# Spoofing Speech Detection using Temporal Convolutional Neural Network

Xiaohai Tian*†, Xiong Xiao‡, Eng Siong Chng*†‡ and Haizhou Li*§
* School of Computer Science and Engineering, Nanyang Technological University (NTU),Singapore
† Joint NTU-UBC Research Center of Excellence in Active Living for the Elderly, NTU, Singapore
‡ Temasek Laboratories, NTU, Singapore
§ Department of Electrical and Computer Engineering, National University of Singapore
Email: {xhtian,xiaoxiong,ASESChng}@ntu.edu.sg, eleliha@nus.edu.sg

***Abstract*—Spoofing speech detection aims to differentiate spoofing speech from natural speech. Frame-based features are usually used in most of previous works. Although multiple frames or dynamic features are used to form a super-vector to represent the temporal information, the time span covered by these features are not sufficient. Most of the systems failed to detect the non-vocoder or unit selection based spoofing attacks. In this work, we propose to use a temporal convolutional neural network (CNN) based classifier for spoofing speech detection. The temporal CNN first convolves the feature trajectories with a set of filters, then extract the maximum responses of these filters within a time window using a max-pooling layer. Due to the use of max-pooling, we can extract useful information from a long temporal span without concatenating a large number of neighbouring frames, as in feedforward deep neural network (DNN). Five types of feature are employed to access the performance of proposed classifier. Experimental results on ASVspoof 2015 corpus show that the temporal CNN based classifier is effective for synthetic speech detection. Specifically, the proposed method brings a significant performance boost for the unit selection based spoofing speech detection.**

## I. Introduction

With the development of speech synthesis (SS) and voice conversion (VC) techniques, the state-of-the-art systems are able to generate high quality and natural speech with small amount of speech data from target speaker [1, 2]. On the other hand, this also imposes a significant threat to automatic speaker verification (ASV) systems, as the ASV system can be easily attacked by the spoofed speech generate by these techniques [3, 4]. Hence, spoofing speech detection is highly demanded to protect the ASV system.

The common way for spoofing speech detection is to build some standalone detection systems using different countermeasures. In previous studies, various feature based countermeasures are proposed. To detect the artifacts of phase, a number of phase-based features have been used, such as modified group delay (MGD) [5, 6, 7, 8], cosine-normalized phase [5, 8], relative phase shift (RPS) [9, 10, 11, 12], cochlear filter cepstral coefficients plus instantaneous frequency (CFCCIF) [13]. To detect the artifacts of magnitude, several magnitude based features are also used, such as Mel-frequency cepstrum coefficient (MFCC) [14, 15], linear frequency cepstral coefficients (LFCC) [16], Subband centroid frequency coefficient (SCFC) [17] and so on. These features, however, are generally low-dimensional to facilitate certain classifiers, such as Gaussian mixture model (GMM) [6, 14, 18]. The dimension reduction process compresses the features based on human perception and may discard the important information for spoofing speech detection. This may degrade the system performance.

In [7], a neural network (NN) based classifier is proposed to handle the high-dimensional features in order to use detailed information for the detection. For the NN based system using high-dimensional features, all the vocoder-based spoofing speech are detected with the equal error rate (EER) of 0%. However, most of the systems mentioned above use frame-based feature, which does not contain enough temporal information to detect the spoofing speech without using vocoder, for example unit selection based spoofing attacks. Due to the frames of unit selection based spoofing attacks are from natural speech and the artifacts occasionally appears in concatenating points. A feature covering long temporal span information, may up to 100 frames, is needed to capture such artifacts for detection. Hence, although multiple frames (51 frames) [7] and dynamic features [17, 19] have been proposed to capture the temporal artifacts, the detection systems still failed to detect the spoofing speech generated by unit selection based techniques.

