JOINT INFORMATION FROM NONLINEAR AND LINEAR FEATURES FOR SPOOFING DETECTION: AN I-VECTOR/DNN BASED APPROACH

Chunlei Zhang, Shivesh Ranjan, Mahesh Kumar Nandwana, Qian Zhang, Abhinav Misra, Gang Liu, Finnian Kelly, John H. L. Hansen

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering, University of Texas at Dallas, Richardson, Texas, U.S.A.
{chunlei.zhang, john.hansen}@utdallas.edu

ABSTRACT

Sustaining automatic speaker verification (ASV) systems from spoofing attacks remains an essential challenge, even if significant progress in ASV has been achieved in recent years. In this study, an automatic spoofing detection approach using an i-vector framework is proposed. Two approaches are used for frame-level feature extraction: cepstral-based Perceptual Minimum Variance Distortionless Response (PMVDR), and non-linear speech-production-motivated Teager Energy Operator (TEO) Critical Band (CB) Autocorrelation Envelope (Auto-Env). An utterance-level i-vector for each recording is formed by concatenating PMVDR and TEO-CB-Auto-Env i-vectors, followed by linear discriminative analysis (LDA) for maximizing the ratio of between-class to within-class scatterings. A Gaussian classifier and DNN are also investigated for back-end scoring. Experiments using the ASVspoof 2015 corpus show that our proposed method successfully detects spoofing attacks. By combining the TEO-CB-Auto-Env and PMVDR features, a relative 76.7% improvement in terms of EER is obtained compared with the best single-feature system.

Index Terms— Spoofing detection, i-vector, TEO-CB-Auto-Env, PMVDR, DNN

1. INTRODUCTION

A spoofing attack on an automatic speaker verification (ASV) system is a situation where an impostor attempts to masquerade as an enrolled person by falsifying speech data traits [1]. ASV systems were initially designed to distinguish between enrolled speakers and zero-effort impostors. Although advancements achieved in channel variability modeling and noise compensation have greatly improved the reliability of ASV systems, studies have shown that ASV systems remain vulnerable to intentional spoofing attacks [2–5]. Spoofing attacks are emerging as a problem due to the maturing process of speech technologies such as speech synthesis (SS), voice conversion (VC), which lower the cost of non-expert spoofing attacks and increase vulnerabilities of ASV systems.

In this study, we focus on developing new approaches to spoofing detection, i.e., given input speech, we identify it either as genuine or spoofed speech. Spoofing detection can be incorporated into an ASV system to reduce false acceptance rates. This study effort was based around ASVspoof 2015 Challenge, the First Automatic Speaker Verification Spoofing and Countermeasures Challenge [5].

One challenge in building a robust spoofing detection system is choosing suitable features. An obvious first step is to adopt the same features used in ASV systems, for example MFCCs, in general, MFCCs perform well in discriminating genuine and spoofed speech [4, 6]. However, performance degrades significantly for attacks that only some coefficients at the feature level are modified. For example, in the ASVspoof 2015 corpus, the spoofed speech category ‘S2’ is generated only by modifying the first coefficient of Mel-Cepstral coefficients. MFCCs derived from converted speech are very similar to the genuine speech (1/13 in difference if MFCCs are 13 dimensional), the EER for spoofing detection is almost 50% in our experiments, which suggests that detection is at a random decision level. Other studies indicated that modified group delay (MGD) or phase features could be used to detect spoofed speech [4, 7]. An explanation for this is that for natural speech, phase information is lost during the analysis-synthesis step in some speech-synthesis techniques, which makes genuine speech different from that which has been synthesized.

In this paper, we employ Perceptual Minimum Variance Distortionless Response (PMVDR) and TEO-CB-Auto-Env as our features [8, 9]. The intuition behind our work is: (a) according to [7, 10], spoofing detection by human listeners outperforms automatic spoofing detection because of the better perceptual ability of humans. PMVDR can accurately model the upper spectral envelope at perceptually important harmonics. By incorporating this perceptual consideration, PMVDR is expected to be suitable for spoofing detection; (b) TEO-CB-Auto-Env models the nonlinear variabilities of speech production introduced by stress/emotion, which makes this feature suitable for irregularity detection [11]. Here, we can treat spoofing attacks as variabilities introduced to genuine speech, thus we employ TEO-CB-Auto-Env in this task.

