Local Features for Speaker Recognition*,**

Roberto Paredes, Enrique Vidal, and Francisco Casacuberta

Instituto Tecnológico de Informática
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera, s/n, 46022 Valencia, Spain

Abstract. An approach combining a *simple local representation method* with a k-nearest neighbors-based *direct voting scheme* is proposed for speaker recognition. This approach rises computational problems that we effectively solved through an approximate fast k-nearest neighbors search technique. Experimental results with the EuTrans and SIVAspeech databases are reported showing the effectiveness of the proposed approach.

Keywords: Speaker Recognition, Local Features, Nearest Neighbor

1 Introduction

Biometric identification (BI) is presently one of the most active areas in pattern recognition. The local features (LF) approach have been used before to solve problems related to the BI field satisfactorily [1, 2].

Local features are explicitly cited mainly in the image database retrieval literature [5–7], where invariances to scale, rotation and illumination changes are needed. A good BI system has to show a high degree of robustness with respect to the *variability* that the biometrics signals use to reflect. Usually, the biometric signal considered to identify a person can be very different of those used to build the identification system. On the other hand, these signals are often composed by several small objects that bear discriminative information by themselves, almost independently of the other parts. In this context the use of the LF approach is clearly recommended.

Among BI systems *speaker recognition* is one of the most unobtrusive methods, well tolerated by users, and with a wide range of applications. In this paper we propose a combination of a simple local representation along a specially devised feature extraction technique and a direct decision scheme based on k-nearest neighbors. This approach has shown to be most effective if the set of local-feature vectors from the training data is large. This entails a high computational cost, which is avoided by using an approximate fast k-nearest neighbors search technique.

* Work partially supported by the Spanish "Ministerio de Ciencia y Tecnología" under grants TIC2003-08496-C04-02 and DPI2001-0880-CO2-02.

** The authors would like to thank the FUB - Fondazione Ugo Bordoni, for providing the SIVA corpus.

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 1087–1095, 2004.

2 Proposed Approach

2.1 Preprocessing and Feature Extraction

The audio speech signal is sampled and represented into vector sequences of Mel-cepstral coefficients. To this end, short-term spectral analysis is performed on short overlapping signal segments (frames). The resulting power spectrum is warped according to the Mel-scale and 11 cepstral-coefficients are derived from each log power spectrum [3].

Figure 1 shows a speech sentence. In the top of the figure, the speech signal of an utterance is represented in the temporal domain. In the middle, the corresponding spectrogram is shown, with frequency presented in the vertical axis, and time in the horizontal axis. Finally, the cepstral coefficients appear in the bottom of the figure. The coefficient number is presented in the vertical axis, and time in the horizontal axis. These cepstral-coefficient vectors are normalized between 0 and 255 in all the experiments carried out in this work.

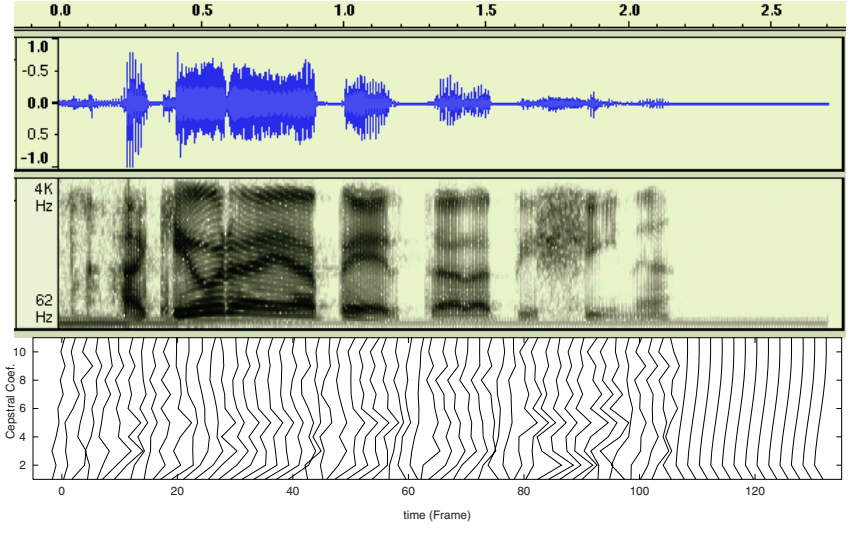

Fig. 1. Top, the original speech signal of a utterance. Middle, the corresponding spectrogram. Bottom, the temporal evolution of the Mel-cepstral coefficients.

