Simultaneous Utilization of Spectral Magnitude and Phase Information to Extract Supervectors for Speaker Verification Anti-spoofing

Yi Liu1, Yao Tian1, Liang He1, Jia Liu1, Michael T. Johnson2

1Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
2Speech and Signal Processing Lab, Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI 53233, USA

{liuyil2, tianyaol1}@mails.tsinghua.edu.cn, {heliang, liuj}@tsinghua.edu.cn, michael.johnson@marquette.edu

Abstract

Protection from spoofing attacks is an essential component of speaker verification systems. This paper proposes a novel approach to detect such attacks by utilizing supervectors derived from spectral magnitude and phase information. Three countermeasures are chosen to represent these important information. To combine different countermeasures, score fusion and an anti-spoofing supervector (ASSV) are used. Experiments conducted on ASVspoof 2015 show that the combination of magnitude and phase information obtains relative 90% improvement in terms of the equal error rate (EER) compared to the best subsystem in the development set. The two systems can also be fused to further improve the performance. In addition to accuracy improvements, the new supervector framework is extensible and allows for a more flexible interface to the back-end classifier design.

Index Terms: speaker verification, anti-spoofing supervector, spectral magnitude and phase information, ASVspoof 2015

1. Introduction

Text-independent automatic speaker verification (ASV) plays an important role in biometric authentication. By virtue of new modeling methods, such as Joint Factor Analysis (JFA) and i-vector representations, ASV systems have become less susceptible to noise or channel effects. However, even state-of-the-art ASV systems are still quite vulnerable to deliberate spoofing attacks. Classical spoofing methods, such as impersonation, replay, speech synthesis, voice conversion (including artificial signal generation), can significantly increase the false acceptance rate of ASV systems [1].

To address this situation, recently several countermeasures have been proposed to make ASV systems more robust from spoofing [2]. Most current anti-spoofing algorithms are built on assumptions related to specific spoofing approaches. For instance, impersonators are considered to exhibit larger acoustic parameter variation than original speakers [3] and different presentation of channel noise and reverberation can be a mechanism to recognize recordings [4]. Speech synthesis and voice conversion are two of the most easily accessible and effective spoofing approaches, which have received the most attention [5].

Dozens of countermeasures have been proposed, including detecting phase information [6], modeling the pattern of long-term features like F0 statistics [7], and observing the texture of the spectrogram [8, 9]. These anti-spoofing algorithms have been shown to work well under individual experimental configurations. However, testing has been done across a variety of databases, which makes performance comparison very difficult, if not meaningless. Moreover, most proposed countermeasures just focus on one specific type of attack. This is a significant limitation, because in practical applications a wide variety of attacks would always be expected. One way to address this is through score fusion, which is able to combine the discriminative capabilities of different features. For example, short-term and long-term features have been fused together to improve performance [10].

In this paper, three countermeasures based on spectral magnitude and phase information are extracted as supervectors and can be classified by support vector machine (SVM). In addition to score fusion, a new anti-spoofing supervector (ASSV) approach is presented to combine these countermeasures to detect diverse spoofing, including voice conversion and speech synthesis. We evaluate our method using the ASVspoof 2015 challenge [11] so that the results are comparable across participants.

The remainder of this paper is organized as follows. The three types of countermeasures we use are briefly introduced in Sections 2. Section 3 presents the details of supervector extraction and our ASSV structure. Experimental work is described in Section 4. Finally, Section 5 concludes the paper.

2. Spoofing countermeasures based on magnitude and phase information

Spoofing is a mechanism to trick an ASV system by imitating a target speaker. Although the imitation is not acoustically exact, spoofing can target the features and models used for speaker identification. Since such synthesized, converted or artificially generated utterances unavoidably change important parameters, making the speech acoustically “unnatural”, it should be possible to explicitly discriminate such spoofed speech. Both magnitude and phase information in the frequency domain can be effectively utilized together to identify such differences. In this work we use three features representative of these important information. They are local binary patterns, modified group delay feature and cosine normalized phase feature.

2.1. Local binary patterns

The spectro-temporal structure is an important property of an utterance. We hypothesize that many spoofing methods would alter the spectro-temporal structure, often referred to as "tex-
frequency” [8]. To detect this disturbance, a well-established approach called Local Binary Patterns (LBP) is used.