Nowadays, convolutional neural network (CNN) is widely used and shows the potential to in achieving good performance in classification and recognition tasks, such as video classification [20], image classification [21], face recognition [22], speech recognition [23] and so on. Inspired by the success of these works, in this paper, we propose a temporal CNN [24] based classifier for spoofing speech detection. Especially, our focus is on unit selection based spoofing speech detection. There are two major advantages of using temporal CNN based classifier for spoofing speech detection. First, as a NN-based classifier, temporal CNN is capable to deal with the high-dimensional features. Second, the max-pooling layer, which outputs the maximum value within long term span, can capture long term temporal information. This temporal information is especially useful to detect the unit selection based spoofing attacks, which the artifacts only appear in concatenation point with a long time span.

## II. System Architecture

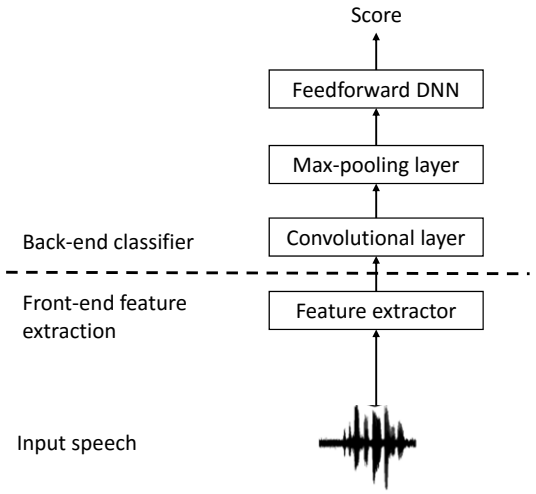


Fig. 1. System architecture for spoofing speech detection.

The temporal CNN based detection system, as shown in Figure 1, contains a front-end feature extraction and a back-end classifier. The details of these parts are introduced as follows.

### A. Front-end Feature Extraction

Similar to our previous system [7, 19], five types of feature, including log magnitude spectrum (LMS), instantaneous frequency derivative [25] (IF), baseband phase difference [26] (BPD), group delay [27] (GD) and modified group delay [27] (MGD), are extracted.

For each frame, the magnitude and phase spectrum are first obtained by applying short-time Fourier transform (STFT) on the speech signal using analysis window of 25ms with 15ms overlap. The Hamming window and direct current (DC) offset is applied on each analysis frame. Then, the magnitude-based feature, LMS, is derived from magnitude spectrum. IF, BPD, GD and MGD are derived from phase spectrum. In our implementation, the FFT length is chosen to be 512 and the dimension of all the original features are 256. The details of feature extraction can be found in [7, 28].

### B. Back-end Classifier

As shown in Figure 2, the temporal CNN based classifier consists of three types of layer: a convolutional layer, a max-pooling layer and a feedforword layer. The convolution layer applies a set of filters that process small regions of the input features along time axis. The max-pooling layer takes the maximum filter responses from a sliding window to form a lower resolution representation of the convolution layer output. Then, the outputs of the max-pooling layer are fed to the feedforword layer for further prediction.

*1) Convolutional Layer:* In conventional CNN, a convolutional layer consist of a set of filters. During the forward pass, each filter slides across the width and height of the input features and the filter is convolved with a subregion of the input features. Each filter is a weight matrix, and all the input features share the same weights to compose a feature map. Thus, the convolutional layer composes many feature maps generated by different filters. In practice, we random initialize the weight matrix, and the weights are learned via backpropagation.

In temporal CNN, the height of the filter is the same as the dimension of input feature. Thus the filter slides and convolves in one direction. Specifically, for speech application, the filter height is cover the whole frequency axis, and the convolution is only applied along the time axis. Hence, in this sense, the output of convolutional layer are feature trajectories substituted to feature maps, as shown in figure 2.

*2) Max-pooling Layer:* A max-pooling layer is usually added on top of the convolution layer. It divides the outputs of the convolutional layer into local regions. Each local region will be down-sampled with the maximum value from that region.

For temporal CNN, as the outputs of the convolutional layer are one dimensional feature trajectories, the major function of max-pooling process is capturing the important local structures distributed over time. In our spoofing speech detection task, it is effective to detect the long time span temporal artifacts, such as the unit selection based spoofing attacks.

