A Fishervoice-based Speaker Identification System

Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo

Department of Signal Theory and Communications, Universidade de Vigo, Spain
{plopez, ldocio, carmen}@gts.tsc.uvigo.es

Abstract

In this paper, a novel approach for speaker identification called Fishervoice is proposed. It was inspired by the two-dimensional (2D) Fisherface technique, which is a method that combines a two-stage "PCA+LDA" strategy and two-dimensional discrimination techniques. Experimental results on the BANCA database demonstrate that the Fishervoice approach is effective for speaker identification tasks, in particular when there are mismatched conditions. The reduction on the number of parameters needed to describe each speaker model achieved with the Fishervoice technique is remarkable, thereby causing a reduction of computational and memory costs.
Index Terms: fishervoices, speaker identification, dimensionality reduction

1. Introduction

Despite the great advances made recently in the field of speaker recognition systems, they still lack robustness, i.e. their performance degrades dramatically when the acoustic training data differs from the given test conditions. Robustness is currently the major challenge in speaker recognition for real-world applications [1]. For example, in telephone services, users may call in under all kinds of acoustic environments (in the office, on the street, in the car) and use different telephone networks (land-line or cellular). Therefore, mismatched conditions may be found at any time, which makes robustness one of the critical factors that decide the success of speaker recognition technology in these applications.

Most of the state-of-art speaker recognition systems use Gaussian mixture models (GMM) as statistical models to represent the speakers in terms of the probability distribution of low-level acoustic features. These systems achieve very high accuracies on high-quality data when training and test conditions are well controlled, but their performance is significantly degraded under adverse and mismatched conditions. Nowadays, an interesting area of research is the use of discriminant analysis techniques in speaker recognition [2], in order to reduce intra-speaker and channel variability. In this way, speaker recognition techniques based on speaker subspace modeling have been proposed, such as the "eigenvoice approach" [3][4], and more recently kernel learning methods are arising a great interest. The work presented in this paper is inspired by the work described in [5] on face recognition.

Given the similarities between the study of faces and voices, the Fishervoice method (analogous to the Fisherface method) is presented in this paper, which takes a two-stage "PCA+LDA" strategy. It first uses two-dimensional Principal Component Analysis (PCA) to reduce the dimensionality, and then performs Linear Discriminant Analysis (LDA) to extract a discriminative subspace. Thus, in this paper a research to analyze the robustness and effectiveness of this dimensionality reduction and subspace learning technique is performed, to find out if this technique used in face recognition tasks to minimize the computational cost and alleviate the curse of dimensionality can also be helpful in speaker identification (SID) tasks.

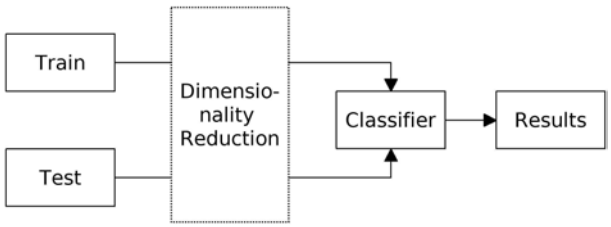

Figure 1: *Fishervoice SID system*

The outline of the paper is as follows. In Section 2, the Fishervoice method and the speaker identification algorithm are presented. In Section 3 the experimental framework is described. In Section 4 experimental results are presented. Finally Section 5 explains the conclusions of the paper.

2. Proposed SID System

The algorithm proposed in this paper to perform speaker identification is very simple, as shown in Fig. 1. There are two data inputs: a train dataset, used to model the speakers in the system, and a test dataset, where each of its elements has to be assigned to a speaker model. To perform this assignment, a classifier is used to decide which speaker segment of the train dataset is more similar to a given speech segment S from the test dataset, thus deciding that the speaker of S is the one that spoke the speech segment in the model that is more similar to S.

It can be seen in figure 1 that a dimensionality reduction step is performed before the classification step. It is at this point when the Fishervoice approach that is proposed in this paper is applied. If this step is skipped, the SID system will be equivalent to a GMM identification system, where a speaker segment is modeled by a GMM and GMMs are compared. This GMM identification system will be used as a baseline to make a comparison between its performance and the performance of the Fishervoice identification system.