For system development, spoofing detection is still a relatively new field of research, and spoofing types are not guaranteed exhaustively or known; no single system has been established as the best to adopt. Given this, we employ an i-vector framework along with a system based on a Deep Neural Network (DNN) in this paper [3, 12]. The i-vector PLDA system from speaker identification domain does not perform well for spoofing attacks under mismatched conditions [13, 14], which is situation in the challenge. Instead, a Gaussian classifier, and a DNN are employed as back-end classifiers [15–18]. Part of results in this paper(i-vector-Gaussian Classifier system) could also be found in [5]. Here, i-vector/DNN results are added to make our research a more complete work on spoofed speech detection for ASV.

Section 2 describes the experimental corpus. Section 3 is the system description which includes details of feature extraction, the
i-vector framework and back-end classifier development. We report results and present further discussions in Section 4. Section 5 concludes our work.

2. CORPUS
The ASVspoof 2015 database contains 3 datasets: training, development and evaluation [5]. The training set has genuine and spoofed speech from 25 speakers (15 female, 10 male), and the spoofed speech utterances were generated from genuine speech using 3 VC and 2 SS based algorithms. The development set has genuine and spoofed speech from 35 speakers (20 female, 15 female), with the same 5 spoofing algorithms used as in the training set. The evaluation set has 193404 test utterances of genuine and spoofed speech from 46 speakers (26 female, 20 male). The evaluation data also contains 5 kinds of spoofed speech created using unknown spoofing algorithms (not seen in the training/development set) to gauge system performance with unknown spoofed utterances.

3. I-VECTOR SYSTEM FOR SPOOFING DETECTION
The i-vector system for this task exploits the concept of total variability modeling and i-vector extraction, which is extensively adopted in speaker identification tasks. By constraining the total variability into a lower dimensional total variability space, the i-vector is capable of effectively representing the variability factors within each speech utterance. In this work, we attempt to model spoof-specific variability across different speakers using i-vectors.

3.1. Feature extraction
3.1.1. TEO-CB-Auto-Env [9]
The TEO profile obtained from the critical band based Gabor bandpass filter output is segmented on a short-term basis. Next, autocorrelation is applied after framing. Once the auto-correlation response is found, the area under the autocorrelation envelope is obtained and normalized. One area coefficient is obtained for each filter bank. 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). We use 18 Gabor filter banks, meaning that 18 dimensional features is extracted from each frame. Fig. 1 shows a flow diagram of TEO-based feature extraction.

3.1.2. PMVDR [8]
PMVDR features were first proposed by Yapanel and Hansen. PMVDR computes cepstral coefficients by incorporating perceptual warping of FFT power spectrum, 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. Previous studies have shown that perceptual knowledge can differentiate between genuine and spoofed speech [4, 7, 10]. 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. 2. 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 20 ms window with 10 ms shift.

3.2. Utterance level i-vector framework
In our utterance level spoofing detection system, each utterance in i-vector modeling is represented by a GMM supervector:

$$ M_u = m + T x_u, \quad (1) $$

Where $M_u$ is the GMM supervector obtained from utterance $u$, $m$ is the speaker, channel, spoofing-independent supervector constructed from UBM. The total variability matrix $T$ is a low-rank projection matrix obtained from all training data by factor analysis training [19]. The i-vector is given by a normally distributed vector $x_u$ containing the total factors. The complete i-vector system is shown in Fig. 3.

3.3. i-vector level fusion
Two i-vectors can be derived from each utterance (i.e. PMVDR-based i-vector and TEO-CB-Auto-Env based i-vector). By concatenating these two i-vectors together, we expect to use genuine-spoofing discriminative information provided by both features simultaneously. After whitening and length normalization, the dimensionality is reduced to the original length by linear discriminative analysis (LDA).

