INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Novel Subband Autoencoder Features for Detection of Spoofed Speech

Meet H. Soni, Tanvina B. Patel and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology, India

{meet_soni, tanvina_bpatel, hemant_patil}@daiict.ac.in

Abstract

Deep Neural Network (DNN) have been extensively used in Automatic Speech Recognition (ASR) applications. Very recently, DNNs have also found application in detecting natural *vs*. spoofed speech at ASV spoof challenge held at INTERSPEECH 2015. Along the similar lines, in this work, we propose a new feature extraction architecture of DNN called the subband autoencoder (SBAE) for spoof detection task. The SBAE is inspired by the human auditory system and extracts features from the speech spectrum in an unsupervised manner. The features derived from SBAE are compared with state-of-the-art Mel Frequency Cepstral Coefficient (MFCC) features. The experiments were performed on ASV spoof challenge database and the performance was evaluated using Equal Error Rate (EER). It was observed that on the evaluation set, MFCC features with *36*-dimensional (static+Δ+$\Delta\Delta$) features gave *4.32%* EER which reduced to *2.9%* when *36*-dimensional SBAE features were used. Further on fusing SBAE features at score-level with MFCC, a further reduction till *1.93%* EER was observed. This improvement in EER was due to the fact that the dynamics of SBAE features captured significant spoof specific characteristics leading to detect significantly even vocoder-independent speech, which is not the case for MFCC.

Index Terms Subband autoncoder, spoof detection, vocoder speech.

1. Introduction

Developing countermeasures for the task of anti-spoofing for Automatic Speaker Verification (ASV) systems has found its application in safeguarding ASV systems against threats to spoofing attacks. ASV systems are known to be vulnerable to spoofing attacks due to replay, impersonation (mimicking), speech synthesis or voice conversion. Detailed analysis of the effect of these attacks on ASV systems is shown in [1]. Synthetic Speech (SS) and Voice Converted (VC) speech have shown to severely affect the performance of ASV systems when used as an attack. Easy availability through open sources and also the use of adapted Hidden Markov Models (HMMs) have made it possible to generate speech for any speaker. This is not the case for replay and impersonation attacks. Previously, the work on developing countermeasures was limited to non-uniform databases and known spoofs. Very recently, the ASV spoof 2015 challenge held as a special session of INTERSPEECH 2015 [2] that used the Spoofing Anti-spoofing (SAS) database [3]. The database provided a common base for evaluating anti-spoofing countermeasures even on unknown attacks. Several countermeasures were proposed at the challenge to tackle known and unknown attacks. The results in terms of Equal Error Rate (EER) were returned by the organizers. Among these countermeasures included phase-based countermeasures [4–6], modified group-delay phase features [7], [8], wavelet-based features [9], [10], linear prediction-based features [11], [12], etc. In addition, to these, with the development of Deep Neural Network (DNN) in the field on Automatic Speech Recognition (ASR) and Speaker Verification (SV), these approaches have found initially its way in detecting spoofed speech.

In [13], DNN-based classifiers were used for spoof detection task using Linear Frequency Cepstral Coefficients (LFCCs) and preprocessed Relative Phase Shift (RPS). This approach achieved as less than *0.1%* EER on vocoded speech (known attacks). However, an EER of *40%* was observed on vocoder independent spoof (unknown attacks). Next, in [14] a supervised DNN was trained using filterbank features on the training data of the ASV spoof challenge database. This approach was able to achieve an average EER of about *0.058%* on known attacks and *5%* on unknown attacks with *22%* EER reported for a vocoder-independent spoof. The combination of several features such as Mel Frequency Cepstral Coefficients (MFCCs), Mel Cepstral Coefficients (MCCs), Band-Aperiodicity (BAP) and pitch (LF_0) was used in this work. Both approaches in [13] and [14] used supervised learning for feature extraction for which, a large amount of labeled data is needed.

