. Convolutional RBM

Convolutional Restricted Boltzmann Machine (ConvRBM) is an unsupervised probabilistic model with two layers, namely, visible layer and hidden layer [25]. Here, we have used ConvRBM to learn an auditory-like subband filters from the utterance-level raw speech signals [22, 23]. The block diagram

http://dx.doi.org/10.21437/Interspeech.2017-1393

of the arrangement of hidden units is shown in Figure 1. The input $\mathbf{x}$ to ConvRBM is an entire speech signal. The hidden layer consists of K groups. Weights ($\mathbf{W}^k$) are shared between visible and hidden units among all the locations in each group ($k = 1, ..., K$). b_k and c are hidden and visible biases that are also shared. For the k^{th} subband, the input to the hidden layer is given as, $\mathbf{I}_k = (\mathbf{x} * \tilde{\mathbf{W}}^k) + b_k$, where $*$ is a convolution operation and $\tilde{\mathbf{W}}$ denote the *flipped* array [25]. With a noisy rectifier linear units (NReLU), the sampling equations for hidden and visible units ($\mathbf{x}_{recon}$ to reconstruct the speech) are given as [22, 23]:

$$\mathbf{h}^k \sim max(0, \mathbf{I}_k + N(0, \sigma(\mathbf{I}_k))),$$
$$\mathbf{x}_{recon} \sim \mathcal{N}\left(\sum_{k=1}^{K}(\mathbf{h}^k * \mathbf{W}^k) + c, 1\right), \tag{1}$$

where $N(0, \sigma(\mathbf{I}_k))$ is a Gaussian noise with mean zero and sigmoid of $\mathbf{I}_k$ as a variance and $\mathbf{x}_{recon}$ is a reconstructed speech signal (visible units). We have used single-step contrastive divergence (CD-1) to train the model [26]. Parameters are updated using Adam optimization method [24]. It was shown that Adam optimization perform better than stochastic gradient-based methods due to use of first and second order moments of the gradient and bias correction terms [24].

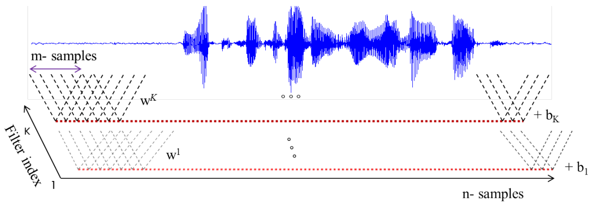

Figure 1: *The arrangement of hidden units in K groups and corresponding weight connections. The filter index-axis is perpendicular to the plane of this paper. Each hidden unit (red dots) in the k^{th} group is wired such that it results in a valid convolution between the speech signal and weights W^k. After [22].*

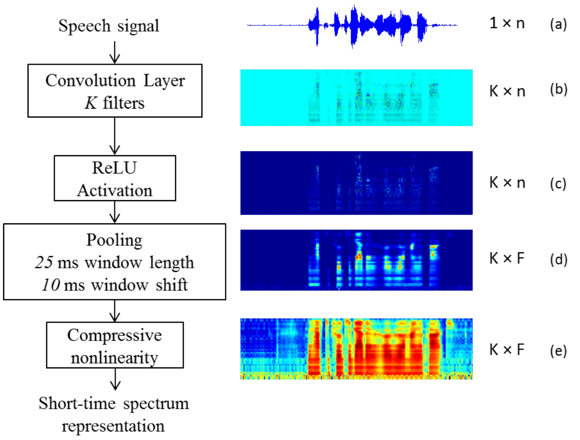

Figure 2: *Feature extraction using trained ConvRBM. (a) speech signal, (b) and (c) responses from the convolutional layer and ReLU nonlinearity, respectively, (d) representation after pooling, (e) logarithmic compression. After [22].*

After ConvRBM is trained, average pooling is applied to reduce the representation of ConvRBM filter responses in the temporal-domain. Here, pooling in the time-domain is equivalent to short-time averaging in spectral features. For a speech signal of the sampling frequency, $Fs = 16$ kHz, pooling is applied using 25 ms window length and 10 ms window shift. Pooling is performed across time and separately for each subband filter. We have experimented with both average and max-pooling and found better results with average pooling. Figure 2 shows the block diagram for feature extraction procedure. During the feature extraction stage, we used the 'same' length convolution and deterministic ReLU nonlinearity $max(0, \mathbf{I}_k)$ as an activation function. The pooling operation reduces the temporal resolution from $K \times n$ samples to $K \times F$ frames. Logarithmic nonlinearity compresses the dynamic range of features.