## III. Experiments

### A. Experimental Setup

*1) Database:* To assess the performance of the temporal CNN based classifier, the ASVspoof 2015 database [29] is used in this work. Ten types of spoofing attack are included in this database, namely S1 to S10. The database consists of three subsets, including training set, development set and evaluation set. The training and development sets only contain the S1 to S5 spoofing attacks. The evaluation set consists of all the ten types of spoofing attack. Refer to our previous results [7, 28], it is observed that the results on the development set are similar to that of the S1 to S5 on the evaluation set. Hence, only the training set and the evaluation set are used in this work.

*2) Detection Systems:* To validate our proposals, the previous MLP based system [28] is used as the reference baseline. We describe the detail information as follows.

- *MLP based system:* Each of the features mentioned above with its delta and acceleration coefficients is used as the input vector to train its own classifier. The MLP consists of one hidden layer with 2,048 sigmoid nodes is used to perform frame-wise classification. Given testing utterance, the posterior probabilities of all the frames are estimated and averaged over the test utterance.
- *Temporal CNN based system:* The convolutional layer contains 512 filters. Each filter has a size of $256\times11$, which means the filter covers all the 256 frequency bins and contains 11 successive frames. Then 512 feature

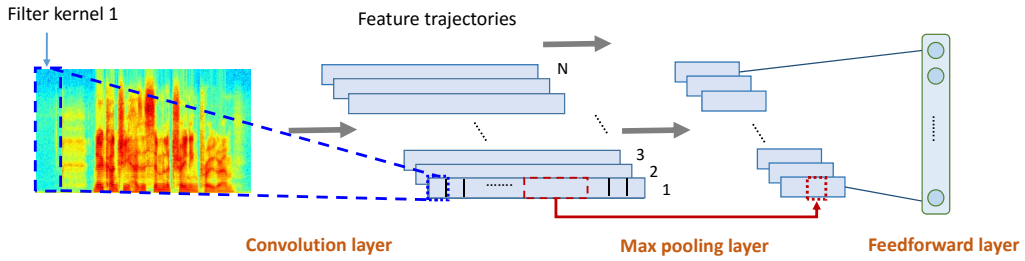


Fig. 2. Architecture of temporal CNN based classifier.

trajectories are generated by convolutional layer. The max-pooling layer down-samples the feature trajectories with the maximum values over a overlapped window of 100 frames with 80 frames overlap. The outputs of the max-pooling layer are fed to the feedforword layer with 2,048 sigmoid nodes. Finally, we use the softmax to predict the posterior probability of the input features.

To avoid the influence of the silence patches, a pitch based voice activity detector (VAD) is used to discard scores of silence at the beginning and the ending of the utterance.

*3) Evaluation metrics and fusion:* The equal error rate (EER), where the false acceptance rate and the miss rejection rate becomes equal, is used to evaluate the system performance.

A system fusion is applied on the feature-based results to benefit the advantages of different features. To avoid the over-fitting problem, the same weight is assigned to all systems. Then the fusion is adopted by averaging the scores of all systems to produce the final score.

In this work, the Bosaris toolkit[1] is used to compute the EERs of each feature and the fused system.

## B. Results and Analysis

*1) Evaluation with mismatched data:* Table I summarizes the experimental results in terms of evaluation with mismatched types of spoofing attacks. Specifically, we use the training set to train the classifiers and perform the evaluation on the evaluation set.

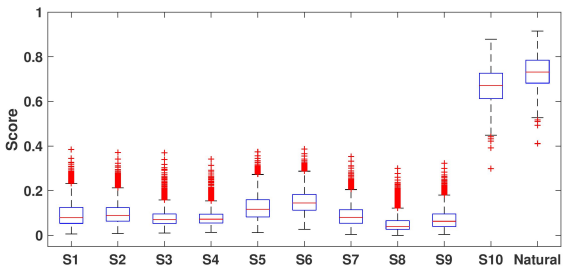


Fig. 3. Boxplot of score distribution of MLP based system using LMS feature on evaluation set. Red lines are medians, box edges are at 25% and 75% quantiles. Score distributed between 0 and 1. 0 indicates the spoofing speech, while 1 indicates natural speech. The red crosses imply the outliers.

