FACTOR ANALYSIS METHODS FOR JOINT SPEAKER VERIFICATION AND SPOOF DETECTION

Dhanush B K[1], Suparna S[2], Aarthy R[4], Likhita C[3], Shashank D[3], Harish H[1], Sriram Ganapathy[1]

[1] LEAP labs, Electrical Engineering, Indian Institute of Science, Bangalore, India.
[2] Indian Institute of Technology Madras, Chennai, India.
[3] National Institute of Technology Karnataka, Surathkal, India.
[4] National Institute of Technology Trichy, Tiruchirappalli, India.

ABSTRACT

The performance of a speaker verification system is severely degraded by spoofing attacks generated from artificial speech synthesizers. Recently, several approaches have been proposed for classifying natural and synthetic speech (spoof detection) which can be used in conjunction with a speaker verification system. In this paper, we attempt to develop a joint modelling approach which can detect the presence of spoofing attacks while also performing the speaker verification task. We propose a factor modelling approach where the spoof variability subspace and speaker variability subspace are jointly trained. The lower dimensional projection in these subspaces are used for speaker verification as well as spoof detection tasks. Several experiments are performed using the speaker and spoofing (SAS) database. We compare the performance of the proposed method with a baseline method of fusing a conventional speaker verification system and a spoof detection system. In these experiments, the proposed approach provides substantial improvements for spoofing detection (relative improvements of 21% in EER over the baseline) as well as speaker verification under spoofing conditions (relative improvements of 17% in EER over the baseline).

Index Terms— Spoof detection, Speaker verification, Joint factor analysis, ivectors.

1. INTRODUCTION

Automatic speaker verification (ASV) systems are widely used in commercial and forensic applications for the binary task of verifying the claimed identity of a speaker. The performance of a typical speaker verification system is severely degraded by the presence of artificial or natural speaker impersonations. In the past, the vulnerability of these systems to various spoof attacks like voice conversion [1], mimicry attacks [2], and synthetic speech [3] has been analyzed. A survey of various spoofing attacks on speaker verification systems can be found in [4].

Recently, the ASVSpoof Challenge 2015 [5] was conducted to enable the development of countermeasures for spoof detection on a variety of speech synthesis methods. The task here was the classification of a speech utterance as natural or synthetic. The majority of the methods developed in the challenge were based on feature extraction approaches including phase spectrum [6], linear prediction error [7] and magnitude spectrum [8]. The best results for this challenge were obtained using a combination of short-term spectral magnitude and frequency modulation features with a simple Gaussian mixture model (GMM) classifier [9]. The use of these countermeasures in the ASV system would require some sort of fusion between the spoof detection system scores and ASV scores.

In this paper, we attempt to jointly model the spoofing attacks within an ASV system. In particular, we propose to model the across speaker variations and within speaker spoof variations in a joint factor model (JFA) [10, 11]. The JFA model is trained to separate the lower dimensional subspaces representing speaker and spoof variability (inter speaker variabilities) and the session variabilities (intra speaker variabilities). The factors representing the inter speaker variabilities alone are used for spoof detection task as well as the speaker verification task. The spoof detection task is achieved by training a support vector machine (SVM) classifier [12] while the speaker verification is achieved by probabilistic linear discriminant analysis (PLDA) scoring [13].

We use the speaker verification and spoofing (SAS) database [14, 5] which contains recordings from several speakers in diverse spoofing conditions. In our spoof detection experiments, we show that the modelling of subspaces using JFA is able to outperform the standalone countermeasure methods. The speaker verification results obtained by the proposed approach is compared with the baseline method of fusing the ASV system scores with the spoof countermeasure scores. In the ASV task, the proposed method improves the baseline significantly (average relative improvement of 17% in equal error rate (EER)).

The rest of the paper is organized as follows. In Sec. 2 we discuss the ivector and JFA modelling methods. Sec. 3 describes the the various approaches for spoof detection and speaker verification. The database, experimental setup and results are described in Sec. 4 followed by a discussion of the results in Sec. 5. In Sec. 6, we provide a brief summary along with potential future directions.