2.1. The datasets

Two datasets are needed to perform speaker identification:

- A train dataset (A_{train}) composed by segments of speech spoken by different known speakers. This data is used to model the target speakers.
- A test dataset (A_{test}) composed by segments of speech that have to be assigned to the most likely speaker in the train set.

A_{train} and A_{test} are tridimensional matrices of dimension $m \times n \times L_{train}$ and $m \times n \times L_{test}$, respectively. L_{train} is the number of speech segments in matrix A_{train}, i.e. it is the number of speaker segments that are available to model the different speakers. Consequently, A_{test} is the number of speaker segments that have to be assigned to a speaker in the model.

Each segment in both sets will be represented by the means of a Gaussian Mixture Model (GMM), where the number of gaussians of the model is m and the dimension of the feature space is n. To obtain this GMMs, a Maximum a Posteriori (MAP) adaptation of a universal background model (GMM-UBM) is performed with the available acoustic features. In this work, the acoustic features are 12 Mel-frequency Cepstral Coefficients (MFCC), extracted using a 25ms Hamming window at a rate of 10ms per frame, and augmented with the normalized log-energy and their delta and acceleration coefficients. Thus, the dimension of the feature space (n) used in this paper is 39.

2.2. Dimensionality reduction: the Fishervoice method

The dimensionality reduction strategy proposed in this paper is based on an adaptation of the procedure presented in [5] for face recognition in order to make it suitable for speaker recognition. In [6] a different, but also called, fishervoice approach has been applied to a speaker clustering task, but in the fishervoice approach described in this paper a two-stage "PCA+LDA" strategy combined with a two-dimensional discrimination technique is applied, while in [6] only LDA is used.

Consider a set A_{train} representing speech segments as explained in 2.1. This dataset will be used to compute two transformation matrices X and Y as explained below.

The between-class D_b, within-class D_w and total D_t scatter matrices are defined as:

$$D_b = \sum_{i=1}^{c} P_i (M_i - M)^T (M_i - M) \quad (1)$$

$$D_w = \sum_{i=1}^{c} \sum_{j,j \in i} (A_{train_j} - M_i)^T (A_{train_j} - M_i) \quad (2)$$

$$D_t = D_b + D_w \quad (3)$$

where c is the number of different speakers in A_{train}, P_i is the a priori probability of the ith class, M_i is the mean matrix of the ith class ($i = 1, 2, \cdots, c$), M is the total mean matrix of A_{train}, and A_{train_j} is the $m \times n$ matrix of the j^{th} segment in A_{train}. So, we can understand M as the mean voice of the whole speaker set, and M_i as the mean voice of speaker i.

After computing these $n \times n$ matrices, the eigenvectors and eigenvalues of D_t are computed, finding a matrix X that maximizes $J(X) = X^T D_t X$. To reduce the dimensionality and make the system less time and memory consuming, an automatic strategy for dimensionality reduction is applied. The proposed selection strategy keeps only a percentage of the energy of the subspace (E_X):

$$E_X = \sum_{i=1}^{n} \lambda_i \quad (4)$$

where λ_i is the i^{th} greatest eigenvalue of X. In the end, matrix X keeps a number u of columns (eigenvectors) equal to the number of eigenvalues needed to absorb a given percentage e_1 of the energy E_X. Hence, X is a $n \times u$ matrix.

After obtaining X, the sample set A_{train} is transformed into a new space with a lower dimensionality by doing $B_{train} = A_{train} X$. Then, new between-class and within-class scatter matrices (R_b and R_w, respectively) are computed:

$$R_b = \sum_{i=1}^{c} P_i (L_i - L)(L_i - L)^T \quad (5)$$

$$R_w = \sum_{i=1}^{c} \sum_{j,j \in i} (B_{train_j} - L_i)(B_{train_j} - L_i)^T \quad (6)$$

where L is the mean voice of the set B_{train}, and L_i is the mean voice of the ith speaker in that set.

Applying the Fisher criterion, a matrix Y that maximizes $J(Y) = \frac{Y^T R_b Y}{Y^T R_w Y}$ is obtained. Again, an automatic strategy for dimensionality reduction is applied as before, causing Y to become a $m \times v$ matrix by keeping the $e_2\%$ of the energy of the subspace E_Y.

Finally, performing the transformation $C_{train} = Y^T B_{train}$ a new sample set C_{train} composed by $v \times u$ matrices is obtained. After this procedure, a new representation of the dataset A_{train} with lower dimensionality is obtained.

