Fisher Vectors in PLDA Speaker Verification System

Zbyněk Zajíc, Marek Hrúz
University of West Bohemia
Faculty of Applied Sciences
NTIS - New Technologies for the Information Society
Univerzitní 8, 306 14 Plzeň, Czech Republic
Email: {zzajic, mhruz}@ntis.zcu.cz

***Abstract*—The goal of this paper is to examine the Fisher Vector and incorporate this vector in the PLDA based speaker verification system. The PLDA based system utilizes the Supervector of Statistics extracted from a Gaussian Mixture Model (adopted from the speaker adaptation task) to collect the information about a speaker from a dataset. We compare the efficiency of the PLDA based speaker verification system using Supervector of Statistics and the same system with Fisher vector. The experimental results of these two approaches to the verification task and the fusion of these two systems indicate that the Fisher Vector brings almost the same information to the PLDA verification process as the Supervector of Statistics when sufficient data are available.**

***Index Terms*—Speaker Verification, PLDA, Fisher Vector, Supervector, iVector.**

I. Introduction

General model based on Probabilistic Linear Discriminant Analysis (PLDA) [1] and EigenVectors (EVs) descriptors [2], used originally in image processing for face recognition, was successfully integrated into speaker verification system which is nowadays considered as a state-of-the-art approach [3]. A new method for automatic face verification utilizing Fisher Vectors (FVs) as high-dimensional descriptors was introduced in [4]. In this paper, we introduced FVs in a speaker verification system and compared it with a system that works with Gaussian Mixture Model (GMM) based supervectors of statistics [5].

Supervector is, in fact, a high-dimensional feature vector obtained by the concatenation of lower-dimensional vectors containing speaker dependent parameters - in our case the first and zeroth statistical moments of speaker data related to a Universal Background Model (UBM) based on GMM [6]. This Supervector of Statistics can be seen as the new Maximum Likelihood (ML) estimate of speaker identity and has roots in the task of speaker adaptation [7]. On the other hand, FV is based on the Fisher Information which measures the amount of information that an observable random variable $\boldsymbol{O}$ carries about an unknown parameter of a distribution that models $\boldsymbol{O}$.

This paper is organized as follows: The PLDA based speaker verification system is described in Section II, where in Subsection II-A the Gaussian Mixture Model (GMM) based Supervector of Statistics is described. In Section III the FV and the replacement of the Supervector of Statistics in the verification process is introduced. The results of two speaker verification systems based on Supervector of Statistics and the Fisher Vectors and the fusion of both system can be found in Section IV.

II. iVectors-PLDA System

The iVector-PLDA framework, a state-of-the-art system for speaker verification [8], is based on the extraction of features from the speech and accumulation of the statistics of these features into supervectors. This supervector is high-dimensional (tens of thousands), and hence it is suitable to find a latent space of a much lower dimension which represents the speakers. Vectors from this space are called iVectors. The iVectors extraction (the dimensionality reduction of the supervector) is based on Factor Analysis (FA).

Two iVectors can be compared with each other using cosine distance. However, iVectors obtained by FA still contain some noisy information not relevant to the speaker identity (e.g. influence of the channel). Therefore a PLDA model trained on a huge amount of structured data (several representations of each speaker from different sources - sessions) is used for decomposing information from iVector into the speaker and session domain. Then, only the speaker domain is used for comparison of two speaker representations. Moreover, PLDA model itself can be used as a powerful tool for identity verification instead of cosine distance [9]. A diagram of this verification system can be seen in Figure 1. Each step is described in detail in the following subsections.

A. Statistics Extracted on GMM

Supervector of Statistics containing the first and zeroth statistical moments of speakers' data related to UBM has origins in the speaker adaptation process, where these statistics are used as a descriptor of a new speaker.

First, a GMM trained on a huge amount of data from different speakers is used as a UBM and consists of a set of parameters $\boldsymbol{\lambda}_{\text{UBM}} = \{\omega_m, \boldsymbol{\mu}_m, \boldsymbol{C}_m\}_{m=1}^{M}$, where M is the number of Gaussians in the UBM, ω_m, $\boldsymbol{\mu}_m$, $\boldsymbol{C}_m$ are the weight, mean and covariance of the m^{th} Gaussian, respectively. In our case, the covariance matrix $\boldsymbol{C}_m$ is diagonal with vector σ_m on diagonal.

