Interspeech 2018
2-6 September 2018, Hyderabad

Auditory Filterbank Learning for Temporal Modulation Features in Replay Spoof Speech Detection

Hardik B. Sailor, Madhu R. Kamble, and 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

In this paper, we present a standalone replay spoof speech detection (SSD) system to classify the natural *vs.* replay speech. The replay speech spectrum is known to be affected in the higher frequency range. In this context, we propose to exploit an auditory filterbank learning using Convolutional Restricted Boltzmann Machine (ConvRBM) with the pre-emphasized speech signals. Temporal modulations in amplitude (AM) and frequency (FM) are extracted from the ConvRBM subbands using the Energy Separation Algorithm (ESA). ConvRBM-based short-time AM and FM features are developed using cepstral processing, denoted as AM-ConvRBM-CC and FM-ConvRBM-CC. Proposed temporal modulation features performed better than the baseline Constant-Q Cepstral Coefficients (CQCC) features. On the evaluation set, an absolute reduction of 7.48 % and 5.28 % in Equal Error Rate (EER) is obtained using AM-ConvRBM-CC and FM-ConvRBM-CC, respectively compared to our CQCC baseline. The best results are achieved by combining scores from AM and FM cues (0.82 % and 8.89 % EER for development and evaluation set, respectively). The statistics of AM-FM features are analyzed to understand the performance gap and complementary information in both the features.

Index Terms: ConvRBM, amplitude and frequency modulations (AM-FM), replay spoof speech detection

1. Introduction

Among all the spoofing attacks, the replay attacks (also known as presentation attack [1]) are a major threat to the ASV systems since they can be easily performed (using playback of recorded voice) [2]. To promote the research in development of countermeasures for the replay Spoof Speech Detection (SSD), ASVspoof 2017 Challenge was organized as a part of the special session at INTERSPEECH 2017 [3]. Baseline for the Challenge was developed using Constant-Q Cepstral Coefficients (CQCC) features using the Gaussian Mixture Model (GMM) classifier [3], [4]. The Instantaneous Frequency (IF)-based features were also explored in [5], [6]. The high resolution temporal-based features known as Single Frequency Filtering (SFF) were used in [7]. It was found that high frequency content in the spectrum is more useful for detecting the replayed speech and thus, authors in [8] proposed High Frequency Cepstral Coefficients (HFCC). The high frequency-band selection in CQCC also performed better compared to the fullband CQCC [9]. Some of the approaches using deep learning along with feature normalization also proposed in [10–12].

Three key observations from ASVspoof 2017 Challenge are the use of high-frequency spectral information, representation learning, and feature normalization. The previous approaches manually search for various spectral features or learn features using very complex classifiers. In this paper, we propose to use temporal modulation features obtained by combining studies from machine learning and signal processing. We have used an auditory filterbank learning using Convolutional Restricted Boltzmann Machine (ConvRBM) [13, 14]. Earlier, the ConvRBM was applied in the ASVspoof 2015 Challenge database and found better results compared to the baseline system [15]. Here, we use ConvRBM filterbank to extract the temporal modulations in amplitude (AM) and in frequency (FM). The AM and FM are two important physical aspects of the speech signal. There is also an evidence of AM-FM demodulation in the auditory cortex [16]. The AM-FM model of the speech describes the dynamic changes in the envelope (AM) and carrier frequency (FM) [17].

The objective of this paper is to use ConvRBM for auditory filterbank learning and extract temporal modulation features (AM-FM) for the replay SSD task. We propose to use pre-emphasized speech signals to learn subband filters that represent high frequency regions more effectively. The proposed approach represents an amalgamation of filterbank learning using ConvRBM and AM-FM features for improving performance.

