Ensemble approach in Speaker Verification

Leibny Paola Garcia Perera[1], *Bhiksha Raj*[2], *Juan Arturo Nolazco Flores*[1]

[1]Computer Science Department, Tecnologico de Monterrey, Monterrey Nuevo Leon, Mexico
[2]Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA, USA

paola.garcia@itesm.mx, bhiksha@cs.cmu.edu, jnolazco@itesm.mx

Abstract

The speech signal is a combination of attributes that contain information of the speaker, channel and noise. Conventional speaker verification systems train a single generic model for all cases, and handle all variations from these attributes either by factor analysis, or by not considering the variations explicitly.

We propose a new methodology to partition the *data* space according to these factors and train separate models for each partition. The partitions may be obtained according to any attribute. We train models for the partitions discriminatively to maximize the separation between them. For classification we suggest multiple ways of combining scores from partitions. Experiments performed on the database NIST2008 show that our method improves the performance with respect to conventional methods when partitions are formed according to speakers. On noisy speech, partitions by noise result in the best performance.
Index Terms: Speaker verification, minimum verification error, discriminative training, robustness, clustering.

10.21437/Interspeech.2013-409

1. Introduction

Of the variety of biometric signals that can be used for identification of human subjects, speech is probably the most convenient. It is the most natural form of human communication, and can be used easily to communicate with remote applications over the network and the telephone, the infrastructure for which is ubiquitous. As a result, not only is the number of subjects who use speech to interact with machines increasing, the variety of recording conditions under which they do so is also becoming more unconstrained and diverse. It has therefore become important to develop robust voice-based biometric techniques that can operate on speech that has been recorded under diverse, and noisy conditions.

In this research we specifically consider Speaker Verification (SV), but the principles we present can also be used in speaker identification. In speaker verification, an identity is claimed for the speaker. The speech recording is used to confirm if it is true or not.

The problematic effects of noise and recording conditions on speech applications in general have been known for a long time, and a large number of solutions have been proposed to deal with them, such as denoising the signal [1], reducing the noise in speech feature vectors [2, 3, 4, 5, 6], or developing robust feature representations that are naturally less sensitive to noise [7].

In addition, speaker *verification* systems have also generally dealt with recording noise and channel variations by directly considering them in the pattern recognition paradigms they employ.

Traditionally, speaker verification is performed through a likelihood ratio test: the likelihood of the recording computed from a "target" model, representing the distribution of feature vectors from the claimed speaker, is compared to the likelihood obtained from a "background" model, representing the distribution of data from possible imposters. Robustness to noise and recording channel variations is obtained from the manner in which these distributions are learned. Both distributions are usually characterized by Gaussian mixture models (GMM). The background model is usually learned from data from a large number of speakers, collected over diverse conditions. The target model is obtained by adapting the background model to training data from the target speaker [8, 9, 10, 11]. In order to deal with noise and recording channel variations the model parameters are decomposed via factor analysis into factors that correspond to speaker identity, channel and noise [12]. Channel and noise factors are marginalized out of the computation when performing the actual likelihood ratio test [13, 14] to verify a test recording. Alternately, the factor analysis may be employed directly to eliminate undesired factors from the features computed from the speech signal themselves [15]. Other training strategies for enhancing robustness to noise and channel variations have also been proposed. For instance [16] shows that a model training scheme that minimizes the area under the detection-error-tradeoff curve obtained with the models naturally also enhances the robustness of the system to noise in the test data.

In this paper we propose an alternate framework for enhancing the robustness of the system to variations induced by different phenomena. Speaker verification systems, like all statistical pattern classifiers, work best when the statistics of the test data match those represented by the distributions in the classifier. Considering this fact, instead of training a single background model to represent all possible imposter signals, we *partition* the space of signals according to the specific factor such as noise that introduces extraneous variations. A separate background model is trained for each of these partitions. Corresponding to each partition we train a separate target model to distinguish between signals from the target speaker and the imposter data within the partition.

