PCA, SFS or LDA: What is the Best Choice for Extracting Speaker Features?

Abdelghani Harrag[1, 2]
[1] Department of Electronics,
Faculty of technology, Mohamed Boudiaf
University, BP 166 Ichbilia Msila 28000 Algeria

Tayeb Mohamadi[2]
[2] Department of Electronics,
Faculty of technology, Ferhat Abbas University,
Bejaia Road Setif 19000 Algeria

ABSTRACT
Feature extraction is the process of deriving new weakly correlated features from the original features in order to reduce the cost of feature measurement, increase classifier efficiency, and allows higher classification accuracy. The selection and quality of the features representing each pattern have considerable bearing on the success of subsequent pattern classification. In this paper, we supply a comparative study for best feature extraction method for speaker recognition system. A Linear Discriminant Analysis (LDA) method is compared to two well-known feature extraction techniques, namely Principal Component Analysis (PCA) and Sequential Forward Search (SFS). Evaluation is carried out on Arabic speech database using four acoustic representations combined with prosodic features. We show that LDA-based feature outperformed PCA and SFS in acoustic alone as well as for acoustic and prosodic combined features.

General Terms
Pattern Recognition, Speaker Recognition

Keywords
Speaker Recognition, Speaker Features, Feature Extraction, Linear Discriminant Analysis.

1. INTRODUCTION
Feature extraction is a key issue for efficient speaker recognition system. Redundant and harmful information should be removed from speech, retaining only those features relevant to classification. An optimal set feature should have the following properties: i) high inter-speaker variation, ii) low intra-speaker variation, iii) easy to measure, iv) robust against mimicry, v) robust against noise and vi) independent of other features. Unfortunately, no single feature fulfils these requirements. High-level speaker features require a lot of data to estimate acoustic and language models and low-level acoustic features are easily corrupted by background noise and other distortion sources. Most of the current implementations use some kind of spectral envelope features to parameterize the voice achieving a great performance [1]. But recent researches are trying to include complementary information into the system, in order to reduce error rates. The examples of such complementary information are pitch [2], residual phase [3], prosody [4-5], dialectical features [6] etc.

One reasonable approach to improve speaker recognition system is to extract from the data several feature vectors (assumedly independent) and then extract from the concatenated features a reduced size fused vector of enhanced separability. The reduced feature set would allow best classification and less computational resources would be required. This paper illustrates the value of feature selection when combining features from different spaces. We study the problem of choosing an optimal feature set and demonstrate that, by removing features that do not encode important speaker information; the error rate can be reduced significantly. Linear Discriminant Analysis (LDA) method is compared to two well-known feature extraction techniques, namely Principal Component Analysis (PCA) and Sequential Forward Search (SFS) applied to four acoustic representations (LPC, PARCOR, LPCC and MFCC) [7] concatenated with prosody (energy, pitch and duration). The concatenated vectors are used in speaker recognition based on K-Nearest Neighbor (KNN) classifier. All experiments were performed using QSDAS speech database [8]. The rest of the paper is organized as follows: Section 2 briefly reviews some feature extraction methods. Next we outline the used speech database and report the experimental results in Section 3. Finally, Section 4 draws the principal conclusions of the paper.

2. FEATURE EXTRACTION
Depending on the acoustic front-end, the resulting feature vectors may have from 20 to 50 components. In real-time speaker applications using low-resource devices, 50-dimensional feature vectors do not seem suitable, so a further feature set and reduction is needed. Several selection procedures are discussed in the pattern recognition literature, among them:

2.1 Exhaustive Search (ES)
Exhaustive search is an optimal method for selecting a subset of k best features among the entire set K. It considers all the combinations of k out of K. Implementation of such a search require an enormous amount of computation. For example, with k=20 and K=50, the number of searches is $\sim 4.712 \times 10^{13}$. Therefore, there is a need for some more effective procedures to avoid the exhaustive search.

2.2 Best Individual Feature (BIF)
In this technique, classification performance of each feature point is calculated separately, that is, on individual basis, and the features giving rise to highest correct recognition rate are selected. The best subset of *k* features is composed of the *k* best features considered one at a time. However, a set of the best individual *k* features is not necessarily the best set of *k* features.

2.3 Sequential Forward Search (SFS)
In the SFS [9] method, features are selected successively by adding the locally best feature point that provides the highest

incremental discriminatory information, to the existing feature subset. The SFS technique starts as the BIF by identifying the first feature that has the highest discrimination power. It proceeds, however, by adding sequentially to the selected subset, those features that contribute most to the classification performance on top of the already selected ones.