2. Temporal Modulation Features

2.1. ConvRBM for Auditory Filterbank Learning

ConvRBM is a probabilistic undirected graphical model that has two layers, the visible and hidden layer [13]. The speech signals are given as an input $\mathbf{x}$ to ConvRBM. In order to learn more subband filters for representing high frequency components, we used pre-emphasized speech signals in ConvRBM training [18]. The hidden layer is divided into K number of groups (i.e., no. of subband filters). The Weights ($\mathbf{W}^k$) are shared between visible and hidden units among all the locations in each group ($k = 1, 2, ..., K$). The hidden and visible biases are also shared denoted as b_k and c, respectively. 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 [13]. With a noisy leaky rectifier linear units (NLReLU), the sampling equations for the hidden and visible units are given as [13, 14]:

$$\mathbf{h}^k \sim \max(0, \mathbf{z}_k) + \alpha_l \cdot \min(0, \mathbf{z}_k), \tag{1}$$

$$\mathbf{x} \sim \mathcal{N}\left(\sum_{k=1}^{K}(\mathbf{h}^k * \mathbf{W}^k) + c, 1\right), \tag{2}$$

where $\mathbf{z}_k = \mathbf{I}_k + N(0, \sigma(\mathbf{I}_k))$, $N(0, \sigma(\mathbf{I}_k))$ is a Gaussian noise with mean zero and sigmoid of $\mathbf{I}_k$ as a variance, and $\mathbf{x}$ is a reconstructed speech signal. The α_l is a parameter controlling the slope in the negative part which is chosen to be 0.01 as suggested in [19]. Compared to our earlier work in [13], [14], we

10.21437/Interspeech.2018-1651

have used noisy leaky rectifier linear units (NLReLU) proposed in [19] to avoid the limitations of ReLU. Annealing dropout is applied in the ConvRBM training with the annealing schedule chosen in [20]. The ConvRBM training is performed using contrastive divergence (CD) [21]. Additional details ConvRBM training on speech signals are given in [14]. ConvRBM parameters are updated using Adam optimization method [22].

2.2. Temporal Modulations in Speech

The acoustic, neurophysiological and psycholinguistic analysis of speech signals demonstrate that there exist the perceptual units of analysis at very different time scales [23], [24]. For subband filtered speech signal, amplitude and frequency modulations in each band are collectively known as temporal modulations. The slow temporal modulations (AM) at the coarsest scale roughly correlate with the different syllabic segments of an utterance. At the finest scale, the fast temporal modulations are due to the frequency component driving the subband at its center frequency. These modulations are also called as Temporal Fine Structure (TFS) of the speech [25]. To extract the temporal modulation features, the speech signal is converted to the subbands using ConvRBM filterbank. The AM-FM are estimated using Energy Separation Algorithm (ESA) [26]. The ESA algorithm estimates the AM and FM using the Teager Energy Operator (TEO) applied on the subband filtered signals [27]. The discrete version of the TEO ($\Psi_D\{\cdot\}$) applied on the i^{th} subband $s_i[n]$ of the filterbank is defined as follows [26]:

$$\Psi_D\{s_i[n]\} \coloneqq s_i^2[n] - s_i[n-1]s_i[n+1]. \tag{3}$$

The discrete ESA algorithm is used to extract the AM $a_i[n]$ and FM $f_i[n]$ for the i^{th} subband and it is given as [26]:

$$a_i[n] \approx \frac{2\Psi_D\{s_i[n]\}}{\sqrt{\Psi_D\{s_i[n+1] - s_i[n-1]\}}}, \tag{4}$$

$$f_i[n] \approx \arcsin\left(\sqrt{\frac{\Psi_D\{s_i[n+1] - s_i[n-1]\}}{4\Psi_D\{s_i[n]\}}}\right). \tag{5}$$

The block diagram for the short-time AM-FM feature extraction is shown in Figure 1. Each ConvRBM subband signals are passed through ESA block for AM-FM demodulation. The short-time AM-FM features are extracted using windowing operation with a Hamming window. The power-law nonlinearity with an exponent 1/15 is applied for dynamic range compression. The Discrete Cosine Transform (DCT) is applied to decorrelate the feature vectors. The feature normalization is performed using Cepstral Mean Normalization (CMN) [28].