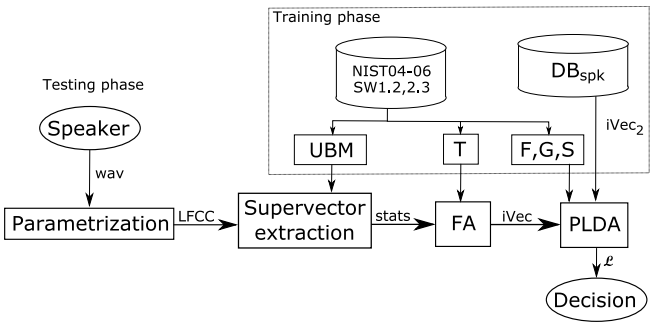

Fig. 1. Diagram of the verification process with four steps: parametrization, supervector extraction, dimensionality reduction by FA and finally verification with another speaker (represented as iVector) using PLDA model. The output from the PLDA system is likelihood ($\mathcal{L}$) that two iVectors (new constructed iVec and iVec$_2$ from databases DB$_{spk}$) representing the same speaker.

Let $\boldsymbol{O}_s = \{\boldsymbol{o}_{st}\}_{t=1}^{T_s}$ be the set of T_s feature vectors $\boldsymbol{o}_{st}$ of dimension D belonging to the s^{th} speaker, and

$$\gamma_m(\boldsymbol{o}_{st}) = \frac{\omega_m \mathcal{N}(\boldsymbol{o}_{st}; \boldsymbol{\mu}_m, \boldsymbol{C}_m)}{\sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{o}_{st}; \boldsymbol{\mu}_m, \boldsymbol{C}_m)} \tag{1}$$

be the posterior probability of m^{th} Gaussian given a feature vector $\boldsymbol{o}_{st}$. The soft count of the m^{th} Gaussian (zeroth statistical moments of feature vectors) is

$$n_m^s = \sum_{t=1}^{T_s} \gamma_m(\boldsymbol{o}_{st}), \tag{2}$$

and the sum of the first statistical moments of feature vectors with respect to the m^{th} Gaussian is

$$\boldsymbol{b}_m^s = \sum_{t=1}^{T_s} \gamma_m(\boldsymbol{o}_{st}) \boldsymbol{o}_{st}. \tag{3}$$

The speaker's supervector for given data $\boldsymbol{O}_s$ is a concatenation of the zeroth and first statistical moments of $\boldsymbol{O}_s$.

Note: the origin of these statistics can be seen if we rearrange the zeroth and first statistics into partial supervectors (of size $DM \times 1$):

$$\begin{aligned} \boldsymbol{n}_s &= \sum_{t=1}^{T_s} \left([\gamma_1(\boldsymbol{o}_{st}), \ldots, \gamma_m(\boldsymbol{o}_{st}), \ldots, \gamma_M(\boldsymbol{o}_{st})]^{\mathrm{T}} \otimes \boldsymbol{1}_D \right) \\ \boldsymbol{b}_s &= \sum_{t=1}^{T_s} \left[\gamma_1(\boldsymbol{o}_{st}) \boldsymbol{o}_{st}^{\mathrm{T}}, \ldots, \gamma_m(\boldsymbol{o}_{st}) \boldsymbol{o}_{st}^{\mathrm{T}}, \ldots, \gamma_M(\boldsymbol{o}_{st}) \boldsymbol{o}_{st}^{\mathrm{T}} \right]^{\mathrm{T}} \end{aligned} \tag{4}$$

where $\otimes$ is the Kronecker product, and $\boldsymbol{1}_D$ is a D-dimensional vector of ones. If we denote $\boldsymbol{N}_s$ a diagonal matrix containing $\boldsymbol{n}_s$ as its diagonal, then

$$\boldsymbol{m}_s = \boldsymbol{N}_s^{-1} \boldsymbol{b}_s \tag{5}$$

can be seen as a new Maximum Likelihood (ML) estimation of $\boldsymbol{m}_0 = [\boldsymbol{\mu}_1^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_m^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_M^{\mathrm{T}}]^{\mathrm{T}}$ (a supervector composed of UBM means) for given $\boldsymbol{O}_s$. The Maximum Aposteriory Probability (MAP) adaptation [10] of UBM means (according to $\boldsymbol{O}_s$) is given by

$$\boldsymbol{m}_{\text{MAP}} = \tau \boldsymbol{m}_s + (1 - \tau) \boldsymbol{m}_0, \tag{6}$$

where τ is an empirically determined factor of data relevance.