Preprocessing consists of two steps. The first step aims at *selecting* those parts of the speech signal with high information content. We have chosen a simple and fast method: the local variance in a small window of the speech signal is measured. Those parts of the signal having local variance above a certain global threshold are selected. The proposed approach based on local variance is applied to these cepstral-coefficient vectors. Around each selected vectors a small window of size w is applied, obtaining a local representation vector of $11 \times w$ cepstral-coefficient components.

The second step aims at reducing the resulting dimensionality using *principal component analysis* (PCA). In this way, a compact local representation of a region of the object is obtained. Finally, we label each vector with an identifier of the class, i.e. the person who uttered the speech sentence considered.

In a classical classifier, each object is represented by a single feature vector, and a discrimination rule is applied to classify a test object that is also represented by a single feature vector. Local features approach, however, implies that each test object is scanned to compute many feature vectors. Each of these vectors can be classified into a different class, and therefore a consensus scheme is required to finally decide a single class for the whole test object.

2.2 Classification through a *k*-NN Based Voting Scheme

The classification procedure used in this work is closely related to a family of techniques often referred to as *"direct voting schemes"* [6]. It is in fact based on applying the well known k-nearest neighbor rule to the set of vectors representing a test utterance, using the local-feature vectors obtained from the training utterances as reference or training set. More formally, we can present the proposed classification technique under the statistical framework of "*classifier combination*" [9].

Let Y be a test speech signal. Following the conventional probabilistic framework, Y can be optimally classified in a class $\hat{w}$ having the maximum posterior probability among all the C classes:

$$\hat{w} = \arg\max_{1 \leq j \leq C} P(\omega_j|Y) \tag{1}$$

By applying the feature extraction process described in the previous section to Y, a set of m_Y feature vectors, $\{\mathbf{y}_1, \ldots, \mathbf{y}_{m_Y}\}$ is obtained. Thus, we can see the classifier (1) as a *combination of* m_Y *classifiers*, each for every feature vector of Y. Assuming independence between each $\mathbf{y}_i$, $P(\omega_j|Y)$ could be written as the product of the posterior probabilities associated to every feature vector and (1) becomes:

$$\hat{w} = \arg\max_{1 \leq j \leq d} \prod_{i=1}^{m_Y} P(\omega_j|\mathbf{y}_i)$$

This is commonly called the "*product rule*" for classifier combination [9].

In order to *smooth* the (poorly estimated) small probabilities the so called "*sum rule*" can alternatively be used for classifier combination [2]:

$$\hat{w} = \arg\max_{1 \leq j \leq d} \sum_{i=1}^{m_Y} P(\omega_j|\mathbf{y}_i) \tag{2}$$

In our case, posterior probabilities are directly estimated by k-nearest neighbors. Let k_{ij} be the number of neighbors of $\mathbf{y}_i$ belonging to class ω_j. Assuming the average number of reference vectors representing all the training local features of each class is more or less constant, a well known estimate of $P(\omega_j|\mathbf{y}_i)$ is:

$$\hat{P}(\omega_j|\mathbf{y}_i) = \frac{k_{ij}}{k}$$

and, using this estimate in (2), our classification rule becomes:

$$\hat{w} = \arg\max_{1\leq j\leq d} \sum_{i=1}^{m_Y} k_{ij} \tag{3}$$

That is, a class $\hat{\omega}$ is selected with the largest number of *"votes"* accumulated over all the vectors belonging to the test biometric signal. This justifies why techniques of this type are often referred to as *"voting schemes"*.