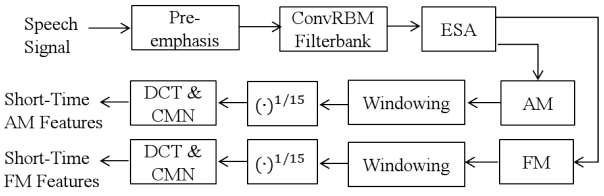

Figure 1: *The AM-FM feature extraction using ConvRBM.*

3. Experimental Setup

3.1. ASVspoof 2017 Challenge Database

The ASVspoof 2017 Challenge database is based on the RedDots corpus and its replayed speech, which is text-dependent database [29]. The spoofed data was recorded through a variety of different environments in the ongoing H2020-funded OCTAVE project2 [3]. The replay corpus was developed through different replay configurations consisting of the varied playback device, recording devices, and the loudspeakers. The statistics of database are given in [3].

3.2. Training of ConvRBM and Feature Extraction

We trained ConvRBM on the training set of ASVspoof 2017 Challenge database. The pre-emphasis filter is applied to the speech signals followed by the utterancewise mean-variance normalization. The ConvRBM is trained with subband filter length $m = 128$ samples (8 ms) and number of subband filters, $K = 40, 60, 80$. The learning rate was empirically chosen to be 0.001 and decayed during each epoch according to the learning rate scheduling in [22]. The moment parameters of Adam optimization was chosen to be β_1=0.5, and β_2=0.999. The annealing dropout probability was chosen to be 0.3 based on our earlier experiments in the ASR [20] and environmental sound classification [30]. After the model was trained, the features were extracted from the speech signal as discussed in Section 2.2. The delta and double-delta features were also appended along with the static feature vectors. The number of feature dimension is chosen based on the experimental results. The notations for different cepstral feature sets are, ConvRBM-CC, AM-ConvRBM-CC, and FM-ConvRBM-CC for ConvRBM and its corresponding AM and FM components.

3.3. Model Training and Baseline Features

We used the GMM classifier with 512 Gaussian components for modeling the two classes, namely, natural and replay spoof. The GMMs are trained with the training set of the database. The final scores are represented in terms of the Log-Likelihood Ratio (LLR). The GMM baseline systems were built with CQCC (90-dimensional (D)) by applying the pre-emphasis on speech signals and feature normalization using CMN. Since the frequency scale obtained after pre-emphasis training is piecewise linear, we also compared our results with Linear Frequency Cepstral Coefficients (LFCC) [31] (90-D). To obtain the complementary information in AM-FM features, we also performed score-level fusion of AM-ConvRBM-CC and FM-ConvRBM-CC. The performance of system is measured using % equal error rate (EER) and Detection Error Trade-off curve (DET).

4. Analysis of the ConvRBM Filterbank

The subband filters (ConvRBM weights) learned from the ASVspoof 2017 Challenge training database are shown in Figure 2. We have compared the subband filters trained with pre-emphasized speech signals (denoted as ConvRBM-PEtraining) and without pre-emphasis. The difference between learned subband filters can be clearly seen in Figure 2. The filterbank learned without pre-emphasized speech signals contains many irregular low frequency subband filters (Figure 2 (a)). Since pre-emphasis increases the intensity of high frequency components, the filterbank learned with the pre-emphasized speech contains relatively few low frequency filters, while many filters represent high frequency components in the spectrum (Figure 2 (b)).

A comparison of frequency scales obtained for the pre-emphasized speech signals from the ASVspoof 2017 Challenge database is shown in Figure 3. The frequency scale obtained from the ConvRBM-PEtraining model is significantly different from other auditory scales as well as ConvRBM trained with-

