An Investigation of Deep Learning Frameworks for Speaker Verification Anti-spoofing

Chunlei Zhang, *Student Member, IEEE,* Chengzhu Yu, *Student Member, IEEE,*
John H.L. Hansen, *Fellow, IEEE*

***Abstract*—In this study, we explore the use of deep learning approaches for spoofing detection in speaker verification. Most spoofing detection systems that have achieved recent success employ hand-craft features with specific spoofing prior knowledge, which may limit the feasibility to unseen spoofing attacks. We aim to investigate the genuine-spoofing discriminative ability from the back-end stage, utilizing recent advancements in deep learning research. In this work, alternative network architectures are exploited to target spoofed speech. Based on this analysis, a novel spoofing detection system which simultaneously employs Convolutional Neural networks (CNNs) and Recurrent Neural Networks (RNNs) is proposed. In this framework, CNN is treated as a convolutional feature extractor applied on the speech input. On top of the CNN processed output, recurrent networks are employed to capture long-term dependencies across the time domain. Novel features including Teager Energy Operator Critical Band Autocorrelation Envelope (TEO-CB-Auto-Env), Perceptual Minimum Variance Distortionless Response (PMVDR) and a more general spectrogram are also investigated as inputs to our proposed deep learning frameworks. Experiments using the ASVspoof 2015 Corpus show that the integrated CNN-RNN framework achieves state-of-the-art single system performance. The addition of score-level fusion further improves system robustness. A detailed analysis shows that our proposed approach can potentially compensate for the issue due to short duration test utterances which is also an issue in the evaluation corpus.**

***Index Terms*—Spoofing detection, convolutional neural networks, recurrent neural networks, TEO-CB-Auto-Env, PMVDR, spectrogram.**

I. Introduction

SPEAKER verification, serving as a popular and flexible solution for biometric authentication, has drawn more attention in recent years [1]. It also represents one of the core scientific concentrations in the U.S. OSAC (Organization for Scientific Area Committees)[1]. In a speaker verification system, a decision whether to reject or accept a claimed identity is made based on a speaker's known utterance. Typical applications include log-in for smart devices, door access control, online information access, telephone banking, etc. [2]. While recent advancements in channel variability modeling, short train/test duration, context mismatch and noise compensation have greatly improved the reliability of speaker verification systems, studies have shown that speaker verification systems remain vulnerable to intentional spoofing attacks [2]–[6]. Earlier studies have also highlighted the vulnerability of GMM-UBM speaker recognition solutions to computer altered speech [7]. In the context of voice biometrics, spoofing refers to when an impostor attempts to masquerade as an enrolled speaker by falsifying speech data traits. Previous studies have shown that the false acceptance rate of state-of-the-art speaker verification systems has been significantly increased with replayed speech, impersonation, synthesized speech, voice conversion and artificial signals as spoofing attacks [2], [6], [8]–[14].

Countermeasures have been investigated since the vulnerability caused by spoofing attacks have been identified by the research community recently. The most common and effective strategy is to build a stand-alone spoofing detection system before the speaker verification system [2], [6]. As an emerging field in speaker verification, standard large-scale datasets including protocols are still being developed for evaluation of spoofing countermeasures [6], [15]. Most countermeasures are developed using closed, specific spoofing datasets, where prior knowledge about the specific spoofing type plays an important role for detection [16], [17]. In the First Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015), which was held as a special session in INTERSPEECH 2015, the challenge result confirms this observation. In the challenge, phase-based features or prosodic features such as Relative Phase Shift (RPS), or modified group delay (MGD) achieved good performance [6], [18], [19]. Careful analysis of the effectiveness of these features in detecting spoofing reveals that, phase information is lost/changed during the analysis-synthesis step in some speech-synthesis techniques, which makes genuine speech different from that which has been synthesized [16]. This represents the main case in ASVspoof 2015: all 10 spoofed speech categories in the challenge corpus are from Voice Conversion (VC) or Speech Synthesis (SS) algorithms. However, such prior knowledge is unrealistic in practice, thus these features are not guaranteed to be effective to attacks which have unchanged phase information [20]. A robust spoofing detection solution is therefore required to generalize well for unknown attacks from different spoofed types.

For this consideration, efforts have also been made to explore features that do not depend on strong prior knowledge for spoofing detection. In [21], the authors employ Local Binary Patterns (LBPs), which is a particular case of texture features adopted in many face verification or face spoofing

This project was funded in part by AFRL under contract FA8750-15-1-0205 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

The authors are with the Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: chunlei.zhang@utdallas.edu; john.hansen@utdallas.edu).

Manuscript received August 15, 2016.

[1]http://www.nist.gov/forensics/osac.cfm

detection [22], [23]. The main advantage of LBP feature is that they capture the differences in the spectro-temporal texture of genuine and spoofed speech, but relies on a genuine speech model. By doing this, the LBP feature combined with a one-class SVM is expected to generalize well to unseen attacks. In our submission to ASVspoof 2015 [24], the Teager Energy Operator Critical Band Autocorrelation Envelope (TEO-CB-Auto-Env) and Perceptual Minimum Variance Distortionless Response (PMVDR) features were investigated for this task. Motivated by a non-linear speech production model, the TEO based features were initially designed to capture variabilities introduced to stressed speech. In the context of spoofing detection, TEO-CB-Auto-Env features are also expected to be able to model the differences between genuine and spoofed speech [25]. While for PMVDR [26], this feature can accurately model the upper spectral envelope at perceptually important harmonics. By incorporating this perceptual consideration, PMVDR is expected to be suitable for detecting spoofed speech. More generally, 40 dimensional filter-bank features with a deep neural network (DNN) as a back-end classifier was reported to be effective in spoofing detection [24]. This is an interesting observation, indicating that the back-end advancement for the speaker verification anti-spoofing task is also a good direction to explore.

In addition to front-end feature investigation, different back-end classifiers have also been examined in recent studies. Gaussian Mixture model (GMM) or a simple Gaussian Classifier (GC) can achieve good performance with respect to alternative features [19], [24], [27], [28]. Linear Discriminative Analysis (LDA), Probabilistic Linear Discriminative Analysis (PLDA) were both reported to be effective in certain conditions [16], [24], [29]. One observation obtained from those systems is that the back-end effectiveness depends on the front-end features. For example, PLDA improves performance of a joint spoofing detection and speaker verification system with Mel-cepstral features and linear predictive coding based features [29], but did not show much advancement in our TEO-CB-Auto-Env or PMVDR based i-vector spoofing detectors [24]. To consider the problem of generality in the context of multiple, unknown spoofing attacks, one-class SVM was investigated [18], [21]. By incorporating a one-class SVM, it is expected to overcome the main drawback of multi-class classification, which over-fits the known attacks and loses the generalization to attacks that are not previously observed. As mentioned above, DNN frameworks achieved competitive results in the ASVspoof 2015 Challenge. For spoofing detection, DNN can be utilized either as a back-end classifier [18], [24] or a feature extractor [30].