Recently, deep learning methods are gaining popularity for feature extraction from the raw data in an unsupervised manner. The Autoencoder (AE) is such a network which uses DNN or Restricted Boltzmann Machine (RBM) to extract low-dimensional information from high-dimensional raw data [15–18]. The AE have been used in various applications such as denoising front-end for such ASR task [19] [20], in finding mapping between noisy and clean speech spectrum for noise reduction in ASR system [21], speech enhancement task in [22] and speech coding [23]. Very recently, authors in [24] used AE for noise reduction in speaker verification system. Deep AE was used in [25] for noise-aware training for noisy ASR. Features learned by deep AE were used for Statistical Parametric Speech Synthesis (SPSS) using DNN in [26]. Despite these properties, AE features are not popular as the acoustic features in most of the speech technology applications. The inability to control the form of the representation which is learned by AE leads some researchers to criticize them as uninterpretable black boxes [27].

To overcome this limitation, many variants of the AE have been proposed. A new architecture called transforming AE was used in [27] to detect acoustic events in speech signal for ASR task. Phone recognition task was done using mean-covariance RBM in [28]. In [29], authors proposed an architecture of AE in which decoding block was constrained for stretching and compressing frequency-domain for ASR task. In this paper, we propose a new architecture of AE, namely, subband AE (SBAE) for feature extraction from speech spectrum. Proposed architecture

http://dx.doi.org/10.21437/Interspeech.2016-668

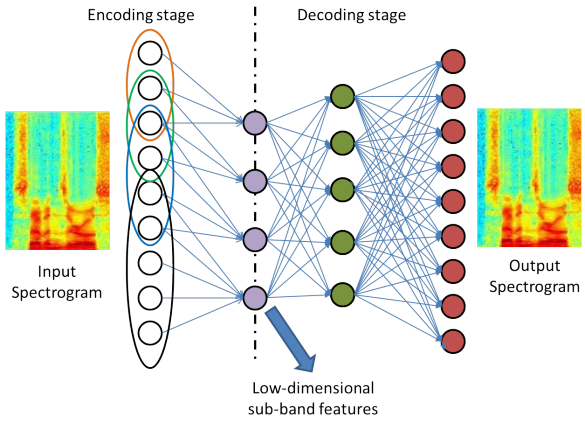

Figure 1: *Architecture of proposed SBAE.*

uses specific domain knowledge about speech processing and incorporates it in the architecture of AE. Inspired by Human Auditory System (HAS), speech is generally processed in subbands. We have restricted the connectivity of units in AE in such a way that each unit in first hidden layer captures the information about a particular band of the speech spectrum. This property of our architecture makes it more suitable for speech technology application. We have used features extracted by SBAE for spoof detection task. To the best of authors' knowledge, this is the first attempt to use features learned by unsupervised machine learning algorithm for spoof speech detection task.

2. Proposed Subband Autoencoder (SBAE)

2.1. Architecture of subband AE

Figure 1 shows the architecture of proposed SBAE. The main difference between proposed SBAE architecture and existing architecture of AE as in [16] is the connectivity of neurons or units immediately after the input layer. In AE, each unit in the layer immediately after input layer is connected with all the units of the previous layer. While in the case of proposed SBAE, the connectivity is restricted. In the proposed architecture, each unit of the first hidden layer is connected with a particular frequency band of input spectrogram. Hence, each unit in the first layer will encode the information about that particular frequency band with which it is connected. The decoding structure is same as a traditional AE with full connectivity [16]. The band structure of restricted connectivity for neurons is same as Mel filterbank, implying one neuron in the first layer is connected with the frequencies of one Mel filterbank. This architecture is nearer to HAS and provides more meaningful information than AE in the case of speech. Mathematically, operation of the subband layer can be represented as follows:

$$a_i = f(\sum_j W_{ij}^1 \times x_j), \quad (1)$$

where a_i is i^{th} subband feature, x_j is short-time power corresponding to j^{th} filterbank frequencies and W_{ij}^1 are weights corresponding to i^{th} subband feature. f represents nonlinear activation function of the neuron. The functionality of preceding layers of SBAE is same as of a traditional AE [16]. Proposed SBAE architecture can be trained by back-propagation similarly as an AE. The a_j learned by SBAE can be used as low-dimensional features for other speech technology tasks, too. These features are different from filterbank energies in following ways: First difference is in the method of extracting features, i.e., MFCCs or filterbank energies are handcrafted features while SBAE features are learned by a machine learning approach. The second difference is that filterbank energies are extracted in a linear way, while SBAE features are extracted in a nonlinear manner. The latter property of SBAE features may provide some more information about speech spectrum which cannot be captured by linear processing.

2.2. Analysis of SBAE features for spoof detection

Figure 2 shows speech waveform, Mel filterbank energies and proposed SBAE features for natural speech and various spoofing attacks. It can be observed that both features show variations for different spoof attacks, hence, they both can be used for spoof detection task. Moreover, both features are invertible, implying speech spectrum can be reconstructed using both features. To quantify reconstruction ability of both the features, average Log Spectral Distortion (LSD) between reconstructed spectrum and the original spectrum was calculated for *50* natural utterances of ASVspoof 2015 database. LSD in case of proposed features was *5.01* dB and using filterbank energies, it was *9.04* dB. Hence, proposed features provide better reconstruction and they can be believed to capture underlying information of speech spectrum for different conditions. However, it is noticeable that proposed features do not show much variations in low-frequency regions for different conditions. It is also evident that proposed features are more sensitive to small variations in the spectrum due to nonlinear processing. This effect can be seen by observing features of two consecutive frames. Unlike filterbank energies, proposed features vary more for consecutive frames (in the time-domain). Thus, SBAE features may capture more dynamic information of speech spectrum. Similar findings for AE features were observed in [26].

3. Spoof Detection System

3.1. Parameterization

For feature extraction, the speech signals were divided into frames with *25* ms frame duration and *50%* overlap. The STRAIGHT spectrum was used for feature extraction using SBAE [30]. The configuration of the network was *513-40-250-513*, implying *513* units in input layer, *40* units in subband layer, *250* units in second layer, and *513* units in output layer. The input and output data was normalized between *0-1* for training. The SBAE trained on training data was used for feature extraction from validation and evaluation datasets. Here, *40* units in subband layer gives *40-D* (dimensional) subband features. To compare the performance of proposed features with the *12-D* MFCCs, *40-D* SBAE features were converted to *12-D* features by following process. As it can be observed from Figure 2, not all *40* SBAE features vary significantly for different types of speech. The SBAE features corresponding to lower bands have almost constant values for natural and spoofed speech. The SBAE features for first *16* bands were removed and features corresponding to rest of the *24* bands were used. Hence, SBAE features corresponding to higher bands are considered for discrimination task. For further dimensionality reduction, the average value of two consecutive subband features was taken. Hence, by this method, *12-D* feature vector was generated to compare with *12-D* MFCCs. As a similarity check, our preliminary experiments suggested that EERs on development set using *40-D* features and reduced *12-D* features were almost similar.