In order to verify the speaker in a new recording, we must first identify the partition against which to evaluate it. The likelihood test may then be performed using the target and background models for that partition. If the characteristic of the recording, in terms of the factor by which the partition is affected, is known *a priori*, this may be used to select the partition. Otherwise, the appropriate partition to use must be chosen according to some other criterion. Empirically, we find that selecting the partition that most favours acceptance of the speaker results in the best overall performance.

Since we consider an ensemble of background models, we will refer to the proposed method as the *ensemble* model approach for speaker verification.

The proposed method is related in principle to similar methods have previously been proposed in the literature to decompse the space of speech signals for improved speech recognition under adverse conditions, *e.g.* [17]. The use of multiple background models for speaker verification has also been explored in [18, 19], who attempt to segregate speakers in terms of vocal tract length, as a natural extension of gender-dependent modeling for verification. However, unlike these, in our scheme partitions are created according to any effect that introduces variations in the signal that we wish to compensate for. So while these could indeed be formed based on *intrinsic* factors such as speaker variations as in [20, 19], they may also be formed according to *extrinsic* factors such as channel or noise. Further, we attempt explicitly to make our background models *specific* to the partition they represent, by training them discriminatively to not only represent their own partition, but also to discriminate against other partitions.

The partitions, in turn may be obtained in many ways. Where the value of the factors by which we are partitioning the signals (*e.g.* the SNR of the signal) are known *a priori*, partitions may be obtained based only these values. In the absence of such information, partitions may be obtained by clustering individual recordings using methods such as k-means. Experiments show that the proposed method can result in significantly better performance than other conventional speaker verification techniques when the partitions are formed by speaker, and significantly more *robust* performance when partitions are formed according to noise conditions in the data. Moreover, the best results are obtained when partitions are based on *a priori* information about factor being considered. We note that the proposed technique does not exclude the possiblity of also including other noise-compensation techniques including many of those mentioned earlier, although we have not explicitly considered these in this paper. The interaction between these and the proposed method will be investigated in future work.

The rest of the paper is as follows. In Section 2 we outline the general problem of SV and the state of the art solution. In Section 3 we describe our ensemble model of paritioning the signal space for speaker verfication. Section 4 explains how we adapt the approach for partitioning the space by speaker and by noise. Section 5 shows some experiments and results. Finally, Section 6 presents our conclusions and discusses directions for future work.

2. Speaker Verification

The objective of speaker verification is to accurately verify if a recorded spoken phrase χ was indeed uttered by the registered speaker S that the recording is claimed to be from. This is generally performed through a likelihood ratio test. A parametric model with parameters Λ_S is defined for the distribution of data from the target speaker S. A *background* model with parameters $\lambda_{\hat{S}}$ is specified for the class of recordings that do *not* belong to the target speaker. The actual likelihood-ratio is as follows:

$$\theta_S(\chi) = \log\left(P(\chi;\Lambda_S)\right) - \log\left(P(\chi;\Lambda_{\hat{S}})\right)$$
$$\text{accept speaker} \quad \text{if } \theta_S(\chi) > \tau \qquad (1)$$
$$\text{reject speaker} \quad \text{otherwise}$$

In order to compute $P(\chi;\Lambda_S)$ and $P(\chi;\Lambda_{\hat{S}})$, each recording χ is transformed into a sequence of feature vectors $\chi = X_1, X_2, \cdots, X_T$, typically mel-frequency cepstral coefficient vectors, augmented by their delta (velocity) and double delta (acceleration) coefficients. The vectors X_i are assumed to be IID and have a Gaussian mixture distribution given by, $P(X;\Lambda_C) = \sum_k w_k^C \mathcal{N}(X,\mu_k^C,\Sigma^C,k)$, where C is either S or $\hat{S}$, and w_k^C, μ_k^C and Σ_k^C are the mixture weight, mean and covariance (usually assumed to be a diagonal matrix) of the k^{th} Gaussian in the mixture, *i.e.* $\Lambda_C = \{w_k^C, \mu_k^C, \Sigma_k^C \ \ \forall k\}$. The likelihood of the complete recording χ is computed as $P(\chi;\Lambda_C) = \prod_t P(X_t;\Lambda_C)$.