In this study, we focus on developing a spoofing detector solution from the back-end level. More specifically, we examine different deep learning frameworks (i.e., Deep neural networks(DNNs), Convolutional Neural networks (CNNs) and Recurrent Neural Networks (RNNs)) for spoofing detection [31]–[34]. Based on that, a novel deep learning architecture which integrates both CNN and RNN is proposed:

DNN: DNN are reported to be effective in different studies [18], [24], [30]. In this work, we provide DNN results with a range of features, for the purpose of system comparison.

TABLE I
EER(%) ON ASVSPOOF 2015 DEVELOPMENT DATA. THE PERFORMANCE IS EVALUATED WITH TEO-CB-AUTO-ENV AND PMVDR, RESPECTIVELY. THE RESULTS FROM TWO SYSTEMS CONFIRM THAT SAD DOES NOT HELP ON SPOOFING DETECTION.

Spoofing type	S1	S2	S3	S4	S5
TEO/SAD	0.52	3.16	0.37	0.40	1.28
TEO/NO SAD	**0.34**	**2.42**	**0.25**	**0.18**	**1.07**
PMVDR/SAD	0.37	2.43	0.29	0.18	0.87
PMVDR/NO SAD	**0.28**	**2.23**	**0.09**	**0.12**	**0.98**

CNN: Although the idea of employing CNN for spoofing detection is not new, most studies focus on face spoofing detection [35], [36]. In the context of speaker verification anti-spoofing, to the best of our knowledge, no studies have been reported that successfully apply CNN as a spoofing detector. In this study, we intend to investigate CNNs for speaker verification anti-spoofing.

RNN: RNN is also investigated. The intuition behind our investigation with RNN is that RNN is able to capture the long-term dependencies along a consecutive sequence (i.e., time in speech application). In fact, as indicated in TABLE I, we found that speech activity detection (SAD) does not improve performance of our i-vector systems which were submitted to the challenge. Based on this observation, we conclude that speech alteration algorithms, such as voice conversion and speech synthesis, have a consistent influence on the speech utterance, even for the background noise which are always discarded in speech recognition and speaker recognition. For this reason, RNN is considered to be a proper model for consistent "spoofing".

CNN+RNN: To further explore the advancement of deep learning frameworks for spoofing detection, we propose a combined CNN+RNN architecture for this task. Here, CNN plays the role of a feature extractor, and RNN is employed to capture the consistent "spoofing" properties. Through back-propagation, the feature extractor (CNNs) and final classification network (RNNs) are optimized simultaneously.

In addition to TEO-CB-Auto-Env and PMVDR features that we adopted in a previous work, we propose a modified spectrogram, a more general feature without any design, as input to our proposed deep frameworks. A detailed feature description is presented in Section II. while the proposed deep learning frameworks are introduced in Section III.

Although more detailed explanations and analysis can be found throughout this paper, let us first summarize the novel contributions and findings below:

1) The features that we propose do not rely on spoofing prior knowledge, which is expected to generalize well to unknown spoofing attacks that were not observed during the training phase.

2) Our proposed systems do not require that we optimize features and classifiers separately. Such an end-to-end approach may have several benefits. For example, modeling from the utterance level reduces the overall system complexity (one vs. number of frames evaluations per utterance). Moreover, this approach often results in considerably simplified systems requiring fewer concepts and heuristics.

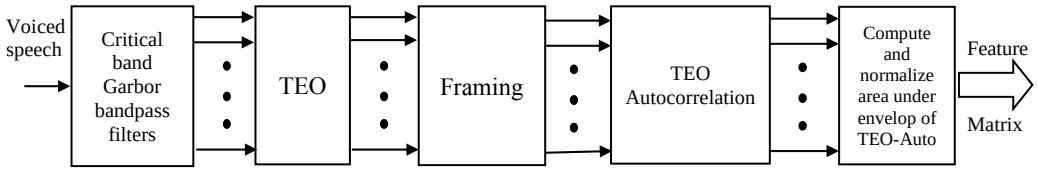


Fig. 1. TEO-CB-Auto-Env feature extraction. We use 18 Gabor filter banks, the bandwidth is partitioned with critical band, the 18^{th} band correspond to 4 kHz. In this manner, 18 dimensional features is extracted from each frame.

3) As one of the earliest studies which employs CNNs and its combination with RNNs for spoofing detection, we demonstrate that deep learning methods, without expertise for spoofing, could also achieve state-of-the-art anti-spoofing performance.

The remainder of this paper is organized as follows. Section II outlines the features used in this study; Section III presents details of our spoofing detection models; Section IV reports experimental setups, evaluations and results; Section V discusses key observations stemming from the experiments. Finally, we conclude in Section VI with a look ahead towards future work.

II. Features for Spoofing Detection

In this section, the following features that are considered for spoofing detection are presented: TEO-CB-Auto-Env, PMVDR and spectrogram features. In addition, a feature preparation method with cropping or padding to help unify different duration lengths of utterances is investigated [37].

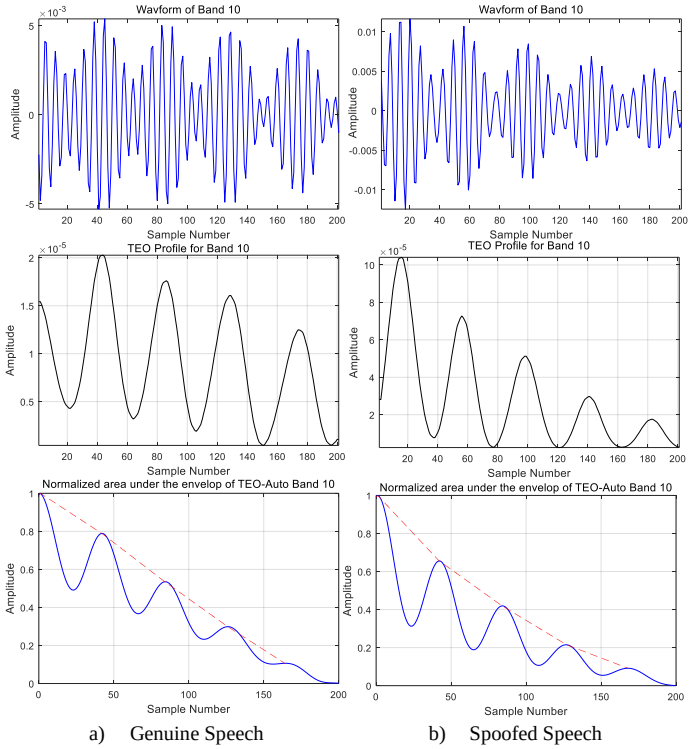
A. TEO-CB-Auto-Env

Fig. 1 shows a flow diagram of TEO-based feature extraction [25]. The TEO profile obtained from the critical band[2] based Gabor bandpass filter output is initially segmented on a short-term basis. Next, an auto-correlation is applied after framing. Once the auto-correlation response is found, the area under the auto-correlation envelope is obtained and normalized. One area coefficient is obtained for each filter band. It has been shown to be large for genuine speech and low for spoofed speech, corresponding to large area coefficient for neutral speech and small coefficient for stressed speech in stress detection tasks [25], [38]. To show the difference, we select one frame with the same time coordinate from one genuine utterance and one spoofed utterance "S1" respectively, and compute the TEO-profiles and corresponding feature coefficients. In order to have a fair comparison, we carefully select utterances with the same context. As shown in Fig.2, the differences can be found in TEO profiles and normalized areas, which suggests that TEO-CB-Auto-Env is a sensitive and effective candidate for spoofing detection tasks.