2. FACTOR ANALYSIS FRAMEWORK

The techniques outlined here are derived from the previous work on joint factor analysis (JFA) and ivectors [10, 11, 15]. We follow the notations used in [10]. The training data from all the speakers is used to train a GMM with model parameters $\lambda = \{\pi_c, \boldsymbol{\mu_c}, \boldsymbol{\Sigma_c}\}$ where π_c, $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ denote the mixture component weights, mean vectors and covariance matrices respectively for $c = 1, .., C$ mixture components. Here, $\boldsymbol{\mu}_c$ is a vector of dimension F and $\boldsymbol{\Sigma}_c$ is of assumed to be diagonal matrix of dimension $F \times F$.

2.1. I-vector Representations

Let $\boldsymbol{\mathcal{M}_0}$ denote the UBM supervector which is the concatenation of $\boldsymbol{\mu}_c$ for $c = 1, .., C$ and is of dimension of $CF \times 1$. Let $\boldsymbol{\Sigma}$ denote the block diagonal matrix of size $CF \times CF$ whose diagonal blocks

are $\mathbf{\Sigma}_c$. Let $\mathcal{X}(s) = \{\boldsymbol{x}_i^s, i = 1, ..., H(s)\}$ denote the low-level feature sequence for input recording s where i denotes the frame index. Here $H(s)$ denotes the number of frames in the recording. Each $\boldsymbol{x}_i^s$ is of dimension $F \times 1$.

Let $\boldsymbol{\mathcal{M}(s)}$ denote the recording supervector which is the concatenation of speaker adapted GMM means $\boldsymbol{\mu}_c(s)$ for $c = 1, .., C$ for the speaker s. Then, the ivector model is,

$$\boldsymbol{\mathcal{M}}(s) = \boldsymbol{\mathcal{M}_0} + \boldsymbol{V}\boldsymbol{y}(s) \tag{1}$$

where $\boldsymbol{V}$ denotes the total variability matrix of dimension $CF \times M$ and $\boldsymbol{y}(s)$ denotes the ivector of dimension M. The ivector is assumed to be distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

In order to estimate the ivectors, the iterative EM algorithm is used [10].

2.2. Joint Factor Analysis

The JFA approach attempts to capture the additional channel factors that represent intraspeaker variability [11]. These factors represent the variability in the recording environment for different segments from the same speaker. For this case, we assume that for speaker s, there are $q = 1, .., Q(s)$ sessions, each with $H_q(s)$ frames. The JFA model is

$$\begin{aligned} \boldsymbol{\mathcal{M}}(s) &= \boldsymbol{\mathcal{M}_0} + \boldsymbol{V}\boldsymbol{y}(s), \\ \boldsymbol{\mathcal{M}}_q(s) &= \boldsymbol{\mathcal{M}}(s) + \boldsymbol{U}\boldsymbol{x}_q(s), \end{aligned} \tag{2}$$

where $\boldsymbol{V}$ denotes the speaker variability matrix of size $CF \times M$, $\boldsymbol{U}$ denotes the channel/session variability matrix of size $CF \times N$. Here, $\boldsymbol{\mathcal{M}}(s)$ and $\boldsymbol{\mathcal{M}}_q(s)$ represent supervectors for the entire data from speaker s and for the session q from speaker s respectively. The factors $\boldsymbol{y}(s)$ and $\boldsymbol{x}_q(s)$ are speaker factors and channel factors of dimension M and N respectively. The subspace $\boldsymbol{VV}^*$ captures the interspeaker variability while the subspace $\boldsymbol{UU}^*$ captures the intraspeaker channel variability. Let $\underline{\boldsymbol{Y}}(s)$ denote the collection of factors for each speaker s. $\underline{\boldsymbol{Y}}(s) = [\boldsymbol{x}_1^*(s)\ \boldsymbol{x}_2^*(s)\ ...\ \boldsymbol{x}_{Q(s)}^*(s)\ \boldsymbol{y}^*(s)]^*$. Also, let

$$\underline{\boldsymbol{V}} = \begin{bmatrix} \boldsymbol{U} & & & \boldsymbol{V} \\ & \ddots & & \vdots \\ & & \boldsymbol{U} & \boldsymbol{V} \end{bmatrix} \tag{3}$$