Most commonly, the background model is a single, usually gender-specific "universal background model" trained from a large collection of recordings from many speakers. Alternately, a separate model may be trained for each of a set of "cohort" speakers who are chosen to provide the best contrast to the target (registered) speaker. In this case the likelihoods computed using the multiple cohort models must be combined [21, 22, 23] to produce the overall likelihood score for the background in Equation 1.

The *target* model Λ_S must be trained from example recordings of the target speaker. Typially, the amount of data available to compute the target speaker models is small. So, the background model is adapted to the target speaker by through MAP (maximum a posteriori) adaptation [8] to obtain target models. When noise or recording channel mismatches are expected between test and enrollment data for the speaker, the state of the art uses various forms of *factor analysis* (JFA) [9, 10] to train Λ_S. This approach decomposes the parameters of the distribution into two sets of factors – one representing the speaker and the second representing extraneous factors. The extraneous factors, which contain no information about the speaker, are marginalized out when performing the likelihood ratio test.

3. Ensemble Model for Robust Verification

We now describe our proposed variation to the basic likelihood-ratio test based framework described above. Instead of estimating a general background model from all available data, we *partition* the space of background signals according to a variation-inducing factor we wish to account for. We then train a separate *specific* background model for each of the partitions, and corresponding to each of the background models we train a target model for the speaker. When a new recording arrives, we must choose the appropriate partition, and use the corresponding background and target models for the likelihood ratio test. Since the approach effectively employs an *ensemble* of background models, we refer to it as the *Ensemble model* approach.

Below we outline our procedure for training background models for each of the steps of our procedure.

We begin by assuming that we have a large collection of recordings from which to train the background models. We assume that the space of signals is partitioned into P non-overlapping regions $\Omega_1, \Omega_2, \cdots, \Omega_P$, such that together these partitions cover the entire space. Correspondingly we assume that the recordings are clustered into P groups, each corresponding to signals that fall into one of these partitions. We defer the description of exactly *how* these partitions are obtained until the next section. For now, we will assume that these are available.

3.1. Training the Model Ensemble

Corresponding to each of the partitions $\Omega_1, \cdots, \Omega_P$ we train a separate partition specific background model. All background models are Gaussian mixtures. In principle these can be trained separately for each partition using the Expectation Maximization algorithm. However, we require each of the background models to be highly *specific* to the partition they represent, and not generalize to other partitions. In order to do so, we train

all of them together using the following *discriminative* training procedure [17].

Let Λ_C represent the model for a partition Ω_C. Let χ_C represent all (training) recordings assigned to Ω_C. For any partition Ω_C, let $\Omega_{\bar{C}} = \bigcup_{C' \neq C} \Omega_{C'}$ represent the *complement* of Ω_C, *i.e.* the union of all partitions that are not Ω_C.

Let $g(X;\Lambda_C) = \log P(X;\Lambda_C)$ represent the log-likelihood of any recording X computed with the distribution for partition Ω_C. We can now define $d(X,\Omega_C)$, a misclassification measure for how likely it is that a data $X \in \chi_C$ from Ω_C will be misclassified as belonging to $\Omega_{\bar{C}}$ as

$$d(X,\Omega_C) = -g(X;\Lambda_C) + G(X,\Omega_{\bar{C}}), \tag{2}$$

$G(X,\Lambda_{\bar{C}})$ represents the combined score obtained from the models for partitions in $\Omega_{\bar{C}}$.