B. PMVDR

PMVDR features were first proposed by Yapanel and Hansen [26]. PMVDR computes cepstral coefficients by incorporating perceptual warping of a FFT power spectrum,

[2]http://www.sfu.ca/sonic-studio/handbook/Appendix_E.htmL



a) Genuine Speech b) Spoofed Speech

Fig. 2. Plots of wavforms, TEO profiles and Normalized areas for genuine and spoofed frames. For simplicity, we only display critical band 10. The higher value for genuine speech highlights the more natural regularity than that of spoofed speech

replacing the Mel-scaled filter bank with the minimum variance distortionless response (MVDR) spectral estimator. These features have better spectral modeling ability of speech signals compared to traditional feature extraction methods [26]. Previous studies have shown that perceptual knowledge can differentiate between genuine and spoofed speech. Since PMVDR incorporates perceptual warping of the spectrum, we used PMVDR for this task. A schematic diagram of the PMVDR front-end is shown in Fig. 3.

C. Spectrogram

As an image representation of the power spectrum, generation of a traditional spectrogram is straightforward. After applying Short-time Fourier transform (STFT) to pre-processed frames (a 256 point Hanning Window with 0.5 skip rate), the spectrogram is formulated with Equation (1):

$$Spectrogram(t,\omega) = |STFT(t,\omega)|^2, \quad (1)$$

For this study, an additional step is performed where we convert the conventional spectrogram into a log scale with

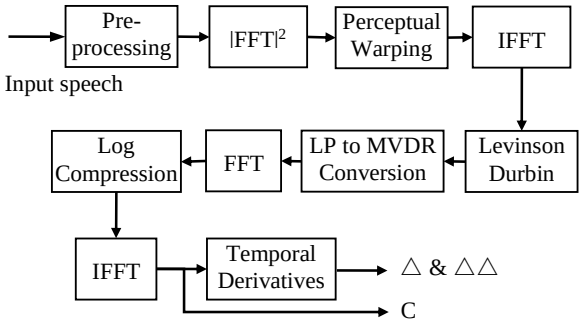


Fig. 3. Flow diagram of PMVDR feature extraction. Pre-processing includes pre-emphasis, frame-blocking and Hamming windowing. For window size and shift, we use the same configuration as TEO-CB-Auto-Env feature, which is a 25 ms analysis window with 10 ms frame shift.

Equation (2), so the decibel (dB) scale is used instead of amplitude.

$$Spectrogram(t,\omega)|_{dB} = 20\log_{10}\left(\frac{|STFT(t,\omega)|}{2\times 10^{-5}}\right), \quad (2)$$

D. Forced alignment for feature matrix

For the ASVspoof 2015 Corpus, the duration of each utterance is not fixed, where the average length is typically 3.5s. Based on this observation, we create a unified 4s feature matrix by incorporating either data padding or cropping. Although it is not necessary to submit fixed-length inputs into RNNs, we maintain this unified feature format unless specified. In fact, a similar operation process could be found in text-dependent speaker recognition as well [37]. Actually, experimental analysis in Section V will show that this simple solution to be presented has the benefit in addressing the issue of short duration.

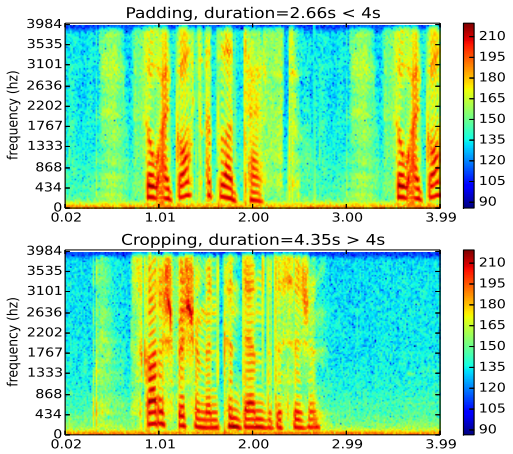


Fig. 4. Illustration of padding and cropping. We concatenate repeat data for the utterance which is less than 4s; we select 4s consecutive data for utterance which is more than 4s. By this operation, we create a unified format for deep learning frameworks.

III. Deep Learning Frameworks

In this section, we outline the essentials of deep learning frameworks for spoofing detection and highlight the architecture of the proposed **CNN+RNN** solution in this study. All models are implemented with with the toolkit: Theano [39].

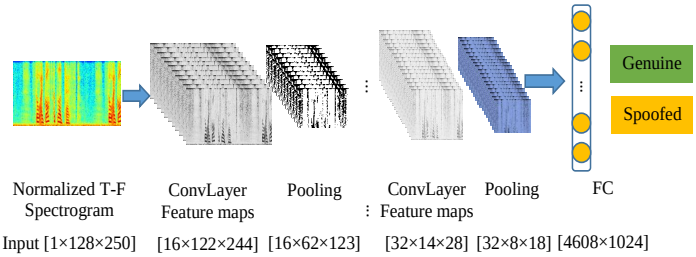


Fig. 5. Diagram of CNN architecture for spoofing detection. ConvLayer stands for convolutional layer; we apply maxpooling for downsampling. The output of CNNs (i.e., [32×8×18]) is vectorized and then submitted to a fully-connected (FC) network.

A. Deep neural networks

DNNs for spoofing detection are kept in similar structures with our previously proposed model [24]. We use 4 hidden layers with 1024 hidden nodes per layer, where the final softmax layer consists of two or six nodes (it depends on how many labels of spoofed speech used for training the spoofing detection networks), which represent the genuine and spoofing class probability respectively. We also compare different activation functions, Rectified Linear Units $ReLU$ performs best [40].

$$f(x) = \max(0, x), \quad (3)$$

where x is the input to a neuron. To address overfitting, a 50% Dropout is applied to every hidden layer [41]. For features with different dimensions, the only difference is the input layer. For example, 18 dimensional TEO-CB-Auto-Env features are derived for each frame. Subsequently, 11 consecutive frames of TEO features (198 dimension in total) are vectorized as the first layer input. Since there is no speech context information involved in this classification task, only the "genuine" or "spoofing" label is assigned for each input acoustic feature set. Therefore, for the decoding part, an average score across the whole utterance is computed as the final classification probability.