[1]https://sites.google.com/site/bosaristoolkit/

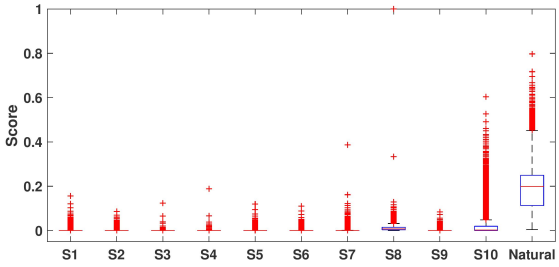


Fig. 4. Boxplot of score distribution of temporal CNN based system using LMS feature on evaluation set. Red lines are medians, box edges are at 25% and 75% quantiles. Score distributed between 0 and 1. 0 indicates the spoofing speech, while 1 indicates natural speech. The red crosses imply the outliers.

We first exam the effectiveness of the temporal CNN based system across different types of spoofing attack. For better analysis, we divide the spoofing attacks into two categorise: S1-S9, using vocoders and low-dimensional features used for generation; S10, generated by waveform-based unit selection.

From the preliminary results, some observations are found. First, although S6-S9 are only available during evaluation, the results of the MLP based system are comparable to that of S1-S5. This is due to the vocoders and low-dimensional features used for S6-S9 attacks generation are similar as S1-S5. As expected, similar phenomenons are observed in temporal CNN based system.

Second, in general, for S1-S9 attacks, the MLP based system slightly outperform to the temporal CNN based system. An example of the score distributions of two systems using LMS feature are showed in Figure 3 and Figure 4. This is due to the analysis and synthesis process of vocoder is based on frame. Thus the artificial cues of vocoder-based spoofing attacks will appear in each frames. The max-pooling layer of the temporal CNN based system, however, tends to down-sample the features in a long time span, i.e. 100 frames in our implementation, so some information are discarded. On the contrary, the MLP based system is capable to preserve all the feature details. Hence lead to a better performance on vocoder-based spoofing attacks.

Third, for both the MLP and the temporal CNN based systems, S10 performs worse than S1-S9. However, it is observed that in general the temporal CNN based system is notably outperform the MLP based system. Take LMS feature based

TABLE I

*EERs (%) of MLP and CNN based system on evaluation set. The classifiers are trained by training set (including S1-S5 attacks).*

| | | Evaluation set | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Classifier | Feature | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | Average |
| MLP | LMS | 0.02 | 0.02 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.00 | 0.01 | 35.24 | 3.53 |
| | IF | 1.30 | 1.30 | 1.30 | 1.30 | 1.34 | 1.32 | 1.30 | 1.30 | 1.30 | 25.56 | 3.73 |
| | BPD | 0.06 | 0.05 | 0.00 | 0.00 | 0.41 | 0.23 | 0.05 | 0.04 | 0.03 | 30.67 | 3.16 |
| | GD | 0.02 | 0.01 | 0.00 | 0.00 | 0.20 | 0.09 | 0.00 | 0.01 | 0.00 | 33.90 | 3.42 |
| | MGD | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.06 | 0.03 | 0.00 | 0.00 | 38.54 | 3.87 |
| | **Fusion** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 28.72 | 2.87 |
| Temporal CNN | LMS | 1.55 | 0.92 | 0.32 | 0.36 | 1.06 | 0.95 | 1.62 | 4.02 | 0.89 | 13.65 | 2.53 |
| | IF | 0.28 | 0.29 | 0.07 | 0.06 | 0.58 | 0.47 | 0.20 | 0.29 | 0.09 | 20.08 | 2.24 |
| | BPD | 0.66 | 0.91 | 0.13 | 0.16 | 1.54 | 1.25 | 0.75 | 0.65 | 0.39 | 38.19 | 4.46 |
| | GD | 0.16 | 0.12 | 0.13 | 0.12 | 0.27 | 0.37 | 0.10 | 0.16 | 0.08 | 12.50 | 1.40 |
| | MGD | 0.29 | 0.25 | 0.13 | 0.13 | 0.58 | 0.45 | 0.24 | 0.17 | 0.10 | 27.20 | 2.95 |
| | **Fusion** | 0.10 | 0.08 | 0.01 | 0.03 | 0.13 | 0.10 | 0.06 | 0.06 | 0.04 | 13.51 | 1.41 |