$$G(X,\Omega_{\bar{C}}) = \log\left\{\frac{1}{|\Omega_{\bar{C}}|}\sum_{C':\Omega_{C'}\in\Omega_{\bar{C}}}\exp\left[\eta g(X,\Lambda_{C'})\right]\right\}^{\frac{1}{\eta}}. \tag{3}$$

where $|\Omega_{\bar{C}}|$ is the number of partitions included in $\Omega_{\bar{C}}$, and η is a positive parameter. Now, we can define a new objective function for discriminative training of Λ_C. This function takes the of the form,

$$\ell(\Lambda_C) = \frac{1}{|\chi_C|}\sum_{X\in\chi_C}\frac{1}{1+\exp\left[-\gamma\left(d(X,\Omega_C)+\theta\right)\right]} \tag{4}$$

where $|\chi_C|$ represents the number of recordings in χ_C, and γ and θ are control parameters. Finally, the objective function in Equation 4 can be optimized by applying the following generalized probabilistic descent (GPD) update rule for Λ_C: $\Lambda_C^{t+1} = \Lambda_C^t - \epsilon\nabla\ell(\Lambda_C)|_{\Lambda_C^t}$.

Since all background models are GMMs, $\Lambda_C = \{w_k^C, \mu_k^C, \Sigma_k^C\}$, where w_k^C, μ_k^C and Σ_k^C are the mixture weight, mean and covariance matrix of the k-th Gaussian of the GMM for Λ_C. To obtain the update rules for these individual parameters, $\frac{\partial\ell(\Lambda_C)}{\partial w_k^C}$, $\frac{\partial\ell(\Lambda_C)}{\partial \mu_k^C}$ and $\frac{\partial\ell(\Lambda_C)}{\partial \Sigma_k^C}$ must respectively be plugged in for $\nabla\ell(\Lambda_C)$ in the update rule of Equation 3.1.

Once background models Λ_C are obtained for all partitions, we can also train *partition-specific* target-speaker models, Λ_S^C by fixing the background model Λ_C and using the same discriminative approach as above to train Λ_S^C.

3.2. Scoring for classification

Given $\{\Lambda_{C_1}, \Lambda_S^{C_1}\}, \{\Lambda_{C_2}, \Lambda_S^{C_2}\}, \cdots \{\Lambda_{C_P}, \Lambda_S^{C_P}\}$, the set of background models for all P partitions and their corresponding partition-specific target speaker models for any claimed speaker S, we can compute the score $\theta^S(X)$ to be employed in the likelihood ratio test for any recording X in one of several ways. Let $\theta_C^S(X) = \log P(X|\Lambda_S^C) - \log P(X|\Lambda_C)$ be the likelihood ratio computed in the log domain from the models for partition Ω_C. The options for obtaining the final score $\theta^S(X)$ are:

a) *Partition Selection*: We first assign the recording to the most likely partition: $\hat{C}(X) = \arg\max_C \log P(X|\Lambda_C)$. We then compute the score from the assigned partition: $\theta^S(X) = \theta^S_{\hat{C}(X)}$. This is a conservative score that selects the partition with signals most likely to be confused with the target speaker.

b) *A priori*: If the correct partition Ω_C for X is known *a priori*, then we can simply set $\theta^S(X) = \theta_C^S(X)$.

c) *Best score*: We select the largest score: $\theta^S(X) = \max_C \theta_C^S(X)$

d) *Combination*: Here we simply combine the scores from the different partitions: $\theta^S(X) = \sum_C w_C^S \theta_C^S(X)$. In our work we trained an SVM to classify the speaker; in this case w_C^S are simply the weights assigned by the SVM. This is equivalent to learning the weights discriminatively.

3.3. Finding the partitions

We now return to the problem of how to obtain the partitions $\Omega_1, \cdots, \Omega_P$ of signal space, in the first place.

3.3.1. Supervised

When *a priori* knowledge of the factors by which we wish to partition the space is available for the background-model training data, it may be used to obtain the partitions. Below we consider two mechanisms: partition by *noise* and partition by *speaker*. Partitions may similarly be obtained by other factors such as channel variations. Hierarchical partitioning strategies that consider multiple factors concurrently may also be used.

