# A Real-Time Text- Independent Speaker Identification System*

L. P. Cordella°, P. Foggia°, C. Sansone°, M.Vento[+]
*(°) Dipartimento di Informatica e Sistemistica*
*Università di Napoli "Federico II" - Via Claudio, 21 I-80125, Napoli, Italy*
*{cordel, foggiapa, carlosan}@unina.it*
*([+]) Dipartimento di Ingegneria dell'Informazione e di Ingegneria Elettrica*
*Università di Salerno - Via P.te Don Melillo, 1 I-84084, Fisciano (SA), Italy*
*mvento@unisa.it*

## Abstract

*The paper presents a real-time speaker identification system based on the analysis of the audio track of a video stream. The system has been employed in the context of automatic video segmentation. It uses features evaluated in both domains of time and frequency. Their combined use significantly improved the performance of the system.*

*Experiments have been carried on a database extracted from over one hour of television news, including 10 speakers. The obtained results confirm the effectiveness of the approach, showing an error rate less then 1% when the time interval used for identifying a speaker is about 1.5 seconds.*

## 1. Introduction

During the past few years, much work has been done on the problem of video segmentation, as a basic step for facing the most general problem of indexing and retrieval by content. Segmentation implies, at a preliminary stage, the partition of the video into footage segments, i.e. sequences of frames obtained by detecting abrupt transitions, typically associated to camera changes.

Up to now, the majority of the approaches to video segmentation has been based on the extraction of information from the video frames, using image analysis techniques. Recently, the relevance of audio as an alternative source of information for video segmentation, has been demonstrated [1-5].

This is particularly true in some domains, e.g. the TV news, where the audio track is very effective for correctly segmenting the video stream. In some cases, the audio track information allows us to perform a correct segmentation, while the video track information is not adequate.

Let us consider the case in which the images of a television news report flow on the screen, while the voice of an anchorman, not visible on the screen, comments on them. In this case, the analysis based only on the visual content would generally fail to consider the report as a unique segment. On the contrary, a system that analyzes the audio track and identifies the speaker could provide a useful mean to properly segment the video.

Starting from these considerations, in this paper we propose a real-time speaker identification system that can be used for automatically segmenting TV news.

Speaker identification is a special case of the more general problem of speaker recognition. In case of speaker identification the goal is to determine which one among a group of known voices best matches the input voice samples. If it is not required that the speaker to be identified pronounces a specific set of phrases, as in our case, the system is said to be *text-independent*.

The speaker identification problem has been addressed in the literature by representing the audio signal using different features, calculated in the frequency or in the time domain. Moreover, different classification paradigms have been employed, basically neural networks and gaussian mixture models.

The paper by Reynolds and Rose [6] illustrates one of the most frequently quoted systems. It uses features evaluated in the frequency domain by using the cepstral analysis [7] (the so-called Mel-scale cepstral coefficients). Different classification paradigms are proposed and the best results are obtained with a gaussian mixture model. The system achieved over 94% recognition rate on the KING Speech Corpus database [8].

Several authors have proposed features calculated starting from the ones described by Reynolds and Rose. In particular, in [9] a transform was applied to the Mel cepstral features in order to compensate the noise components of the audio channel. From the transformed domain, the so-called *formants* features were then calculated and used for classification. Results on the NIST 95 database provided a 90% recognition rate.

In [10] the authors used also the wavelets. They

* This work has been partially supported by the Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) in the framework of the FIRB Project "Middleware for advanced services over large-scale, wired-wireless distributed systems (WEB-MINDS)".

0-7695-1948-2/03 $17.00 © 2003 **IEEE**

proposed a system architecture with a neural classifier for each speaker to be recognized. Identification was performed by assigning the input sample to the speaker whose associated classifier exhibited the highest output.

In order to reduce the computational complexity of the classification phase, in [11] the principal component analysis was used on the features proposed by Reynolds and Rose. In this way, the authors achieved a 90% recognition rate on a set of 50 speakers.

Other authors have proposed the use of acoustic features directly obtainable from the time domain, such as pitch, speech rate, voice quality and temporal variation of the audio signal. This is the case of [12], where a system devoted to recognize only voiced segments of the audio track is proposed and of [13], in which the authors try to identify the speakers of the KING Speech Corpus. In both cases Gaussian mixture models were employed at the classification stage.