LBP was first proposed in texture recognition to represent local properties of a gray scale image. When dealing with a pixel \( g_r \) in an image, the LBP operator characterizes \( P \) equally spaced pixels \( (g_0, \ldots, g_{P-1}) \) on a circle of radius \( R \) around \( g_r \). The central pixel compares its gray-scale value with other neighbors and the sign of each comparison comprises the corresponding LBP. The LBP operator can be expressed as:

\[
LBP_{P,R} = \sum_{p=0}^{P-1} \text{sign}(g_p - g_r) \cdot 2^p
\]  

(1)

where \( \text{sign}(x) \) denotes the sign function which equals 1 when \( x \geq 0 \) and 0 otherwise. It is straightforward to see there are \( 2^P \) LBP indices with each pixel mapped to one unique index. Indices that can be equivalently represented by a circular bit-wise right shift are grouped together to remove the effect of rotation. Examples for \((P = 8, R = 1.0)\) and \((P = 16, R = 2.0)\) are shown as Figure 1.

Modified group delay features (MGDF) have been used in phoneme recognition for a long time [14]. Derived from the modified group delay function, MGDF provides meaningful phase information. Results have shown that MGDF is well suited to defend voice conversion attacks in contrast to conventional cepstral features [10].

Given a speech signal \( x(n) \), the modified group delay function is expressed as:

\[
\tau_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^2}
\]  

(2)

\[
\tau_{\alpha,\gamma}(\omega) = \frac{\tau_p(\omega)}{|\tau_p(\omega)|} \cdot \left| \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^2} \right|^\alpha
\]  

(3)

where \( X(\omega) \) and \( Y(\omega) \) are the short-time Fourier transform (STFT) of \( x(n) \) and \( ax(n) \) respectively, and the subscripts \( R \) and \( I \) denote the real and imaginary parts of a complex number. \( |S(\omega)|^2 \) is the smoothed version of \( |X(\omega)|^2 \), which can be obtained by simple cepstral smoothing or median filtering. Two parameters \( \alpha \) and \( \gamma \) are introduced to further suppress the spiky nature of group delay spectrum. The difference between the power spectrum and the modified group delay spectrum is shown in Figure 2.

The discrete cosine transform (DCT) is applied to decorrelate the coefficients and obtain the final MGDF. The zero-th coefficient \( c_0 \) is ignored per [14].

2.3. Cosine normalized phase feature

Unlike MGDF, cosine normalized phase features (CNPF) use a more direct approach to extract phase information [15]. After the STFT is applied to a speech frame, the short-time phase spectrum \( \psi(\omega) \) is first unwrapped to eliminate discontinuity. Cosine normalization is necessary to constrain the dynamic range to \([-1, 1]\). Then, a CNPF is extracted using a DCT applied to the cosine normalized phase spectrum.

3. Anti-spoofing supervector extraction

Motivated by the success of supervectors in speaker recognition [16], we use a supervector-based structure in our anti-spoofing system. Each utterance is converted to a single supervector, allowing flexibility for different classifiers.
To make the LBP features described in Section 2 more appropriate for spectrogram spoofing detection, we note some differences between traditional LBPs and those applied to spectrograms:

- Unlike images, spectral texture does not need rotation invariance, since rotation does not occur in speech-based spectrograms.
- The rows of feature texture are physically significant. If we use FBank features, each row denotes the spectral character of a certain frequency range. Computing the histogram over the whole image would lose this useful information.

Taking these two points into account, we use unique LBPs without rotation invariance. In addition, the histogram over each coefficient is derived separately as shown in Figure 3, rather than being combined. Each histogram is normalized to limit the dynamic range and concatenated together to form the normalized unique LBP (NULBP) supervector $SV_{NULBP}$.