Environment-based partitions: Here we partition the space of signals by noise in them. In this paper we assume that partitions are formed based on the SNR of signals. We divide the range of all possible SNR values into P intervals. Each interval represents a partition of the signal space. Let SNR_{min}^C and SNR_{max}^C represent the minimum and maximum SNR associated with partition Ω_C. A signal X with SNR SNR_X is assigned to a partition C such that $SNR_{min}^C < SNR_X \leq SNR_{max}^C$. Note that partitions may also be formed based on noise type, or other known characteristics of the noise. In this paper, however, we have only considered SNR.

Speaker Partitions: When speaker identity is known for all recordings in the training set, partitions are obtained by clustering them by speaker. We first compute a universal background model (UBM) from unpartitioned data. We then use an agglomerative clustering procedure to cluster speakers. Initially each speaker forms their own cluster. At each stage of the clustering, the UBM is adapted via MAP adaptation to learn a model Λ_C for each new cluster C. The distance between any two clusters is the empirical cross entropy: $d(C_1,C_2) = \frac{1}{|\chi_{C_1}|}\log\frac{P(\chi_{C_1};\Lambda_{C_1})}{P(\chi_{C_1};\Lambda_{C_2})} + \frac{1}{|\chi_{C_2}|}\log\frac{P(\chi_{C_2};\Lambda_{C_2})}{P(\chi_{C_2};\Lambda_{C_1})}$, where χ_{C_i} is the set of all recordings in cluster C_i. Agglomerative clustering iteratively merges the closest clusters until the desired number of clusters (and consequently, partitions) is obtained. Other clustering mechanisms may also be employed.

3.3.2. Unsupervised

When *a priori* knowledge about the training recordings is unavailable, partitions may be formed by clustering them using unsupervised methods, *e.g.* k-means. The factor by which partitions are formed can be controlled by using an appropriate distance function. Generic clustering based on Euclidean distances or likelihoods will cluster the data by the dominant factor.

4. Experiments and Results

We employed the NIST Speaker Evaluation 2004, 2005, 2010 and 2008 database [24] to complete this study. We followed the evaluation rules (*e.g.* not considering subjects as imposters for other subjects). 49-dimensional feature vectors, including delta and double-delta terms, were extracted from the audio using a 25ms analysis window with a 10ms frame shift. We included a frame removal criterion to eliminate low energy frames that do not provide information about the identity of the person.

Condition	EER						minDCF					
Baseline MAP	16.9						6.9					
Baseline JFA	15.2						6.3					
Baseline MVE	16.3						6.5					
	4-clusters				8-clusters				16-clusters			
Condition	A	B	C	D	A	B	C	D	A	B	C	D
k-means	16.2	15.5	17.1	14.9	15.2	13.8	16.0	13.2	14.8	13.2	15.4	12.5
minDCF	6.45	6.35	6.65	6.23	6.28	5.85	6.37	5.74	6.10	5.32	6.36	5.71
k-B	14.9	14.4	15.4	13.7	13.5	11.9	13.9	11.3	12.1	10.9	12.8	9.5
minDCF	6.24	5.93	6.34	5.67	5.77	5.58	5.92	5.21	5.61	5.08	5.62	4.83

Table 1: EER and minDCF for different clusters on clean condition (speaker ensemble).

	5c-mve				5c-map				5c-jfa			
Condition	A	B	C	D	A	B	C	D	A	B	C	D
Baseline	28.3				28.6				25.0			
minDCF	15.32				15.60				13.23			
k-means	27.0	24.1	27.7	23.1	28.3	25.8	29.5	24.2	26.0	25.1	26.8	24.5
minDCF	14.21	10.48	14.34	8.52	16.89	13.48	17.43	10.57	13.70	13.22	14.17	10.66
k-B	23.7	22.2	24.4	20.9	25.2	23.4	26.4	22.3	24.8	24.2	25.2	23.7
minDCF	9.21	8.35	10.59	7.78	13.34	8.97	13.68	8.45	9.35	9.16	13.28	9.32

Table 2: EER and minDCF for different clusters on a noisy task (babble noise).