where $\underline{\boldsymbol{V}}$ is of dimension $[Q(s)CF \times (Q(s)N + M)]$. If we also have $\underline{\boldsymbol{\mathcal{M}}}(s)$ as the splicing of all $\boldsymbol{\mathcal{M}}_q(s)$ for $q = 1, .., Q(s)$ and $\underline{\boldsymbol{\mathcal{M}}}_0$ as the splicing of the same vector $\boldsymbol{\mathcal{M}}_0$ repeated $Q(s)$ times, then we can rewrite Eq. (2) as

$$\underline{\boldsymbol{\mathcal{M}}}(s) = \underline{\boldsymbol{\mathcal{M}}}_0 + \underline{\boldsymbol{V}}\ \underline{\boldsymbol{Y}}(s) \tag{4}$$

which is similar to Eq. (1). Thus, the parameters of the JFA model can be computed in a similar fashion to the EM formulation described in Sec. 2.1. For the ivector and the JFA framework, we use the minimum divergence formulation and orthogonalization after every iterative step [10].

3. APPROACHES FOR SPEAKER VERIFICATION AND SPOOF DETECTION

We highlight three approaches that we have experimented for joint speaker verification and spoof detection. The first approach is to have two stand alone systems - one for spoof detection and one for speaker verification. These stand alone systems are fused to perform speaker verification under spoof conditions. This represents our baseline system. We also develop two systems which can jointly perform these two tasks - based on ivector modelling and joint factor analysis model as shown in Fig. 1. We use the MSR Identity toolbox [16] for the ivector and factor analysis modelling and HTK Speech Recognition toolkit [17] for feature extraction and GMM training.

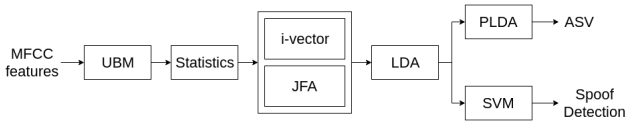

Fig. 1. Block diagram showing the steps involved in the joint speaker verification and spoof detection system

3.1. Fusion of standalone systems

A spoof detection system is developed to separate human and spoofed speech (similar to approaches used for ASVChallenge [18]. Separately, an automatic speaker verification (ASV) system is trained on human speech using the state-of-the-art approaches consisting of ivector with linear discriminant analysis (LDA) and length normalization [19] with probabilistic LDA (PLDA) scoring [20]. These scores are combined with the spoof detection system to perform speaker verification under spoofing conditions.

3.2. Combined ASV and spoof detection - ivector

Here, we use the model represented by Eq. 1 and consider that the speaker, session and spoofing variabilities are all represented by the total variability space $\boldsymbol{V}$. The ivector-PLDA system is then used for speaker verification. The PLDA model is used to separate speakers and to reject spoof trials. The approach is similar to the S-PLDA system of [21].

3.3. Combined ASV and spoof detection - JFA

In this approach, we try to separate the inter-session variabilities from the inter-speaker variabilities and spoof variabilities according to Eq. 2. Using the formulation described in Sec. 2.2, the estimation of the $\boldsymbol{U}$, $\boldsymbol{V}$ subspaces is done using natural and spoofed utterances. This process is intended to separate the inter-speaker and spoof variations represented by factors $\boldsymbol{y}(s)$ and the unwanted session variations represented by the factors $\boldsymbol{x}_q$. The $\boldsymbol{y}(s)$ factors are alone used for speaker verification and spoof detection.

4. EXPERIMENTS

A. Databases – For all the three approaches described in Sec. 3, the evaluation set consists of a separate set of 46 speakers corresponding to genuine speaker recordings and samples generated from all 10 spoofing techniques (Table 1). For speaker verification, we use 100 target trials and 1000 imposter trials for each of the 46 speakers.
(i) *Standalone ASV system* – The clean utterances of Wall Street Journal (WSJ0 and WSJ1) and Resource Management (RM) databases are used for the baseline ASV system on human speakers. This consists of $97,000$ speech recordings totalling about 93 hours of speech. The data is used for training GMM-UBM, total variability matrix, LDA and PLDA models.

