ROBUST TEXT-INDEPENDENT SPEAKER IDENTIFICATION USING SHORT TEST AND TRAINING SESSIONS

Christos Tzagkarakis and Athanasios Mouchtaris

Department of Computer Science, University of Crete and
Institute of Computer Science (FORTH-ICS)
Foundation for Research and Technology - Hellas
Heraklion, Crete, Greece
{tzagarak, mouchtar}@ics.forth.gr

ABSTRACT

In this paper two methods for noise-robust text-independent speaker identification are described and compared against a baseline method for speaker identification based on the Gaussian Mixture Model (GMM). The two methods proposed in this paper are: (a) a statistical approach based on the Generalized Gaussian Density (GGD), and (b) a Sparse Representation Classification (SRC) method. The performance evaluation of each method is examined in a database containing twelve speakers. The main contribution of the paper is to investigate whether the SRC and GGD approaches can achieve robust speaker identification performance under noisy conditions using short duration testing and training data, in relevance to the baseline method. Our simulations indicate that the SRC approach significantly outperforms the other two methods under the short test and training sessions restriction, for all the signal-to-noise ratios (SNR) cases that were examined.

1. INTRODUCTION

Speaker recognition systems are essential in a variety of security and commercial applications, such as information retrieval, control of financial transactions, control of entrance into safe or reserved areas and buildings, *etc*. [1]. Speaker recognition can be based on both the separate or combined use of several biometric features [2] (voice, face, fingerprints, *etc*.). In our study, we focus on speaker recognition using only voice patterns.

Speaker recognition can be categorized into speaker verification and speaker identification. In speaker verification, a speaker claims to be of a certain identity and his/her voice is used to verify this claim. On the other hand, speaker identification is the task of determining an unknown speaker's identity. Generally speaking, speaker verification is a one-to-one match where one speaker's voice is matched to one template ("voice print" or "voice model") whereas speaker identification is a one-to-*N* match where the voice is compared against *N* templates. Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former require the speaker to provide utterances of keywords or sentences, the same text being used for both training and recognition. In text-independent recognition, the decision does not rely on a specific text being spoken. In our study we focus on text-independent speaker identification.

In order to correctly identify a person, each speaker in the database is usually assigned a specific speaker model consistently describing the extracted speech features. During the identification process, the system returns the speaker's identity based on the closest matching of the test utterance against all speaker models. This procedure has proven to be effective under acoustic conditions in matched training and testing [3]. However, in practical applications where speech signals are corrupted by noise due to either the environment in which the speaker is present (*e.g.* the user is crossing a busy street) or due to the voice transmission medium (*e.g.* the user is speaking through a cell-phone), robust identification is a challenging problem.

The most popular approach for speaker identification is based on Gaussian Mixture Models (GMM) [3] (a brief description is given in Section 2.1). Other classifiers such as Support Vector Machines (SVM) [4] have also been used for this task. Recently, the focus of the speaker recognition research community has been given both on the study of features that are more robust in noise environments and on finding more robust and efficient identification algorithms. Specifically, in [5] robust features based on mel-scale frequency cepstral coefficients (MFCCs [6]) are proposed, in combination with a projection measure technique for speaker identification. In [7], the speech features are based on a harmonic decomposition of the signal where a reliable frame weighting method is adopted for noise compensation. In [8], the descriptors introduced are based on the AM-FM representation of the speech signal, while in [9] the proposed features are derived from auditory filtering and cepstral analysis (in both cases a GMM is used to model the feature space). In [10, 11] the noise robust speaker identification problem under mismatched testing and training conditions is studied. In [10], the identification is performed in the space of adapted GMMs where Bhattacharyya shape is used to measure the closeness of speaker models, while in [11] a multicondition model training and missing feature theory is adopted to deal with the training and testing mismatch, where this model is incorporated into a GMM for noise robust speaker identification.