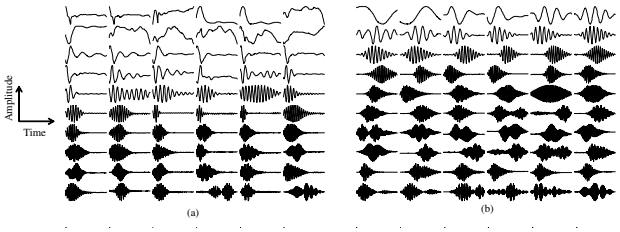

Figure 2: *The subband filters in temporal-domain (a) without and (b) with pre-emphasis, respectively.*

out pre-emphasized speech signals. Since the pre-emphasis performs flattening of the spectrum, the frequency scale is piecewise linear compared to the nonlinear scale obtained without pre-emphasis. It uses progressively more subband filters to represent higher frequencies. For frequencies above 2 kHz, ConvRBM-PEtraining model uses double the number of subband filters (45 *vs*. 20) compared to the other auditory scales and ConvRBM trained without pre-emphasized speech signals.

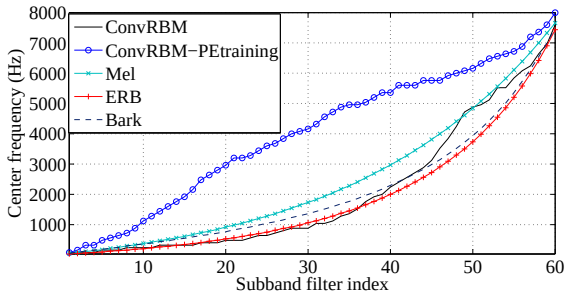

Figure 3: *The comparison of ConvRBM filterbank scales with standard auditory frequency scales.*

5. Experimental Results

5.1. The Effect of Pre-emphasis in Filterbank Learning

It is observed from the previous studies that the replay SSD task requires more cepstral coefficients. Hence, we compare different ConvRBM configurations with the 90-D feature vector (30 static+ Δ+ $\Delta\Delta$) using 60 subband filters. The 90-D was chosen to compare the results with baseline 90-D CQCC features. The ConvRBM-CC obtained from the pre-emphasized filterbank resulted in an improved performance compared to the ConvRBM-CC feature set without a pre-emphasized filterbank learning. Using proposed approach, there is a relative improvement of 10.85 % and 2.98 % in EER compared to ConvRBM-CC (without preemphasis) and CQCC, respectively. The ConvRBM-CC did not perform well with 40 and 80 subband filters. The variance normalization (CMVN) also did not help to reduce EER. Hence, our proposed idea of using pre-emphasized speech signals along with CMN performed better. We used this approach for rest of the experiments, however, the number of subband filters and feature dimension were chosen based on the performance in the SSD task.

5.2. Results using Temporal Modulation Features

The results of replay SSD task using AM-ConvRBM-CC feature set are shown in Table 2. Based on our initial experiments, AM-ConvRBM-CC required Δ and $\Delta\Delta$ along with static features. We then try to see the effect of feature dimension after

Table 1: *The effect of pre-emphasis and feature normalization on CQCC and ConvRBM-CC on the dev set (in % EER)*

Feature Set	Filters	Pre-emphasis	Normalization	Dev
CQCC	-	No	No	10.35
CQCC	-	Yes	CMN	9.06
ConvRBM-CC	60	No	No	12.39
ConvRBM-CC	60	No	CMN	9.86
ConvRBM-CC	60	Yes	CMN	**8.79**
ConvRBM-CC	40	Yes	CMN	11.93
ConvRBM-CC	80	Yes	CMN	9.80
ConvRBM-CC	60	Yes	CMVN	13.94

DCT for ConvRBM with 60 subband filters. AM-ConvRBM-CC performed well with an EER of 2.65 % on dev and 12.76 % on eval set using 120-D features (i.e., 40 static+Δ+ $\Delta\Delta$). Increasing the number of filters did not help to reduce EER. We also changed the nonlinearity from power-law 1/15 to 1/3 and logarithm. However, the power-law nonlinearity with exponent 1/15 performed well. Here, larger exponents and logarithm may suppress the noise characteristics (as observed in [32]) that is actually required to distinguish natural and replayed speech. Hence, the power-law nonlinearity with exponent 1/15 performed well compared to 1/3 and logarithm.