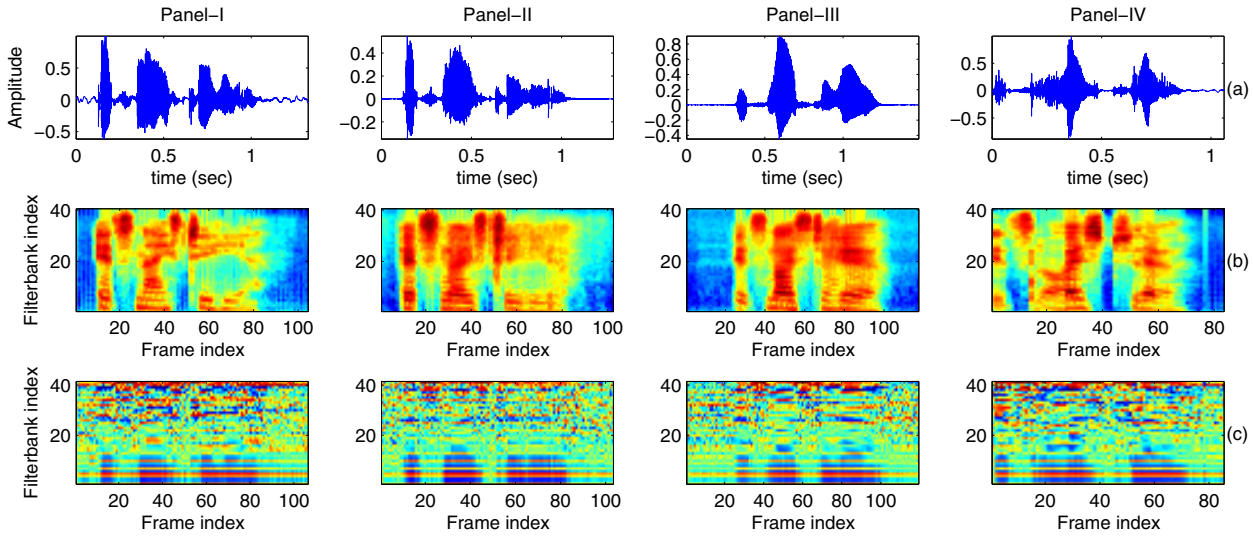

Figure 2: *(a) Waveform of an utterance, (b) mel filterbank energies and (c) SBAE features for panel-I: Natural speech, panel-II: vocoder-based VC speech, panel-III: vocoder-based SS and panel-IV: SS using vocoder-independent (Unit Selection Synthesis (USS)).*

3.2. Details of Database

The database provided for the ASVspoof 2015 challenge is used for this study [3]. Brief details of the database are given in Table I. Details of the spoofing algorithms (S) are provided in [2]. The training and development dataset consisted of utterance generated by five spoofing algorithms (S1-S5) while evaluation data was based on S1-S10, i.e., both known and previously unseen (i.e., unknown) attacks. The S3, S4 and S10 are SS spoof and remaining are VC spoofs. Spoofing algorithm S5 uses Mel Log Spectrum Approximation (MLSA) filter [31] and S10 is implemented with open-source MARY TTS system [32] that uses the FESTIVAL framework [33] for speech synthesis. The remaining spoofs were generated by STRAIGHT vocoder.

Table 1: Details of the ASV spoof 2015 challenge database

	No. of Speakers		No. of Utterances	
Dataset	Male	Female	Genuine	Spoofed
Training set	10	15	3750	12625
Development set	15	20	3497	49875
Evaluation set	20	26	9404	184000

3.3. Performance Measures

A stand-alone detector system may falsely reject a genuine trial to the ASV system or falsely accept a spoof or impostor trial and allow it to pass through an ASV system. The error rates are expressed as False Acceptance Rate (FAR), i.e., ratio of FA to an actual number of positives (natural) and False Rejection Rate (FRR), i.e., ratio of FR to an actual number of negatives (spoofed). Based on the FRR and FAR, the Detection Error Tradeoff (DET) curve is used to measure the performance of various features [34]. It gives uniform treatment to both FAR and FRR for evaluation of system performance. In the DET curve, the operating point where FAR and FRR becomes equal is referred to as EER and is used as a performance measure.