An important aspect in speaker identification is that in real-time applications the system should be able to respond within a short time duration about the identity of the speaker. However, when the number of the enrolled speakers in the database grows significantly, it is quite difficult for the system to quickly assign the speaker with a specific identity. For addressing such real-time efficiency concerns, in [12] a method based on approximating GMM likelihood scoring with an approximated cross entropy is proposed. In [13], the GMM-based speaker models are clustered using a *k*-means algorithm so as to select only a small proportion of speaker models used in likelihood computations. These approaches achieve a more efficient operation compared to state-of-the-art, without degrading the identification performance in large population databases.

In this paper, we study the problem of noise-robust text-independent speaker identification under the assumption of short testing and training sessions. There are two reasons for following this approach: (i) it is often not feasible to have large amounts of training data from all the speakers and (ii) in order to speed up the identification process, the testing data (*i.e.*, the speaker utterance to be identified) should be as short as possible. Towards this direction, two methods are proposed and tested under noisy conditions (additive white Gaussian noise), and compared to a baseline GMM method [3]. The first approach is adopted from the music classification task [14], while the second method is based on sparse representation classification which was recently proposed and applied on robust face recognition [15].

This work has been funded in part by the PEOPLE-IAPP "AVID-MODE" grant, within the 7th European Community Framework Program, and in part by the University of Crete Research Committee, Grant KA2739.

2. CLASSIFICATION METHODS

In the current section a brief description of the methods used to perform the identification process is given. For the feature extraction task it is assumed that the speech signal/utterance is segmented into overlapping frames. In this paper we use the MFCC features [6].

2.1 Gaussian Mixture Model

Gaussian Mixture Models (GMMs) have been applied with great success in the text-independent speaker identification problem [3]. The approach is to model the probability density function (PDF) of the feature space of each speaker in the dataset (training phase) as a sum of Gaussian functions, and then use the maximum a-posteriori rule to identify the speaker. A Gaussian mixture density is a weighted sum of M multidimensional Gaussian densities, where the mixture density can be represented as

$$\lambda_i = \left\{ p_m^i, \mu_m^i, \mathbf{\Sigma}_m^i \right\}, \; m = 1, \ldots, M, \tag{1}$$

where for the i^{th} speaker, p_m^i is the weight of the m^{th} mixture (prior probability), μ_m^i is the corresponding mean vector, $\mathbf{\Sigma}_m^i$ is the covariance matrix, and M is the total number of Gaussian mixtures. Each speaker is represented by a GMM and the corresponding model λ, whose parameters are computed via the Expectation-Maximization (EM) algorithm applied on the training features. For the speaker identification task (testing phase), the estimated speaker identity (speaker index) is obtained based on the maximum a-posteriori probability for a given sequence of observations as follows

$$S_q = \arg \max_{1 \leq i \leq S} p(\lambda_i | X) = \arg \max_{1 \leq i \leq S} \frac{p(X|\lambda_i) p(\lambda_i)}{p(X)}. \tag{2}$$

In the above equation, $X = [\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_n]$ denotes the sequence of n feature vectors, and S is the total number of speakers. For equally likely speakers and since $p(X)$ is the same for all speaker models the above equation becomes

$$S_q = \arg \max_{1 \leq i \leq S} p(X|\lambda_i). \tag{3}$$

For independent observations and using logarithms, the identification criterion becomes

$$S_q = \arg \max_{1 \leq i \leq S} \sum_{t=1}^{n} \log p(\mathbf{x}_t|\lambda_i), \tag{4}$$

where

$$p(\mathbf{x}_t|\lambda_i) = \sum_{m=1}^{M} \frac{p_m^i}{(2\pi)^{K/2} |\mathbf{\Sigma}_m^i|^{1/2}} \exp\Big\{ -\frac{1}{2} (\mathbf{x}_t - \mu_m^i)^T {\mathbf{\Sigma}_m^i}^{-1} (\mathbf{x}_t - \mu_m^i) \Big\},$$

K being the dimension of each feature vector.

2.2 Statistical Modeling based on Generalized Gaussian Density

In this subsection, we briefly describe a statistical approach which treats the speaker identification problem as a multiple hypothesis problem. We previously proposed this approach within the context of music genre classification in [14], and in this paper we are interested to test its applicability for the speaker identification task.

