Anti-spoofing System: An Investigation of measures to Detect Synthetic And Human Speech

*Abhinav Misra, Shivesh Ranjan, Chunlei Zhang, John H. L. Hansen**

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
The University of Texas at Dallas (UTD), Richardson, TX 75080-3021, USA
{abhinav.misra, shivesh.ranjan, chunlei.zhang, john.hansen}@utdallas.edu

Abstract

Automatic Speaker Verification (ASV) systems are prone to spoofing attacks of various kinds. In this study, we explore the effects of different features and spoofing algorithms on a state-of-the-art i-vector speaker verification system. Our study is based on the standard dataset and evaluation protocols released as part of the ASVspoof 2015 challenge. We compare how different features perform while detecting both genuine and spoofed speech. We observe that features that contain phase information (Modified Group Delay based features) are better in detecting synthetic speech, and give comparable performance when compared to standard MFCCs. We report an anti-spoofing system that performs well both on known as well as unknown spoofing attacks.

Index Terms: speaker verification, anti-spoofing, countermeasures, i-vector

1. Introduction

Automatic speaker verification systems, just like other biometric systems, are prone to spoofing. However, as compared to other branches of biometrics like face recognition or finger print recognition, spoofing of speaker recognition/verification is yet to be thoroughly addressed. Some of the previous works [1, 2] include developing anti-spoofing measures that require prior knowledge of the spoofing attacks to be targeted at the ASV system. However, generalized countermeasures, that have no knowledge of the nature of attack are much more relevant in a practical scenario. In this paper, we focus on such generalized countermeasures.

With the aim of facilitating study of generalized spoofing countermeasures, ASVspoof 2015 challenge was designed [3]. It provides a platform to compare and evaluate different anti-spoofing techniques against a set of standard database and evaluation metrics. This study is based on the training and development subset of the standard dataset released by the organizers of this challenge.

Present state-of-the-art ASV system is based on an i-vector [4] representation of speech utterance modelled by a Probabilistic Linear Discriminant Analysis (PLDA) back-end [5, 6]. An i-vector contains both speaker and channel information, but channel component is suppressed by compensation techniques like Linear Discriminant Analysis (LDA), Within-Class Covariance Normalization (WCCN) [7] or length normalization [8]. Thus, at the scoring stage, there is little channel component left. All this makes an i-vector a very attractive feature to be used in an anti-spoofing system. It is assumed that an i-vector would be able to model natural speech characteristics in a much better way and hence, would be less prone to attacks by synthetic speech generating algorithms. In this study, we develop an i-vector based anti-spoofing classifier to discriminate between human and synthetic speech. It may be noted here that in [9], authors have introduced the use of i-vectors in a generalized countermeasure scenario, but they have integrated it with the ASV system. Our approach here, is to design an anti-spoofing system independently of an i-vector-PLDA ASV system. The designed system can later be added to an ASV system through score fusion.

Though, synthetic speech increases the false alarms of an ASV system, several studies have shown that comparitively humans are much better in differentiating synthetic and genuine speech [10, 11, 12]. One of the reasons behind this is that phase spectrum, which plays an important role in human speech perception [13, 14], is generally not taken into account in most of the speech conversion/synthesis algorithms. All this motivated us to consider features that contain phase information. In [15], authors have demonstrated the usefulness of phase based features in speech recognition. They compute group delay functions from the speech signals and then convert them to cepstral coefficients. We study i-vector extraction based on Mel-Frequency Cepstral Coefficients (MFCCs) and group delay based features and compare their effect on an anti-spoofing system. Based on our analysis, we offer new directions at addressing a challenging problem which has received limited attention in the speaker recognition community.

The remainder of this paper is organized as follows: Section 2 explains the standard corpus and evaluation metrics, Section 3 describes the anti-spoofing system, Section 4 details the experiments conducted and analyses the results, Section 5 concludes the paper and discusses future work.