After computing X and Y, the test data matrix A_{test} is projected to this new low dimensionality subspace by doing:

$$B_{test} = A_{test} X \quad (7)$$

$$C_{test} = Y^T B_{test} \quad (8)$$

2.3. Classifier

After reducing the dimensionality of the datasets, two tridimensional matrices C_{train} and C_{test} of dimensions $v \times u \times L_{train}$ and $v \times u \times L_{test}$ respectively are obtained. A transformation to bi-dimensional matrices is performed, obtaining two matrices C'_{train} and C'_{test} of dimensions $vu \times L_{train}$ and $vu \times L_{test}$ respectively. This transformation consists on stacking the means of the GMMs, i.e. concatenating the rows of each of the L_{train} (L_{test}) matrices in C_{train} (C_{test}) to obtain a matrix of supermean vectors. This transformation is not really necessary, but it makes the classification task easier, because now vectors are compared instead of matrices.

To decide which of the L_{train} speakers spoke one of the segments S in the test set, the following expression is evaluated:

$$T = \min_i d(C'_{test_S}, C'_{train_i}) \quad (9)$$

where $d(.,.)$ is the euclidean distance between two vectors. The speaker of the segment T that minimizes the euclidean distance to the segment S is chosen as the speaker of S. Experiments were performed with different distance measures (for example, Mahalanobis distance), but the best results were achieved with the euclidean distance.

3. Experimental framework

3.1. Description of the database

The speaker identification system proposed in this paper was tested using the BANCA database [7] [8]. This database includes 52 English speakers (26 males and 26 females) each of whom recorded 12 sessions divided into 3 different scenarios: controlled, degraded and adverse.

Each session was recorded using two different-quality microphones. In each session the speaker recorded two different utterances, hence there are eight utterances per speaker in each scenario. A partition of the data in three groups has to

Table 1: Summary of the datasets used in the experiments.

	Experiment	GMM-UBM	Train	Test
GT	1	All	All	All
Matched	1	Controlled	Controlled	Controlled
	2	Degraded	Degraded	Degraded
	3	Adverse	Adverse	Adverse
	4	All	Controlled	Controlled
	5	All	Degraded	Degraded
	6	All	Adverse	Adverse
Mismatched	1	Cont./Deg.	Controlled	Degraded
	2	Cont./Adv.	Controlled	Adverse
	3	Degr./Cont.	Degraded	Controlled
	4	Degr./Adv.	Degraded	Adverse
	5	Adv./Cont.	Adverse	Controlled
	6	Adv./Deg.	Adverse	Degraded
	7	All	Controlled	Degraded
	8	All	Controlled	Adverse
	9	All	Degraded	Controlled
	10	All	Degraded	Adverse
	11	All	Adverse	Controlled
	12	All	Adverse	Degraded

be done, in order to have different data for training the GMM-UBM, computing the matrices X and Y and testing. The eight utterances per speaker are divided as follows: two are used to train the GMM-UBM, three to train the matrices, and three are used for testing.

Three different groups of experiments are described in this paper. The first group consists of only one experiment, and is called Grand Test (GT) because of its similarity to the one with the same name in [7]. This is an experiment that uses data from all the scenarios both for training and for testing. The second group are experiments in matched conditions, i.e. experiments where the data used for training is from the same scenario as the data used for testing. Finally, the third group are experiments in mismatched conditions, where the data used for training is from a different scenario as the data used for testing.

Table 1 describes the different datasets used for the experiments. GT experiment is a global recognition test, using the three different scenarios for training and testing. Experiments 1, 2 and 3 in matched conditions are scenario-based, i.e. training and testing are performed using only data from a given scenario. Experiments 4, 5 and 6 in matched conditions use a scenario-independent GMM-UBM (trained with data from the three scenarios), but matrices X and Y are obtained using data from the same scenario as the Test set.

In the experiments in mismatched conditions, Train comes from scenario i, while Test comes from scenario j, where $i \neq j$. Note that there are two different GMM-UBM sets, separated by a slash (/), i.e. T_i/T_j. This means that T_i is the GMM-UBM used in Train, and T_j is the GMM-UBM used in Test. This means that Train and Test use a GMM-UBM trained with data that belongs to their respective scenarios. GMM-UBMs of 16, 32, 64 and 128 gaussians are going to be used in the experiments. The number of gaussians in the GMM-UBM will be referred to as M.

