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 imposter 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].

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.

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

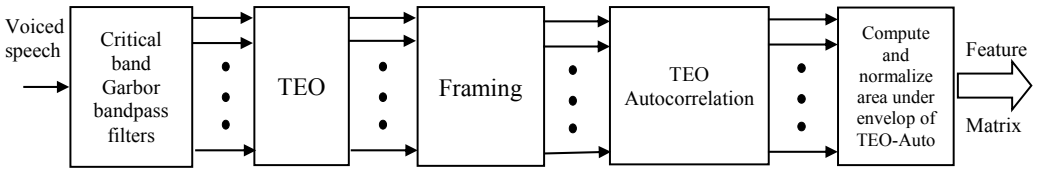

Fig. 1. TEO-CB-Auto-Env feature extraction.

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, auto-correlation 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.

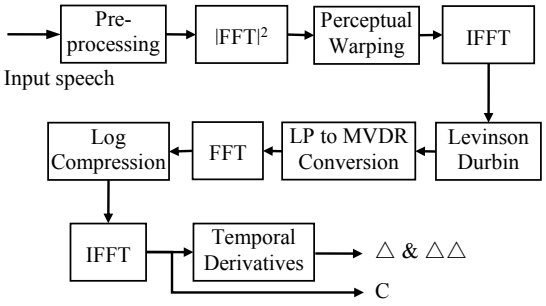

Fig. 2. Flow diagram of PMVDR feature extraction.

For each feature extracted, a Universal Background Model (UBM) is trained using all available training data. In the context in the spoofing detection i-vector system outlined of this paper, we expect UBM to roughly model the acoustic structure represented by TEO-CB-Auto-Env and PMVDR features.

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:

$$M_u = m + Tx_u, \tag{1}$$

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).

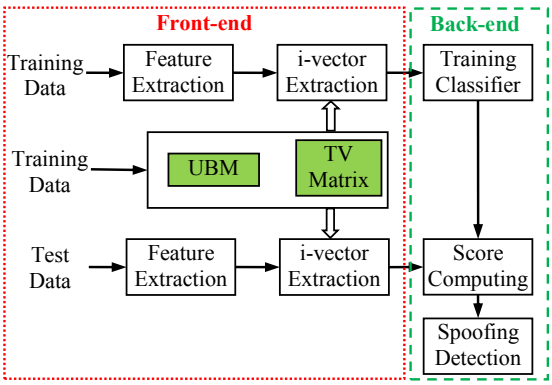

Fig. 3. Flow diagram of i-vector system.

3.4. Back-end classifiers

3.4.1. Gaussian Classifier

A Generative Gaussian Classifier(GC) is investigated here for spoofing detection. GC is a classical classifier which is more commonly adopted in language identification. For the GC in our spoofing detection system, i-vectors of each class (genuine and spoofing) are modeled by a Gaussian distribution, where the covariance matrix is the same for both categories. For each i-vector x corresponding to a test utterance, we evaluate the log-likelihood for each category:

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

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, and set $P(Y_{genuine}) = P(Y_{spoof}) = 0.5)$).

3.4.2. Deep Neural Network

Two deep neural networks(DNNs) are trained for back-end scoring. The first DNN system for spoofing challenge is implemented based on KALDI and PDNN [12, 20]. Firstly, 40 dimensional filter bank feature are derived for each frame of every utterance. Subsequently, 11 consecutive frames of f-bank features (440 dimension in total) are utilized as the first layer input. Here, we use 4 hidden layers with 1024 hidden nodes per layer, the final softmax layer consists of two nodes, which represent the genuine and spoofing class probability respectively. Since there is no speech content information involved in this classification task, only two labels assigned for each input acoustic feature set. Therefore, on the decoding part, an average score across the whole utterance is employed as final classification probability.

The second DNN is trained for each feature-based i-vector (TEO-CB-Auto-Env and PMVDR). The input layer of the DNN consists of i-vectors derived from the i-vector system described above. We use a sigmoid activation function, 2 hidden layers with 1024 hidden nodes per layer, and a final softmax layer consisting of two nodes, representing the genuine and spoofed class probability respectively.

In the end, the output scores for all neural networks (f-bank feature, TEO-CB-Auto-Env and PMVDR) are fused using logistic regression as the final output result.

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.

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

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.

Feature	System	EER(%)	Accuracy
TEO	i-vector/GC	8.80	93.61
PMVDR	i-vector/GC	7.17	92.88
TEO+PMVDR	i-vector/GC	**1.67**	**98.85**
f-bank	DNN	6.14	96.00
TEO	i-vector/DNN	4.83	96.27
PMVDR	i-vector/DNN	4.58	95.53
f-bank+TEO+PMVDR	fusion	**0.71**	**99.33**

Table 2. EER(%) and Accuracy(%) for ASVspoof 2015 on development data.

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%.

known attacks	S1	S2	S3	S4	S5
Primary	0.24	2.16	**0.07**	0.10	0.78
Flexible Primary	0.26	**1.22**	0.35	0.38	0.77
Fusion/DNN	**0.17**	2.03	0.19	**0.08**	**0.67**
Unknown attacks	S6	S7	S8	S9	S10
Primary	2.31	0.24	**0.14**	0.31	**27.20**
Flexible Primary	2.47	**0.22**	0.35	0.32	32.65
Fusion/DNN	**2.01**	0.35	0.40	**0.30**	27.94

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

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.