B. iVectors extraction

For iVectors extraction the Factor Analysis (FA) approach [11] (or extended Joint Factor Analysis (JFA) [3] to handle more sessions of each speaker) is used for dimensionality reduction of the supervector. The generative iVector model has the form

$$\boldsymbol{\psi}_s = \boldsymbol{m}_0 + \boldsymbol{T} \boldsymbol{w}_s + \boldsymbol{\epsilon}, \;\; \boldsymbol{w}_s \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \;\; \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{7}$$

where $\boldsymbol{T}$ (of size $D \times D_w$) is called the total variability space matrix, $\boldsymbol{w}_s$ is the s^{th} speaker's iVector of dimension D_w having standard Gaussian distribution, $\boldsymbol{m}_0$ is the mean vector of $\boldsymbol{\psi}_s$, however often the UBM's mean supervector $\boldsymbol{m}_0$ is taken instead as a good approximation, and $\boldsymbol{\epsilon}$ is some residual noise with a diagonal covariance $\boldsymbol{\Sigma}$ constructed from covariance matrices $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_m$ of the UBM ordered on the diagonal of $\boldsymbol{\Sigma}$. The iVectors are also length-normalised [12]. Details about training of total variability space matrix $\boldsymbol{T}$ can be seen in [13] or [14].

C. Probabilistic Linear Discriminant Analysis (PLDA)

In the iVector extraction phase by FA, no distinction between session space and speaker space were made (in contrast with JFA). If structured training data (more than one session - source for each speaker) are available, PLDA can be trained to model speaker and session variability separately. PLDA is a generative model [3] of the form:

$$\boldsymbol{w}_{sh} = \boldsymbol{m}_w + \boldsymbol{F} \boldsymbol{z}_s + \boldsymbol{G} \boldsymbol{r}_{sh} + \boldsymbol{\epsilon}, \; \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}) \tag{8}$$

where $\boldsymbol{m}_w$ is the mean of $\boldsymbol{w}_{sh}$, columns of $\boldsymbol{F}$ span the speaker identity space, $\boldsymbol{z}_s$ of dimension D_z are coordinates in this space and they do not change across sessions of one speaker, columns of $\boldsymbol{G}$ span the channel space, $\boldsymbol{r}_{sh}$ of dimension D_r are the session dependent speaker factors, and $\boldsymbol{\epsilon}$ is some residual noise with diagonal covariance $\boldsymbol{S}$ and a zero mean. Further restrictions are placed on distributions of latent variables $\boldsymbol{z}_s$ and $\boldsymbol{r}_{sh}$, namely that both follow a standard Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Hence, $\boldsymbol{w}_{sh} \sim \mathcal{N}(\boldsymbol{m}_w, \boldsymbol{F}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{G}\boldsymbol{G}^{\mathrm{T}} + \boldsymbol{S})$. It is a common and reasonable assumption that $D_z << D_w$ and that $D_z + D_r \approx D_w$. To train the PLDA model parameters $\boldsymbol{F}$, $\boldsymbol{G}$ and $\boldsymbol{S}$ the system of equations must be solved [15] which leads to the standard FA problem (for more details see [1]).

III. Fisher Vectors

Fisher Vectors based on Fisher Kernel [16] and Fisher information (which measures the amount of information that an observable random variable $\boldsymbol{O}$ carries about an unknown parameter of a distribution that models $\boldsymbol{O}$) was recently used in face recognition [4] as an effective encoding of the feature space structure. If we assume that $\boldsymbol{O}$ can be modeled by a

probability density function u_λ with parameter $\boldsymbol{\lambda}$, then $\boldsymbol{O}$ can be described by the gradient vector [17]:

$$\boldsymbol{G}_\lambda^O = \frac{1}{T}\nabla_\lambda \log(u_\lambda(\boldsymbol{O})), \tag{9}$$