2. Corpus and Protocols

In this study, we work on the standard Spoofing and Anti-Spoofing (SAS) corpus [1]. The corpus was divided into three subsets by the organizers of ASVspoof 2015 challenge [3]. The labels of evaluation subset are not yet released, so we base our

This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

[1] https://wiki.inf.ed.ac.uk/CSTR/SASCorpus

study on the training and development subset. Both training and development subsets contain human and synthetic speech from male and female speakers. There is no overlap between the speakers in these two subsets. The synthetic speech is generated from the following voice conversion and speech synthesis algorithms:

- S1: A simplified frame selection based voice conversion algorithm [16]
- S2: A voice conversion technique that changes only the first coefficient of Mel-Cepstral coefficients.
- S3: Speaker adapted speech synthesis based on HMM [17]
- S4: Same synthesis system as in S3, but using more data for speaker adaptation.
- S5: Voice conversion algorithm based on a joint Gaussian mixture model with maximum likelihood parameter estimation taking global variance into account. [18]

All the files are 16 kHz, mono channel with most of them around 5 seconds long. Training dataset has human and synthetic speech from 25 speakers (10 male and 15 female), while the development dataset has human and synthetic speech from 35 speakers (15 male and 20 female). Both the datasets comprise of a list of trials consisting of spoofed and genuine speech audio files. Corresponding to each trial, a score has to be assigned determining whether it's a genuine or spoofed speech. Going on similar lines as NIST Speaker Recognition Evaluations (SREs), these scores are then used to compute Equal Error Rate (EER). Genuine speech belongs to target trials, while spoofed speech belongs to non-target trials. Training dataset has 3750 genuine speech audio files (trials) and 12625 spoofed speech audio files/trials. Similarly, development dataset has 3497 genuine speech audio files and 49875 spoofed speech audio files.

3. System Description

3.1. I-vector extraction

An i-vector allows a fixed low dimensional representation of an utterance while preserving the utterance-specific information. Let M be a speech-category specific (i.e. genuine or spoofed speech) GMM mean supervector, m be the speech-category independent supervector, and T be a low rank total variability matrix. Then, the speech-category specific GMM mean supervector M can be expressed as a linear combination of m, and the columns of T as,

$$M = m + Tw. \tag{1}$$

In (1), w is a random vector with standard normal distribution, and the columns of T are weighted by its elements. The *total variability matrix* T is learned by using large amounts of labeled training data. The i-vector of an utterance can also be viewed as the corresponding coordinates in the *total variability space* (whose basis are given by the columns of T), and are extracted as the maximum a posteriori (MAP) point estimates of w, using the utterance [4].

3.2. PLDA based scoring

Currently, state-of-the-art speaker recognition systems employ one of the many variants of PLDA based scoring techniques. Using the i-vectors extracted from a collection of labeled-training data, a PLDA model computes the within-class and across-class variabilities using an Expectation Maximization (EM) algorithm. In this study, we used a Gaussian PLDA (G-PLDA) of the form described in [8]. The G-PLDA model parameters $\{m, \Phi, \Sigma\}$ are estimated using EM algorithm on the labeled training data. Given a test utterance, we compute scores corresponding to genuine and spoofed speech, and decide in favor of the category with a higher score. Let ψ_e, ψ_t be the enrollment and test i-vectors respectively. To calculate the log-likelihood score between these two i-vectors, we use the following equation:

$$S(\psi_e, \psi_t) = \log \frac{p(\psi_t, \psi_e|\theta)}{p(\psi_t|\theta)p(\psi_e|\theta)} \tag{2}$$

where,

$p(\psi_t, \psi_e|\theta)$ is the probability that i-vectors ψ_e, ψ_t are coming from the same class, whereas $p(\psi_t|\theta)p(\psi_e|\theta)$ is the probability that i-vectors ψ_e, ψ_t are from different classes.

3.3. Gaussian scoring (GS)

In our experiments, we observed that PLDA approach did not work well in spoofing detection under mismatched conditions [19]. So, we further investigated a Generative Gaussian Classifier in our work. GS is a classical classification technique which is more commonly used in language identification (LID).