2.3 Efficient Approximate Search for Matching Feature Vectors

The nearest neighbor search is performed by a fast approximate nearest neighbor search algorithm [4]. This algorithm uses a *kd*-tree structure to store the set of local features from the training objects. In a *kd*-tree, the search of the nearest neighbor of a test point is performed starting from the root, which represents the whole space, and choosing at each node the sub-tree that represents the region of the space containing the test point. When a leaf is reached, an exhaustive search of the b prototypes contained in the associated region is performed. Since the closest point may also be a member of some other region, the algorithm needs to backtrack until all possible regions are checked.

If a guaranteed exact solution is not needed, as can be assumed in our case, the backtracking process can be aborted as soon as a certain criterion is met by the current best solution. In [4], the concept of $(1+\epsilon)$-approximate nearest neighbor query is introduced. A point p is a $(1+\epsilon)$-approximate nearest neighbor of q if the distance from p to q is less than $1+\epsilon$ times the distance from p to its nearest neighbor. This concept is used here to obtain an efficient approximate search that can easily cope with very large sets of reference vectors at significantly lower runtime.

3 Experiments

The speaker recognition experiments were carried out with two different corpus, the EuTrans and SIVA. Both experiments were carried out varying the variance threshold t used to decide if a cepstral-coefficient vector is selected as a center of a local feature, the window w of cepstral-coefficient vectors considered, and the PCA dimensionality reduction applied. A lower value of t means a large number of local features obtained from the speech signal. Figure 2 shows the relation between the total number of training local features and the parameter t for EuTrans and SIVA.

The parameter w is related with the dimensionality of the original representation space while the PCA parameter is the final dimensionality of the feature vectors.

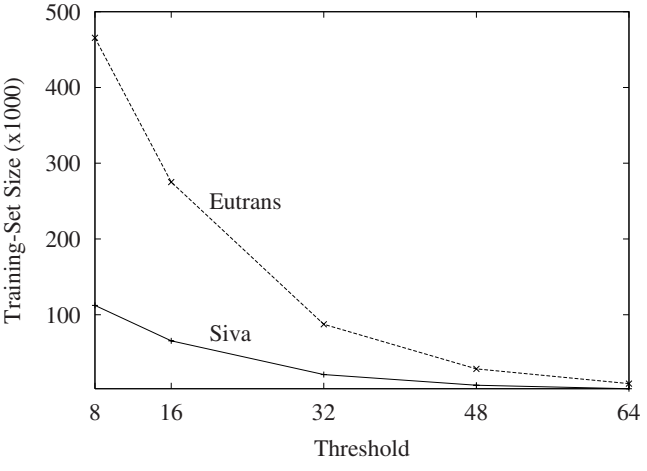

Fig. 2. Training set size with respect to the variance threshold.

3.1 EUTRANS

Experiments were carried out with Italian speech sentences acquired in the EuTrans project [17]. This database has a total of 2,757 sentences and 213 speakers (approximately 7.9h of speech). The database was splited into 2,161 sentences for training and 596 sentences for test. The speech corpus consisted of acquisitions of real phone calls to the front desk of a hotel, simulated using *Wizard of Oz* techniques [16]. This corpus is highly spontaneous and contains many non-speech artifacts. The speech signal was sampled at 8 kHz.

Figure 3 shows the results obtained for $pca = 40$. Similar results were obtained for $pca = 20$ and $pca = 30$. In the experiments the threshold value, t, was varied from 8 to 64. We obtained 465,527 local vectors from the training speech sentences by setting the variance threshold $t = 8$. The window size w, was varied from 5 to 11.

For a window size of $w = 9$ samples and a variance threshold of $t = 8$ the best result is obtained leading to an 85.9% of accuracy. This can be considered a good result taking into account the large number of speakers and the high degree of spontaneity of the corpus. This results show again that this approach is more effective if the set of local-feature vectors obtained from the training data is large.

3.2 SIVA

The Italian speech database SIVA (Speaker Identification and Verification Archives) [18], contains the recordings of more than 2,000 speakers. This corpus consists of four speaker categories: male users, female users, male impostors and female impostors.