Fig. 1. TEO-CB-Auto-Env feature extraction.

Fig. 2. Flow diagram of PMVDR feature extraction.
4. RESULTS AND DISCUSSION

For our i-vector system, the UBM utilized in all experiments was trained on all training data provided in the challenge database (both genuine and spoofed dataset). The number of components here is set to 512. The rank of the total variability matrix $T$ defines the i-vector dimensionality. We set this to 100 in our experiments empirically.

For comparison, we use the same ‘threshold-free’ equal error rate (EER) metric as in ASVspoof 2015, which is implemented using the Bosaris toolkit [21]. Performance of all systems outlined in Section 3 are evaluated on development and evaluation data.

4.1. Results on development data

As mentioned in the Introduction, the spoofing detection performance of specific feature is heavily related with VC and SS algorithms. Results from GC of different spoofing attacks are summarized in Table 1. The fusion system always gives better performance for all 5 different spoofing attacks. TEO-CB-Auto-Env and PMVDR are doing well with different spoofing categories, which also inspires us to combine them together to boost performance. Spoofing category S2 is the worst among all attacks. This is not surprising because we use amplitude based features, while S2 is obtained only by changing a very small amplitude part of the Mel-cepstral coefficient.

<table>
<thead>
<tr>
<th>Spoof type</th>
<th>Feature</th>
<th>EER(%)</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>MFCC</td>
<td>6.59</td>
<td>93.72</td>
</tr>
<tr>
<td></td>
<td>TEO</td>
<td>4.80</td>
<td>95.58</td>
</tr>
<tr>
<td></td>
<td>PMVDR</td>
<td>1.94</td>
<td>98.30</td>
</tr>
<tr>
<td></td>
<td>TEO+PMVDR</td>
<td>0.79</td>
<td><strong>99.27</strong></td>
</tr>
<tr>
<td>S2</td>
<td>MFCC</td>
<td>47.77</td>
<td>42.87</td>
</tr>
<tr>
<td></td>
<td>TEO</td>
<td>15.71</td>
<td>84.32</td>
</tr>
<tr>
<td></td>
<td>PMVDR</td>
<td>24.63</td>
<td>75.56</td>
</tr>
<tr>
<td></td>
<td>TEO+PMVDR</td>
<td>12.73</td>
<td><strong>86.30</strong></td>
</tr>
<tr>
<td>S3</td>
<td>MFCC</td>
<td>1.32</td>
<td>99.06</td>
</tr>
<tr>
<td></td>
<td>TEO</td>
<td>4.48</td>
<td>97.34</td>
</tr>
<tr>
<td></td>
<td>PMVDR</td>
<td>0.82</td>
<td>99.42</td>
</tr>
<tr>
<td></td>
<td>TEO+PMVDR</td>
<td>0.27</td>
<td><strong>99.78</strong></td>
</tr>
<tr>
<td>S4</td>
<td>MFCC</td>
<td>1.97</td>
<td>98.37</td>
</tr>
<tr>
<td></td>
<td>TEO</td>
<td>4.43</td>
<td>97.16</td>
</tr>
<tr>
<td></td>
<td>PMVDR</td>
<td>0.69</td>
<td>99.48</td>
</tr>
<tr>
<td></td>
<td>TEO+PMVDR</td>
<td>0.21</td>
<td><strong>99.86</strong></td>
</tr>
<tr>
<td>S5</td>
<td>MFCC</td>
<td>20.78</td>
<td>79.47</td>
</tr>
<tr>
<td></td>
<td>TEO</td>
<td>6.79</td>
<td>93.49</td>
</tr>
<tr>
<td></td>
<td>PMVDR</td>
<td>9.79</td>
<td>90.43</td>
</tr>
<tr>
<td></td>
<td>TEO+PMVDR</td>
<td>4.85</td>
<td><strong>94.22</strong></td>
</tr>
</tbody>
</table>

Table 1. EER(%) and Accuracy(%) across all 5 spoofing attacks using i-vector systems.