3.4. Model Training and Score-level Fusion

Here, we use a binary Gaussian Mixture Model (GMM) classifier with *128* mixtures for modeling the classes corresponding to natural and spoofed speech on the training set. GMM for natural speech (λ_{nat}) is built using genuine utterances and GMM for spoofed speech (λ_{syn}) is built with spoofed utterances. Final scores on a test sequence Y are represented in terms of log-likelihood ratio (LLR) obtained from the likelihood values of natural and spoofed speech model. The decision of the test speech being human or spoof is based on the LLR, i.e.,

$$LLR = \log(p\,(Y|\lambda_{nat})) - \log(p\,(Y|\lambda_{syn})),$$

where $(Y|\lambda_{nat})$ and $(Y|\lambda_{syn})$ are the likelihood scores from the GMM for the human speech and spoofed speech, respectively. To utilize possible complementary information between features, their score-level fusion is preferred, i.e.,

$$LLk_{combine} = (1-\alpha_f)LLk_{feature1} + \alpha_f LLk_{feature2},$$

where $LLk_{combine}$ is the combined log-likelihood score of two scores feature1 and feature2. The weights of the scores are decided by fusion factor α_f and are optimized w.r.t performance of system after fusion. We consider score-level fusion to know the contribution of the individual set of features and to avoid the higher dimensionality due to the feature-level fusion of features.

4. Experimental Results

4.1. Results on the Development Set

Using the feature extraction process mentioned in Section 4, the features are extracted and GMMs are built on the training set. The results on the development set for MFCC and SBAE are shown in Table 2. It is observed from Table 2 that for the static features the MFCC features gave an EER of *3.3 %* while for the SBAE features an EER of *5.37%* is obtained. On using the Δ features for MFCC and SBAE, the EER is almost similar, i.e.,

Table 2: The results on development set in % EER for MFCC, SBAE and their score-level fusion at various α_f

feature1	Fusion Factor											feature2
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	
SBAE: s (static)	5.38	4.38	3.68	3.23	2.88	2.71	**2.60**	2.72	2.86	3.06	3.26	MFCC: s (static)
SBAE: s+Δ	2.37	1.86	1.54	**1.37**	**1.37**	1.40	1.46	1.57	1.72	1.92	2.17	MFCC: s+Δ
SBAE: s+Δ+$\Delta\Delta$	1.49	1.06	0.83	**0.71**	**0.71**	0.77	0.86	1.00	1.14	1.34	1.60	MFCC: s+Δ+$\Delta\Delta$

Table 3: The results in % EER in terms of individual attacks, average known attacks (Kn.) and average unknown attacks (Unkn.) on the evaluation data for MFCC, SBAE and their score-level fusion for $\alpha_f = 0.3$

	VC S1	VC S2	SS S3	SS S4	VC S5	VC S6	VC S7	VC S8	VC S9	SS S10	Kn.	Unkn.	Average
MFCC: s (static)	0.11	3.78	**0.00**	**0.00**	4.51	3.64	0.40	0.03	1.01	41.47	1.68	9.31	5.50
MFCC:s+Δ	0.02	1.58	**0.00**	**0.00**	1.63	1.75	0.10	**0.00**	**0.00**	49.58	0.65	10.29	5.47
MFCC: s+Δ+$\Delta\Delta$	**0.01**	0.99	**0.00**	**0.00**	0.83	0.90	0.05	**0.00**	**0.00**	39.72	0.37	8.13	4.25
SBAE: s	0.41	4.71	**0.00**	**0.00**	14.60	10.38	0.39	0.34	3.05	56.96	3.94	14.22	9.08
SBAE: s+Δ	0.04	3.04	**0.00**	**0.00**	3.57	4.07	0.23	0.26	1.01	34.88	1.33	8.09	4.71
SBAE: s+Δ+$\Delta\Delta$	0.03	2.99	**0.00**	**0.00**	2.26	2.97	0.11	0.52	0.91	**15.09**	1.06	3.92	2.49
SBAE+MFCC: s+Δ+$\Delta\Delta$	**0.01**	**0.93**	**0.00**	**0.00**	**0.82**	0.88	**0.05**	0.02	0.13	16.52	**0.35**	**3.52**	**1.93**