B. Convolutional neural networks

Here, the classification task consists of determining whether input speech utterances are genuine or spoofed. With padding or cropping, we have created features (e.g., spectrogram) for each utterance with a unified time-frequency (T-F) shape. The classifier maps feature matrices to either genuine/spoofed class probabilities using several convoluational/pooling layers as feature extractors, followed by a fully connected network with a softmax layer as the final classification layer. The architecture is illustrated in Fig.5. Here, we use a normalized T-F spectrogram with [1×128×250] dimensionality as input to the CNNs. In this case, a height 128 represents the number of frequency bins, width 250 is the length along time axis. ConvLayer will compute the output of neurons that are connected to local regions in the input. A ConLayer $C_{m\to n}^{k\times l}$

computes $m \times n$ convolutions between m input frames an n output frames, with convolution filters of size $k \times l$. Thi may result in a total volume of [16×122×244] if we decide to use 16 filters, 7×7 convolutions. Pooling will perform downsampling operation along the spatial dimensions (width height), resulting in a smaller volume as [16×62×123]. Th FC (i.e. fully-connected) layer will compute the class scores resulting in a volume of size [1×1×2], which represents th probabilities of both genuine and spoofed speech.

C. Recurrent neural networks

As the extension of the conventional feedforward neura network, an RNN is designed to address a variable length inpu sequence. This is particularly suitable for modeling speech By having a recurrent hidden state whose activation at each time is dependent on that from the previous time, an RNN could learn the long-term dependencies of the sequence. For example, given a sequence $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)$, the RNN updates its recurrent hidden state $\mathbf{h}_t$ by

$$\mathbf{h}_t = \begin{cases} 0, & t = 0 \\ \phi(\mathbf{h}_{t-1}, \mathbf{x}_t), & otherwise \end{cases} \quad (4)$$

where ϕ is a nonlinear function. To prevent the problem of gradient vanishing, a recently proposed gated recurrent unit (GRU) activation function is employed in this study [42]. The proposed RNN architecture is shown as Fig. 6. Although it is not necessary to force align the feature dimensions, we simply use our proposed padding or cropping method to keep the same input size format as the CNN or CNN+RNN model.

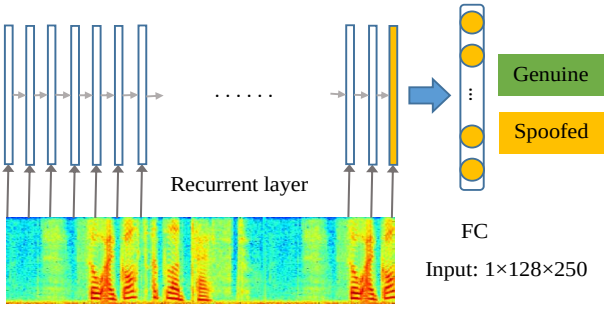


Fig. 6. Diagram of RNN architecture for spoofing detection. As shown in the figure, we employ many-to-one recurrent model for the classification task. The output of the recurrent layer is followed by a fully-connected hidden layer and a final classification softmax layer, similar to that of CNNs.

D. Integration of CNNs and RNNs

As noted in Section I, the introduction of SAD to the spoofing detection pipeline did not help to improve our i-vector based systems previously submitted to ASVspoof 2015. This observation suggests that "spoofing" may have a consistent effect on genuine speech. Inspired by the fact that the CNN plays a role for extracting genuine/spoofing discriminative features, and that an RNN is capable of modeling the long-term dependencies (in this study, "spoofing" is the factor which we want to model with the RNN) across the long sequence

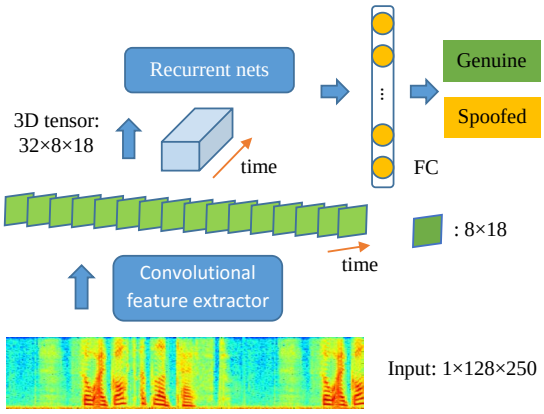


Fig. 7. Diagram of CNN+RNN architecture for spoofing detection. From the 3D tensor, the 32×8 snippet is vectorized to feed into RNN layers.

instead of short frames, we propose to employ CNNs and RNNs simultaneously for spoofing detection.

In the CNN+RNN framework, the output of the CNN is a set of channels (i.e., feature maps), as illustrated in Fig. 7. For example, here the feature maps are formulated as a 3D tensor, where 18 is the number of time steps mapped from the 250 time steps in the original spectrogram. This means 18 recurrent layers should be constructed in the RNNs. Similar with other frameworks, the RNN output is followed by a fully-connected network with a softmax layer for final classification.

Exploiting recent advancements in deep learning research, Batch Normalization is implemented for CNN and RNN model [43]. This significantly reduces training time. Also, 50 % Dropout is applied to the final fully-connected layer to address overfitting.

IV. Experiments

In this section, we first provide an overview of the corpus used in our experiments, and present our initial results on development data and evaluation data, respectively.

A. ASVspoof 2015 Corpus

ASVspoof 2015 provides a database which consists of both genuine and spoofed speech, with the aim to boost research for developing generalized countermeasures to spoofing attacks. The spoofed speech is generated from the original genuine speech with different speech synthesis (SS) and voice conversion (VC) algorithms. For more details about spoofing techniques, please see [6].

The whole dataset is partitioned into three subsets: training, development and evaluation. The training set is provided to train spoofing and genuine speech models. The development set is used to test models constructed from training data, as well as for score calibration when we want to fuse different spoofing detection systems. Finally, spoofing detection performance is measured on the evaluation set. In the corpus, 10 spoofing algorithms were used to generate spoofed utterances. All three subsets contain spoofing types S1-S5, which are denoted as *known* attacks; while S6-S10 only appear in the

TABLE II
DEEP ARCHITECTURES PROPOSED FOR SPOOFING DETECTION. $Conv$, $Pooling$, $filts$ WITH PARAMETERS $[7 \times 7, 3 \times 3, 16]$ STANDS FOR A $[7 \times 7]$ 2D CONVOLUTION LAYER, A $[3 \times 3]$ MAXPOOLING LAYER, WITH 16 FILTERS. 4 SUCH $Conv$, $Pooling$, $filts$ LAYERS ARE INVESTIGATED AS CONVOLUTIONAL FEATURE EXTRACTORS. FOR RNNS, A $Recurrent$ LAYER IS SET TO HAVE 300 NODES. FC= FULLY CONNECTED LAYER. THE INPUT OF THESE NETWORKS IS THE $[1\times128\times250]$ SPECTROGRAM.

layer	Conv, Pooling, filts	Conv, Pooling, filts	Conv, Pooling, filts	Conv, Pooling, filts	Recurrent	FC	output
CNNs	7×7, 3×3, 16	5×5, 3×3, 32	3×3, 3×3, 32	3×3, 3×3, 32	/	1024	6
RNNs	/	/	/	/	300	1024	6
CNNs+RNNs	7×7, 3×3, 16	5×5, 3×3, 32	3×3, 3×3, 32	3×3, 3×3, 32	300	1024	6

evaluation part and are denoted as $unknown$ attacks. The purpose of adding unknown attacks is to test the generalization ability of the spoofing countermeasures to previously unseen attacks. A Summary of statistics of the ASVspoof 2015 Corpus is shown in TABLE III.