Table 1. Definition of Spoof conditions in SAS Database [4]

	Type	Techniques
S1 - S5	Known	VC_FS, VC_EVC, SS_SMALL, SS_LARGE, VC_FESTVOX
S6 - S10	Unknown	VC_GMM, VC_LSP, VC_TVC, VC_KPLS, SS_MARY_LARGE

(ii) *Standalone Spoof Detection* – The SAS database consists of genuine speaker samples and spoofed speech samples corresponding to each speaker generated using ten different spoofing techniques. There are utterances from 106 speakers – 45 male and 61 female. Each utterance has a duration of 2-3 seconds. The database is divided into known and unknown attacks as shown in Table 1. The training portion of SAS database, consisting of 25 speakers is used for developing spoof detection system. We use the development portion of the SAS database consisting of 35 speakers for benchmarking the spoof detection algorithms. The training part contains 8969 genuine speaker recordings and 117713 spoof recordings. The development part consists of 13182 human recordings and 208847 spoof recordings.

(iii) *Combined ASV Spoof Detection* – We use a training set consisting of genuine and spoofing utterances from 60 speakers from the SAS database (training and development). The spoofing utterances are taken from the 5 known techniques (Table 1). This training set is used for both GMM-UBM model training as well as the ivector/JFA subspace training.

B. Feature Extraction – We use 13 mel frequency cepstral coefficients extracted using a Hamming window of 25 ms and a frame shift of 10 ms along with delta and acceleration coefficients. A voice activity detection (VAD) [22] and cepstral mean variance normalization (CMVN) are applied on the features to remove silences and suppress channel artefacts.

C. Spoof Detection – We compare two baseline methods for spoof detection. (i) A GMM loglikelihood ratio based system where two separate GMMs are trained on genuine and spoof speech and a likelihood ratio score is used for the detection task. Here, we compare the performance of diagonal covariance 1024 mixture component GMM trained on 39 dimensional MFCC features with a full covariance 64 mixture component GMM trained on 40 dimensional mel filter bank energies. (ii) An ivector system is developed using a single GMM-UBM (1024 mixture components) trained on both genuine and spoof recordings which is followed by a support vector machine (SVM) based scoring. Here, 200D ivectors are extracted from 1024 mixture component diagonal GMM trained on MFCC features. For the SVM model, 6000 human and spoof utterances are chosen for training. The ivectors are used as features for the SVM training with radial basis function (RBF) kernels. The evaluation set for spoof detection consists of 17000 spoof utterances per spoof condition and 2558 human utterances. The spoof detection results on the development set using the known spoofing conditions is reported in Table 2. As seen here, the full covariance approach with filter bank energy features significantly improves the spoof detection performance compared to the diagonal covariance GMMs. Further, the ivector-SVM approach improves the spoof detection results and the scores from this system are used for fusion with the standalone ASV system.

D. Standalone ASV Setup – The WSJ and RM databases are used for creating a gender independent genuine speaker UBM. The UBM consists of 512 mixture components with diagonal covariance. A 400 dimensional total variability matrix V is trained using the UBM supervectors. The ivectors are reduced to 200 dimensions using LDA and subsequently scored using the PLDA. A length normalization of the ivectors is also performed before the PLDA training [19].

Table 2. Spoof detection EER results (%) on SAS development data using the GMM-diag-1024 system trained on MFCC features, GMM-full-64 system trained on log mel features and ivec-SVM system trained on 200D ivectors from GMM-UBM-diag-1024 (MFCC).

Cond.	GMM-diag-1024	GMM-full-64	ivec-SVM
S1	6.17	0.15	0.45
S2	0.12	0.75	0.22
S3	0.14	0.59	0.22
S4	2.98	0.18	0.42
S5	1.18	0.28	0.30
Avg.	3.41	0.41	0.33

Table 3. ASV performance (Average EER %) in spoofing conditions comparing the standalone system, fusion of standalone systems and the combined system using the ivector/JFA approach.