3. Analysis of the ConvRBM filterbank

3.1. Analysis of subband filters

The ConvRBM is trained using the training set of ASVspoof Challenge 2015 database. Figure 3 shows the subband filters learned using ConvRBM trained on entire training set (denoted as ConvRBM-TrainingAll), synthetic speech from the training set (denoted as ConvRBM-TrainingSyn) of ASVspoof 2015 database and TIMIT database (denoted as ConvRBM-TIMIT). We have analyzed the model with $K = 40$ subband filters for all the cases. We found the center frequencies (CFs) of subband filters as described in [22]. Filters were arranged according to their increasing order of CFs. Weights of the model called as impulse responses of the subband filters in time-domain shown in Figure 3 (a)-(c) and the corresponding frequency responses are shown in Figure 3 (d)-(f). The time-domain subband filters of ConvRBM-TrainingAll and ConvRBM-TrainingSyn are different than ConvRBM-TIMIT whose subband filters resemble more to the Gammatone impulse responses. However, the filterbanks of ConvRBM-TrainingAll and ConvRBM-TrainingSyn includes more lower frequency subband filters with many of the subband filters are wavelet-like basis functions (see the short duration impulses responses in Figure 3 (a), (b)). We can also see that all the subband filters are localized in the frequency-domain with different CFs except in the synthetic speech case as shown in Figure 3 (e). The training set of ASVspoof 2015 database contains 3750 utterances of the natural speech and 12625 utterances of the synthetic speech. Hence, the ConvRBM subband filters trained on training set (that includes both natural and synthetic speech) adapted more towards representing the synthetic speech signals. From the frequency-domain representation of filters, we can see that it also limits the model to represent higher frequencies, which are difficult to model in the synthetic speech signals, such as, fricative and transient sounds.

3.2. Filterbank scale analysis

In order to compare the learned filterbank with the standard auditory filterbanks, we have shown a CFs *vs.* subband filter index plot in Figure 4. We have also shown the frequency scale of ConvRBM-TIMIT that follows the Equivalent Rectangular Bandwidth (ERB) scale. The filterbank learned using ConvRBM-TrainingAll, ConvRBM-TrainingSyn and ConvRBM-TrainingNat (natural speech from the training set) use more number of lower frequency subband filters compared to rest of the filterbanks. The frequency scale of ConvRBM-TrainingNat is slightly different in the frequency range 1-3 kHz compared to ConvRBM-TrainingAll and ConvRBM-TrainingSyn. However, the frequency scales of ConvRBM-TrainingNat and ConvRBM-TIMIT alone are sig-

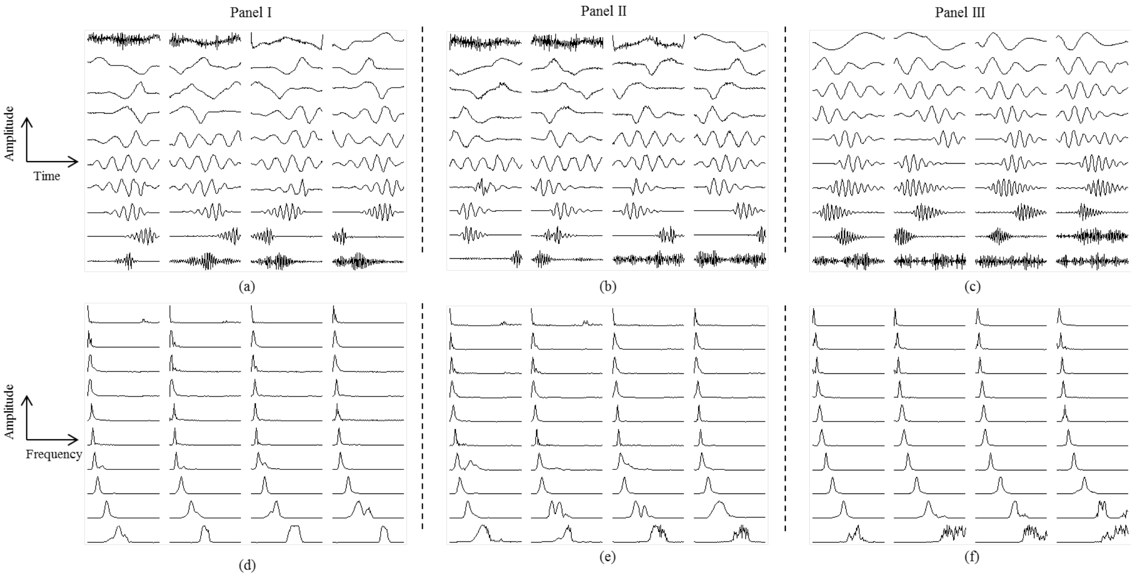

Figure 3: *Examples of subband filters trained on entire training set of ASVspoof 2015 (Panel I), synthetic speech of ASVspoof 2015 (Panel II) and TIMIT (Panel III) databases, respectively:(a)-(c) subband filters in time-domain (i.e., impulse responses), (d)-(f) subband filters in frequency-domain (i.e., frequency responses).*