To use GS in our spoofing detection system, i-vectors of each speech type (genuine and spoofed) are modeled by a Gaussian distribution. A full covariance matrix is shared across both the classes of speech. For each i-Vector x corresponding to a test utterance, we evaluated the log-likelihood for each category as:

$$\log P(\mathbf{x}|Y_i) = \mu_i^T \Sigma^{-1} \mathbf{x} - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_\mathbf{i} - \frac{1}{2} \mathbf{x^T} \Sigma^{-1} \mathbf{x} + c \tag{3}$$

Where μ_i is the mean vector for speech type Y_i, Σ is the common covariance matrix, c is a constant related to training data and the prior distribution of test data (we assume this is unknown in the challenge, set $P(Y_{genuine} = P(Y_{spoof}) = 0.5)$).

3.4. Gaussian scoring with LDA (GS + LDA)

This back-end is same as Gaussian scoring except that it also has a layer of LDA before it.

3.5. Gaussian cosine distance scoring (GCDS)

The classical cosine distance scoring (CDS) for i-vector based system is given by the following equation:

$$k(w_1, w_2) = \frac{(A^T w_1)^T (A^T w_2)}{\sqrt{(A^T w_1)^T (A^T w_1)} \sqrt{(A^T w_2)^T (A^T w_2)}} \tag{4}$$

where A is a projection matrix, which may come from within class covariance normalization (WCCN) or linear discriminative analysis (LDA) projection, and w denotes the i-vector of corresponding speech utterance. The operations are generally performed in a cascade fashion: where the i-vector is first projected through WCCN matrix and then LDA transformation is applied, both of which are estimated from a background data set. We note that performance of classical LDA-WCCN-CDS methods highly depend on the WCCN projection,

which is usually difficult to estimate (especially in noisy and/or channel mismatch conditions). Therefore, WCCN with background data based Gaussianization, called Gaussianized CDS (GCDS) is used in this work. The algorithm is outlined below [20]:

- Average the i-vectors of the j^{th} enrollment speaker;
- Calculate the mean and variance of the background data, which are used to Gaussianize the i-vectors from above step;
- Apply length normalization on all the data [8];
- Apply LDA on all the data to reduce the dimensionality;
- Repeat Step 3;
- Perform cosine distance scoring;
- Score normalization (calculate the mean and variance of scores involved in the i th test file, which are then Gaussianized by the derived mean and variance).

3.6. Modified group delay based features

Let $x(n)$ be a given speech signal, then its short time fourier transform (STFT) is:

$$X(\omega) = |X(\omega)|^{e^{j\phi(\omega)}} \tag{5}$$

where $|X(\omega)|$ is short-time magnitude and $\phi(\omega)$ is short-time phase of the speech signal. Group delay is defined as the negative derivative of phase with respect to ω :

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2} \tag{6}$$

where, $Y(\omega)$ is the STFT of $nx(n)$, $X_R(\omega)$, $Y_R(\omega)$ are real parts of $X(\omega)$, $Y(\omega)$ while $X_I(\omega)$, $Y_I(\omega)$ are their imaginary parts.

To reduce the spiky nature of group delay function, the power spectrum ($|X(\omega)|^2$) is smoothed and two new parameters, namely, α and γ are added. The modified group delay function is now defined as:

$$\tau_\gamma(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|S(\omega)|^{2\gamma}} \tag{7}$$

where, $S(\omega)$ is the smoothed version of $X(\omega)$.
Further,

$$\tau_{\alpha,\gamma}(\omega) = \frac{\tau_\gamma(\omega)}{|\tau_\gamma(\omega)|}|\tau_\gamma(\omega)|^\alpha \tag{8}$$

where, $\tau_{\alpha,\gamma}(\omega)$ is the final modified group delay function. Its value can be adjusted for any given environment by varying the parameters α and γ. For the experiments done in this study, $\alpha = 0.4$ and $\gamma = 1.2$.

After computing the modified group delay function, it is converted to cepstra using Discrete Cosine Transform (DCT). We consider only first 12 coefficients, after dropping the 0th coefficient.