2.17% and *2.37%* respectively. Furthermore, the use of $\Delta\Delta$ features the EER of SBAE reduces to *1.49%* which is slightly better than MFCC with an EER of *1.6%*. Thus, on adding the dynamic features the % EER for SBAE features reduces significantly as compared to MFCC. This happens due to the fact that proposed features, along with the dynamic features (Δ and $\Delta\Delta$) capture more spectral variation than only static features. On the other hand, when traditional *36-D* AE features were used an EER of *7.9%* was obtained. Hence, in this work, we do not consider AE features for further analysis. To use the possible complementary information present in SBAE and MFCC, a score-level fusion was carried out. It is observed that with a fusion factor of around α_f*= 0.3*, the EER achieved was *0.715%* which is almost half as compared to using MFCC and SBAE features alone. Thus, SBAE features captured complementary information as compared to MFCC features.

4.2. Results on the Evaluation Set

The utterances in the development set consisted of the same type of spoof as used in the training. However, the anti-spoofing ability of the countermeasure depends on the performance for unknown spoofing attacks. To that effect, the performance of the features is tested on the evaluation set which consists of unknown vocoder-based spoofs and one vocoder independent spoof. The results in % EER are shown in Table 3. It is observed that for known attacks (S1-S5) the average EER of MFCC and SBAE features (static+Δ+$\Delta\Delta$) are *0.37%* and *1.06%*, respectively. These evaluations do not include the S10 spoof. Considering the average EER for all the unknown attacks the EER with MFCC is *8.13%* as compared to *3.92%* for SBAE features. This is because the MFCC with static and the dynamic features gives an EER of around *40%* for S10 spoof, while SBAE features give as low as *15%*. Hence, the average performance of the SBAE features for unknown attacks is much better. This suggests that proposed features capture finer variations in speech spectrum in more precise way than MFCCs due to nonlinear processing. It was observed on the development set that fusion of MFCC and SBAE gave better performance than both the features used individually. Thus, the score-level fusion was applied on the evaluation set using α_f*= 0.3*. The better performance of MFCC on known attacks and SBAE on unknown attacks is

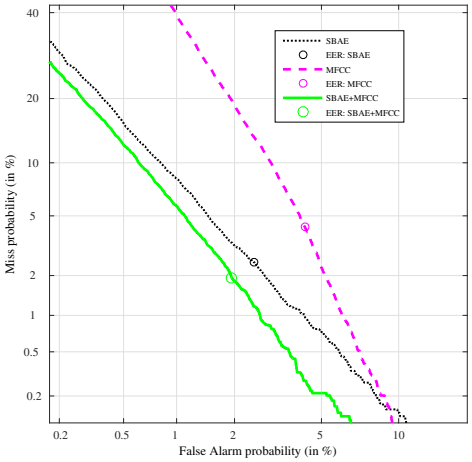

Figure 3: *The DET curve on the evaluation set for MFCC features, SBAE features and score-level fusion of MFCC and SBAE features at* α_f*=0.3.*

combined resulting in the reduction of average EER to *1.93 %* as compared to *4.25 %* for MFCC and *2.49 %* for SBAE. The DET curve for the MFCC, SBAE and fusion of MFCC an SBAE is shown in Figure 3. It is observed that the MFCC had large FRR than SBAE features and slightly better FAR than SBAE at low FRR. However, on fusing both MFCC and SBAE, the DET curve shows better performance over all the operating points.

5. Summary and Conclusions

In this paper, we propose novel SBAE features for the spoof detection task. The results using SBAE features are compared with MFCC features. Due to the fact that the SBAE features capture more dynamic information as compared to MFCC, the relative decrease in % EER is more for SBAE features compared to MFCC features. The MFCC features performed well for known attacks and the SBAE features performed much better for the unknown attacks. Therefore, the combination of both the SBAE and MFCC features at score-level reduces the average % EER. It was also observed that the SBAE features work well on vocoder-independent spoof as compared to MFCC.