4.1. Experiment Setup

We conducted two experiments, one clean data and the second on noise-corrupted data. For all experiments, we used 100 male registered users chosen randomly from the NIST2008 SRE database as targets. Following NIST2008 Evaluation rules, the probability of being a target, P_{target}, is 0.01 and the probability of being a impostor, $P_{impostor}$, is 0.99 in this data set. For the noise condition experiments, babble noise, extracted from the Aurora 2 database, was added to the training and test files at different SNRS: 0,5,10,15 and 20 dB. The training data were randomly partitioned into equal parts, one for each noise condition; the same procedure was applied for the test set.

The objective in the clean experiment was to partition the space by speaker. For the noisy data, we evaluated partitioning by noise condition. In both cases we evaluated partitions formed from *a priori* information about the data (labelled as "K-B" in the tables), as well as by unsupervised k-means clustering. The four methods for partition selection: *partition selection*, *a priori*, *best score* and *combination*, labelled A,B,C and D respectively in the tables, were evaluated. We used a 256-Gaussian GMM for background models in all cases.

In the clean experiment we evaluated the performance obtained with speaker-based partitions with different numbers of partitions. As a baseline we also evaluated the conventional verification framework, where we trained generic gender-dependent background models (UBM) and adapted these to the target speaker using MAP [8], Joint Factor Analysis (JFA) [9, 10] and Minimum Verification Error (MVE) [25] training. The UBM did not include data from any target speaker. The code for JFA was obtained from the implementation by [26].

For the noise experiment we employed 5 partitions, one corresponding to each SNR level. The proposed method only establishes a mechanism for defining background models. The method described in Section 3.1 actually uses MVE to learn partition-specific target models. Partition-specific target models may also be trained via JFA or MAP. For the noise experiments these were also evaluated in addition to UBM-based baselines.

4.2. Results

Table 1 shows the results for the clean experiments with different numbers of partitions. The EER and minDFC (minimum detection cost function) results are shown. The ensemble model improves performance in every case. Morever, partitions based on *a priori* knoweldge consistently outperform partitions obtained from unsupervised clustering. The best result is obtained when scores from the partitions are combined; this result significantly outperforms JFA, and is the best result we have ever obtained on this test set. Increasing the number of partitions beyond 16 resulted in degradation of performance. Discriminative training of background models to make them specific also turns out to be key. The alternative is to simply train each of them using maximum likelihood. This was consistently worse than discriminative refinement of partition-specific background models in all experiments.

Table 2 shows results for noisy data. Once again, the ensemble method results in improvements in every case. Supervised partitions obtained from *a priori* knoweldge of SNR result in the best performance. Interestingly, identifying the partition for a test utterance through *a priori* knowledge of its SNR did not result in the best performance, although it does outperform other methods of selecting partitions. The best results are obtained, once again, by combining scores from the partitions. Moreover the best results are obtained by MVE, rather than JFA. This is an inversion with respect usual results with generic models, where JFA consistently outperforms MVE training.

5. Conclusions

We have proposed an *Ensemble* approach to partition the signal space to perform speaker verification. We find that partitioning and refining the space by speaker or environment both improve performance over the baseline significantly. The primary drawback is that the best results are obtained when *a priori* knoweldge of the factor considered in partitioning is available. We conjecture that much of this benefit may be obtained if this information is only *estimated* for the training data. We will investigate this in future work. We also propose to investigate methods for identifying partitions when *multiple* factors must be considered concurrently, nad particularly when they must be estimated. We will also investigate methods for formally optimizing the partitions for verification.

6. Acknowledgements

Thanks to Catedra de Seguridad de la Informacion, Tecnologico de Monterrey Campus Monterrey

7. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, 1979.