4. Experiments

The entire training subset was first used to train a Universal Background Model (UBM) whose mean supervector is represented by M in equation 1. We then train a T matrix using the same data. Next, i-vectors corresponding to each trial, are extracted from the UBM and T matrix. Further, mean of all the human speech as well as synthetic speech training i-vectors are computed. These two mean i-vectors represent our target (human speech) and non-target (synthetic speech) classes.

Now, development set i-vectors are computed the same way from T matrix and UBM. They are scored against the mean human speech i-vector and mean synthetic speech i-vector. Scoring is done based on different back-ends as explained in previous section. All these back-ends are trained using training set i-vectors having two classes, namely, human and synthetic speech. Scores obtained against mean human speech i-vector correspond to target scores, while those against synthetic speech i-vector correspond to non-target scores. These target and non-target scores are then employed to compute EER of the system. Table 1, report the results obtained for two systems, one using MFCCs while the other using modified group delay (MGD) based features.

Table 1: *Results for MFCC and MGD features.*

Back-ends	MFCC		MGD	
	EER%	Accuracy%	EER%	Accuracy%
GS	**15.58**	83.73	**17.96**	82.10
GS + LDA	27.46	72.51	29.20	70.77
GCDS	36.09	63.91	19.39	80.61
PLDA	36.06	63.91	19.39	80.61

It can be observed from the table, that MGD features give comparable performance as MFCC features. MFCC gives 15.58% EER with Gaussian back-end, while MGD gives 17.96%. Also, MGD performs much better than MFCC with PLDA back-end. If we look at the confusion matrix plotted in fig 2, we can easily observe that MGD features are much better in detecting synthetic speech. In contrast, MFCC features are better in detecting genuine speech. Y-axis in figure corresponds to percentage of trials detected as spoofed/genuine speech. MFCC was able to detect only around 60% of spoofed speech, while MGD was able to detect more than 80% of spoofed speech. Thus, both have complementary information, that motivated us to further fuse the two systems.

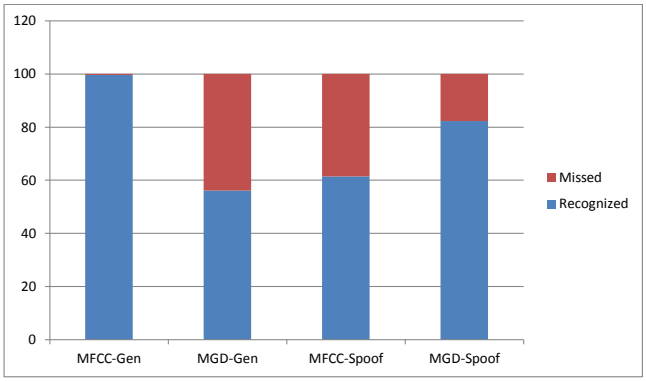

Figure 2: Plot of accuracies of MFCC and MGD features in detecting genuine and spoofed speech

Fusion: Score level fusion of MFCC and MGD based systems is done. Out of total 35 speakers in development dataset, 17 (7 male and 10 female) are taken out and used to form a train set for fusion. Logistic regression is used to learn combination

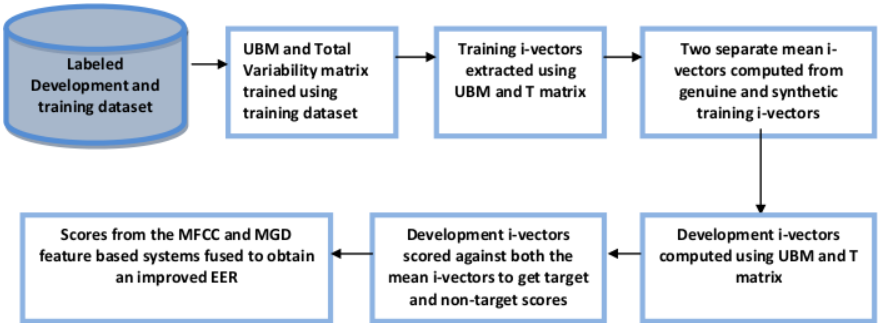

Figure 1: Flow diagram of the proposed anti-spoofing system.

weights from these train set scores. These weights are then used to fuse individual MFCC and MGD based system scores. From table 1, the best EER of our anti-spoofing system is 15.58% obtained using an MFCC based system. After fusion, we got a **7.03%** absolute improvement in this EER bringing it down to **8.55%**. It can be noted that in fig 3, Detection Error Trade-off (DET) curves for MFCC and MGD systems are straight lines. This happened as there was not much fluctuation in the scores we got for these systems.