TABLE III
STATISTICS OF ASVSPOOF 2015 CORPUS. S1 TO S5 ARE KNOWN ATTACKS, S6 TO S10 ARE UNKNOWN ATTACKS SEEN ONLY IN THE EVALUATION SET. DEV=DEVELOPMENT; EVA=EVALUATION. WE ALSO PROVIDE THE MEAN DURATION OF EACH SPEECH TYPE FOR EVALUATION SET.

Spoofing	Train	Dev	Eva	Eva MD / s
Genuine	3750	3497	9404	3.58
S1	2525	9975	18400	3.50
S2	2525	9975	18400	3.50
S3	2525	9975	18400	2.63
S4	2525	9975	18400	2.63
S5	2525	9975	18400	3.50
S6	0	0	18400	3.50
S7	0	0	18400	3.50
S8	0	0	18400	2.63
S9	0	0	18400	3.50
S10	0	0	18400	2.48

B. Evaluation metric

While more details regarding the evaluation metric can be found in [44], we provide a brief overview in this section. The evaluation metric provided by the ASVspoof 2015 Challenge treats spoofing detection as a verification task: test whether the utterance belongs to the genuine speech class. We use the false alarm rate, $P_{fa}(\theta)$ and the miss rate, $P_{miss}(\theta)$, in a way similar to the evaluation of speaker recognition, which are defined in the challenge as:

$$\begin{aligned} P_{fa}(\theta) &= \frac{\#\{spoofed\ trials\ with\ score > \theta\}}{\#\{total\ spoofed\ trials\}}, \\ P_{miss}(\theta) &= \frac{\#\{genuine\ trials\ with\ score \leq \theta\}}{\#\{total\ spoofed\ trials\}}. \end{aligned} \quad (5)$$

The Equal Error Rate (EER) is the primary metric for the challenge. EER corresponds to the threshold $\theta|_{EER}$ at which the two detection error rates are equal (i.e., EER = $P_{fa}(\theta|_{EER})$ = $P_{miss}(\theta|_{EER})$). In this study, we use the same metric for ease in comparison with other systems for the same task.

C. Experimental setup

The sample rate of the corpus is 16 kHz. In fact, networks that use frequency content up to 8 kHz do not help in spoofing detection compared with that on 4 kHz frequency content in our experiments (probably because the higher frequencies do not contain much useful genuine-spoofing information and potentially increases overfitting). For this consideration, we only extract features corresponding to 4 kHz frequency content. So, in this study, the TEO-CB-auto-Env feature is 18 dimensional, PMVDR feature is 36 dimensional, and the number of frequency bins for the spectrogram feature is set to 128.

For DNNs with TEO-CB-Auto-Env and PMVDR features, we follow our previous DNN setup. The only difference is the input layer: for TEO-CB-Auto-Env, the input layer has 18×11 nodes; and for PMVDR, the input layer has 36×11 nodes. The output layer is a softmax of dimension 6, with one output for the human hypothesis, and one output for each of the five types of spoof in the training set. Actually, a softmax layer with only 2 nodes (i.e., genuine and spoofing labels) does not perform well in our experiments, mainly because of overfitting and imbalanced data (3750 genuine utterances vs. 12625 spoofed utterances). We compute the log-likelihood ratio (LLR) given the proposed networks:

$$LLR = \log P(genuine|\mathbf{F}) - \log(1 - p(genuine|\mathbf{F})), \quad (6)$$

where $\mathbf{F}$ is the feature vector, $P(genuine|\mathbf{F})$ is the output posterior w.r.t. the genuine model.

For CNNs, RNNs and CNNs+RNNs with spectrogram features as input, detailed architecture hyper-parameters are illustrated in TABLE II. We also explored adding TEO-CB-Auto-Env and PMVDR features to these more advanced networks, but no better results were achieved due to overfitting.

With the experimental setup, results from 5 newly proposed systems (i.e., DNNs with TEO and PMVDR, CNNs, RNNs and CNNs+RNNs with spectrogram features) are reported in the following sections.

D. Results on development set

The EER performance from different proposed systems and their fusion on the development set are listed in TABLE V. The DET curves are also illustrated in Fig.8. From the experimental results, we confirm that all three proposed features are effective in spoofing detection. Compared with ether the TEO or PMVDR i-vector system in our previous submissions, the single feature TEO or PMVDR with DNN back-end achieves better performance, which shows the effectiveness of the discriminative model such as DNNs in spoofing detection task.

By converting speech into a spectrogram feature set as the input feature, CNN achieves the best single system performance. This is not surprising because CNNs were initially

TABLE IV
EER(%) FOR DIFFERENT SPOOFING ATTACKS ON EVALUATION DATA. KNOWN ATTACKS INCLUDE S1-S5; UNKNOWN ATTACKS INCLUDE S6-S10. WE ALSO REPORT THE AVERAGE EER OF ALL ATTACKS, IN THE COLUMN OF "ALL".

Spoofing attacks	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	known	unknown	all
Spectro/CNN	**0.08**	**0.19**	**0.02**	**0.03**	**1.26**	1.48	**0.68**	**0.01**	**0.16**	26.83	**0.31**	5.83	3.07
Spectro/RNN	1.21	0.79	0.24	0.39	1.77	0.87	0.96	0.04	0.41	17.97	0.87	4.05	2.46
Spectro/CNN+RNN	0.16	0.50	0.03	**0.03**	1.38	**0.85**	0.91	0.03	0.59	**14.27**	0.40	**3.33**	**1.86**
fusion	0.09	0.29	0.00	0.00	0.99	0.64	0.71	0.00	0.29	11.67	0.27	2.66	1.47

designed for tasks such as image classification. System fusion further improves robustness for our proposed systems. In the experiments, a greedy fusion and a selective fusion are compared. The relatively poor performance of the 5-way (System No. a-e) system greedy fusion compared with the 3-way (System No. c-e) system selective fusion shows that it is not always a good idea to integrate all systems. Also, integrating multiple systems is very computational expensive. Careful selective fusion could lead to the best performance.

Since features with a DNN back-end did not show much performance improvement for development set, we only report our results with more advanced deep architectures such as CNNs, RNNS and CNNs+RNNs on evaluation set.

TABLE V
EER% OBTAINED FROM DIFFERENT SYSTEMS AND THEIR FUSED SYSTEMS ON DEVELOPMENT SET.