Let us assume that there are S speakers and that we have represented the speaker to be identified as S_q, given a set X of feature vectors $\mathbf{x}_t = (x_1, x_2, ..., x_K)^T$. Each speaker is assigned a hypothesis H_i. The goal is to select one hypothesis out of S, which best describes the data from S_q. Under the common assumption of equal prior probabilities of the hypotheses, the optimal rule resulting in the minimum probability of classification error is to select the hypothesis with the highest likelihood among the S. Thus, S_q is assigned to the speaker corresponding to the hypothesis H_j if

$$p(\mathbf{x}_t|H_j) \geq p(\mathbf{x}_t|H_i), \; i \neq j, \forall \; i = 1, ..., S. \tag{5}$$

For solving this problem, a parametric approach is adopted where each conditional probability density $p(\mathbf{x}|H_i)$ is modeled by a member of a family of PDFs, denoted by $p(\mathbf{x}; \theta_i)$ where θ_i is a set of model parameters. Under this assumption, the extracted features for the i^{th} speaker are represented by the estimated model parameter $\hat{\theta}_i$, computed in the feature extraction stage. For assigning S_q to the closest speaker identity:

1. Compute the Kullback-Leibler Divergence (KLD) between the density of the speaker to be identified $p(\mathbf{x}; \theta_q)$ and the density $p(\mathbf{x}; \theta_i)$ associated with the i^{th} speaker identity in the database, $\forall i = 1, \ldots, S$:

$$D(p(\mathbf{x}; \theta_q) \| p(\mathbf{x}; \theta_i)) = \int p(\mathbf{x}; \theta_q) \log \frac{p(\mathbf{x}; \theta_q)}{p(\mathbf{x}; \theta_i)} d\mathbf{x}. \tag{6}$$

2. Assign S_q to the identity corresponding to the smallest value of the KLD.

A chain rule holds for the KLD and is applied in order to combine the KLDs from multiple data sets or dataset dimensions. This rule states that the KLD between two joint PDFs, $p(\mathbf{X}, \mathbf{Y})$ and $q(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}, \mathbf{Y}$ are assumed to be independent data sets, is given by

$$D(p(\mathbf{X}, \mathbf{Y}) \| q(\mathbf{X}, \mathbf{Y})) = D(p(\mathbf{X}) \| q(\mathbf{X})) + D(p(\mathbf{Y}) \| q(\mathbf{Y})). \tag{7}$$

The proposed method is based on fitting a Generalized Gaussian Density (GGD) on the PDF of the data set (features). In fact, independence among MFCC vector components is assumed, thus a GGD for each scalar component is estimated. This task can be achieved by estimating the two parameters of the GGD (α, β), which is defined as

$$p(x; \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|x|/\alpha)^\beta}, \tag{8}$$

where $\Gamma(\cdot)$ is the Gamma function, and the GGD parameters are computed using Maximum Likelihood (ML) estimation. Substitution of (8) into (6) gives the following closed form for the KLD between two GGDs

$$D(p_{\alpha_1,\beta_1} || p_{\alpha_2,\beta_2}) = \log \Big(\frac{\beta_1 \alpha_2 \Gamma(1/\beta_2)}{\beta_2 \alpha_1 \Gamma(1/\beta_1)} \Big) + \Big(\frac{\alpha_1}{\alpha_2} \Big)^{\beta_2} \frac{\Gamma(\frac{\beta_2+1}{\beta_1})}{\Gamma(\frac{1}{\beta_1})} - \frac{1}{\beta_1}. \tag{9}$$

Based on the independence assumption for the MFCC coefficients, (7) yields the following expression for the overall normalized distance between two test utterances U_1, U_2

$$D(U_1 \| U_2) = \frac{1}{K} \sum_{k=1}^{K} D(p_{U_1,k} \| p_{U_2,k}), \tag{10}$$

where K is the order of the MFCCs (dimension of a feature vector).