Table 2: *The experimental results using AM-ConvRBM-CC on the dev and eval set in % EER*

Filters	Nonlinearity	Dim	Dev	Eval
60	1/15	90	3.87	15.54
60	1/15	120	**2.92**	**12.76**
60	1/15	160	2.65	14.01
60	1/15	180	2.97	19.05
60	1/15	120	3.93	17.10
80	1/15	120	3.10	16.88
60	Log	120	1.74	18.15
60	1/3	120	4.23	24.93

The results of using FM-ConvRBM-CC feature set are shown in Table 3. Experiments on the dev set suggests that more subband filters are required for better performance with FM-ConvRBM-CC. With 80 subband filters, FM-ConvRBM-CC performed well on dev set compared to 60 filters. We also explore only using static (S), Δ, and S+Δ features. Without using $\Delta\Delta$, FM-ConvRBM-CC with S+Δ features reduce EER to 5.44 % on dev set and 14.96 on eval set similar to AM-ConvRBM-CC. In FM-ConvRBM-CC also power-law nonlinearity 1/15 performed well compared to 1/3 and logarithm. The AM-ConvRBM-CC give lowest % EER compared to FM-ConvRBM-CC on both the dev and eval sets.

5.3. Baseline Comparison and Score-Level Combination

The comparison of proposed features with the baseline features is reported in Table 4. The ConvRBM-CC (without AM-FM demodulation) performed better than CQCC and LFCC, specifically on eval set. The AM-ConvRBM-CC significantly reduce EER (compared to ConvRBM-CC) with an absolute reduction of 5.87 % and 2.1 % on the dev and eval sets, respectively. The FM-ConvRBM-CC reduce EER only on dev set with an absolute reduction of 3.35 % compared to ConvRBM-CC. Interest-

Table 3: *The experimental results using FM-ConvRBM-CC in % EER on the development set*

Filters	Nonlinearity	Type	Dim	Dev	Eval
60	1/15	S+Δ+ΔΔ	180	10.45	-
60	1/15	S+Δ	120	7.03	-
80	1/15	S+Δ+ΔΔ	240	5.74	15.47
80	1/15	S	80	5.63	18.90
80	1/15	Δ	80	6.02	18.44
80	1/15	S+Δ	160	**5.44**	**14.96**
80	Log	S+Δ	160	5.50	15.10
80	1/3	S+Δ	160	6.47	15.79

S=static, Δ=delta, and ΔΔ=double-delta features

ingly, the score-level combination of AM-ConvRBM-CC and FM-ConvRBM-CC (S1⊕S2 in Table 4) achieves the best performance in this study. The S1⊕S2 drop the EER to 0.82 % on dev set and 8.89 % on the eval set. Hence, AM-ConvRBM-CC and FM-ConvRBM-CC contains remarkable complementary information that resulted in the better performance.

The DET curves of all the feature sets along with the score fusion S1⊕S2 is shown in Figure 4. From the dev set, it is observed that AM-ConvRBM-CC has lower False Acceptance Rate (FAR) , (i.e., % false alarm probability) while higher False Rejection Rate (FRR), (i.e., % miss probability). The FM-ConvRBM-CC shows reverse characteristics, i.e., higher FAR and lower FRR. However, DET curve for eval set shows opposite trend compared to DET curve on the dev set. Hence, score combination of AM-ConvRBM-CC and FM-ConvRBM-CC shows the significant reduction in EER.