2.4 Principal Component Analysis (PCA)

Principal Component Analysis is an old technique of multivariate statistical analysis [10], consisting of computing the eigenvectors of D*D covariance matrix Σ, then sorting them according to the corresponding eigenvalues, in descending order, and finally building the projection matrix A (called Karhunen-Loeve Transform, KLT) with the largest K eigenvectors (i.e. the K directions of greatest variance). Each feature vector X is then pre-processed according to the expression $Y = A(X - \mu)$, where μ represents the mean feature vector. KLT decorrelates the features and provides the smallest possible reconstruction error among all linear transforms, i.e. the possible mean-square error between data vectors in the projection K-feature space.

2.5 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis [11] can be summarized as a two phase procedure: in phase one, a class-dependent normalization function collects statistical information, and in phase two a discrimination function is derived from classes so that the resultant elements of the LDA feature are less correlated, and ranked according to an objective criterion computed from the inter-class covariance matrix B and the averaged intra-class matrix W, I.e. $W^{-1}B$.

3. EXPERIMENTAL RESULTS

For our study, we use the QSDAS speech database dedicated for Arabic speaker recognition. We use four parameterizations LPC, PARCOR, LPCC and MFCC as acoustic parameters used alone and in fusion with prosodic features $LogF_0$, LogE and Duration D as defined in Table 1. For each vowel and for the seven feature sets defined, we calculate the Recognition Rate (RR) for each feature set using KNN classifier. Fig. 1 to Fig. 4 bring the corresponding results for features alone, PCA-features, LDA-features and SFS-features respectively. Among the 20 available figures, we decide to present only one figure for each parameterization (LPC, PAR, LPCC and MFCC).

Table 1. Table captions should be placed above the table

Set	Conf 1	Conf 2	Conf 3	Conf 4	Size
1	LPC	PAR	LPCC	MFCC	5
2	LPC	PAR	LPCC	MFCC	10
3	LPC	PAR	LPCC	MFCC	15
4	LPC	PAR	LPCC	MFCC	20
5	LPC+E	PAR+E	LPCC+E	MFCC+ E	22
6	LPC+E+ F_0	PAR+E+ F_0	LPCC+E +F_0	MFCC+ E+F_0	24
7	LPC+E+ F_0+D	PAR+E+ F_0+D	LPCC+E +F_0+D	MFCC+ E+F_0+D	27