The system we propose, starting from the one proposed in [6], uses features calculated both in time and frequency domains. In order to find the best configuration to be used in a real-time system, we performed an experimental analysis by considering three sets of features, three different training set sizes and two neural classifiers. The combined use of all the proposed features significantly improved the performance of the system with respect to the use of a single set of features; this happened for each training set size and for each classifier architecture.

On a database of 10 speakers, the system we propose is able to reach a 99% recognition rate when a time period of about 1.5 seconds is used to assign the audio signal to one of the considered speakers. This time period is surely reasonable for building a real-time application suitable for the considered domain (i.e. TV news).

The rest of the paper is organized as follows: in the next Section, we present the architecture of the proposed system together with a description of the features used. Section 3 is devoted to the description of the audio database. Finally an experimental analysis of the system is presented in Section 4.

## 2. System Architecture

A sketch of the system is shown in Fig. 1.

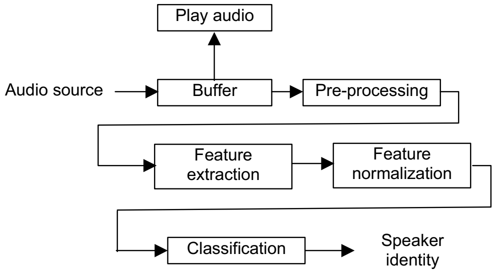


**Figure 1. Block diagram of system architecture.**

In the following subsections, some details will be provided on the basic processing blocks.

### 2.1. Pre-processing

After the input signal has been acquired (e.g. by using a microphone, or from the audio portion of an MPEG stream) it is stored in a circular buffer, needed to overlap the acquisition with the processing and, optionally, with the play-back of the signal. The buffered samples are then fed to a pre-processing module, whose aim is to prepare the data for feature extraction. The pre-processing phase consists of three operations: *frame blocking*, *pre-emphasis* and *windowing.*

*Frame-blocking:* the first step of the signal processing chain is the partition of the audio samples into fixed-length partially overlapping frames. The overlap between adjacent frames is used to smooth the frame-to-frame transitions and to provide a better handling of the correlation existing between successive parts of the voice signal.

The frame length has been fixed to about 23 msec, yielding 1024 samples at a 44.1 KHz sampling rate. This choice has been made because several studies [7] suggest that the most adopted features can provide significant information over an interval of about 30 msec, and the FFT used during feature extraction requires that the number of samples is a power of 2.

The overlap between two adjacent frames has been fixed to one third of the frame length.

*Pre-emphasis*: the voice signal has a limited band (80 Hz to 5 KHz). Thus, the use of a low-pass filter can reduce significantly the high-frequency components that are due to environmental noise or to the presence of non-vocal sounds in the audio source, enhancing the signal-to-noise ratio.

*Windowing*: in order to reduce the effect of the discontinuities at the boundaries of a frame, a windowing operation is used to give a greater weight to the samples that are in the central segment of the frame. In particular, we have used a Hamming window.

### 2.2. Feature extraction

After filtering and windowing, the frames undergo the process of feature extraction. We have adopted three different feature sets: Linear Predictive Cepstral Coefficients, Post Filter Linear and Mel Filtered Cepstral Coefficients. 14 coefficients are computed from each feature set, obtaining a feature vector made up of 42 components.

*Linear Predictive Cepstral Coefficients (LPCC)*: they are obtained by means of a linear prediction analysis. We used the Levinson-Durbin method [14] in order to calculate them.

*Post Filter (PF)*: the Post Filter coefficients [15] are also based on a linear prediction analysis. They have been used in order to improve the performance obtainable with the LPCC at the low frequencies.

*Mel Filtered Cepstral Coefficients. (MFCC)*: the Mel Cepstrum coefficients [6] are computed by means of the inverse Fourier transform applied to the logarithm of the input signal spectrum.

### 2.3. Feature normalization

Before being passed to the classification stage, the feature vector needs to be normalized in order to avoid that the classifier gets biased towards only a subset of the features. In particular, the classifier that we have used is based on the Euclidean distance between feature vectors, so it is important that each feature lies in a range having the same order of magnitude than the others. For this purpose the components of the feature vector are scaled using a set of fixed coefficients that have been determined off-line using a subset of the training samples.

### 2.4. Classification

In our system the classification step is performed by means of an LVQ neural classifier trained with the FSCL algorithm [17]. In particular we have tested networks having 50 and 100 prototypes per class respectively. We have trained the networks under the assumption of “closed world”, that is each input is classified as belonging to one of the assigned classes. However, the LVQ architecture would make relatively easy adding a reject option [16], to correctly deal with inputs that are extraneous to the speakers used for training.