Table 4: *The comparison of various feature sets*

Feature Set	Dev	Eval
CQCC	9.06	20.24
LFCC	10.28	16.62
ConvRBM-CC	8.79	14.86
S1: AM-ConvRBM-CC	2.92	12.76
S2: FM-ConvRBM-CC	5.44	14.96
S1⊕S2	**0.82**	**8.89**

⊕ indicates score-level fusion

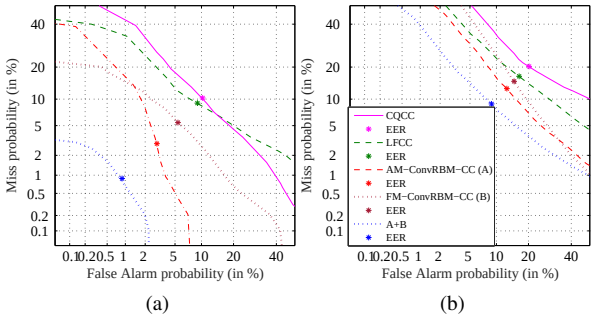

Figure 4: *The DET curves for (a) dev set and (b) eval set.*

5.4. Analysis of AM-FM Statistics

To investigate the performance gap in AM-FM features and how their fusion scores improve the performance, we analyzed the statistics of AM and FM subband features (without DCT) obtained for the natural and replay speech signals from the training set. Inspired from the study in [33], we calculated ensemble averages of mean and standard deviations of AM and FM features. The mean and standard deviation of AM (μ_{AM}, σ_{AM}) and FM (μ_{FM}, σ_{FM}) features are calculated across time for each subband. The ensemble average of mean and standard deviations is calculated for all the utterances in both class and shown in Figure 5. It can be observed that μ_{AM} for natural and replay class are different in most of the subbands. The difference is due to the fact that AM features are significantly affected by the background noise, recording, replay device channel mismatch, reverberation, etc. [34], [35]. There is less difference in μ_{FM} for genuine and replay class (except in the last subbands) since FM features are less affected by noise [35]. However, when we analyzed σ_{FM} and σ_{AM}, we found that σ_{FM} also shows discrimination between natural and replay class. Hence, the analysis of AM-FM statistics shows that both the AM and FM features are affected distinctively in the replayed speech and that might be resulted in an improved performance when AM-FM scores are combined. However, this needs further investigation since there is also an evidence that some of the auditory neurons jointly analyze the AM-FM signals instead of separating AM and FM via demodulation technique [17].

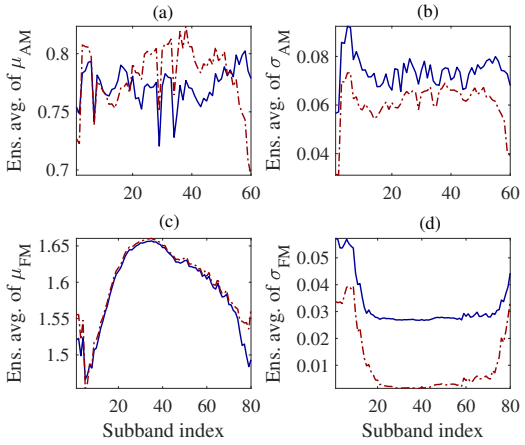

Figure 5: *Analysis of (a)-(b) AM and (c)-(d) FM statistics for natural (solid line) and replay speech (dot-dashed line).*

6. Summary and Conclusions

We presented the use of temporal modulation features extracted from the ConvRBM auditory filterbank learning. The ConvRBM was trained using pre-emphasized speech signals to learn filters that represent high frequency information much better way. The AM-FM features were extracted from ConvRBM filterbank using ESA algorithm. Both the AM-ConvRBM-CC and FM-ConvRBM-CC performed well than the baseline features. The score combination of AM-FM features significantly improved the performance. The statistics of AM-FM features were also analyzed to investigate characteristics of AM-FM features from natural and replay speech and their performance gap. Future works includes using deep learning classifiers and investigation of all the statistics (as suggested in [33]) of AM-FM features.

7. Acknowledgments

Authors would like to thank the MeitY, Govt. of India, for two sponsored projects (1) TTS Phase-II and (2) ASR Phase-II.

8. References

[1] ISO/IEC Information Technology Task Force (ITTF), "Information technology – biometric presentation attack detection," URL: https://www.iso.org/standard/53227.html, 2016, {Last Accessed: 22 March 2018}.