No.	System	S1	S2	S3	S4	S5	all
a	TEO/DNN	2.43	2.82	0.49	0.52	5.27	2.31
b	PMVDR/DNN	1.52	1.07	0.76	0.85	2.98	1.44
c	Spectro/CNN	**0.10**	**0.08**	**0.03**	**0.03**	1.60	**0.36**
d	Spectro/RNN	1.13	0.88	0.22	0.36	2.62	1.04
e	Spectro/CNN+RNN	0.11	0.27	0.27	0.26	**1.19**	0.42
a-e	fusion	0.48	0.58	0.22	0.13	1.39	0.49
c-e	fusion	**0.09**	**0.19**	**0.00**	**0.01**	**0.68**	**0.19**

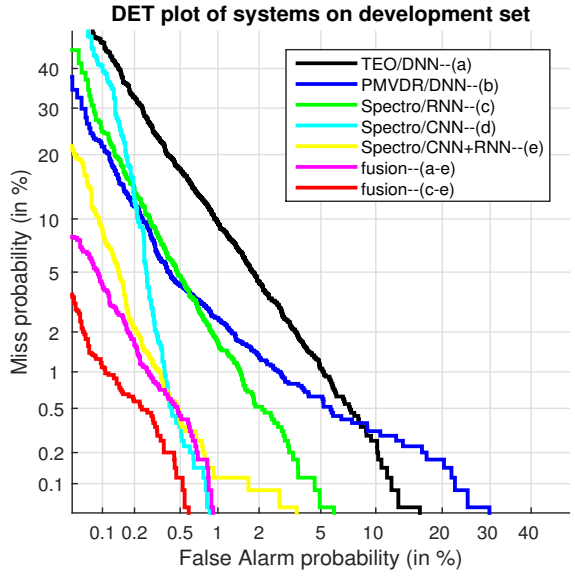


Fig. 8. DET plots for different system on developemnt set. Spectro/CNN stands for Spectrogram with CNN back-end.

E. Results on evaluation set

In this section, we report results on the evaluation set. TABLE IV details the EER for different spoofing attacks for our proposed systems. For a single system, it appears that Spectro/CNN achieves effective performance in all spoofing classes S1-S9, expect in S10. While on an alternative aspect, RNNs or CNNs+RNNs could be a better spoofing detector for case S10. For consideration in system development, although Spectro/CNN+RNN does not have the best performance in every spoofing attack, it achieves state-off-the-art single system performance for overall attacks. From our perspective, Spectro/CNN+RNN acts in a manner that provides a balanced trade-off between CNN and RNN systems. As shown in TABLE IV, fusion with three systems further improves the robustness of our proposed spoofing detector.

V. DISCUSSION

Thus far, we have seen a simple general feature (i.e., spectrogram) with deep learning frameworks can yield state-off-the-art performance for spoofing detection. Unfortunately, we still see a large gap between S10 and other spoofing attack cases. In this section, we discuss some observations we find in the experiments. Particularly, we focus on the variability introduced by duration mismatch, and propose some possible solutions as directions for future work.

A. Does duration matter?

The duration mismatch is one of the major issues in speaker verification [45], [46]. In a conventional i-vector system, duration mismatch can also be interpreted as context mismatch, because short duration always results in insufficient data for effective MAP adaption.

Motivated by the problem in speaker verification, we explore how duration influences spoofing detectors. First, we split each spoofing set into two subsets according to a determined duration size. As shown in TABLE III, S10 has the smallest mean duration–2.48s (This could be the reason that S10 does not perform well in most spoofing detection systems.). We use 2.5s as the criteria for a data set partition. Next, we compute the EER for long and short duration subsets, as illustrated in TABLE VI. For demonstration simplicity, we only show results from two systems (i.e., TEO/PMVDR i-vector system and Spectro/CNN system).

From TABLE VI, it is obvious that S10 has the largest number of short duration utterances, which might be a reason for relatively high EER in S10 spoofing detection. It is noted that spoofing S3, S4, and S8 have similar numbers of short

TABLE VI
EER(%) FOR LONG/SHORT DURATION SUBESTS ON EVALUATION DATA. 2.5S DURATION IS SELECTED AS THE PARTITION CRITERIA. UTTERANCE NUMBER FOR EACH SUBSET IS ALSO GIVEN.

Spoofing attacks	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
utt number (short)	2848	2848	9298	9309	2868	2740	2848	9359	2848	**10425**
utt number (long)	15552	15552	9102	9085	15532	15660	15552	9041	15552	7975
EER% (short/i-vector)	0.50	2.75	0.07	0.14	1.17	2.84	0.44	0.17	0.47	26.72
EER% (long/i-vector)	0.19	2.04	0.07	0.08	0.69	2.23	0.19	0.11	0.28	27.66
EER% (difference/i-vector)	0.31	0.71	0.00	0.06	0.48	0.61	0.25	0.06	0.19	**-0.86**
EER% (short/CNN)	0.09	0.31	0.02	0.02	1.48	2.28	1.15	0.01	0.19	24.86
EER% (long/CNN)	0.07	0.15	0.02	0.03	1.22	1.32	0.60	0.01	0.15	29.31
EER% (difference/CNN)	0.02	0.16	0.00	**-0.01**	0.26	0.96	0.55	0.00	0.04	**-4.45**