nificantly different after 1 kHz. Since ConvRBM is a statistical model, it better learns the subband filters with more diverse database and *encode* the statistical properties of the underlying database (such as 462 speakers in TIMIT vs. 25 speakers in the training set of ASVspoof 2015). We also observe that model is biased towards the synthetic speech due to large number of examples compared to natural speech in the training set.

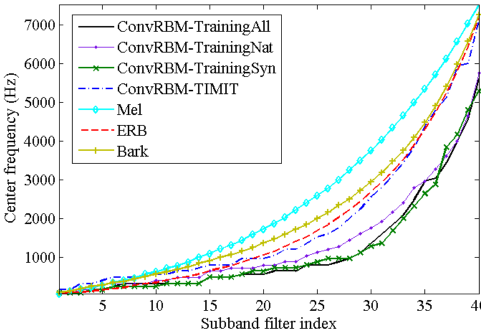

Figure 4: *Comparison of the filterbank learned using ConvRBM with auditory filterbanks.*

4. Experimental Setup

4.1. ASVspoof 2015 database

The experiments are conducted on the ASVspoof Challenge 2015 database [1]. It consists of speech data without channel or background noise collected from 106 speakers (45 male and 61 female). It is divided into three subsets: training, development and evaluation set. More detailed description of the database is discussed in [27].

4.2. Training of ConvRBM and feature extraction

We have trained ConvRBM on training subset of ASVspoof Challenge 2015 database for the SSD task. Each speech signal after mean-variance normalization was applied to ConvRBM. The filter length is chosen to be m=128 samples (i.e., 8 ms) similar to as in [22]. The learning rate was chosen to be 0.0001 and decayed at each epoch according to the learning rate scheduling as suggested in [24]. The moment parameters of Adam optimization chosen to be β_1=0.5 and β_2=0.999. We have trained the model with a different number of ConvRBM filters, with average and max-pooling. After the model was trained, the features were extracted from speech signal with details shown in Figure 2. To reduce the dimension and compare the proposed features with MFCC features, the Discrete Cosine Transform (DCT) was applied and only first 13-D were retained. Delta and delta-delta features were also appended resulting in 39-D cepstral features (denoted as ConvRBM-CC).

Table 1: *The results of different parameters of ConvRBM-CC features on the development set in % EER*

Features	No. of filters (K)	Pooling	% EER
MFCC	40	-	6.14
ConvRBM-CC	60	average	3.71
A:ConvRBM-CC	40	average	3.18
B:ConvRBM-CC	40	max	2.53
A+MFCC	40	-	2.80
B+MFCC	40	-	2.31

4.3. Model training and score-level fusion

We have used Gaussian Mixture Model (GMM) with 128 mixtures for modeling the two classes, in which the classes correspond to the genuine and impostor in ASVspoof 2015 database. The GMMs are trained with the training set of the database. The use of GMM classifier has been shown to perform best in the detection of genuine *vs.* impostor speech in the ASVspoof 2015 challenge [28]. Final scores are represented in terms of the log-likelihood ratio (LLR).

The decision of test speech being genuine or impostor and the score-level fusion of features are done as in [10].

Table 2: *Results on evaluation dataset for each spoofing attack in terms of % EER*

Feature Set	Known Attacks						Unknown Attacks						All
	S1	S2	S3	S4	S5	Avg.	S6	S7	S8	S9	S10	Avg.	Avg.
MFCC	0.78	9.68	**0.00**	**0.00**	7.42	3.57	7.45	1.82	0.17	1.80	57.57	13.76	8.66
A:ConvRBM-CC (avg)	**0.00**	5.68	**0.00**	**0.00**	3.97	1.93	3.26	1.60	**0.00**	1.88	**22.64**	5.87	3.90
B:ConvRBM-CC (max)	**0.00**	**3.61**	**0.00**	**0.00**	2.82	1.26	2.15	1.69	**0.00**	1.45	33.20	7.69	4.47
A+B	0.35	3.70	0.16	0.21	**2.50**	**1.13**	**2.13**	1.13	**0.00**	1.20	24.49	**5.79**	**3.46**
MFCC+A	**0.00**	4.13	**0.00**	**0.00**	2.79	1.38	2.39	**0.68**	**0.00**	**0.00**	54.16	11.44	6.41