[2] J. Gaka, M. Grzywacz, and R. Samborski, "Playback attack detection for text-dependent speaker verification over telephone channels," *Speech Communication*, vol. 67, no. Supplement C, pp. 143 – 153, 2015.

[3] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 Challenge: Assessing the limits of replay spoofing attack detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1–6.

[4] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language, Elsevier*, vol. 45, pp. 516–535, 2017.

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

[6] S. Jelil, R. K. Das, S. M. Prasanna, and R. Sinha, "Spoof detection using source, instantaneous frequency and cepstral features," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 22–26.

[7] K. R. Alluri, S. Achanta, S. R. Kadiri, S. V. Gangashetty, and A. K. Vuppala, "SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 107–111.

[8] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using DNN for channel discrimination," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 97–101.

[9] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio replay attack detection using high-frequency features," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 27–31.

[10] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 82–86.

[11] W. Cai, D. Cai, W. Liu, G. Li, and M. Li, "Countermeasures for automatic speaker verification replay spoofing attack: On data augmentation, feature representation, classification and fusion," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 17–21.

[12] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and model fusion for automatic spoofing detection," in *INTERSPEECH 2017*, Stockholm, Sweden, 2017, pp. 102–106.

[13] 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)*, Shanghai, China, 20-25 March 2016, pp. 5895–5899.

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

[15] H. B. Sailor, M. R. Kamble, and H. A. Patil, "Unsupervised representation learning using convolutional restricted Boltzmann machine for spoof speech detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2601–2605.

[16] A. Dimitrijevic, M. S. John, P. van Roon, and T. W. Picton, "Human auditory steady-state responses to tones independently modulated in both frequency and amplitude," *Ear and Hearing, American Auditory Society*, vol. 22, no. 2, pp. 100–111, 2001.

[17] H. Luo, Y. Wang, D. Poeppel, and J. Z. Simon, "Concurrent encoding of frequency and amplitude modulation in human auditory cortex: MEG evidence," *Journal of Neurophysiology*, vol. 96, no. 5, pp. 2712–2723, 2006.

[18] L. Deng and D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization Oriented Approach*. Marcel Dekker Inc., First Edition, June 2003.

[19] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. International Conference on Machine Learning (ICML), Atlanta, USA*, vol. 30, 2013, pp. 1–6.

[20] 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.

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

[22] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR), San Diego*, 2015, pp. 1–11.

[23] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations," *Nature Neuroscience*, vol. 15, no. 4, pp. 511–517, 2012.

[24] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America (JASA)*, vol. 118, no. 2, pp. 887–906, 2005.

[25] K. Hopkins and B. C. J. Moore, "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," *The Journal of the Acoustical Society of America (JASA)*, vol. 125, no. 1, pp. 442–446, 2009.

[26] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Trans. on Signal Processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

[27] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Albuquerque, New Mexico, USA, 1990, pp. 381–384.

[28] S. Molau, F. Hilger, and H. Ney, "Feature space normalization in adverse acoustic conditions," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong, China*, vol. 1, April 2003, pp. 656–659.

[29] T. Kinnunen, M. Sahidullah *et al.*, "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *IEEE Int. Conf. on Acoust., Speech and Sig. Process. (ICASSP), New Orleans, LA, USA*, 2017, pp. 1–5.

[30] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised filterbank learning using convolutional restricted Boltzmann machine for environmental sound classification," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3107–3111.

[31] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection results on the ASVspoof 2017 Challenge," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 7–11.

[32] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1315–1329, July 2016.

[33] J. H. McDermott, M. Schemitsch, and E. P. Simoncelli, "Summary statistics in auditory perception," *Nature Neuroscience*, vol. 16, no. 4, pp. 493–498, 2013.

[34] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *Journal of Acoustical Society of America (JASA)*, vol. 95, no. 2, pp. 1053–1064, Feb. 1994.

[35] F.-G. Zeng, K. Nie, G. S. Stickney, Y.-Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proceedings of the National Academy of Sciences, USA*, vol. 102, no. 7, pp. 2293–2298, 2005.