2.3 Sparse Representation Classification

The approach of classification based on sparse representation is described in this subsection. This approach was initially applied in face recognition in [15], and is first applied in speaker identification in this paper.

Let us assume that the n_i training samples corresponding to the feature vectors of the i^{th} speaker are arranged as columns of a matrix

$$\mathbf{V}_i = [\mathbf{v}_{i,1} | \mathbf{v}_{i,2} | \ldots | \mathbf{v}_{i,n_i}] \in \mathbb{R}^{K \times n_i}, \tag{11}$$

where K is the dimension of each (column) feature vector. Given a new test sample (feature vector) $\mathbf{x}_t \in \mathbb{R}^K$ that belongs to the i^{th} class, $\mathbf{x}_t$ can be expressed as a linear combination of the training samples associated with class i

$$\mathbf{x}_t = c_{i,1}\mathbf{v}_{i,1} + c_{i,2}\mathbf{v}_{i,2} + \ldots + c_{i,n_i}\mathbf{v}_{i,n_i} = \mathbf{V}_i\mathbf{c}_i, \tag{12}$$

where $c_{i,j} \in \mathbb{R}$ are scalars. Let us also define the matrix $\mathbf{V}$ for the entire training set as the concatenation of the $N = n_1 + \ldots + n_S$ training samples of all S classes (speakers):

$$\mathbf{V} = [\mathbf{V}_1|\mathbf{V}_2|\ldots|\mathbf{V}_S] = [\mathbf{v}_{1,1}|\mathbf{v}_{1,2}|\ldots|\mathbf{v}_{i,j}|\ldots|\mathbf{v}_{S,n_S}]. \tag{13}$$

The linear representation of $\mathbf{x}_t$ can be rewritten as $\mathbf{x}_t = \mathbf{V}\mathbf{c}$, where

$$\mathbf{c} = [0,\ldots,0,c_{i,1},c_{i,2},\ldots,c_{i,n_i},0,\ldots,0]^T \in \mathbb{R}^N, \tag{14}$$

is a coefficient vector whose elements are zero except those associated with the i^{th} class. As a result, if S is large enough, $\mathbf{c}$ will be sufficient sparse. This observation motivates us to solve the following optimization problem for a sparse solution

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \|\mathbf{c}\|_0, \text{ s.t. } \mathbf{x}_t = \mathbf{V}\mathbf{c}, \tag{15}$$

where $\|\cdot\|_0$ denotes the ℓ_0 norm, which counts the number of non-zero elements in a vector. The optimization problem in (15) is an NP-hard problem. However, an approximate solution can be obtained if the ℓ_0 norm is substituted by the ℓ_1 norm as follows

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \|\mathbf{c}\|_1, \text{ s.t. } \mathbf{x}_t = \mathbf{V}\mathbf{c}, \tag{16}$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm of a vector. The efficient solution of the optimization problem in (16) has been studied extensively. Orthogonal Matching Pursuit (OMP) [16] is a popular solution to this problem, and this method is used in our simulations.

In the ideal case, the non-zero entries in $\hat{\mathbf{c}}$ will be associated with the columns of matrix $\mathbf{V}$ from a single class i, and the test sample will be assigned to that class. However, because of modeling errors and/or noise, there are small non-zero entries in $\hat{\mathbf{c}}$ that correspond to multiple classes. To overcome this problem, we perform an iterative procedure where we classify $\mathbf{x}_t$ to each one of the possible classes and use the training vectors of this class for reconstructing $\mathbf{x}_t$. In other words, in each repetition we retain only the coefficients in $\hat{\mathbf{c}}$ that correspond to a particular class, and use the training vectors of this class as a basis to represent $\mathbf{x}_t$. We introduce for each class a function $\delta_i : \mathbb{R}^N \to \mathbb{R}^N$, which selects the coefficients associated only with the class i. Then, each test feature vector is classified to the class that minimizes the ℓ_2 norm residual

$$\min_i r_i(\mathbf{x}_t) = \|\mathbf{x}_t - \mathbf{V}\delta_i(\hat{\mathbf{c}})\|_2 \tag{17}$$

for $i = 1,\ldots,S$.