system as an instance. As shown in Figure 3 and Figure 4, it is observed that compare to the MLP based system, the temporal CNN based system can easily discriminate S10 from natural speech, and reduces the EER of 21.59% (from 35.24% to 13.65%). This indicates that the temporal CNN based classifier can effectively improve the system performance on unit selection based spoofing speech detection, owing to its capability to capture the long term temporal information.

Then, we exam the temporal CNN based system performance using difference features. We observe that in the MLP based system the best performance on evaluation set is obtained by the BPD feature achieving an averaged EER of 3.16%. While in the temporal CNN based system, the GD feature performs best with an averaged EER of 1.40%. However, the BPD feature in the temporal CNN based system performs much worse than in MLP based system, especially for S10 detection.

By applying the feature fusion, finally the MLP and the temporal CNN based system obtain the averaged EER of 2.87% and 1.41%, respectively. This implies the temporal CNN based system offer overall better performances than the MLP based system.

*2) Evaluation with matched data:* To further validate the performance of both MLP and temporal CNN based system, we adopt an experiment with matched types of spoofing attacks of training and evaluation. Hence, we divide the evaluation set into two parts, each types of spoofing attacks and natural speech are equally assigned into these two part. One part is used to train the classifiers and the evaluation is performed on the other part. Table II summarizes the experimental results.

We first analysis the performance of the MLP based system. It is surprisingly to observe that the performance of the MLP based system is significant degraded by including all the ten types of spoofing attacks in classifier training. Even after system fusion using different features, the averaged EER is 3.60%. Figure 5 shows an example of score distribution using LMS feature to train the MLP based system. We observe that the scores of spoofing attacks are overlapped with that of natural speech. It because the MLP based system performs

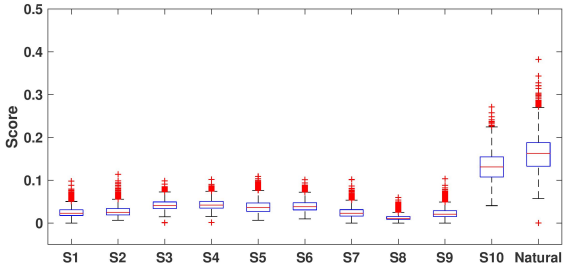


Fig. 5. Boxplot of score distribution of MLP based system using LMS feature on evaluation set. Red lines are medians, box edges are at 25% and 75% quantiles. Score distributed between 0 and 1. 0 indicates the spoofing speech, while 1 indicates natural speech. The red crosses imply the outliers.

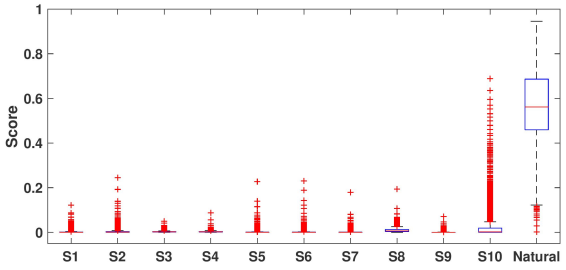


Fig. 6. Boxplot of score distribution of temporal CNN based system using LMS feature on evaluation set. Red lines are medians, box edges are at 25% and 75% quantiles. Score distributed between 0 and 1. 0 indicates the spoofing speech, while 1 indicates natural speech. The red crosses imply the outliers.

frame-wise with short windows for training and classification. For unit selection based spoofing attack (S10), during the short window the training vectors of S10 are highly similar to natural speech, as each frame of S10 is actually from natural speech. This will inevitably decrease the discrimination power of the classifier and degrade the system performance.