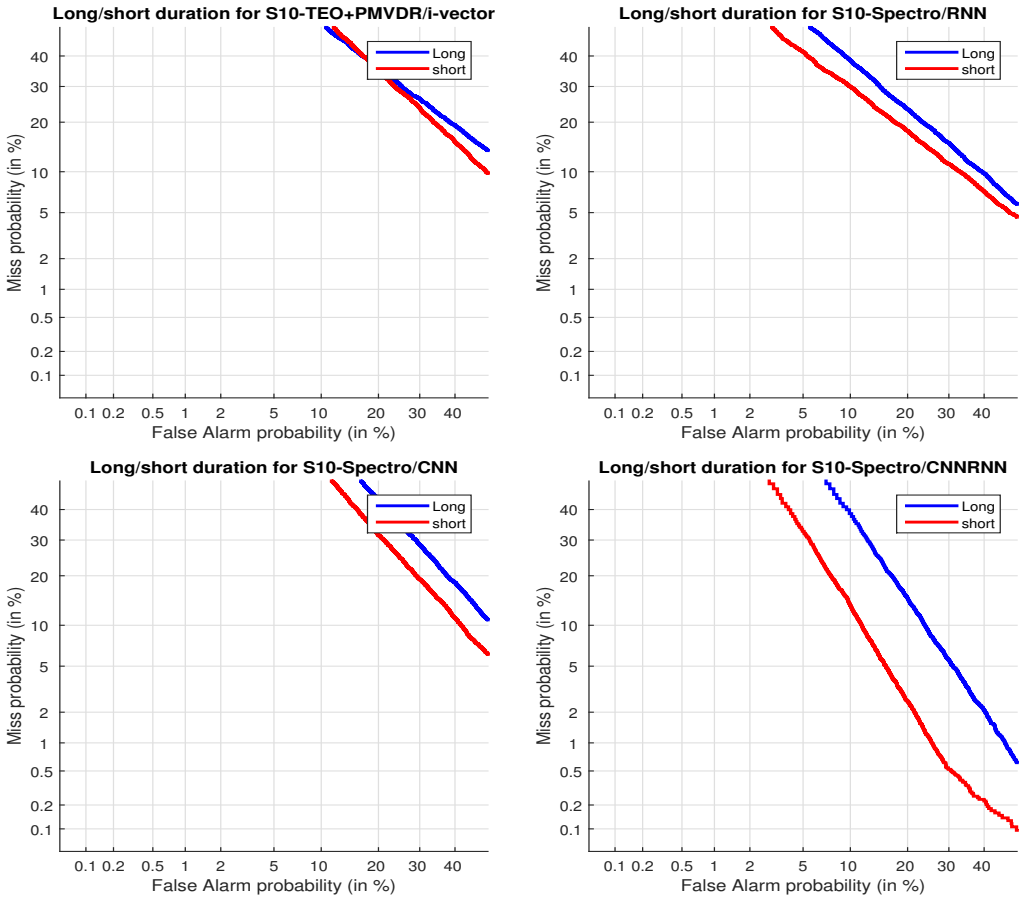


Fig. 9. DET plots for S10 short/long duration subsets with 4 different systems. All 4 different systems have the same pattern, which shows that EER for short duration subset is lower than that for long duration subset. This observation is different from other 9 spoofing attacks, and thus is very hard to explain.

utterances. So, short duration may be one of the reasons for poor spoofing detection performance, but it is not necessary the only reason in explaining the problem.

Compared with the i-vector system, our proposed spectro/CNN system seems to have potential to partially compensate for duration mismatch. As seen in TABLE 6, 7 out of 10 spoofing attacks show decreasing gaps between short and long duration subsets. While the only unusual behavior is found in S10. For S10, it is surprising to see that short duration set performs better than the long duration set. This observation is confirmed by various systems, see Fig. 9. At the same time, our padding/cropping approach for the spectrogram feature set is no longer effective to compensate for duration mismatch, Actually, the EER difference becomes larger in our Spectro/CNN spoofing detector. All these observations are not common in speaker verification research, and thus are very difficult to address. The phenomenon discussed in this section is intended to draw more attention for research of speaker verification anti-spoofing.

B. Extract "deep features" for utilizing the back-end advancements in speaker verification?

In our current systems, the spoofing detection decision is directly made by the probabilities from the final output layer.

Our proposed methods follow an end-to-end manner, where no additional concepts or heuristics are needed to train another back-end classifier.

At the same time, we feel that we have not fully explored the potential of our proposed frameworks. Extracting so-called deep features and then applying back-ends from speaker verification research community is a good direction. In fact, Linear Discriminative Analysis (LDA), Within-class Covariance Normalization (WCCN), or one-class SVM are reported to be effective for spoofing detection as well [18], [47]. Although, it is easy to extend our current models to such ideas, we restrict our contributions in this study to the investigations of different deep learning frameworks for spoofing detection. Therefore, the idea of combined deep features with other back-ends remains as a topic for future work to consider.

VI. Conclusion

In this study, we explored the application of deep learning as a tool for speaker verification anti-spoofing. To do so, we implemented several deep learning frameworks, including DNNs, CNNs and RNNs. Based on this, we proposed a novel architecture for spoofing detection, which integrates the advantages of CNNs for feature extraction and RNNs to model long-term dependencies. All these models were proved to be effective in our experiments, and our proposed combined CNN+RNN model achieved the state-off-the-art single system performance. System combination can further improve system robustness.

For research on feature extraction, three features (i.e., TEO-CB-Auto-Env, PMVDR and more general spectrogram feature) which are expected to generalize well for unseen spoofing attacks were employed. The effectiveness of TEO-CB-Auto-Env and PMVDR feature were analyzed. The spectrogram feature set, proposed as the input to the deep frameworks in this study, proved to be a good feature that does not depend on spoofing prior knowledge.

Extensive experiments were conducted with our propose system. The variability introduced by duration mismatch was also investigated. Using dataset partitioning, duration mismatch in spoofing detection was carefully analyzed. The results showed that our feature preparation method combined with deep back-ends can compensate for duration mismatch to some extent. At the same time, an interesting but challenging observation was found with spoofing attack S10. This S10 case provided relatively poor performance, and it also showed the opposite trends compared with the other nine spoofing attacks.

The study therefore highlights effective methods for spoofing detection, as well as fundamental observations for future work in deep features for spoofing detection advancements.

Acknowledgment

The authors would like to thank Dr. Zhizheng Wu and other ASVspoof 2015 Challenge organizers for providing the data and protocols. The authors would also like to thank Dr. Finnian Kelly, Shivesh Ranjan and other CRSS colleagues for their many insights and helpful discussions in the development of this work.

References

[1] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: a tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] 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.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[4] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2014, pp. 1695–1699.

[5] G. Liu and J. H. L. Hansen, "An investigation into back-end advancements for speaker recognition in multi-session and noisy enrollment scenarios," *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 22, no. 12, pp. 1978–1992, 2014.

[6] 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. Interspeech*, 2015.

[7] B. L. Pellom and J. H. L. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 2, 1999, pp. 837–840.

[8] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *IEEE Int. Carnahan Conf. on Security Technology (ICCST),*, 2011, pp. 1–8.

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

[10] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2014, pp. 1–5.

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

[12] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A. Laukkanen, "I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry." in *Proc. Interspeech*, 2013, pp. 930–934.

[13] Z. Wu, C. E. Siong, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition." in *Proc. Interspeech*, 2012, pp. 1700–1703.

[14] F. Alegre, R. Vipperla, and N. Evans, "Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals," in *Proc. Interspeech*, 2012.

[15] Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, Z. Ling, and S. King, "Sas: A speaker verification spoofing database containing diverse attacks," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2015, pp. 4440–4444.

[16] Z. Wu, T. Kinnunen, E. S. Chng, H. Li, and E. Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *Asia-Pacific Signal & Info. Proc. Asso. Annual Summit and Confe. (APSIPA ASC)*, 2012, pp. 1–5.

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

[18] J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with dnn and one-class svm for the asvspoof 2015 challenge," in *Proc. Interspeech*, 2015.

[19] 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. Interspeech*, 2015, pp. 2052–2056.

[20] D. Matrouf, J. Bonastre, and J. Costa, "Effect of impostor speech transformation on automatic speaker recognition," *Biometrics on the Internet*, p. 37, 2005.

[21] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *IEEE Sixth Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–8.