3. EXPERIMENTAL RESULTS

In this section, we examine the identification performance of the three methods described in Section 2, regarding the correct speaker identification rate. For this purpose, several simulations under noisy conditions were conducted. The speech signals used for the simulations were obtained from the VOICES corpus, available by OGI's CSLU [17], which consists of twelve speakers (seven male and five female speakers). For all simulations, 20-dimensional MFCC coefficients were extracted from the speech utterances in a segment-by-segment basis. The frame duration was kept at 20 msec with 10 msec of frame shift. Before the feature extraction task, the training as well as the test utterances were pre-filtered using a low-pass filter of the form $H(z) = 1 - 0.97z^{-1}$, and then a silence detector algorithm based on the short-term energy and zero-crossing measures of speech segments was applied [1]. All the speech signals in the corpus have a sampling rate of 22050 Hz. For the GMM-based identification results, a GMM with a diagonal covariance matrix was chosen for the simulations. The number of mixtures depended on the amount of training data (see description of Experiment 1 below).

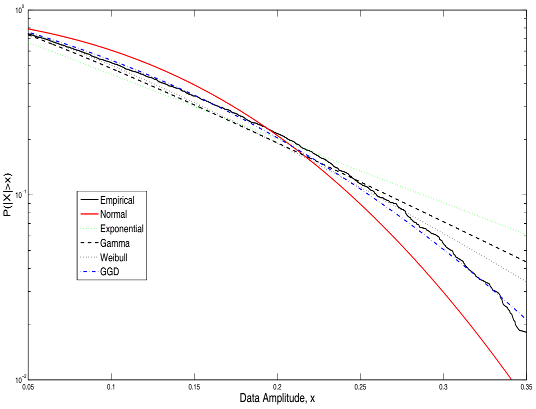

Figure 1: Example Amplitude Probability Density curves of the 8^{th} MFCC coefficient from the training data (20 sec) of the 10^{th} speaker.

For the GGD-based identification case, Amplitude Probability Density (APD) curves $(P(|X| > x))$ are adopted to show that the GGD best matches the actual density of the data. An example for a part of the VOICES corpus is given in Figure 1, where we compare the empirical APD (solid line) against the APD curves obtained for the GGD, Weibull, Gamma, Exponential and the Gaussian models. The results in the figure correspond to the 8^{th} MFCC coefficient of the training data (20 sec duration) corresponding to the 10^{th} speaker (independence among feature vector components is assumed). Clearly, the GGD follows more closely the empirical APD than the other densities. This trend was observed in the majority of the training utterances used in our experiments. Thus, the GGD model is expected to give better results than the other densities when applied directly to the MFCC coefficients of the twelve speakers.

The performance evaluation follows the philosophy as described in [3]: each sequence of feature vectors $\{\mathbf{x}_t\}$ is divided into overlapping segments of L feature vectors, where the first two segments have the following form

$$\underbrace{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\ldots,\mathbf{x}_L}_{1^{st}\,segment},\mathbf{x}_{L+1},\mathbf{x}_{L+2},\ldots$$

$$\mathbf{x}_1,\underbrace{\mathbf{x}_2,\mathbf{x}_3,\ldots,\mathbf{x}_L,\mathbf{x}_{L+1}}_{2^{nd}\,segment},\mathbf{x}_{L+2},\ldots$$

The comparison between the identified speaker of each segment and the actual speaker of the test utterance is repeated for each speaker in the corpus, and the total correct identification rate is computed as the percentage of the correctly identified segments of length L over all test utterances

$$\text{correct ident. rate} = \frac{\#\text{ correctly identified segments}}{\text{total}\#\text{ of segments}} \cdot 100\%. \tag{18}$$

In the previous sections, it was mentioned that in this paper the focus is given on noise robust speaker identification using short training and testing sessions. Towards this direction, white Gaussian noise is added on the test utterances, the SNR taking the values

[1] http://www.mathworks.com/matlabcentral/fileexchange/19298-speechcore

of 10, 15, 20, 25 dB. In addition, the test segment lengths L vary from 10 to 500 with a step size of $L = 40$. Length $L = 10$ corresponds to 0.1 sec, length $L = 50$ corresponds to 0.5 sec, and so forth. The training utterances have a duration of 5, 10, 15 and 20 seconds, corresponding to a quite short training session. The training for all methods is performed using the clean speech data. The testing data have a duration of approximately 20 sec.