Then we analysis the performance of the temporal CNN based system. For S6-S9 attacks, even all these spoofing types attack are observed during training, the performance of temporal CNN based system almost consistent with previous results. This is due to S6-S9 attacks are vocoded speech which is similar as S1-S5. They may not provide additional useful

TABLE II

*EERs (%) of MLP and CNN based system on half of the evaluation set. The classifiers are trained by the other half of the evaluation set (including S1-S10 attacks).*

| | | Evaluation set | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Classifier | Feature | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | Average |
| MLP | LMS | 0.86 | 1.17 | 1.39 | 1.44 | 2.16 | 1.45 | 0.92 | 0.12 | 0.91 | 36.99 | 4.74 |
| | IF | 0.64 | 0.94 | 0.12 | 0.12 | 3.49 | 3.27 | 0.30 | 0.12 | 0.20 | 33.24 | 4.24 |
| | BPD | 0.40 | 0.75 | 0.05 | 0.06 | 2.95 | 1.75 | 0.37 | 0.18 | 0.25 | 34.78 | 4.15 |
| | GD | 0.58 | 0.42 | 1.28 | 1.35 | 5.85 | 3.22 | 0.17 | 0.33 | 0.20 | 38.42 | 5.18 |
| | MGD | 0.96 | 1.82 | 1.00 | 1.22 | 6.79 | 6.99 | 1.90 | 0.22 | 0.63 | 42.86 | 6.44 |
| | **Fusion** | 0.08 | 0.12 | 0.04 | 0.04 | 0.31 | 0.36 | 0.07 | 0.02 | 0.05 | 34.89 | 3.60 |
| Temporal CNN | LMS | 0.42 | 0.77 | 0.18 | 0.22 | 0.53 | 0.64 | 0.36 | 0.42 | 0.20 | 5.40 | 0.91 |
| | IF | 0.33 | 0.36 | 0.15 | 0.10 | 0.96 | 0.85 | 0.19 | 0.40 | 0.13 | 8.42 | 1.19 |
| | BPD | 2.05 | 5.25 | 3.23 | 3.42 | 7.75 | 9.00 | 2.89 | 1.66 | 2.40 | 36.20 | 7.43 |
| | GD | 0.60 | 0.37 | 0.34 | 0.30 | 0.89 | 0.84 | 0.42 | 1.03 | 0.20 | 7.01 | 1.20 |
| | MGD | 0.16 | 0.27 | 0.15 | 0.13 | 0.52 | 0.40 | 0.22 | 0.30 | 0.30 | 6.07 | 0.90 |
| | **Fusion** | 0.07 | 0.06 | 0.06 | 0.04 | 0.13 | 0.11 | 0.04 | 0.09 | 0.05 | 3.45 | 0.41 |

information for the training process. On the contrary, in the temporal CNN based system, long term temporal information will be covered for classification. For unit selection based spoofing attack, the information includes concatenating transitions. Hence adding S10 attacks in the classifier training can effectively improve the performance of temporal CNN based system in unit selection based spoofing attack detection. Figure 6 shows the score distribution using LMS feature to train the temporal CNN based system. It is observed that we can easily discriminate natural speech from spoofing attacks.

By applying the system fusion, the EER of S10 decreases from 34.89% in MLP to 3.45% in temporal CNN. Finally we obtain the averaged EER of 0.41% in temporal CNN, which is much lower than that of the MLP based system (3.60%). This confirms the effectiveness of the temporal CNN based system for spoofing speech detection.