	Standalone ASV	Score Fusion	Comb. -ivec	Comb. -JFA
S1	22.2	1.96	0.72	1.27
S2	11.5	0.41	0.08	0.28
S3	32.8	1.28	0.17	0.16
S4	37.7	1.96	0.17	0.2
S5	33.6	1.97	0.6	1.13
S6	35	1.77	1.77	1.36
S7	11.8	0.08	0.17	0.32
S8	19.8	0.75	0.17	0.42
S9	22.3	0.16	0.28	0.61
S10	50.1	49.9	49.9	44.1
Avg. known	27.56	1.52	**0.35**	0.61
Avg. unknown	27.8	10.73	10.46	**9.36**
Avg. all	27.68	6.02	5.4	**4.98**

E. Fusion of standalone spoof detection and human ASV system – The log probability estimates obtained from the trained SVM model are fused with the PLDA scores from the standalone ASV system. The SVM scores are scaled by a factor in order to match the range of scores coming from the ivector PLDA system. Table 3 shows the ASV results (measured in EER (%)) obtained for known and unknown spoofing types before and after score fusion with the spoof detection system. As seen here, the score fusion has a substantial improvement in the speaker verification performance under all spoofing conditions. The performance of the fused system forms the baseline results for the combined ASV-spoof detection approaches.

F. Combined spoof detection and ASV (ivector) – A joint 1024 mixture component diagonal covariance UBM is trained and 300 dimensional ivectors are extracted. These ivectors are LDA transformed to 200 dimensions. For the ASV scoring, the LDA transformed ivectors are used in a PLDA framework. For the task of spoof detection, the LDA transformed ivectors are used to train SVM with RBF kernel. The results of the ASV system and the spoof detection system using this joint front-end

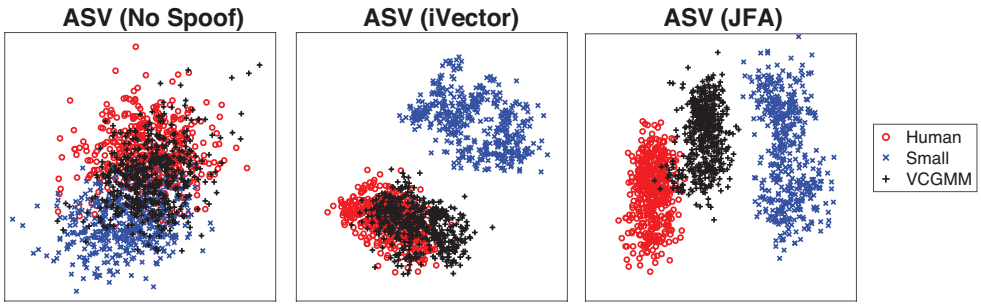

Fig. 2. Scatter plots of first 2 LDA dimensions for standalone ASV, ivector and JFA based joint approaches.

Table 4. Spoof detection performance (Average EER %) comparing the standalone system and the combined system using the ivector/JFA approach.

	Standalone	Comb. -ivec	Comb. -JFA
S1	0.56	0.04	0.12
S2	0.31	0.04	0.16
S3	0.29	0.08	0.18
S4	0.35	0.07	0.16
S5	0.47	0.04	0.12
S6	0.43	0.04	0.14
S7	0.14	0.04	0.08
S8	0.41	0.04	0.14
S9	0.14	0.04	0.08
S10	38.17	31.31	31.45
Avg. known	0.39	**0.05**	0.14
Avg. unknown	7.86	**6.29**	6.38
Avg. all	4.124	**3.17**	3.29

of LDA transformed ivectors are shown in Table 3 and Table 4 respectively.

G. Combined spoof detection and ASV (JFA) – The 1024 mixture component UBM is also used to train two joint factor subspaces corresponding to inter speaker/spoof variability (represented by $\boldsymbol{y}(s)$ and intra speaker session variability (represented by $\boldsymbol{x}_q$. We use only $\boldsymbol{y}(s)$ factors for the ASV system and the spoof detection system. The ASV and spoof detection is done similar to the previous method. The results of the ASV system and the spoof detection system using this joint front-end of LDA transformed JFA factors is also shown in Table 3 and Table 4 respectively.

5. DISCUSSION

The first observation from the results in the previous section points to the substantial drop in performance of the state-of-art ASV system in the presence of spoofing attacks (very high EERs in the first column of Table 3). In order to counter this, the spoof detections need to be integrated with the ASV system (second column of Table 3). The fusion of spoof detection countermeasure scores and ASV scores substantially improves the ASV performance.

The fusion of standalone ASV and spoof detection systems requires the processing of each enroll and test speech utterance through both the systems. This can result in substantially higher level of computational complexity. The framework of having a combined system for performing speaker verification and spoof detection has the advantage of combining the front-end processing for both these tasks using a single pipeline. This also alleviates the need for developing a fusion mechanism in the ASV system. We propose two approaches based on ivectors and JFA models for the purpose of combined ASV and spoof detection.