where $\boldsymbol{G}_\lambda^O$ describes the contribution of these parameters to the generation process. A natural kernel on these gradients is

$$K(\boldsymbol{O}, \boldsymbol{Q}) = \boldsymbol{G}_\lambda^{O^T} \boldsymbol{F}_\lambda \boldsymbol{G}_\lambda^Q, \tag{10}$$

where $\boldsymbol{F}_\lambda$ is the Fisher information matrix defined as:

$$\boldsymbol{F}_\lambda = E_{o\approx u_\lambda}[\nabla_\lambda \log(u_\lambda(\boldsymbol{o}))\nabla_\lambda \log(u_\lambda(\boldsymbol{O})^T)] \tag{11}$$

and this matrix has a Cholesky decomposition $\boldsymbol{F}_\lambda = \boldsymbol{L}_\lambda^T \boldsymbol{L}_\lambda$. From this, Fisher vector of $\boldsymbol{O}$ can be defined as:

$$\boldsymbol{\phi}_\lambda^O = \boldsymbol{L}_\lambda \boldsymbol{G}_\lambda^O. \tag{12}$$

We consider gradient with respect to the parameters of the UBM $\boldsymbol{\lambda}_{UBM} = \{\omega_m, \boldsymbol{\mu}_m, \boldsymbol{C}_m\}_{m=1}^M$. FVs encoding aggregates a large set of vectors into a high-dimensional supervector representation by fitting UBM to the features $\boldsymbol{O}$ and encoding the derivatives of the log-likelihood of UBM. This representation captures the average first (and possibly second) order differences between the features and UBM components:

$$\boldsymbol{\phi}_m^s = \frac{1}{N\sqrt{\omega_m}} \sum_{t=1}^{T} \gamma_m(\boldsymbol{o}_{st}) \left(\frac{\boldsymbol{o}_{st} - \boldsymbol{\mu}_m}{\sigma_m} \right), \tag{13}$$

$$\boldsymbol{\phi}^{s^2}_m = \frac{1}{N\sqrt{2\omega_m}} \sum_{t=1}^{T} \gamma_m(\boldsymbol{o}_{st}) \left(\frac{(\boldsymbol{o}_{st} - \boldsymbol{\mu}_m)^2}{\sigma_m^2} - 1 \right), \tag{14}$$

where $\boldsymbol{\phi}^{s^2}_m$ is the average second order differences of data $\boldsymbol{O}_s$ dependent on UBM model.

FV is obtained by concatenating the differences of all UBM components into one supervector for each speaker s. In this paper, we use FV constructed only from the first-order differences (13) and soft count of occurrences (2). The goal of this paper is the comparison of the efficiency of the verification system with FVs and the system with Supervector of Statistics (where only zeroth and first moment is used in general). The dimensionality of FV is $M*(D+1)$, where M is the number of components in UBM and D is the dimensionality of the feature vector $\boldsymbol{o}_{st}$.

IV. Experiments

In this paper, we try to answer the question if the FVs can bring new information to the speaker verification system compared to the system which uses the Supervector of Statistics. The experiment was carried out on the Czech telephone corpus (cell phone or fixed line) consisting of 2005 speakers each with 2-4 min for training and 2-4 min for testing phase including the silence (which can be considered as sufficient amount of data in speaker recognition task). From all possible 4020025 trials, 10% was used for training the fusion coefficients via the linear logistic regression from the FoCal toolkit [18]. The rest was used for evaluating the verification systems.

TABLE I
Comparison of the state-of-the-art system using supervectors of statistics or Fishers vectors and a combination of these two systems. Results are given as given as EER [%] and minDCF.

system	EER	minDCF
statistics	4,69%	0.3393
fisher vectors	6.23%	0.4789
combination	4,59%	0.3345

The feature extraction was based on Linear Frequency Cepstral Coefficients (LFCCs), Hamming window of length 25 ms with 10 ms shift of the window. There are 25 triangular filter banks which are spread linearly across the frequency spectrum, and 20 LFCCs were extracted. Delta coefficients were added leading to a 40-dimensional feature vector. The Feature Warping (FW) normalization procedure was applied utilizing a sliding window of length 3 seconds. Right before the FW, the Voice Activity Detector (VAD) based on detection of energies in the filter banks located in the frequency domain was used in order to discard the non-speech frames. All the feature vectors were down-sampled by a factor of 2.