For each audio frame, the classifier outputs the most likely speaker identity. However, because of the scarce reliability that can be reached on certain phonemes (that are difficult to associate to a speaker even for a human listener), our system combines the evidence coming out of a fixed-length sequence of frames, that we call a *shot*, before issuing its final response.

## 3. The Database

Since our aim is to use the proposed speaker identification system in the context of automatic segmentation of video news, the most commonly used standard and publicly available database are not adequate to test it. This is the case, for example, of the above-cited KING Corpus [8], which is made up of audio segments captured from telephone talks.

Therefore, we have built a database starting from audio segments extracted from twelve different Italian TV news. Ten different speakers were considered, five male and five female. This number was considered sufficient to represent the variability of the speakers in a given TV network. We assumed that the characteristics of the audio signal of a given speaker were the same, even if the samples were extracted from videos recorded in different days.

Each audio segment in the database has 15 seconds duration. The audio signal has been digitized at 44.1 kHz rate, using 16 bits per sample. Twenty-five segments were extracted for each speaker, thus the whole database has about 1h and 2 min duration. Since the features described in the previous section are calculated over a time interval of 23 msec, the number of feature vectors available in our database is about 470,000; this number is comparable with those of the databases typically used in the literature.

Finally, it is worth noting that, differently from audio material manually recorded, in our case it was not possible to equalize the audio signal, whose characteristics significantly depend on the source used to acquire it.

## 4. Experimental Results

The above-described audio database was first used to assess the performance of the system when classifying audio shots of length ranging from 0.5 to 5 seconds. Then, the best system configuration was chosen to build a real-time application.

In order to perform a classification on an audio shot of fixed length, a majority voting approach was used. All the classification results (votes) on the single feature vectors composing the shot are first collected; then the system attributes the shot to the speaker who has obtained the higher number of votes. This approach allows us to overcome the problem of the “unvoiced segments” [9].

Different training sets (TRS) were used in order to establish the influence of the training set size on the obtainable results. In particular, three different sizes were considered, made up of one (15 sec), two (30 sec) and three (45 sec) audio segments per speaker respectively. Furthermore, both the LVQ classifiers with 50 and 100 prototypes per class were tested.

In all the tests the recognition rate was averaged over five different experiments. The size of the test set (TS) was always fixed to 30 seconds; obviously in each experiment the TS samples were different from those used for training the selected classifier.

Three different sets of features have been considered: the MFCC alone, the LPCC alone, and the MFCC together with LPCC and PF features.

The average results obtained by the system on the TS, considering the only MFCC feature set, are reported in Fig. 2. It can be noted that the error rates are not particularly satisfactory; moreover, the recognition slightly improves as the shot length increases. The use of 100 prototypes per class gives better results with respect to the use of 50 prototypes per class. Surprisingly enough, the results on the TS are better if a smaller TRS is used. This could be explained by considering that an overfitting of the TRS data can occur when the size of the TRS increases.

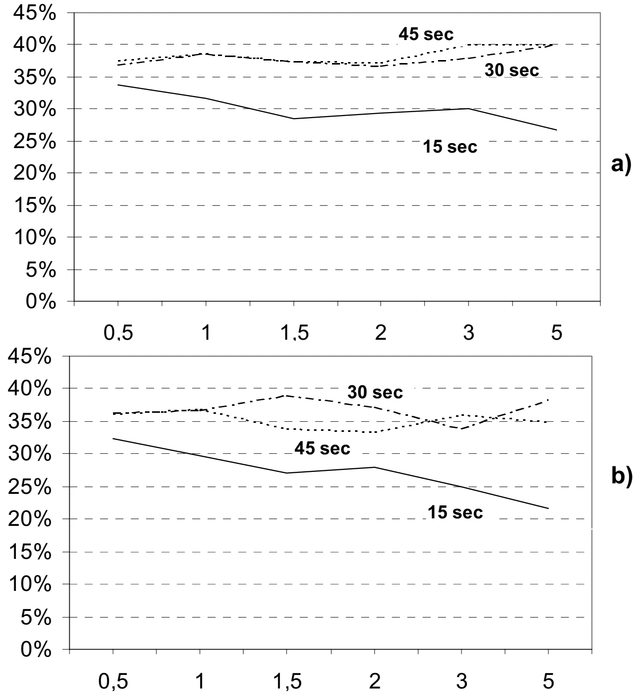


**Figure 2: The error rate on the TS as a function of the shot length (in seconds) for three different TRS sizes. Classification has been performed by means of an LVQ net with a) 50 and b) 100 prototypes per class. Only the MFCC feature set was used.**

