Comparing Genetic Algorithms to Principal Component Analysis and Linear Discriminant Analysis in Reducing Feature Dimensionality for Speaker Recognition

Maider Zamalloa†‡, L.J. Rodríguez-Fuentes†, Mikel Peñagarikano†, Germán Bordel†, and Juan P. Uribe‡

†Grupo de Trabajo en Tecnologías del Software, DEE, ZTF/FCT
Universidad del País Vasco / Euskal Herriko Unibertsitatea
Barrio Sarriena s/n, 48940 Leioa, SPAIN

‡Ikerlan – Technological Research Centre
Paseo J.M. Arizmendiarrieta 2, 20500 Arrasate-Mondragón, SPAIN

maider.zamalloa@ehu.es

ABSTRACT

Mel-Frequency Cepstral Coefficients and their derivatives are commonly used as acoustic features for speaker recognition. Reducing the dimensionality of the feature set leads to more robust estimates of the model parameters, and speeds up the classification task, which is crucial for real-time speaker recognition applications running on low-resource devices. In this paper, a feature selection procedure based on genetic algorithms (GA) is presented and compared to two well-known dimensionality reduction techniques, namely Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Evaluation is carried out for two speech databases, containing laboratory read speech and telephone spontaneous speech, and applying a state-of-the-art speaker recognition system. GA-based feature selection outperformed PCA and LDA when dealing with clean speech, but not for telephone speech, probably due to some noise compensation implicit in linear transforms, which cannot be accomplished just by selecting a subset of features.

Categories and Subject Descriptors: I.5.2 [**Pattern Recognition**]: Design Methodology – Feature Evaluation and Selection

General Terms: Performance

Keywords: Genetic Algorithms, Feature Dimensionality Reduction, Speaker Recognition

1. INTRODUCTION

Mel-Frequency Cepstral Coefficients (MFCC) [2] are commonly used as acoustic features for speaker recognition, since they convey not only the frequency distribution identifying sounds, but also information related to the glottal source and the vocal tract shape and length, which are speaker specific features. Additionally, it has been shown that dynamic information improves the performance of recognizers, so first and second derivatives are appended to MFCC. The resulting feature vector ranges from 30 to 50 dimensions. However, for applications requiring real-time operation on low-resource devices, high dimensional feature vectors do not seem suitable and some kind of dimensionality reduction must be applied, maybe at the cost of performance degradation.

A simple approach to dimensionality reduction is feature selection, which consists of determining an optimal subset of K features by exhaustively exploring all the possible combinations of D features. Most feature selection procedures use the classification error as the evaluation function. This makes exhaustive search computationally infeasble in practice, even for moderate values of D. The simplest method consists of evaluating the D features individually and selecting the K most discriminant ones, but it does not take into account dependencies among features. So a number of suboptimal heuristic search techniques have been proposed in the literature, which essentially trade-off the optimality of the selected subset for computational efficiency [5]. Genetic Algorithms (GA) suitably fit this kind of complex optimization problems. GA can easily encode decisions about selecting or not selecting features as sequences of boolean values, allow to smartly explore the feature space by retaining those decisions that benefit the classification task, and simultaneously avoid local optima due to their intrinsic randomness. GA have been recently applied to feature extraction [1], feature selection [8] and feature weighting [9] in speaker recognition.

Alternatively, the problem of dimensionality reduction can be formulated as a linear transform which projects feature vectors on a transformed subspace defined by relevant directions. Among others, two well-known dimensionality reduction techniques, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), fall into this category. GA-based feature selection projects the original D-dimensional feature space into a reduced K-dimensional subspace by just selecting K features. PCA and LDA not only reduce but also scale and rotate the original feature space, through a transformation matrix A which optimizes a given criterion on the training data. From this point of view, PCA and LDA generalize feature selection, but the criteria applied to compute A (the highest variance in PCA, and the highest ratio of between to within class variances in LDA) do not match the criterion applied in evaluation (the speaker recognition rate). This is the strong point of GA, since feature selection is performed in order to maximize the speaker recognition rate on an independent development corpus.

Copyright is held by the author/owner(s).
GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.
ACM 978-1-60558-130-9/08/07.

2. EXPERIMENTAL SETUP

In this work, MFCC, energy and their first and second derivatives were taken as acoustic features. The distribution of feature vectors extracted from a speaker's speech was represented by a linear combination of M multivariate Gaussian densities, which is known as *Gaussian Mixture Model* (GMM) [7]. Speaker recognition was performed using 32-mixture diagonal covariance GMMs as speaker models.

The well-known *Simple Genetic Algorithm* (SGA) [4], implemented by means of ECJ [3], was applied to search for the optimal feature set. Each candidate was represented by a D-dimensional vector of positive integers $R = \{r_1, r_2, \ldots, r_D\}$, ranging from 0 to 255 (8 bits), the K highest values determining what features were selected. An initial population of N candidate solutions (ranging from 80 to 200 individuals, depending on K) was randomly generated. To evaluate each K-feature subset $\Gamma = \{f_1, f_2, \ldots, f_K\}$, the acoustic vectors of the whole speech database were reduced to the components enumerated in Γ; speaker models were estimated using a training corpus; utterances in a development corpus were then classified by applying the speaker models; finally, the speaker recognition accuracy obtained for the development corpus was used to evaluate Γ. After all the K-feature subsets were evaluated, some of them were selected (according to a fitness-proportional criterion), mixed (one-point crossover) and mutated (mutation probability: 0.01) in order to get the population for the next generation. After 40 generations, the optimal K-feature subset $\hat{\Gamma} = \{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_K\}$ was evaluated on the test corpus. The three datasets used in this procedure: training, development and test, were composed of disjoint sets of utterances.