5. Experimental Results

5.1. Results on Development Dataset

The results on the development set for the individual performance of 39-D (static+Δ+$\Delta\Delta$) MFCC and ConvRBM-CC with different parameters (number of subband filters K and pooling techniques) are shown in Table 1. It is observed that ConvRBM-CC features (3.71-2.53 % EER) gave relatively better performance compared to MFCC features (6.14 % EER). However, increasing number of subband filters (i.e., K=60) does not improve the performance of classification compared to smaller number of subband filters (i.e., K=40). Lowest % EER is achieved using max pooling and 40 number of subband filters. We have used 40 subband filters for rest of the experiments. The score-level fusion of ConvRBM-CC (K=40, average and max pooling) with MFCC further reduces % EER. This shows that ConvRBM-CC features contain complementary information that was not evident from MFCC features alone. The DET curves of MFCC and ConvRBM-CC with different parameters are shown in Figure 5.

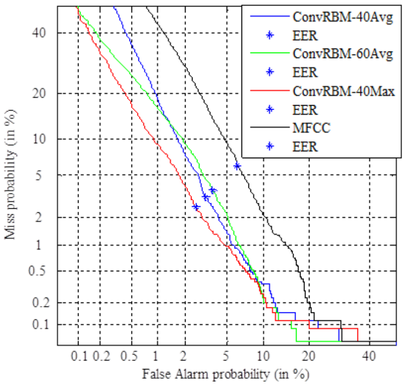

Figure 5: *The DET curve for performance of ConvRBM-CC (60, 40 filters with max and average pooling) and MFCC features on the development set.*

5.2. Results on Evaluation Dataset

Table 2 shows the performance of ConvRBM-CC features for each of the different spoofing attacks grouped into known and unknown attacks with their average EERs for known and unknown spoofing attacks. It is observed that ConvRBM-CC features perform better than MFCC on the evaluation set in all the spoofing attacks. ConvRBM-CC with the max-pooling perform better then average-pooling for known attacks (S1-S5). However, in the case of unknown attacks specifically S10 (unit selection speech synthesis), the average-pooling perform better (22.64 %) than the max-pooling (33.20 % EER). Due to dominance of S10 % EER, the average pooling gave lowest EER of 5.87 % in unknown attacks compared to the max-pooling (7.69 %) and MFCC (13.76 %). Compared to the development set, ConvRBM-CC with the average-pooling perform better than the max-pooling on the evaluation set with lowest EER 3.90 % on an average. To observe whether any complementary information is being captured in the average and max-pooling of ConvRBM-CC, score-level fusion is performed. It resulted in the reduction of % EER in few cases and increased % EER including S10. We have also performed the score-level fusion of MFCC and ConvRBM-CC with the average-pooling. Here, the fusion reduce the % EER for all the attacks except S2, S6 and S10 (with significantly high % EER). Hence, the individual ConvRBM-CC with average pooling performed well in the SSD task. A comparison of the proposed feature representation with the literature (state-of-the-art and feature learning methods) is shown in Table 3. Compared to the supervised features obtained using DNN with LDA/GMM classifiers [18], [19], our unsupervised filterbank learned using ConvRBM perform better in S10 class and similar % EER on an average in unknown attacks. It also perform better than supervised Spectro/CNN [29] in S10 and resulted in similar % EER for unknown attacks and on an average. Given the success of a RNN, we would also like to use it as a back-end instead of GMM. CQCC features gave the lowest results achieved on the ASVspoof 2015 databases.