[22] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proc. IEEE Int. Conf. of the Biome. Special Interest Group (BIOSIG)*, 2012, pp. 1–7.
[23] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. on Patt. Analy. and Mach. intel.*, vol. 28, no. 12, pp. 2037–2041, 2006.
[24] C. Zhang, S. Ranjan, M. K. Nandwana, Q. Zhang, A. Misra, G. Liu, F. Kelly, and J. H. Hansen, "Joint information from nonlinear and linear features for spoofing detection: an i-vector/dnn based approach," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2016, pp. 5035–5039.
[25] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. on Speech and Audio Process.*, vol. 9, no. 3, pp. 201–216, 2001.
[26] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated mvdr-based acoustic front-end (pmvdr) for robust automatic speech recognition," *Speech Communication*, vol. 50, no. 2, pp. 142–152, 2008.
[27] 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. Interspeech*, 2015, pp. 2062–2066.
[28] 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. Interspeech*, 2015, pp. 2072–2076.
[29] E. Khoury, T. Kinnunen, A. Sizov, Z. Wu, and S. Marcel, "Introducing i-vectors for joint anti-spoofing and speaker verification," in *Proc. Interspeech*, 2014.
[30] 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. Interspeech*, 2015, pp. 2097–2101.
[31] G. Hinton, L. Deng, and et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
[32] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. on audio, speech, and lang. process.*, vol. 22, no. 10, pp. 1533–1545, 2014.
[33] G. Montavon, "Deep learning for spoken language identification," in *NIPS Workshop on deep learning for speech recognition and related applications*, 2009, pp. 1–4.
[34] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2013, pp. 6645–6649.
[35] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falcao, and A. Rocha, "Deep representations for iris, face, and fingerprint spoofing detection," *IEEE Trans. on Infor. Foren. and Securi.*, vol. 10, no. 4, pp. 864–879, 2015.
[36] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," *arXiv preprint arXiv:1408.5601*, 2014.
[37] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2016, pp. 5115–5119.
[38] C. Zhang, G. Liu, C. Yu, and J. H. L. Hansen, "I-vector based physical task stress detection with different fusion strategies," in *Proc. Interspeech*, 2015.
[39] T. D. Team, "Theano: A python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, 2016. [Online]. Available: http://arxiv.org/abs/1605.02688
[40] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *14th International Conference on Artificial Intelligence and Statistics*, vol. 15, 2011, pp. 315–323.
[41] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[42] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.
[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[44] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "Asvspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training*, vol. 10, no. 15, p. 3750, 2014.
[45] T. Hasan, R. Saeidi, J. H. Hansen, and D. A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2013, pp. 7663–7667.
[46] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2013, pp. 7649–7653.
[47] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Classifiers for synthetic speech detection: A comparison," in *Proc. Interspeech*, 2015.

Chunlei Zhang received the B.S. degree in Environmental Engineering, M.S. degree in Acoustics from Northwestern Polytechnical University (NPU), Xian, China, in 2011 and 2014, respectively. Currently he is pursuing his Ph.D. degree as a Research Assistant in the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, U.S.A. since August 2014. His research interests focus on robust speaker recognition in train/test mismatched conditions, stress/emotion detection and machine learning.

PLACE PHOTO HERE

Chengzhu Yu received the B.S. degree in electrical engineering from China University of Petroleum, Beijing, China, in 2008. He is currently a Research Assistant at The University of Texas at Dallas where he is pursuing his Ph. D degree in electrical engineering. His research is on speaker recognition, automatic speech recognition, speaker diarization.

John H. L. Hansen (IEEE S'81-M'82-SM'93-F'07) received the Ph.D. and M.S. degrees in Electrical Engineering from Georgia Institute of Technology, Atlanta, Georgia, in 1988 and 1983, and B.S.E.E. degree from Rutgers University, College of Engineering, New Brunswick, N.J. in 1982.

He joined University of Texas at Dallas (UTDallas), Erik Jonsson School of Engineering and Computer Science in 2005, where he presently serves as Jonsson School Associate Dean for Research, as well as Professor of Electrical Engineering and also holds the Distinguished University Chair in Telecommunications Engineering. He previously served as Department Head of Electrical Engineering from Aug. 2005 Dec. 2012, overseeing a +4x increase in research expenditures ($4.5M to $22.3M) with a 20% increase in enrollment and the addition of 18 T/TT faculty, growing UTDallas to be the 8th largest EE program from ASEE rankings in terms of degrees awarded. He also holds a joint appointment as Professor in the School of Behavioral and Brain Sciences (Speech & Hearing). At UTDallas, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Dept. Chairman and Professor of Dept. of Speech, Language and Hearing Sciences (SLHS), and Professor of the Dept. of Electrical & Computer Engineering, at Univ. of Colorado Boulder (1998-2005), where he co-founded and served as Associate Director of the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTDallas. He has been named IEEE Fellow (2007) for contributions in "Robust Speech Recognition in Stress and Noise," Inter. Speech Communication Association (ISCA) Fellow (2010) for contributions on research for speech processing of signals under adverse conditions, and received The Acoustical Society of Americas 25 Year Award (2010) in recognition of his service, contributions, and membership to the Acoustical Society of America. He is currently serving as an elected Vice-President of ISCA and member of the ISCA Board. He was also selected and is serving on the U.S. Office of Scientific Advisory Committees (OSAC) for OSAC-Speaker in the voice forensics domain (2015-2017). Previously he served as IEEE Technical Committee (TC) Chair and Member of the IEEE Signal Processing Society: Speech-Language Processing Technical Committee (SLTC) (2005-08; 2010-14; elected IEEE SLTC Chairman for 2011-2013, Past-Chair for 2014), and elected ISCA Distinguished Lecturer (2011/12). He has also served as member of the IEEE Signal Processing Society Educational Technical Committee (2005-08; 2008-10). Previously, he served as the Technical Advisor to the U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005/06), Associate Editor for IEEE Trans. Speech & Audio Processing (1992-99), Associate Editor for IEEE Signal Processing Letters (1998-2000), Editorial Board Member for the IEEE Signal Processing Magazine (2001-03). He has also served as guest editor of the Oct. 1994 special issue on Robust Speech Recognition for IEEE Trans. Speech & Audio Proc. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000-03), and is serving as a member of the ISCA (Inter. Speech Communications Association) Advisory Council. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human-computer interaction. He has supervised 73 PhD/MS thesis candidates (36 PhD, 37 MS/MA), was recipient of The 2005 University of Colorado Teacher Recognition Award as voted on by the student body, author/co-author of 593 journal and conference papers including 11 textbooks in the field of speech processing and language technology, coauthor of the textbook Discrete-Time Processing of Speech Signals, (IEEE Press, 2000), co-editor of DSP for In-Vehicle and Mobile Systems (Springer, 2004), Advances for In-Vehicle and Mobile Systems: Challenges for International Standards (Springer, 2006), In-Vehicle Corpus and Signal Processing for Driver Behavior (Springer, 2008), and lead author of the report The Impact of Speech Under Stress on Military Speech Technology, (NATO RTO-TR-10, 2000). He also organized and served as General Chair for ISCA Interspeech-2002, Sept. 16-20, 2002, and Co-Organizer and Technical Program Chair for IEEE ICASSP-2010, Dallas, TX. He also served as Co-Chair and Organizer for IEEE SLT-2014, Dec. 7-10, 2014 in Lake Tahoe, NV.