3.1 Experiment 1 – Identification using GMM

In this experiment, during the training process the MFCC coefficients for each speaker are collected. For each speaker, the corresponding MFCC data are modeled using a diagonal GMM. The number of mixtures was chosen to be 4, for the 5 and 10 sec training data, and 8 for the 15 and 20 sec training data. These choices of parameters were found experimentally to produce the best performance for the GMM-based identification. Clearly, the number of mixtures is small due to the small size of the training dataset. During the identification process, the identification rule (4) is used, and the correct identification rate is computed as in (18).

3.2 Experiment 2 – Identification using KLD based on GGD

The same experimental steps as in Experiment 1 are also followed here. Thus, for each speaker the MFCC vectors are collected during the training process. We estimate the GGD parameters (α, β) for each vector component, assuming independence among the MFCC components. During the identification process, a test utterance contains multiple MFCC vectors as explained. For each MFCC component of the test vectors, the GGD parameters (α, β) are estimated. In order to identify a speaker, we compute the KLD between the GGD model of the test data and each of the GGD models of the speakers in the dataset (per vector component). This procedure results in 20 distance values (since each MFCC vector contains 20 components). The final step is to compute the mean of these distances, as in (10). The identity of the speaker whose data result in the minimum distance is identified as the final result. The correct identification rate is computed as in (18).

3.3 Experiment 3 – Identification using SRC

In this subsection, the experimental procedure for the SRC approach is described. First, consider that from the training speech data of each speaker a number of n_i of MFCC vectors is extracted. Consider a test utterance length of L frames. Adopting the notations from the theory of SRC in Section 2.3, the training matrix $\mathbf{V}$ has dimension $20 \times (12 \cdot n_i)$ and the test sample (feature) vector $\mathbf{x}_t$ is a 20×1 vector. The test segment consists of L distinct test samples $\mathbf{x}_t$. Thus, the optimization problem of the form

$$(P_l): \; \hat{\mathbf{c}}_l = \arg\min_{\mathbf{c}_l} \|\mathbf{c}_l\|_1 \,, \text{ s.t. } \mathbf{x}_{t,l} = \mathbf{V}\mathbf{c}_l, \text{ for } l = 1, \ldots, L \qquad (19)$$

is solved L times for each different $\mathbf{x}_{t,l}$. The Orthogonal Matching Pursuit [16] is used to solve this problem. Each solution $\hat{\mathbf{c}}_l$ of the problem (P_l) is used to get an identity i (for $i = 1, \ldots, 12$) of one of the 12 speakers in the dataset. Thus, a segment of length L vectors will provide L identification results. The predominant identity is found based on the majority of the decisions and the identification rate is computed as in (18).

3.4 Discussion

In this subsection, the main observations of the results in Figures (2.a)-(2.d) are discussed. The percentage of correct identification results is given as a function of the length of the test utterance. We are mainly interested to examine the performance of the described methods for short test sessions. The four figures correspond to training data of duration 5, 10, 15, and 20 sec respectively, so as to examine the effect of using a short training dataset. The correct identification rates as a function of the test utterances segment length L are depicted. The black, red and green curves correspond to the SRC, GMM and KLD-GGD method, respectively. There are twelve curves in total, where the first part of each legend name indicates the corresponding method and the last part indicates the SNR value used for this method, *e.g.* "SRC 10dB" means that the black solid curve depicts the identification performance of the SRC approach under noise conditions of 10dB. From the Figures (2.a)-(2.d) we notice that the SRC method is superior over the GMM and KLD-GGD approach, especially for short test and training sessions, and is quite robust to noise. The GMM performance improves as the training and test data duration increases because the large amount of feature vectors increases the accuracy of the GMM model, however its sensitivity to noise is clearly indicated. The KLD-GGD approach does not have high correct identification rates even in the case where the amount of training and test data is 20 and 5 sec, respectively. Based on the results, we can assume that the GGD parameters (α, β) are not well-estimated in the case where the test data have short duration.