Speaker verification PLDA based system was trained using corpora: NIST SRE 2004, NIST SRE 2005, NIST SRE 2006, Switchboard 1 Release 2 and Switchboard 2 Phase 3. The number of Gaussians in the UBM was set to 512. The latent dimension (dimension of iVectors) in the the FA total variability space matrix $\boldsymbol{T}$ in the iVector extraction was set to 400. At last, the dimension of the speaker identity space in the PLDA model was set to 200 and the dimension of the session/channel space was set to 400.

A. Results

The metrics for evaluation are Equal Error Rate (EER) and the Minimum Decision Cost Function (minDCF) [19]. The results are shown as Detection Error Tradeoff (DET) curve [20] in Figure 2 and in Table I.

Recently published paper [21] reported comparison on a similar verification system on NIST2010 with different results. They obtain slightly better results for system with Fisher Vectors than supervectors of statistics and the fusion of these two systems (FV and S) brings improvement. In their paper, authors used iVector model approach to speaker verification system with Linear discriminant analysis (LDA) and Within-Class Covariance Normalization (WCCN). The Fisher Vector consist of the first and second order moments while Supervector of Statistics doesn't. In our paper we used PLDA model instead and only the first and zeroth statistical moments for supervector (in both cases: Fisher Vectors and Supervector of Statistics). We assume that the comparison of supervectors of statisctics and Fisher Vectors attribution to the verification system is more comparable if both contain the same amount of information (only zeroth and first moments).

The experimental results of these two approaches to the verification task and the fusion of these two systems indicates that the Fisher Vector brings almost the same information to

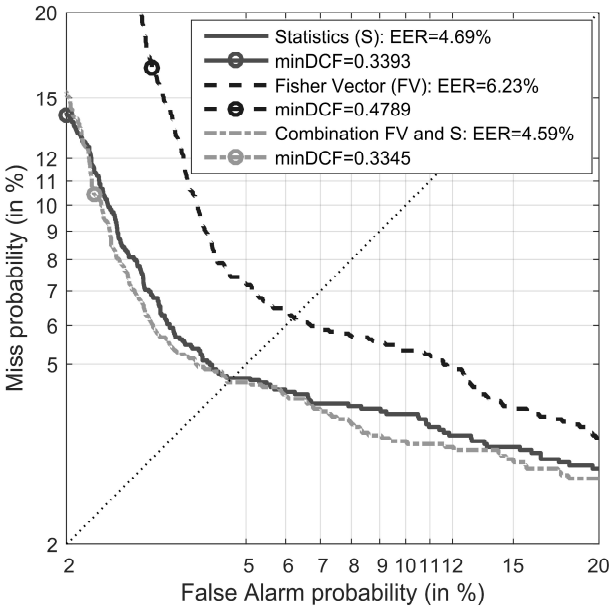

Fig. 2. DET curves of comparison systems: the system using Supervector of Statistics (S) or FV and a combination of these two systems. Circles denote points where minDCF occurred. Dotted line indicates EER.

the PLDA verification process as the Supervector of Statistics when sufficient data are available.

V. Conclusions

In this work, we compared two approaches to preserve the information about a speaker - as Supervector of Statistics and as Fisher Vector - both containing the same amount of information (zeroth and first moments). These representations were used in a state-of-the-art verification system based on PLDA. The experimental results of these two approaches show only a small difference in the EER of these systems. Moreover, the fusion of these two systems indicates that the Fisher Vector brings almost the same information to the PLDA verification process as the Supervector of Statistics. Although, when observing the shapes of the DET curves we can conclude that the fused system is a bit more robust on a larger scale of verification thresholds. Since our results are different from the ones reported recently, we assume that the difference is mainly in the use of PLDA and the amount of the information used for Fisher Vectors.

Acknowledgment

The work was supported by the Ministry of Education, Youth and Sports of the Czech Republic project No. LO1506. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

References

[1] S. J. Prince and J. H. Elder, *Probabilistic Linear Discriminant Analysis for Inferences About Identity*. IEEE 11th International Conference on Computer Vision, 1–8, Rio de Janeiro, 2007.