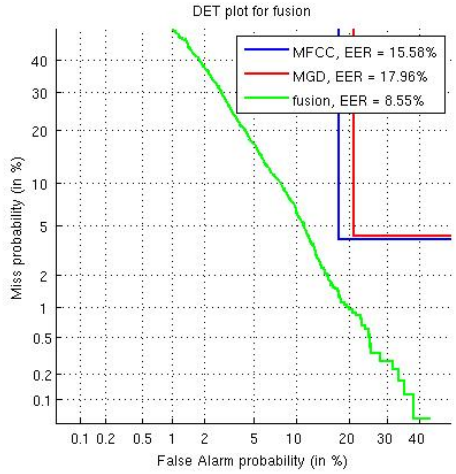

Figure 3: Fusion of MFCC and MGD based systems

5. Conclusion

In this paper, we considered a systematic study on a generalized countermeasure/anti-spoofing system. The system was based on an i-vector framework and experiments were reported on a standard dataset following standard protocols . It was observed that features based on the phase spectrum of speech signal are better in detecting synthetic speech. This happens as voice conversion or speech synthesis algorithms used to generate synthetic speech usually do not contain natural/original phase information. When these phase based features were fused with MFCC features, a relative improvement of **+45.12**% was observed in the system EER, demonstrating the highly complimentary nature of the two features. In future, work would be focussed on seamlessly integrating this i-vector based anti-spoofing system with general state-of-the-art i-vector based ASV system.

6. References

[1] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an hmm based speech synthesis system," in *Proc. Eurospeech*, 2001.

[2] P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet, "Voice forgery using alisp: indexation in a client memory," in *Acoustics, Speech and Signal Processing, 2005. ICASSP 2005. IEEE International Conference on*, 2005.

[3] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagish, "Asvspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," in *Proc. InterSpeech*, 2015.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, May 2010.

[5] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey*, Brno, Czech Republic, 2010.

[6] J. Villalba and N. Brummer, "Towards fully Bayesian speaker recognition: Integrating out the between-speaker covariance," in *Proc. InterSpeech*, Florence, Italy, Oct. 2011.

[7] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. InterSpeech*, Pittsburgh, Pennsylvania, 2006.

[8] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-Vector length normalization in speaker recognition systems," in *Proc. InterSpeech*, Florence, Italy, Oct. 2011.

[9] 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.

[10] Q. Jin, A. Toth, T. Schultz, and A. Black, "Is voice transformation a threat to speaker identification?" in *ICASSP*, 2007.

[11] T. Kinnunen, Z. Wu, K. Lee, F. Sedlak, E. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech," in *ICASSP*, 2012.

[12] J. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Interspeech*, 2007.

[13] K. Paliwal and L. Alsteris, "On the usefulness of stft phase spectrum in human listening tests," *Speech Communication*, 2005.

[14] L. Alsteris and K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication*, 2006.

[15] R. Hegde, H. Murthy, and V. Gadde, "Significance of the modified group delay feature in speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, 2007.

[16] T. Wu, Z. Virtanen, T. Kinnunen, E. Chng, and H. Li, "Exemplar-based unit selection for voice conversion utilizing temporal information," in *Proc. InterSpeech*, 2013.

[17] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, January 2009.

[18] T. Toda, A. Black, and k. Tokuda, "Voice conversion based on maximumlikelihood estimation of spectral parameter trajectory," *Audio, Speech, and Language Processing, IEEE Transactions on*, 2007.

[19] 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.

[20] G. Liu, T. Hasan, H. Boril, and J. H. Hansen, "An investigation on back-end for speaker recognition in multi-session enrollment," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7755–7759.