Table 3: *Comparison of various features in the literature in terms of feature vector dimension (D), classifier, S10 class, unknown attacks and all the attacks*

Features	D	Classifier	S10	Unknown	All
ConvRBM-CC	39	GMM	**22.64**	**5.87**	**3.90**
CQCC [12]	38	GMM	1.07	0.46	0.26
Best DNN [18]	96	LDA	25.5	5.1	2.6
Best RNN [18]	96	LDA	10.7	2.5	1.4
DMCC-BNF [19]	64	GMM	21.47	-	2.15
DPSCC-DNN [19]	60	DNN	12.86	-	2.18
Spectro/CNN [29]	128	CNN	26.83	5.83	3.07
Spectro/RNN [29]	128	RNN	17.97	4.05	2.46
Spectro/CNN+RNN [29]	128	RNN	14.27	3.33	1.86

6. Summary and Conclusions

In this study, we propose to use the unsupervised filterbank learning using ConvRBM for the SSD task. The filterbank learned on training set shows that the frequency scale is different than the one learned using natural signals. It is observed that due to more number of synthetic speech in the training set, ConvRBM filterbank is more biased to the synthetic speech signals. The experimental results on the development and evaluation set shows that ConvRBM-CC features with the average-pooling perform better than the MFCC features. Our future works includes detailed analysis of natural and spoof speech regarding the nature of subband filters and frequency scale. We would also like use our Unsupervised Deep Auditory Model (UDAM) [30] along with TEO [31] for the SSD task.

7. References

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

[2] Y. Stylianou, "Voice trasformation: a survey," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3585–3588.

[3] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[4] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *IEEE Biometrics Special Interest Group (BIOSIG), 2014 International Conference of the*, 2014, pp. 1–6.

[5] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *IEEE Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on*, 2004, pp. 145–148.

[6] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Classifiers for synthetic speech detection: a comparison,," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2057–2061.

[7] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2087–2091.

[8] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilci, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "ASVspoof: the automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, 2017.

[9] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[10] T. B. Patel and H. A. Patil, "Combining evidences from Mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural *vs.* spoofed speech,," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2062–2066.

[11] T. B. Patel and H. A. Patil, "Cochlear filter and instantaneous frequency based features for spoofed speech detection," *accepted in IEEE Journal of Selected Topics in Signal Processing*, 2016.

[12] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop, Bilbao, Spain*, vol. 25, 2016, pp. 249–252.

[13] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, 2017.

[14] R. M. Stern and N. Morgan, "Features based on auditory physiology and perception," in *Techniques for Noise Robustness in Automatic Speech Recognition*. T. Virtanen, B. Raj, and R. Singh, (Eds.) John Wiley and Sons, Ltd, New York, NY, USA, 2012, pp. 193–227.

[15] R. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 34–43, Nov 2012.

[16] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug 2013.

[17] 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 *INTERSPEECH*, 2015, pp. 2097–2101.

[18] Y. Qian, N. Chen, and K. Yu, "Deep features for automatic spoofing detection," *Speech Communication*, vol. 85, pp. 43–52, 2016.

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

[20] H. Dinkel, N. Chen, Y. Qian, and K. Yu, "End-to-end spoofing detection with raw waveform CLDNN," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA*, 2017, pp. 4860–4864.

[21] P. Korshunov, S. Marcel, H. Muckenhirn, A. Gonçalves, A. S. Mello, R. V. Violato, F. Simoes, M. Neto, M. de Assis Angeloni, J. Stuchi *et al.*, "Overview of btas 2016 speaker anti-spoofing competition," in *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*, 2016, pp. 1–6.

[22] H. B. Sailor and H. A. Patil, "Novel unsupervised auditory filterbank learning using convolutional RBM for speech recognition," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 24, no. 12, pp. 2341–2353, 2016.

[23] H. B. Sailor and H. A. Patil, "Filterbank learning using convolutional restricted Boltzmann machine for speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016*, Shanghai, China, March 2016, pp. 5895–5899.

[24] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR), San Diego*, 2015.

[25] H. Lee, R. B. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26^{th} Annual International Conference on Machine Learning, (ICML), Canada, June 14-18*, 2009, pp. 609–616.

[26] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[27] Z. Wu, T. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2037–2041.

[28] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanili, M. Sahidullah, A. Sizov, N. Evans, and M. Todisco, "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, June 2017.

[29] C. Zhang, C. Yu, and J. H. L. Hansen, "An investigation of deep learning frameworks for speaker verification anti-spoofing," *accepted in IEEE Journal of Selected Topics in Signal Processing*, 2017.

[30] H. B. Sailor and H. A. Patil, "Unsupervised deep auditory model using stack of convolutional RBMs for speech recognition," in *INTERSPEECH*, San Francisco, California, USA, September 2016, pp. 3379–3383.

[31] H. B. Sailor and H. A. Patil, "Auditory feature representation using convolutional restricted Boltzmann machine and Teager energy operator for speech recognition," *Journal of Acoustical Society of America Express Letters (JASA-EL)*, vol. 141, no. 6, pp. EL500–EL506, June. 2017.