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

[3] Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, and S. King, "SAS: A speaker verification spoofing database containing diverse attacks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 4440–4444.

[4] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, "The AHOLAB RPS SSD spoofing challenge 2015 submission," in *Proc. of INTERSPEECH*, Dresden, Germany, 2015, pp. 2042–2046.

[5] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proc. of INTERSPEECH*, Dresden, Germany, 2015, pp. 2092–2096.

[6] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. Raitio, "Toward a Universal Synthetic Speech Spoofing Detection Using Phase Information," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 810–820, 2015.

[7] Y. Liu, Y. Tian, L. He, J. Liu, and M. T. Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," in *Proc. of INTERSPEECH*, Dresden, Germany, 2015, pp. 2082–2086.

[8] 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. of INTERSPEECH*, Dresden, Germany, 2015, pp. 2052–2056.

[9] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "STC anti-spoofing systems for the ASVspoof 2015 challenge," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5475–5479.

[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 *Proc. of INTERSPEECH*, Dresden, Germany, 2015, pp. 2062–2066.

[11] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *Proc. of INTERSPEECH*, Dresden, Germany, 2015, pp. 2077–2081.

[12] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015," in *Proc. of INTERSPEECH*, Dresden, Germany, 2015, pp. 2072–2076.

[13] J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with DNN and one-class SVM for the asvspoof 2015 challenge," in *Proc. of INTERSPEECH*, Dresden, Germany, 2015, pp. 2064–2071.

[14] 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. of INTERSPEECH*, Dresden, Germany, 2015, pp. 2097–2101.

[15] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 3377–3381.

[16] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[17] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 4153–4156.

[18] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks." in *Proc. of INTERSPEECH*, Florence, Italy, 2011, pp. 237–240.

[19] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, "Reverberant speech recognition based on denoising autoencoder." in *Proc. of INTERSPEECH*, Lyon, France, 2013, pp. 3512–3516.

[20] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 1759–1763.

[21] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust asr." in *Proc. of INTERSPEECH*, Portland, Oregon, 2012, pp. 22–25.

[22] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Proc. of INTERSPEECH*, Lyon, France, 2013, pp. 436–440.

[23] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-R. Mohamed, and G. E. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder." in *Proc. of INTERSPEECH*, Makuhari, Japan, 2010, pp. 1692–1695.

[24] O. Plchot, L. Burget, H. Aronowitz, and M. Pavel, "Audio enhancing with DNN autoencoder for speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, Shanghai, China, 2016, pp. 5090–5094.

[25] K. H. Lee, S. J. Kang, W. H. Kang, and N. S. Kim, "Two-stage noise aware training using asymmetric Deep Denoising Autoencoder," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, Shanghai, China, 2016, pp. 5765–5769.

[26] S. Takaki and J. Yamagishi, "A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, Shanghai, China, 2016, pp. 5535–5539.

[27] N. Jaitly and G. E. Hinton, "A new way to learn acoustic events," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[28] G. Dahl, A.-r. Mohamed, G. E. Hinton *et al.*, "Phone recognition with the mean-covariance restricted boltzmann machine," in *Advances in Neural Information Processing Systems*, 2010, pp. 469–477.

[29] N. Jaitly and G. E. Hinton, "Using an autoencoder with deformable templates to discover features for automated speech recognition." in *Proc. of INTERSPEECH*, Lyon, France, 2013, pp. 1737–1740.

[30] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[31] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, San Francisco, California, 1992, pp. 137–140.

[32] "The MARY text-to-speech system (MaryTTS)," http://mary.dfki.de/, {Last Accessed: 30^{th} March, 2016}.

[33] A. Black, P. Taylor, and R. Caley, "The festival speech synthesis system," http://festvox.org/festival/, {Last Accessed: 30^{th} March, 2016}.

[34] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," DTIC Document, Tech. Rep., 1997.