By considering only the LPCC feature set (see Fig. 3) the behavior of the system is quite different. The recognition rate is higher than the previous case, and the performance of the system improves passing from a TRS of 15 sec to a TRS of 30 sec. However, an overfitting phenomenon can be observed also in this case, when the TRS size is of 45 sec. In this case, the recognition rate increases as the size of the audio samples becomes longer, as it is expected.

This behavior encourages the combined use of MFCC, LPCC and PF features: by augmenting the feature space the overfitting phenomenon should disappear, while the different properties of the considered feature sets should improve the recognition performance of the system.

This is confirmed by the results shown in Fig. 4, where the MFCC features are used together with LPCC and PF features.

As it can be noted, the recognition performance of the system is now very interesting. A recognition rate over 99% is reached when the length of the audio shots becomes greater than 1.5 seconds. The results obtained with a TRS of 30 sec are now very similar to those obtained with a TRS of 45 sec, as well as the performance achieved by using an LVQ with 50 prototypes is quite the same of that obtained with 100 prototypes.

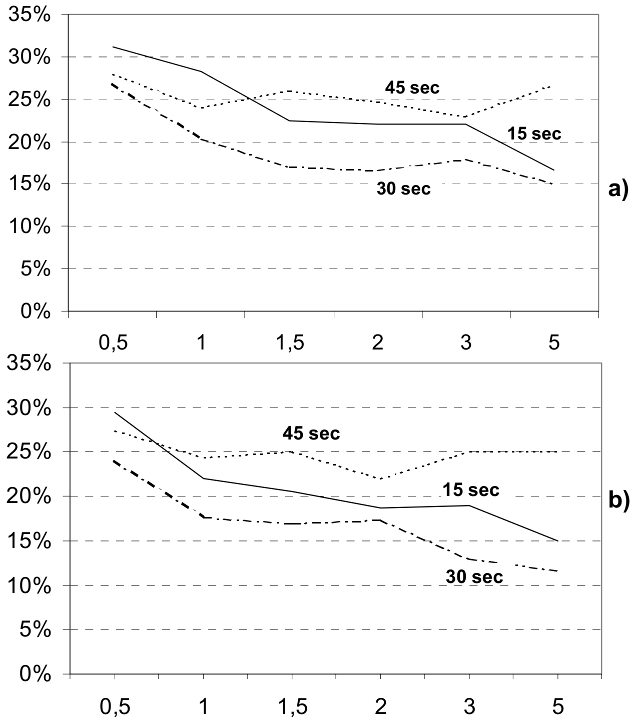


**Figure 3: The error rate on the TS as a function of the shot length (in seconds) for three different TRS sizes. Classification has been performed by means of an LVQ net with a) 50 and b) 100 prototypes per class. Only the LPCC feature set was used.**

Starting from these results, a real-time speaker identification application was built, based on the system trained with a TRS of 30 sec per speaker, and using an LVQ classifier with 50 prototypes. Figs. 5 and 6 show a snapshot of the realized application, that permits to choose the shot length used to recognize the speaker identity.

Moreover, in the real-time application we also implemented the computation of a reliability measure of each performed classification. For each sample to be classified, the reliability $R$ is measured as follows:

$$R = 100 * (1-N_2/N_1)$$

where $N_1$ is the number of feature vectors attributed to the most likely speaker and $N_2$ is the number of feature vectors attributed to the runner-up. If the value of $R$, which ranges from 0 to 100, is greater or equal to 50, the audio sample classification is considered reliable (see Fig. 6).

In order to test the real-time system, one further experiment was performed. A video of 2 min and 30 seconds was created, by collecting five TV news reports of 30 seconds each. Every report is presented by a different speaker, obviously belonging to the set on which the system was trained. By setting the length of the audio shots to 1 sec, the system achieved a recognition rate of 96.46%. However, if we consider only reliable classifications, the recognition rate reached the 99.79%.