4.3. Imbalanced learning for i-vector system

Most learning algorithms assume or expect a balanced data distribution. When faced with imbalanced data, some learning algorithms may fail to properly represent the distribution characteristics of the data classes.

Here, we also experience the imbalanced learning problem; the quantity of spoofed speech is roughly 3 times that of genuine speech in the training set, while for development, the ratio becomes over 14. One simple solution here is resampling, either upsampling or downsampling [22]. For upsampling, synthetic minority oversampling technique (SMOTE) is a powerful method that has shown some success in many applications [23]. Cluster-based downsampling works in the opposite direction to acquire balanced data [24].

We apply downsampling on our training data for the i-vector PLDA system (which does not perform well in spoofing detection as described above). K-means is applied to cluster the spoofing i-vectors. We downsampled 12625 spoofed i-vectors to 3750 to equal the number of genuine speech recordings. Detailed performance

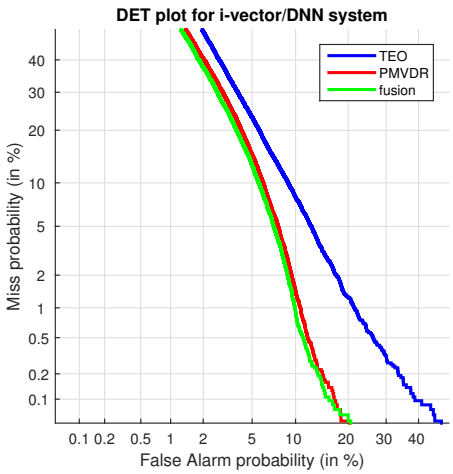

Fig. 4. DET plot for i-vector/DNN on evaluation data.

Before Clustering		After Clustering	
EER	Accuracy	EER	Accuracy
30.71	69.28	10.07	88.92
Confusion matrix		Confusion matrix	
genuine	spoofing	genuine	spoofing
3488	9	**3394**	103
16388	**33487**	5810	**44065**

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

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 [**?**, 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

[1] B.L Pellom and J. H.L Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *IEEE ICASSP*. IEEE, 1999, vol. 2, pp. 837–840.

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

[3] 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 Langu. Proce.*, vol. 22, no. 12, pp. 1978–1992, 2014.

[4] 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.

[5] 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 *ISCA INTERSPEECH*, 2015.

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

[7] 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 *Signal & Info. Proc. Asso. Annual Summit and Confe. (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 2012, pp. 1–5.

[8] 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.

[9] G. Zhou, J. H.L. Hansen, and J. F Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. on Speech and Audio Proce.*, vol. 9, no. 3, pp. 201–216, 2001.

[10] M. Wester, Z. Wu, and J. Yamagishi, "Human vs machine spoofing detection on wideband and narrowband data," in *ISCA INTERSPEECH*, 2015.

[11] C. Zhang, G. Liu, C. Yu, and J. H.L. Hansen, "I-vector based physical task stress detection with different fusion strategies," in *ISCA INTERSPEECH*, 2015.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, et al., "The kaldi speech recognition toolkit," 2011.

[13] E. Khoury, T. Kinnunen, A. Sizov, Z. Wu, and S. Marcel, "Introducing i-vectors for joint anti-spoofing and speaker verification," in *ISCA INTERSPEECH*, 2014.

[14] 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 *IEEE ICASSP*. IEEE, 2015, pp. 4440–4444.

[15] G. Liu, T. Hasan, H. Boril, and J. H.L Hansen, "An investigation on back-end for speaker recognition in multi-session enrollment," in *IEEE ICASSP*. IEEE, 2013, pp. 7755–7759.

[16] G. Liu, C. Zhang, and J. H.L Hansen, "A linguistic data acquisition front-end for language recognition evaluation," in *ISCA Odyssey*, 2012.

[17] M. Penagarikano, A. Varona, M. Diez, L. J. Rodriguez-Fuentes, and G. Bordel, "Study of different backends in a state-of-the-art language recognition system.," in *ISCA INTERSPEECH*, 2012.

[18] Q Zhang, G Liu, and J.H.L. Hansen, "Robust language recognition based on hybrid fusion," in *ISCA Odyssey*, 2014.

[19] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[20] Y. Miao, "Kaldi+pdnn: building dnn-based asr systems with kaldi and pdnn," *arXiv preprint arXiv:1401.6984*, 2014.

[21] N. Brümmer and E. De Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013.

[22] H. He and E. A Garcia, "Learning from imbalanced data," *IEEE Trans. on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[23] N. V Chawla, K. W Bowyer, L. O Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artif. Intel. Resea.*, vol. 16, no. 1, pp. 321–357, 2002.

[24] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 40–49, 2004.

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

[26] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Classifiers for synthetic speech detection: A comparison," in *ISCA INTERSPEECH*, 2015.

[27] A. Misra, S. Ranjan, C. Zhang, and J. H.L. Hansen, "Anti-spoofing system: An investigation of measures to detect synthetic and human speech," in *ISCA INTERSPEECH*, 2015.