In contrast to the NULBP feature, the MGDF and CNPF are extracted following a frame-by-frame fashion as conventional acoustic features. To extract supervectors for this kind of feature, a Gaussian mixture model (GMM) - universal background model (UBM) is first trained across all training speech. For each utterance, a GMM is adapted from the UBM using maximum a posteriori (MAP) adaptation:

$$g(x) = \sum_{i=1}^{N} \lambda_i N(x; \mu_i, \Sigma_i)$$

where $N$ denotes the mixture number, $\lambda_i$ are the weights of the mixtures, $N()$ indicates a Gaussian distribution, and $\mu_i$ and $\Sigma_i$ are the mean and covariance of the $i$-th Gaussian. The means of the GMM are scaled [16]:

$$\hat{m}_i = \sqrt{\lambda_i} \Sigma_i^{-1/2} m_i$$

and stacked to form the MGDF/CNPF supervector.

$$SV_{MGDF/CNPF} = [m_1^T, m_2^T, \ldots, m_N^T]^T$$

Although these three countermeasures are designed to detect a variety of spoofing attacks, they are dissimilar and there is not a simple way to merge them directly. If the parameters are carefully selected, score fusion would improve the performance; however, this approach sometimes requires more classifiers than available. In this paper, we also concatenate these supervectors to yield what we term an anti-spoofing supervector (ASSV). The flowchart of ASSV extraction is demonstrated in Figure 3. Our ASSV is described as

$$ASSV = [SV_{NULBP}^T, SV_{MGDF}^T, SV_{CNPF}^T]^T$$

This approach is extensible, so that an even larger ASSV can be created, if additional anti-spoofing features are available.

4. Experimental work

4.1. ASVspoof 2015 challenge

The ASVspoof 2015 challenge was designed to evaluate the performance of state-of-the-art spoofing countermeasures. The database contains both genuine and spoofed speech. In the training and development set, three voice conversion and two speech synthesis algorithms are presented. To avoid the use of prior knowledge, five more algorithms were added to the evaluation set which was unseen during the development period. Generalized countermeasures would be preferred. The details of ASVspoof 2015 are described in [11].

4.2. Experiment setup

The NULBP is derived from 120-dimension FBank features: 40 static coefficients plus their delta and delta-delta coefficients. $P$ and $R$ are chosen to be 8 and 1.0, respectively. The group with bit-wise transitions larger than 2 is discarded, leaving 58 indices. Histograms for all coefficients except the first and last are computed, giving a total NULBP dimension of $58 \times (120 - 2) = 6844$.

Both MGDF and CNPF use 12-dimension static features. We fix $\gamma = 1.2$ and $\alpha = 0.4$ in MGDF as described in [15]. Cepstral normalization is used on MGDF but not on CNPF. Each GMM consists of 512 mixture components. The UBM s are trained from all utterances in the training set. GMMs for each utterance are adapted using only the UBM means [16]. The means are stacked to form $512 \times 12 = 6144$ length supervectors.

Each type of supervector is classified by a two-class SVM with linear kernel, and is considered as a separate sub-system. The score fusion system is optimized in the development set using grid search. Our ASSV concatenates the above three supervectors thus the dimension would be $6844 + 6144 + 2 = 19132$. A single SVM is able to discriminate the class of each ASSV.

In comparison with our supervector method, the original GMM log-likelihood scoring [10, 15] is also implemented. In this approach, two GMMs are trained separately on genuine and spoofed speech. During the test phase, each utterance scores against these two GMMs to compute the likelihood ratio.
In our experiments, we strictly follow the common training condition of the ASVspoof 2015 challenge. Performance is reported for the development and evaluation sets. The results for the evaluation set were returned by the organizer of ASVspoof 2015. The equal error rate (EER) is used as the official primary metric for ASVspoof 2015.

4.3. Results

We first compare the performance of sub-systems using separate countermeasures individually. Both GMM log-likelihood scoring and supervector systems are applied to MGDF and CNPF. The results are shown in Table 1.

Table 1: The performance of sub-systems.

<table>
<thead>
<tr>
<th>Sub-system</th>
<th>EER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULBP+SVM</td>
<td>0.858</td>
</tr>
<tr>
<td>MGDF+SVM</td>
<td>2.602</td>
</tr>
<tr>
<td>MGDF+GMM</td>
<td>3.713</td>
</tr>
<tr>
<td>CNPF+SVM</td>
<td>4.403</td>
</tr>
<tr>
<td>CNPF+GMM</td>
<td>4.487</td>
</tr>
</tbody>
</table>

Among the three countermeasures, NULBP achieves better performance than MGDF and CNPF. This suggests that the spectrogram texture effectively indicates spoofed speech. Meanwhile phase information, i.e. MGDF and CNPF, is also useful to filter genuine speech. On the other hand, as is also typically the case for speaker verification, a supervector with linear kernel SVM consistently outperforms GMM log-likelihood scoring. We will only use the supervector structure in the remaining experimental comparisons.