Spoofing detection performance on development data using various systems are presented in Table 2. Here, we regard five different spoofing attacks as spoofed speech, and give the overall results. The benefits of fusing two spoofing sensitive features is apparent, resulting in an absolute 5.50% EER performance boost and a 5.24% accuracy improvement for the i-vector/GC system. Also i-vector/DNN achieves the best overall performance.
### 4.2. Results on evaluation data

On the evaluation data, we have 5 additional types of spoofed speech acting as unknown attacks. The EER of known attacks for our primary i-vector/GC system is 0.67%, while the EER for unknown attacks is 6.04%. Overall performance for all attacks is 3.35%.

Table 3 shows the results obtained ASVspoof 2015 as primary (train only using training data) and flexible primary (train using training and development data) submissions. The relatively weak performance for flexible primary shows that it is not necessarily better to use more training data for spoofing detection. In a real word application, a spoofing attack is more likely to be an open set problem, as we always meet unknown attacks. We can’t include all spoofed speech in the training data. It should also be noted that although i-vector/DNN system performs better on the development data, there is no obvious advantage in using DNN classifier compared with simple Gaussian Classifier, as seen in Fig.4.

#### Table 3. EER(%) and Accuracy(%) for ASVspoof 2015 on evaluation data.

<table>
<thead>
<tr>
<th>Feature</th>
<th>System</th>
<th>EER(%)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEO</td>
<td>i-vector/GC</td>
<td>8.80</td>
<td>93.61</td>
</tr>
<tr>
<td>PMVDR</td>
<td>i-vector/GC</td>
<td>7.17</td>
<td>92.88</td>
</tr>
<tr>
<td>TEO+PMVDR</td>
<td>i-vector/GC</td>
<td>1.67</td>
<td>98.85</td>
</tr>
<tr>
<td>f-bank</td>
<td>DNN</td>
<td>6.14</td>
<td>96.00</td>
</tr>
<tr>
<td>TEO</td>
<td>i-vector/DNN</td>
<td>4.83</td>
<td>96.27</td>
</tr>
<tr>
<td>PMVDR</td>
<td>i-vector/DNN</td>
<td>4.58</td>
<td>95.53</td>
</tr>
<tr>
<td>f-bank+TEO+PMVDR</td>
<td>fusion</td>
<td>0.71</td>
<td>99.33</td>
</tr>
</tbody>
</table>

#### Table 4. System performance of imbalanced training based on i-vector PLDA system (%).

<table>
<thead>
<tr>
<th>System</th>
<th>EER</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-bank+TEO+PMVDR</td>
<td>0.71</td>
<td>99.33</td>
</tr>
</tbody>
</table>

#### Fig. 4. DET plot for i-vector/DNN on evaluation data.

Comparison using the PLDA system is shown in Table 4, using the development set.

From Table 4, the PLDA system is greatly improved after K-means clustering. The confusion matrix shows that more genuine speech has been identified as spoofed speech (103 VS 9). The ratio is relatively small compared with the performance gain obtained on spoofing data (16388 VS 5810). For spoofing detection system, the primary goal is to reduce False Acceptance (FA) rates. Although not as good as the i-vector GC system we proposed, it gives some motivation for applying a simple clustering solution as a preprocessing step to i-vectors.

### 5. CONCLUSIONS

This study described a systems for spoofing detection. Two spoofing sensitive features (TEO-CB-Auto-Env and PMVDR) were explored. The results showed our i-vector based system gives competitive overall performance compared with [7, 25, 26]. A relative +76.7% improvement in terms of EER was obtained by fusion. The DNN setup performs well on known attacks, but not well on unknown attacks. The issue of imbalanced training data, a typical feature of spoofing datasets, was demonstrated. A probe solution using resampling showed promise. The low performance for ASVspoof 2015 ‘S10’ condition inspires us to focus more on this spoofing type. However, this does not guarantee good performance for other unseen attacks. The results here show both meaningful advancements, also a point to direction for future research. Thus, our future work will aim to identify spoofing detection features that generalize well.
6. REFERENCES


5039