As reported in Table 3, the combined system improves the average ASV performance compared to the fusion of standalone systems. The JFA based approach provides a better modelling framework to segregate the effects of session and inter-speaker spoof variabilities. This results in an improvement in the average ASV performance (relative EER improvement of 17% over the baseline and 8% over the ivector system). A scatter plot of the first two LDA dimensions for the genuine speaker and spoof utterances, shown in Fig. 2 provides a graphical illustration of various approached experimented in this paper. As seen in this plot, there is significant overlap between the genuine utterances and the spoof utterances in the standalone ASV system. The combined ivector and JFA approaches improve the separation between human and spoof utterances. With the additional subspace training involved in JFA framework, the spoof recordings are further segregated away from the genuine utterances.

In addition to the improvements in the ASV performance with the combined approaches, the spoof detection results reported in Table 4 indicate that the combined approach also provides superior spoofing detection performance compared to the standalone system. This further highlights considerable value provided by the framework of combined ASV and spoof detection in addition to the previously mentioned benefits of reduced computational complexity and improved ASV performance.

6. SUMMARY

In this paper, we have proposed a combined model for performing speaker verification and spoof detection. With a set of experiments on both these tasks, we highlight the advantages of the joint modelling approach. In the future, we would like to progress in the joint modelling framework to have additional subspaces which separate the spoofing variabilities within a given speaker and the inter-speaker variability (JFA with 3 subspaces). In addition, we would also like to incorporate the recent advancements in ASV which include posterior features from a deep neural network.

7. REFERENCES

[1] Tomi Kinnunen, Zhi-Zheng Wu, Kong Aik Lee, Filip Sedlak, Eng Siong Chng, and Haizhou Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4401–4404.

[2] Yee W Lau, Dat Tran, and Michael Wagner, "Testing voice mimicry with the yoho speaker verification corpus," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2005, pp. 15–21.

[3] Phillip L De Leon, Michael Pucher, and Junichi Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," *Proc. IEEE Speaker and Language Recognition Workshop*, 2010.

[4] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[5] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," *Training*, vol. 10, no. 15, pp. 3750, 2015.

[6] Longbiao Wang, Yohei Yoshida, Yuta Kawakami, and Seiichi Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proc. Interspeech*, 2015, pp. 2092–2096.

[7] Artur Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *Proc. Interspeech*, 2015, pp. 2077–2081.

[8] Md Jahangir Alam, Patrick Kenny, Gautam Bhattacharya, and Themos Stafylakis, "Development of crim system for the automatic speaker verification spoofing and countermeasures challenge 2015," in *Proc. 16th Annual Conference of the International Speech Communication Association,(ISCA)*, 2015, pp. 2072–2076.

[9] Tanvina B Patel and Hemant 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.

[10] Patrick Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, 2005.

[11] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[12] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[13] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[14] Zhizheng Wu, Ali Khodabakhsh, Cenk Demiroglu, Junichi Yamagishi, Daisuke Saito, Tomoki Toda, and Simon King, "SAS : A speaker verification spoofing database containing diverse attacks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4440–4444.

[15] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[16] Seyed Omid Sadjadi, Malcolm Slaney, and Larry Heck, "MSR identity toolbox v1. 0: A matlab toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter*, vol. 1, no. 4, 2013.

[17] Steve J Young and Sj Young, *The HTK hidden Markov model toolkit: Design and philosophy*, University of Cambridge, Department of Engineering, 1993.

[18] Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Hong Yu, Tomi Kinnunen, Nicholas Evans, and Zheng-Hua Tan, "Integrated spoofing countermeasures and automatic speaker verification: an evaluation on asvspoof 2015," *Interspeech 2016*, 2016.

[19] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems.," in *Interspeech*, 2011, pp. 249–252.

[20] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[21] Aleksandr Sizov, Elie Khoury, Tomi Kinnunen, Zhizheng Wu, and Sébastien Marcel, "Joint speaker verification and antispoofing in the-vector space," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 821–832, 2015.

[22] Zheng-Hua Tan and Børge Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 798–807, 2010.