[2] P. J. Moreno, B. Raj, and R. M. Stern, "Data-driven environmental compensation for speech recognition: A unified approach," *Speech Communication*, vol. 24, no. 4, pp. 267–285, 1998.

[3] J. Dropo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database (web update)," in *Eurospeech*, vol. 2, Aalborg, Denmark, 2001, pp. 217–220.

[4] L. Deng, A. Acero, J. L. Droppo, and J. X. Huang, "High-performance robust speech recognition using stereo training data," in *Proc. ICASSP*, Salt Lake City, UT, May 2001, pp. 301–304.

[5] A. Moreno-Daniel, J. A. Nolazco-Flores, T. Wada, and B.-H. Juang, "Acoustic model enhancement: An adaptation technique for speaker verification under noisy environments," in *Proc. ICASSP*, Honolulu, USA, April 2007.

[6] J. Nolazco-Flores and S. Young, "Continuous speech recognition in noise using spectral subtraction and hmm adaptation," in *Proc. ICASSP*, Adelaide, Australia, 1994, pp. 409–412.

[7] C. Hanilçi, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, and F. Ertas, "Regularization of all-pole models for speaker verification under additive noise," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[8] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291–299, Apr. 1994.

[9] P. Kenny, G. Boulianne, P. Ouellet, and P. 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.

[10] P. Kenny, P. Oueleet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. ASLP*, vol. 16, pp. 980–988, 2008.

[11] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[12] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot *et al.*, "Promoting robustness for speaker modeling in the community: the prism evaluation set," in *Proceedings of SRE11 Analysis Workshop*, 2011.

[13] F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, and D. Petrovska-Delacretaz, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, 2004.

[14] D. Petrovska-Delacrétaz, A. El Hannani, and G. Chollet, "Text-independent speaker verification: state of the art and challenges," *Progress in nonlinear speech processing*, pp. 135–169, 2007.

[15] T. Hasan and J. Hansen, "Integrated feature normalization and enhancement for robust speaker recognition using acoustic factor analysis," in *Proc. Interspeech*, 2012.

[16] L. P. Garcia-Perera, J. A. Nolazco-Flores, B. Raj, and R. Stern, "Optimization of the det curve in speaker verification," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 318–323.

[17] Y. Tsao and C.-H. Lee, "An ensemble speaker and speaking environment modeling approach to robust speech recognition," *IEEE transactions on Speech*, vol. 43, pp. 781–785, August 1994.

[18] W.-Q. Zhang, Y. Shan, and J. Liu, "Multiple background models for speaker verification," *Odyssey 2010, Brno*, 2010.

[19] A. Sarkar and S. Umesh, "Multiple background models for speaker verification using the concept of vocal tract length and mllr super-vector," *International Journal of Speech Technology*, pp. 1–14, 2012.

[20] Y. Zhang and J. R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4366–4369.

[21] A. Brew and P. Cunningham, "Combining cohort and ubm models in open set speaker detection," *Multimedia Tools and Applications*, vol. 48, no. 1, pp. 141–159, 2010.

[22] T. Isobe and J. Takahashi, "A new cohort normalization using local acoustic information for speaker verification," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 2. IEEE, 1999, pp. 841–844.

[23] T. Kinnunen, E. Karpov, and P. Fränti, "Efficient online cohort selection method for speaker verification," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2004)*, vol. 3, 2004, pp. 2401–2402.

[24] A. Martin and C. Greenberg, "NIST 2008 Speaker Recognition Evaluation: Performance Across Telephone and Room Microphone Channels," in *Proc. Interspeech*, 2009.

[25] A. E. Rosenberg, O. Siohan, and S. Parthasarathy, "Speaker verification using minimum verification error training," *Proc. ICASSP*, pp. 105–108, 1998.

[26] L. Burget, M. Fapso, and V. Hubeika, "BUT system for NIST 2008 speaker recognition evaluation," in *Interspeech*, 2009.