A public domain software developed at the MIT Lincoln Laboratory, *LNKnet* [6], was used to perform PCA. Regarding LDA, a custom implementation was developed in Java.

3. RESULTS

GA-based feature selection, PCA and LDA were tested in speaker recognition experiments for two different databases, *Albayzín* and *Dihana*. *Albayzín* is a phonetically balanced read speech database in Spanish, recorded at 16 KHz in laboratory conditions, containing 204 speakers. *Dihana* is a spontaneous task-specific speech corpus in Spanish, recorded at 8 kHz through telephone lines, containing 225 speakers. First, D-dimensional feature vectors ($D = 39$ for *Albayzín*; $D = 33$ for *Dihana*) were transformed into reduced K-dimensional feature vectors, according to the selection/transformation given by GA, PCA or LDA, then speaker models were estimated on the training corpus and speaker recognition experiments were carried out on the test corpus. Results are shown in Table 1.

In the case of Albayzín, neither PCA nor LDA outperformed GA. Error rates for Dihana were much higher, because it was recorded through telephone channels in an office environment and a large part of it consists of spontaneous speech. The presence of channel and environment noise in Dihana makes PCA and LDA more suitable than GA, because feature selection cannot compensate for noise, whereas linear transforms can do it to a certain extent. This may explain why either PCA or LDA outperformed GA in all cases but for $K = 8$. LDA was the best approach in most cases (for $K = 10$, 11, 12, 13 and 20). GA was the second best approach for $K = 6$, 10, 11, 12 and 13. Finally, the lowest error rate (15.97%) was obtained for $K = 30$ using PCA.

Table 1: Error rates in speaker recognition experiments for read laboratory speech (Albayzín) and spontaneaous telephone speech (Dihana), using the K-dimensional feature sets provided by GA, PCA and LDA, for K = 6, 8, 10, 11, 12, 13, 20 and 30.

K	Albayzín			Dihana		
	GA	PCA	LDA	GA	PCA	LDA
6	**5.71**	14.37	8.11	34.23	**33.23**	35.52
8	**1.81**	5.86	2.64	**23.90**	24.19	25.06
10	**0.94**	2.73	1.21	19.70	20.67	**19.43**
11	**0.35**	1.61	1.12	19.32	20.27	**18.10**
12	**0.30**	0.94	0.79	19.27	19.75	**18.18**
13	**0.33**	0.56	0.88	19.12	19.63	**17.66**
20	**0.16**	0.19	0.39	19.99	17.61	**17.24**
30	**0.13**	0.15	0.33	19.10	**15.97**	18.17

4. CONCLUSION

GA-based feature selection outperformed PCA and LDA when dealing with read laboratory speech, and performed quite well even for spontaneous telephone speech when the target K was small. However, in this latter condition the feature subsets provided by PCA and LDA yielded better performance, probably due to some noise compensation implicit in linear transforms, which cannot be accomplished just by selecting a subset of features. In any case, since applying a linear transform is more costly than selecting a subset of features, depending on the application, the gain in performance provided by linear transforms might not be worth the additional effort they involve, and GA-based feature selection would be a better choice.

5. ACKNOWLEDGEMENTS

This work has been jointly funded by the Government of the Basque Country, under projects S-PE06UN48, S-PE07UN43, S-PE06IK01 and S-PE07IK03, and the University of the Basque Country, under project EHU06/96.

6. REFERENCES

[1] C. Charbuillet, B. Gas, M. Chetouani, and J. L. Zarader. Filter Bank Design for Speaker Diarization Based on Genetic Algorithms. In *Proceedings of the IEEE ICASSP'06*, Toulouse, France, 2006.

[2] S. B. Davis and P. Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 28(4):357–366, 1980.

[3] ECJ 16. http://cs.gmu.edu/ eclab/projects/ecj/.

[4] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, 1989.

[5] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, January 2000.

[6] R. P. Lippmann, L. Kukolich, and E. Singer. LNKnet: Neural Network, Machine Learning and Statistical Software for Pattern Classification. *Lincoln laboratory Journal*, 6(2):249–268, 1993.

[7] D. A. Reynolds and R. C. Rose. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, January 1995.

[8] M. Zamalloa, G. Bordel, L. J. Rodriguez, and M. Peñagarikano. Feature Selection Based on Genetic Algorithms for Speaker Recognition. In *IEEE Speaker Odyssey: The Speaker and Language Recognition Workshop*, pages 1–8, Puerto Rico, June 2006.

[9] M. Zamalloa, G. Bordel, L. J. Rodríguez, M. Peñagarikano, and J. P. Uribe. Using Genetic Algorithms to Weight Acoustic Features for Speaker Recognition. In *Proceedings of the ICSLP'06*, Pittsburgh (USA), September 2006.