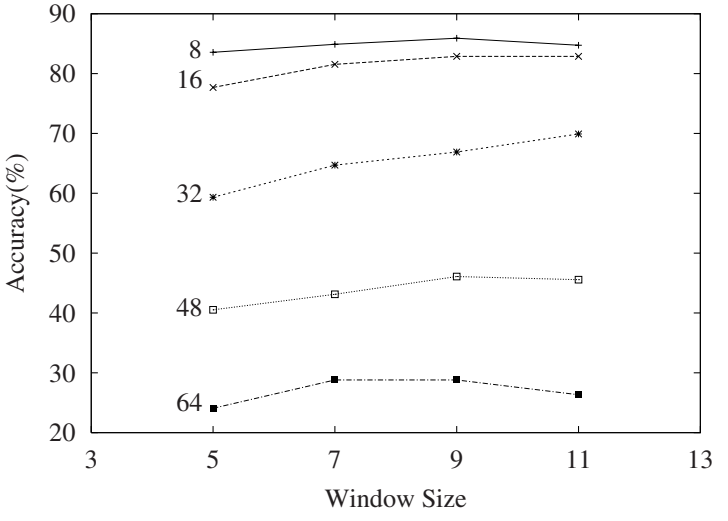

Fig. 3. Accuracy on the EuTrans data base for different window sizes and variance thresholds.

The speakers access the recording system calling a 'toll free' number. An automatic answering system guides them along the three sessions that complete the recording. This corpus has a controlled utterance scenario involving non spontaneous sentences.

For speaker recognition experiments a reduced part of this corpus was selected. We used only utterances from the female and male impostors sets and only of those speakers having more than one utterances. Under these restrictions our final set has 102 different speakers and 612 utterances. We used half utterances for training and the other half error estimation.

In the experiments the threshold value, t, was varied from 8 to 64 and the window size, w, from 5 to 11.

Figure 4 shows the results obtained. As in the previous experiment, it is important to remark that the most important parameter of this approach is the size of the local feature set obtained from the training data, controlled by the variance threshold, t. The best accuracy, 99.7%, is obtained for the largest data set extracted with the variance threshold $t = 8$ and with a window size $w = 7$, using $pca = 40$. Again, similar results were obtained for different pca values tested. This result is significantly better than the results obtained for the EuTrans corpus, mainly it is due to the lower degree of spontaneity of the utterances involved.

Table 1 shows comparative results for the EuTrans and SIVA corpora. The accuracy is reported for different values of the variance threshold, with window sizes $w = 7$ and $w = 9$ for SIVA and EuTrans respectively.

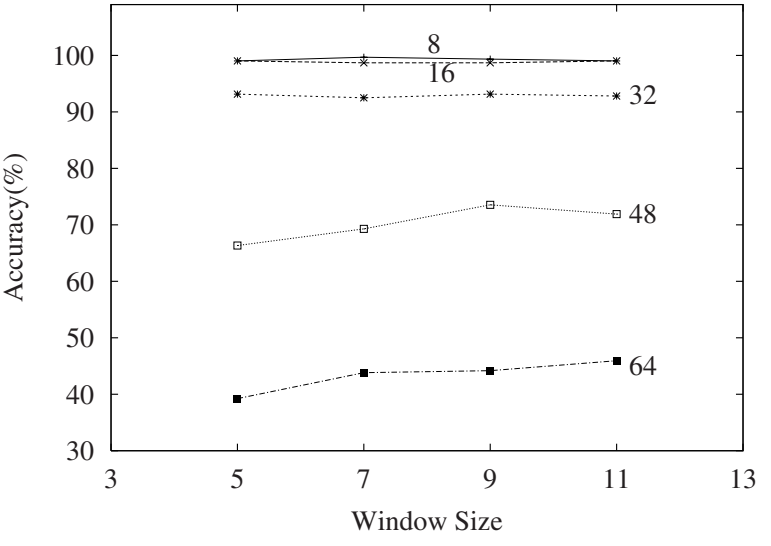

Fig. 4. Accuracy on the SIVA data base for different window sizes and variance thresholds.

Table 1. Comparative results for EuTrans and SIVA corpora. Results in boldface are the best results obtained for each corpus.