4. Results

4.1. Baseline

Table 2 shows the accuracies obtained by performing speaker identification without using dimensionality reduction techniques, i.e. comparing GMMs directly. The results obtained for the GT and the experiments in matched conditions show that the baseline achieves acceptable accuracies, but the error rate in mismatched conditions is, in general, excessively high.

For each experiment, Table 2 also indicates which number of gaussians obtained the highest accuracy (M), choosing the lowest one when the same result was obtained with different GMMs. The dimensionality of the data is always $M \times n$.

Table 2: Baseline results.

	Experiment	Accuracy	M
GT	1	91.9872	128
Matched	1	96.1538	32
	2	94.2308	64
	3	94.2308	32
	4	96.1538	64
	5	94.8718	128
	6	97.4359	64
Mismatched	1	10.2564	32
	2	3.8462	32
	3	13.4615	32
	4	4.4872	32
	5	5.1282	32
	6	6.4103	32
	7	94.2308	128
	8	32.6923	128
	9	88.4615	32
	10	31.4103	64
	11	33.3333	64
	12	33.9744	64

4.2. Fishervoice approach

Figure 2 shows the accuracies obtained in the GT experiment, using different values of the energy percentages e_1 and e_2, and GMMs with 16, 32, 64 and 128 gaussians. The values in blue are the lowest, and the ones in red are the highest. The aim is to obtain the highest accuracy with the lowest dimensionality, and in this case it is an accuracy of 99.0385%, with a subspace of dimension 33×39. Nevertheless, an accuracy of 98.7179% can be reached with a subspace of dimension 32×9, which will be computationally better.

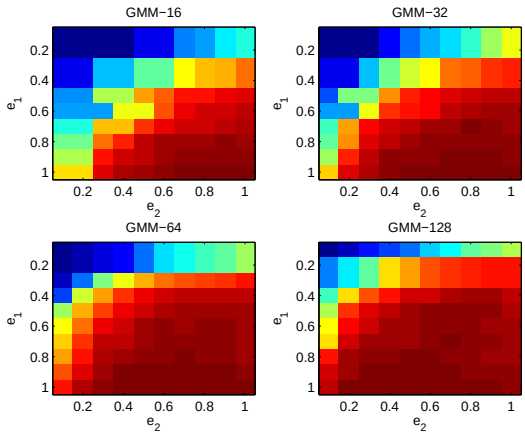

Figure 2: *Results for experiment 1.*

As one of the aims of this system is to obtain high accuracies with low dimensionality subspaces, the accuracies obtained

using $e_1 = 100\%$ or $e_2 = 100\%$ will be discarded. Table 3 shows the maximum accuracies achieved in the experiments in Table 1 (using the best-quality microphone) and the lowest dimensionality subspaces that achieve them with a GMM-UBM of M gaussians.

Table 3: Experimental results with the Fishervoice approach.

	Experiment	Accuracy	Dimension	e_1	e_2	M
GT	1	98.7179	32×9	90	80	64
Matched	1	100	5×8	90	70	16
	2	100	20×6	80	60	64
	3	99.359	37×8	80	70	128
	4	100	3×9	90	30	32
	5	100	10×6	70	50	64
	6	100	18×10	90	50	128
Mismatched	1	28.2051	2×9	90	10	64
	2	9.6154	1×6	70	10	64
	3	39.7436	1×5	60	10	16
	4	10.2564	2×9	90	20	32
	5	30.7692	1×9	90	10	32
	6	31.4103	2×5	60	20	32
	7	100	15×9	90	60	64
	8	70.5128	15×9	90	80	32
	9	100	8×9	90	60	32
	10	74.359	22×9	90	90	32
	11	79.4872	21×9	90	90	32
	12	78.2051	11×9	90	90	16

The error rate in the GT experiment, which is the most general, is approximately 1.3%. Experiments in matched conditions obtain an accuracy of 100% with the different GMM-UBMs, except in experiment 3, which corresponds to a degraded scenario, where the error rate is about 0.7%. It can also be appreciated in table 3 that the dimensionality of the subspace needed to obtain these results is lower in experiments 4,5,6 than in experiments 1,2,3.