[2] M. Turk and A. Pentland, *Face Recognition Using Eigenfaces*. Journal of Cognitive Neuroscience 3(1), 72 – 86, 1991.

[3] P. Kenny, *Joint factor analysis of speaker and session variability: Theory and algorithms*. Tech. rep. , 2006.

[4] K. Simonyan and O. Parkhi and A. Vedaldi and A. Zisserman, *Fisher Vector Faces in the Wild*. Procedings of the British Machine Vision Conference 2013, 8.1–8.12. Bristol, 2013.

[5] W. M. Campbell and D. E. Sturim and D. A. Reynolds, *Support vector machines using GMM supervectors for speaker verification*. IEEE Signal Processing Letters 13(5), 308–311, 2006.

[6] J. Vaněk and J. Trmal and J. V. Psutka, *Optimization of the Gaussian Mixture Model Evaluation on GPU*. 12th Annual Conference of the International Speech Communication Association (INTERSPEECH), 1748–1751, Firenze, 2011.

[7] Z. Zajíc and L. Machlica and L. Müller, *Robust Statistic Estimates for Adaptation in the Task of Speech Recognition*. Lecture Notes in Computer Science, vol. 6231, 464–471, 2010.

[8] W. Rao and M. w. Mak and K. a. Lee, *Normalization of Total Variability Matrix for i-Vector/PLDA Speaker Verification*. Acoustics, Speech and Signal Processing (ICASSP), 4180–4184, 2015.

[9] L. Machlica and Z. Zajíc and L. Müller, *On Complementarity of State-of-the-art Speaker Recognition Systems*. IEEE International Symposium on Signal Processing and Information Technology, 164–169, Ho Chi Minh City, 2012.

[10] D. A. Reynolds and T. F. Quatieri and R. B. Dunn, *Speaker Verification Using Adapted Gaussian Mixture Models*. Digital Signal Processing 10(1-3), 19–41, 2000.

[11] P. Kenny and P. Dumouchel, *Experiments in Speaker Verification using Factor Analysis Likelihood Ratios*. Odyssey - Speaker and Language Recognition Workshop, 219–226, Toledo, 2004.

[12] D. Garcia-Romero and C. Y. Espy-Wilson, *Analysis of I-vector Length Normalization in Speaker Recognition Systems*. Interspeech 2011, 249–252, Florence, 2011.

[13] L. Machlica and Z. Zajíc, *Factor Analysis and Nuisance Attribute Projection Revisited*. Interspeech 2012, 1570–1573, Portland, 2012.

[14] P. Kenny and P. Ouellet and N. Dehak and V. Gupta and P. Dumouchel, *A Study of Interspeaker Variability in Speaker Verification*. IEEE Transactions on Audio, Speech, and Language Processing, vol. 16(5), 980–988, 2008.

[15] L. Machlica and Z. Zajíc, *Analysis of the Influence of Speech Corpora in the PLDA Verification in the Task of Speaker Recognition*. Lecture Notes in Computer Science, vol. 7499, 464–471, 2012.

[16] T. S. Jaakkola and D. Haussler, *Exploiting generative models in discriminative classifiers*. Proceedings of the 1998 conference on Advances in neural information processing systems II. vol. 1, 487–493, 1999.

[17] F. Perronnin and J. Sánchez and T. Mensink, *Improving the Fisher kernel for large-scale image classification*. Lecture Notes in Computer Science, vol. 6314, 143–156, 2010.

[18] N. Brummer *FoCal: Tools for fusion and calibration of automatic speaker detection systems*, 2006. https://sites.google.com/site/nikobrummer/focal

[19] N. Brummer and E. de Villiers, *The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifier Score Processing*. Tech. rep., 2011.

[20] A. Martin and G. Doddington and T. Kamm and M. Ordowski and M. Przybocki, *The DET Curve in Assessment of Detection Task Performance*. Eurospeech. vol. 4, 1899–1903. Rhodes, 1997.

[21] Y. Tian, and L. He and Z. Y. Li and W. L. Wu and W. Q. Zhang and J. Liu, *Speaker verification using Fisher vector*. In: Proceedings of the 9th International Symposium on Chinese Spoken Language Processing (ISCSLP), 419–422, Singapore, 2014.