In order to combine the ability of the three countermeasures, both score fusion and the proposed ASSV approach are evaluated. We also combine the two fusion systems to further boost performance. The results in the development and evaluation sets are shown in Table 2.

Table 2: The performance of combined systems.

<table>
<thead>
<tr>
<th>Combined System</th>
<th>EER(%)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dev</td>
<td>eval known</td>
<td>eval unknown</td>
<td>eval avg</td>
</tr>
<tr>
<td>Score fusion (Grid Search)</td>
<td>0.058</td>
<td>0.104</td>
<td>6.775</td>
<td>3.439</td>
</tr>
<tr>
<td>ASSV</td>
<td>0.117</td>
<td>0.159</td>
<td>6.227</td>
<td>3.193</td>
</tr>
<tr>
<td>Score fusion+ASSV</td>
<td>0.025</td>
<td>0.059</td>
<td>6.114</td>
<td>3.086</td>
</tr>
</tbody>
</table>

From Table 2, we find that although the MGDF and CNPF systems do not perform as well as NULBP, their inclusions in the score fusion and ASSV systems reduce the EER from the best subsystem’s 0.858% to 0.058% and 0.117%, leading to 93% and 86% relative improvements in the development set. This phenomenon shows the importance of the complementary information contained in spectral magnitude and phase.

For known attacks in the development and evaluation sets, score fusion achieves better results than the proposed ASSV approach. Our hypothesis for this is that ASSV simply concatenates all supervectors together without any weighting technologies, even though the discriminative abilities of different supervectors are unbalanced. The final system with both the fusion and ASSV combined together gives additional performance improvement.

However, when it comes to unknown attacks, the two systems both perform considerably worse, and achieve basically the same EERs. Table 3 illustrates the performance of different attacks in the evaluation set. S1–S3 appear in the training set while S6–S10 stand for unknown attacks [11]. The results come from our original primary submission before the ASVspoof 2015 deadline.

Table 3: Attack-dependent EER results in the evaluation set.

<table>
<thead>
<tr>
<th>Known</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0.173</td>
<td>0.610</td>
<td>0.319</td>
<td>0.289</td>
</tr>
<tr>
<td>S2</td>
<td>0.906</td>
<td>0.242</td>
<td>0.417</td>
<td>0.246</td>
</tr>
</tbody>
</table>

It seems that S6–S9 are relatively easy to detect because they are voice conversion algorithms using the same STRAIGHT vocoder with S1–S4 [17]. The main error comes from S10 which dwarfs all others by 2 orders of magnitude. S10 denotes a text-to-speech technology that does not use a vocoder. The fact that no vocoder involved makes it sound more natural than other spoofed speech. The magnitude texture and phase information of such speech are unfamiliar to our classifiers as well, making it difficult to recognize these utterances in the evaluation set. New approaches need to be developed to overcome this problem.

5. Conclusions

This paper presents a novel anti-spoofing system that combines several unique features based on both spectral magnitude and phase information. The supervector structure is effective at detecting attacks, and outperforms the conventional GMM log-likelihood scoring method. Experimental results show that, compared with the subsystems, score fusion and ASSV greatly reduce the EER, by 93% and 86% in the development set, respectively. For known attacks, the fusion of these two systems lowers the EER further, while more countermeasures are needed to detect advanced unknown spoofing attacks. Overall the combination of spectral magnitude and phase features produces a significant improvement in spoofing detection, and the proposed ASSV framework is extensible and has the potential for a simple but flexible back-end design.

Future work includes incorporating a weighting algorithm to emphasize specific dimensions of ASSV in proportion to their discriminative ability, and enhancing ASSV with other powerful supervectors.

6. Acknowledgements

The authors want to thank Zhizheng Wu in University of Edinburgh for his timely reply to our ASVspoof 2015 submissions.

The work is supported by National Natural Science Foundation of China under Grant No. 61370034, No. 61403224 and No. 61273268.
7. References