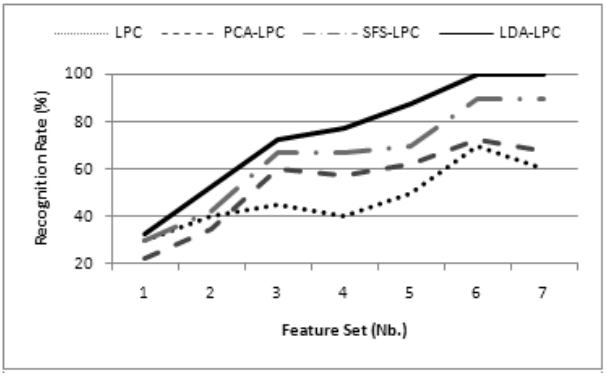

Fig 1. LPC features

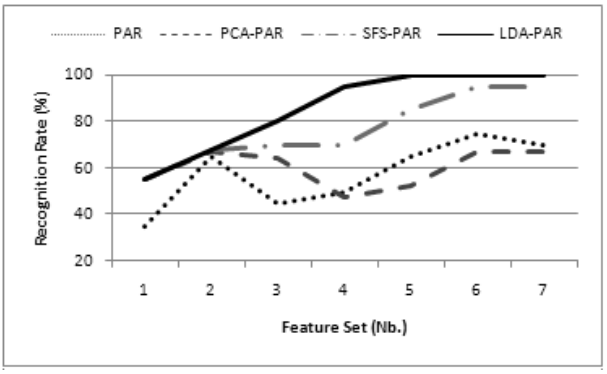

Fig 2. PAR features

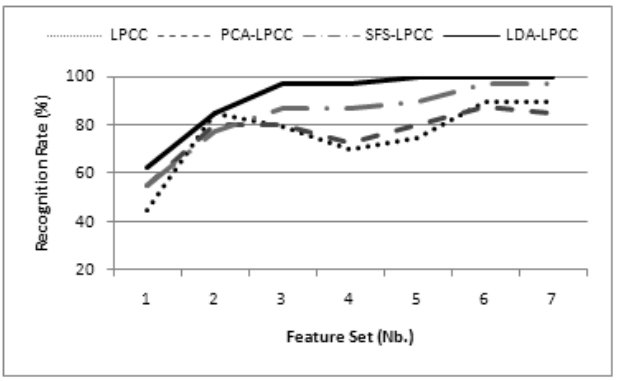

Fig 3. LPCC features

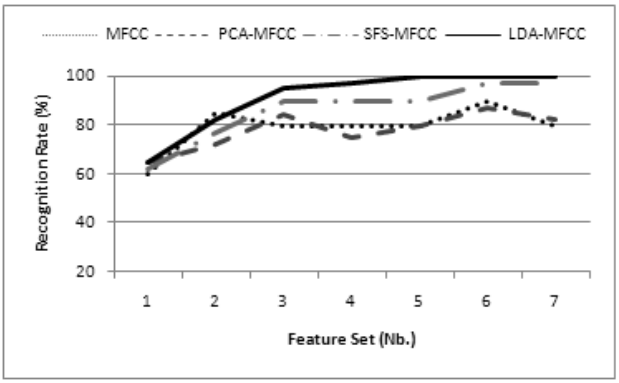

Fig 4. MFCC features

4. CONCLUSION

In this paper we supply a comparative study between three feature extraction methods that improve automatic speaker recognition system accuracy: LDA, PCA and SFS. Four acoustic parameterizations are examined: LPC, PARCOR, LPCC and MFCC fused with prosodic features: energy, pitch and duration. All experiments carried out in this work were held on Arabic speech database known as QSDAS and dedicated for Arabic speaker recognition. Seven feature set combinations are defined and examined using acoustic alone (K=5, 10, 15 and 20) and acoustic combined with prosodic features (K=22, 24 and 27). We use a simple KNN to calculate the Recognition Rate (RR) for concatenated features, PCA-features, LDA-features and SFS-features respectively. We demonstrate that for concatenated vectors without fusion using PCA or LDA, the inclusion of additional features did not necessarily improve performance. In fact it even may degrade the performance of speaker recognition system. The same features combined by PCA do not necessarily give better results due to the fact that mean-square error between data vectors in the projection K-feature PCA space does not necessarily equal minimizing classification error. For against, we get better results using LDA and SFS with best score for LDA. LDA-features are less correlated who is not the case of SFS. Additionally, SFS algorithm suffers from the inability to correct previous additions of features selected successively. The results demonstrate LDA as viable technique for improving system accuracy; the recognition rate is increased for each inclusion of new feature which is not the case for PCA, SFS or concatenated features alone.

5. REFERENCES

[1] Przybocki, M. A. and Martin, A. F. 2004. NIST Speaker Recognition Evaluation Chronicles, Odyssey, Toledo Espana.

[2] Harrag A., Mohamadi T. , Serignat J.F. 2005. LDA Combination of Pitch and MFCC Features in Speaker Recognition, INDICON Chennai, India, (Dec 2005). pp. 237-240.

[3] K. Sri Rama Murty and B. Yegnanarayana, Combining evidence from residual phase and MFCC features for speaker recognition, IEEE Signal Processing Letters, 13 (2006), pp. 52-55.

[4] Yegnanarayana, B., Prasanna, S.R.M., Zachariah, J.M., and Gupta, C.S. 2005. Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system, IEEE Trans Speech and Audio Processing, (2005), 13, pp. 575-582.

[5] Avinash B., Guruprasad S. and Yegnanarayana B. 2010. Exploring subsegmental and suprasegmental features for a text-dependent speaker verification in distant speech signals, Interspeech 2010, Makuhari, Chiba, Japan. pp. 1073-1076.

[6] Chakroborty, S. and Saha, G. 2009. Improved Closed set Text-Independent Speaker Identification by Combining MFCC with Evidence from Flipped Filter Banks. International Journal of Signal Processing, (2009), 5, 1-2, pp. 11-19.

[7] Rabiner, L. and Huang, B. H. 1993. Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs.

[8] Harrag A. and Mohamadi T. 2010. QSDAS: New Quranic Speech Database for Arabic Speaker Recognition, AJSE Journal, Theme Issue on Arabic Computing, Vol. 35(2C), (December 2010), pp. 7-19.

[9] Whitney, A. 1997. A direct method of nonparametric measurement selection, IEEE Trans. Comput., (1997), 20, pp. 1100-1103.

[10] Jolliffe, I.T. 2002. Principal Component Analysis, Springer.

[11] Duda, R.O., Hart, P.E., Stork, D.G. 2000. Pattern Classification, Wiley Interscience.