In experiments 1 to 6 in mismatched conditions, where Train and Test are adapted using the GMM-UBM corresponding to their own scenario, the error rate is too high. Nevertheless, in experiments 7 to 12, where a global GMM-UBM was used for the adaptation, higher accuracies are obtained, mainly in experiments 7 and 9 where Train and Test belong to the least degraded scenarios.

Comparing tables 3 and 2, it can be observed that the use of the Fishervoice approach achieves higher accuracies in all the experiments. Not only accuracies are higher, but lower dimensionality of data is handled in all cases, making the Fishervoice dimensionality reduction method effective for speaker identification.

Figure 3 compares the results obtained in the experiments using the two available microphones, where the green bars represent the results obtained with microphone 1, and the yellow bars represent the results obtained with microphone 2. It can be seen that, in general, better accuracies are obtained using microphone 1, due to its better quality. Nevertheless, results obtained with microphone 2 are acceptable in matched conditions.

5. Conclusions and Future Work

A PCA-LDA based speaker identification system is presented in this paper with two goals: obtain a good performance even in mismatched conditions, and reduce the dimensionality of the data in order to reduce the computational load. Table 3 shows that the speaker identification is almost perfect in matched conditions, and acceptable in mismatched conditions. Moreover, it outperforms the baseline, where no dimensionality reduction techniques are applied, and reduces the dimensionality of the data to be handled. It is also noticeable that the best accuracies are obtained when the GMM-UBM employed is trained with both clean and degraded data, thus helping the recognition system to work with degraded samples. In addition, a substantial reduction of the dimensionality is achieved, allowing the system to be less time and memory consuming, as the vectors and matrices that represent the speaker segments are smaller, and the computing time for classification is reduced.

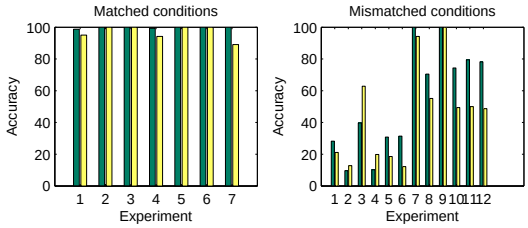

Figure 3: *Results with different microphones.*

The main problem of this method is the selection of the best values for e_1, e_2 and M, being necessary to perform some research in the future to find an automatic manner to choose the most suitable values for these parameters.

The GT experiment, which is the most interesting because it does not matter the quality of the samples used for training and testing, has an error rate of 1.3%, making this method useful for real applications.

In future work, the validity of the Fishervoice method for speaker verification will be tested, going into the method in depth to improve it both in speaker identification and verification.

6. References

[1] X.-H. Ren, Y.-F. Zhang, Y.-J. Xing, M. Li, "Application of KPCA and PNN for Robust Speaker Identification", Proceedings of the 2008 Congress on Image and Signal Processing, vol. 4, pp. 533–536, 2008.

[2] A. Errity and J. McKenna, "A Comparative Study of Linear and Nonlinear Dimensionality Reduction for Speaker Identification", Proc. of the 15^{th} International Conference on Digital Signal Processing (DSP), pp. 587–590, Cardiff, Wales, 2007

[3] R. Kuhn, J.-C. Junqua, P. Nguyen and N. Niedzielski, "Rapid Speaker Adaptation in Eigenvoice Space", IEEE Trans. Speech and Audio Processing, vol. 8, n. 6, pp. 695–707, 2000

[4] O. Thyes, R. Kuhn, P. Nguyen and J.-C. Junqua, "Speaker Identification and Verification Using Eigenvoices", International conference on Spoken Language Processing, pp. 242–245, Beijing, China, October 2000.

[5] X.Y. Jing, H.S. Wong and D. Zhang, "Face Recognition Based on 2D Fisherface Approach", Pattern Recognition, vol. 39, n. 4, pp. 707–710, 2006.

[6] S.M. Chu, H. Tang, T.S. Huang, "Fishervoice and Semi-supervised Speaker Clustering", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4089–4092, Los Alamitos, CA, USA, 2009

[7] E. Bailly-Baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Mariethoz, J. Matas, K. Messer, F. Poree, B. Ruiz, "The BANCA Database and Evaluation Protocol", 2003.

[8] The BANCA Database Website, Online: http://www.ee.surrey.ac.uk/CVSSP/banca/