The main point regarding the SRC method that has to be highlighted is that even in the case where the training data duration is 5 sec and the test utterance segments length is as low as 2 sec, the performance is greater than 80% for SNR values 15, 20 and 25 dB. Even in the extreme case of 10 dB SNR, the correct identification rate is above 70% for at least 2 sec test utterance segments length. Additionally, for lower test sessions than 2 sec the identification results for SRC are significantly better than the baseline method. For example, for 20 sec training data and 1.5 sec of test data, the SRC method gives correct identification above 70% for all SNR values. For the same case, for 10 dB SNR, GMM results in correct identification of slightly more than only 20%. This is important for applications where a decision must be made using a small amount of test data, without having enough training data for a given number of speakers, and the speaker is located in a noisy environment.

4. CONCLUSIONS

In this paper, we presented two methods for noise robust speaker identification using short-time training and testing data. They were both compared to a baseline GMM-based system. The first method was previously proposed for music genre classification, based on modeling the MFCC coefficients of the speakers using the GGD model. The second identification method was based on the recently proposed SRC algorithm. It was shown through experimental evaluation that the SRC approach performs significantly better than the other two methods when the amount of testing and training data is small, and is very robust to noise. Our future research plans include testing the SRC method with a larger set of speakers and a wide variety of noise types.

REFERENCES

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

[2] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Trans. on Multimedia*, vol. 9(7), pp. 1396–1403, November 2007.

[3] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3(1), pp. 72–83, January 1995.

[4] J. C. Wang, C. H. Yang, J. F. Wang, and H. P. Lee, "Robust speaker identification and verification," *IEEE Comp. Intelligence Magazine*, vol. 2(2), pp. 52–59, May 2007.

[5] K. H. Yuo, T. H. Hwang, and H. C. Wang, "Combination of autocorrelation-based features and projection measure technique for speaker identification," *IEEE Trans. on Speech and Audio Processing*, vol. 13(4), pp. 565–574, July 2005.