Corpus/Threshold	8	16	32	48	64
SIVA $w = 7$	**99.7%**	98.7%	92.5%	69.3%	43.8%
EuTrans $w = 9$	**85.9%**	82.9%	66.9%	46.1%	28.8%

Both experiments show that the most important parameter is the number of local features used to represent the speech sentences. This behavior, under the proposed probability estimation scheme, reflects the importance of using a fast k-nearest neighbors search technique. In our experiments the system takes less than one second to recognize a speaker in a conventional PC computer.

4 Conclusions

A local feature approach is proposed for speaker recognition which combines a simple local representation method with a direct voting scheme based on k-nearest neighbors. The results confirm that the most important parameter is the number of local features extracted from the speech signals. This number of local features is controlled by a variance-threshold parameter in the proposed feature selection approach. Large number of local features implies slow k-nearest neighbors searches, this problem is effectively solved through an approximate fast k-nearest neighbors search technique.

Current work is under way to test the proposed approach on other public-domain databases. We are also interested in studying other voting schemes and local feature extraction methods.

References

1. K. Messer, J. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, J. Czyz, L. Vandendorpe, S. Srisuk, M. Petrou, W. Kurutach, A. Kadyrov, R. Paredes, B. Kepenekci, F. B. Tek, G. B. Akar, F. Deravi, N. Mavity. Face Verification Competition on the XM2VTS Database. *Proc. 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, Guildford, 2003.
2. R. Paredes, J. C. Perez-Cortes, A. Juan, and E. Vidal. Local Representations and a Direct Voting Scheme for Face Recognition. In *Workshop on Pattern Recognition in Information Systems*, Setúbal, Portugal, July 2001.
3. L.R. Rabiner, and R.W. Shafer. Digital processing of speech signals. *Ed. Prentice Hall.* 1978
4. S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *JACM*, 45:891–923, 1998.
5. C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. on PAMI*, 19(5):530–535, 1997.
6. R. Mohr, S. Picard, and C. Schmid. Bayesian decision versus voting for image retrieval. In *Proc. of the CAIP-97*, 1997.
7. C. Shyu et al. Local versus Global Features for Content-Based Image Retrieval. In *Proc. of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 30–34, 1998.
8. R. Deriche and G. Giraudon. A Computational Approach to Corner and Vertex Detection. *Int. Journal of Computer Vision*, 10:101–124, 1993.
9. R.P Duin J. Kittler, M. Hatef and J. Matas. On combinig classifiers. *IEEE Trasn. on PAMI*, 1998.
10. R. Liao and S. Z. Li. Face Recognition Based on Multiple Facial Features. In *Proc. of the 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2000.
11. Z. Zhang, R. Deriche, O. Faugeras, and Q. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78:87–119, 1995.
12. M. Lhuillier and L. Quan. Robust Dense Matching Using Local and Global Geometric Constraints. In *Proc. of ICPR-2000*, volume 1, pages 968–972, 2000.
13. K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The Extended M2VTS Database. In *Second International Conference on Audio and Video-based Biometric Person Authentication*, pages 964–966, March 1999.
14. F. Samaria and A. C. Harter. Parameterisation of a Stochastic Model for Human Face Identification. In *Proc. of the 2nd IEEE Workshop on Applications of Computer Vision*, pages 138–142, 1994.
15. J. Ben-Arie and D. Nandy. A volumetric/iconic frequency domain representation for objects with application for pose invariant face recognition. *IEEE Trans. on PAMI*, 20:449–457, 1998.
16. D. Aiello, L. Cerrato, C. Delogu, and A. Di Carlo. The acquisition of a speech corpus for limited domain translation. In *Proceedings of the European Conference on Speech Communication and Technology*, Budapest, 1999.

17. EuTrans. Example-based language translation systems. Final report. Technical report, Instituto Tecnológico de Informática, Fondazione Ugo Bordoni, Rheinisch Westfälische Technische Hochschule Aachen Lehrstuhl für Informatik VI, Zeres GmbH Bochum: Long Term Research Domain, Project Number 30268, 2000.
18. Falcone M., Gallo The SIVA speech database for speaker verification: description and evaluation *ICSLP'96*, Philadelphia, USA, October, pp. 1902–1905.