## IV. Conclusions

In this paper, we investigate the used of the temporal CNN based classifier to discriminate synthetic spoofing attacks and natural speech. The experiments results show that,

- compare to the MLP based classifier, the temporal CNN based classifier can effectively boost the performance of unit selection based spoofing attack detection;
- when including the S10 in training, the performance of the temporal CNN based system for unit selection based spoofing attack is improved significantly.

However, there is a degradation of the temporal CNN based system for vocoder based spoofing attacks detection. This will become a research point of our following work. Moreover, as inconsistent performance is observed using BPD feature in two systems, in future, we will investigate the robustness of features for spoofing speech detection.

## Acknowledgment

This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its IDM Futures Funding Initiative. This work is also supported by the DSO funded project MAISON DSOCL14045, Singapore.

## References

[1] Heiga Zen, Keiichi Tokuda, and Alan W Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] Yannis Stylianou, "Voice transformation: a survey," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[3] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[4] Zhizheng Wu, Phillip L De Leon, Cenk Demiroglu, Ali Khodabakhsh, Simon King, Zhen-Hua Ling, Daisuke Saito, Bryan Stewart, Tomoki Toda, Mirjam Wester, et al., "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[5] Zhizheng Wu, Chng Eng Siong, and Haizhou Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. INTERSPEECH*, 2012.

[6] Zhizheng Wu, Tomi Kinnunen, Eng Siong Chng, Haizhou Li, and Eliathamby Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2012, pp. 1–5.

[7] Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Eng Siong Chng, and Haizhou Li, "Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge," in *Proc. INTERSPEECH*, 2015.

[8] Yi Liu, Yao Tian, Liang He, Jia Liu, and Michael T Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," in *Proc. INTERSPEECH*, 2015.

[9] Phillip L De Leon, Inma Hernaez, Ibon Saratxaga, Michael Pucher, and Junichi Yamagishi, "Detection of synthetic speech for the problem of imposture," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4844–4847.

[10] Jon Sanchez, Ibon Saratxaga, Inma Hernaez, Eva Navas, and Daniel Erro, "The AHOLAB RPS SSD spoofing challenge 2015 submission," in *Proc. INTERSPEECH*, 2015.

[11] Longbiao Wang, Yohei Yoshida, Yuta Kawakami, and Seiichi Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proc. INTERSPEECH*, 2015.

[12] Jon Sanchez, Ibon Saratxaga, Inma Hernaez, Eva Navas, Daniel Erro, and Tuomo 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.

[13] Tanvina B Patel and Hemant A Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Proc. INTERSPEECH*, 2015.

[14] Tomi Kinnunen, Zhi-Zheng Wu, Kong Aik Lee, Filip Sedlak, Eng Siong

Chng, and Haizhou Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4401–4404.

[15] Takayuki Satoh, Takashi Masuko, Takao Kobayashi, and Keiichi Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," in *Proc. EUROSPEECH*, 2001, pp. 759–762.

[16] Jean-François Bonastre, Driss Matrouf, and Corinne Fredouille, "Artificial impostor voice transformation effects on false acceptance rates.," in *Proc. INTERSPEECH*, 2007, pp. 2053–2056.

[17] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. INTERSPEECH*, 2015.

[18] Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li, "Synthetic speech detection using temporal modulation feature," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7234–7238.

[19] Xiaohai Tian, Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li, "Spoofing detection from a feature representation perspective," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

[20] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei, "Large-scale video classification with convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*. 2012.

[22] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back, "Face recognition: A convolutional neural-network approach," *IEEE Transactions onNeural Networks*, 1997.

[23] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[24] Yann LeCun and Yoshua Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, 1995.

[25] Leigh D Alsteris and Kuldip K Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.

[26] Michal Krawczyk and Timo Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, 2014.

[27] Bayya Yegnanarayana and Hema A Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Transactions on Signal Processing*, vol. 40, no. 9, pp. 2281–2289, 1992.

[28] Xiaohai Tian, Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li, "Spoofing detection from a feature representation perspective," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

[29] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. INTERSPEECH*, 2015.