[6] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

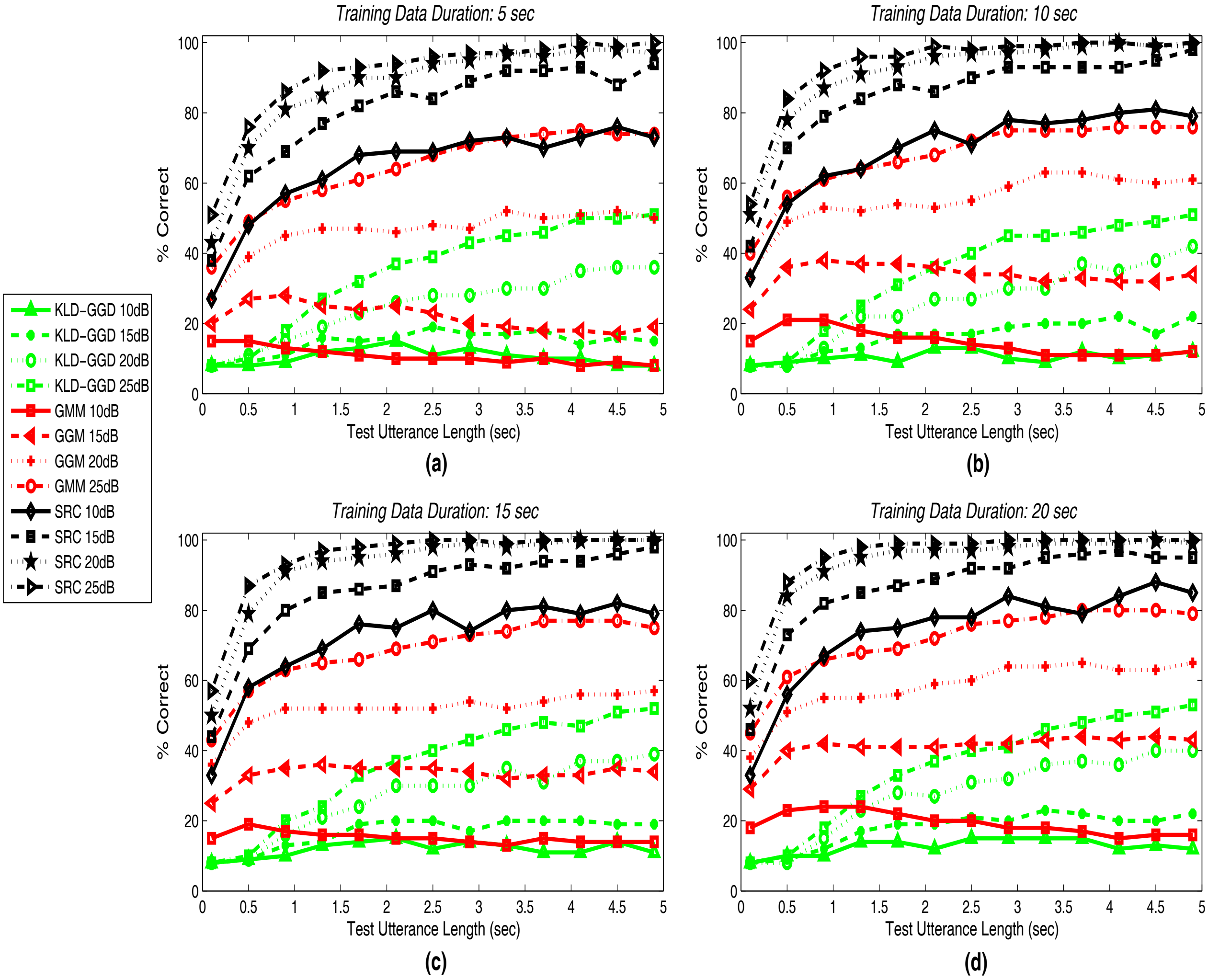

Figure 2: Speaker identification performance as a function of the test data duration for different number of SNR values. The duration of the training data is: (a) 5 sec, (b) 10 sec, (c) 15 sec and (d) 20 sec.

[7] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, “Speaker identification under noisy environments by using harmonic structure extraction and reliable frame weighting,” in *in Proc. Int. Conf. on Spoken Language Proc. (INTERSPEECH)*, Pittsburgh, Pennsylvania, USA, September 2006.

[8] M. Grimaldi and F. Cummins, “Speaker identification using instantaneous frequencies,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16(6), pp. 1097–1111, August 2008.

[9] Y. Shao and D. Wang, “Robust speaker identification using auditory features and computational auditory scene analysis,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, April 2008.

[10] K. Kumar, Q. Wu, Y. Wang, and M. Savvides, “Noise robust speaker identification using Bhattacaryya distance in adapted gaussian models space,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Lausanne, Switzerland, August 2008.

[11] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, “Robust speaker recognition in noisy conditions,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15(5), pp. 1711–1723, July 2007.

[12] H. Aronowitz and D. Burshtein, “Efficient speaker recognition using approximated cross entropy (ACE),” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15(7), pp. 2033–2043, September 2007.

[13] V. R. Apsingekar and P. L. D. Leon, “Speaker model clustering for efficient speaker identification in large population applications,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17(4), pp. 848–853, May 2009.

[14] C. Tzagkarakis, A. Mouchtaris, and P. Tsakalides, “Musical genre classification via Generalized Gaussian and Alpha-Stable modeling,” in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, vol. 5, May 2006.

[15] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 31(2), pp. 210–227, February 2009.

[16] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Trans. on Information Theory*, vol. 53(12), pp. 4655–4666, December 2007.

[17] A. Kain, *High Resolution Voice Transformation*. PhD thesis, OGI School of Science and Engineering at Oregon Health and Science University, October 2001.