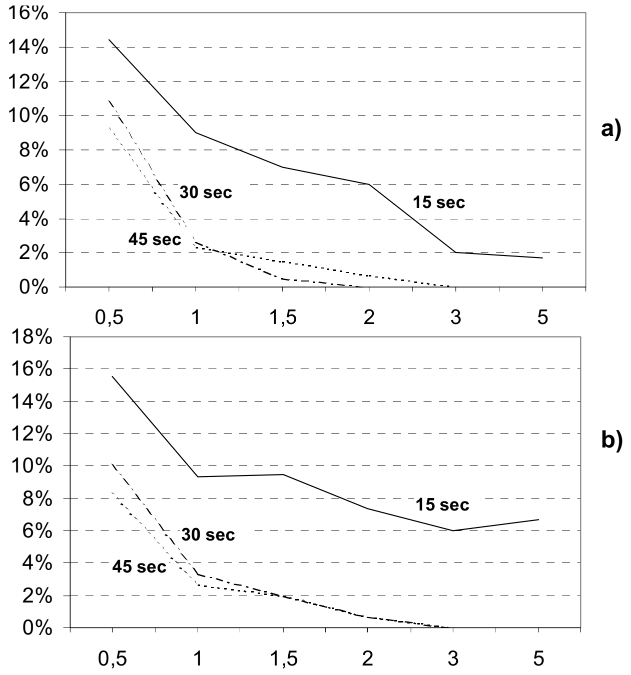


**Figure 4: The error rate on the TS as a function of the shot length (in seconds) for three different TRS sizes. Classification has been performed by means of an LVQ net with a) 50 and b) 100 prototypes per class. MFCC, LPCC and PF features were used.**

## 5. Conclusions

In this paper we presented a real-time system for speaker identification that uses features extracted from both the frequency and the time domain. By using a majority voting approach, when classifying an audio frame, it is able to cope with the presence of "unvoiced segments".

Experimental results on a database of audio segments extracted from TV news demonstrate the effectiveness of the system in identifying speakers in real-time.

Such system can also be of great help in implementing an application that uses the information connected to the audio track for automatically segmenting a video stream.

## References

[1] J. Nam, E. Cetin, A. Tewfik, "Speaker Identification and Analysis for Hierarchical Video Shot Classification", *IEEE Int. Conf. on Image Processing*, Santa Barbara, CA, 1997, vol. 2, pp. 26-29.

[2] C. Saraceno, R. Leonardi, "Audio as a Support to Scene Change Detection and Characterization of Video Sequences", *Proc. of ICASSP*, Munich, 1997, vol. 4, pp. 2597-2600.

[3] S. Tsekeridou, I. Pitas, "Speaker Dependent Video Indexing Based on Audio-Visual Interaction", *IEEE Int. Conf. on Image Processing*, Chicago, IL, 1998, vol. 1, pp. 358-362.

[4] J. Foote, "An Overview of Audio Information Retrieval", *ACM Multimedia Systems*, vol. 7, 1999, pp. 2-10.

[5] M. De Santo, G. Percannella, C. Sansone, M.Vento, "Cooperating Experts for Soundtrack Analysis of MPEG Movies", *Information Fusion*, vol. 3, no. 3, 2002, pp. 225-236.

[6] D.A. Reynolds, R.C. Rose, "Robust text-independent speaker identification using Gaussian Mixture speaker models", *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, 1995, pp. 72 -83.

[7] J. Picone, "Signal Modeling Techniques in Speech Recognition", *IEEE Proceedings*, vol. 81, no. 9, 1993, pp. 1215-1247.

[8] J. Godfrey, D. Graff, A. Martin, "Public databases for speaker recognition and verification", *Proc. of the ESCA Workshop Automatic Speaker Recognition, Identification, Verification*, 1994, pp. 39-42.

[9] H.A. Murthy, F. Beaufays, L.P. Heck, M. Weintraub, "Robust text-independent speaker identification over telephone channels", *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 5 , 1999, pp. 554 -568.

[10] H.M. Torres, H. Rufiner, "Automatic speaker identification by means of Mel cepstrum, wavelets and wavelet packets", *Proc. of the 22nd Annual IEEE Intern. Conf. on Engineering in Medicine and Biology*, vol. 2, 2000, pp. 978–981.

[11] C. Seo, K.Y. Lee, J. Lee, "GMM based on local PCA for speaker identification", *Electronics Letters* , vol. 37, no. 24, 2001, pp. 1486 -1488.

[12] C. Miyajima, Y. Hattori, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Speaker identification using Gaussian mixture models based on multi-space probability distribution", *Proc. of ICASSP*, vol. 1, 2001, pp. 433 -436.

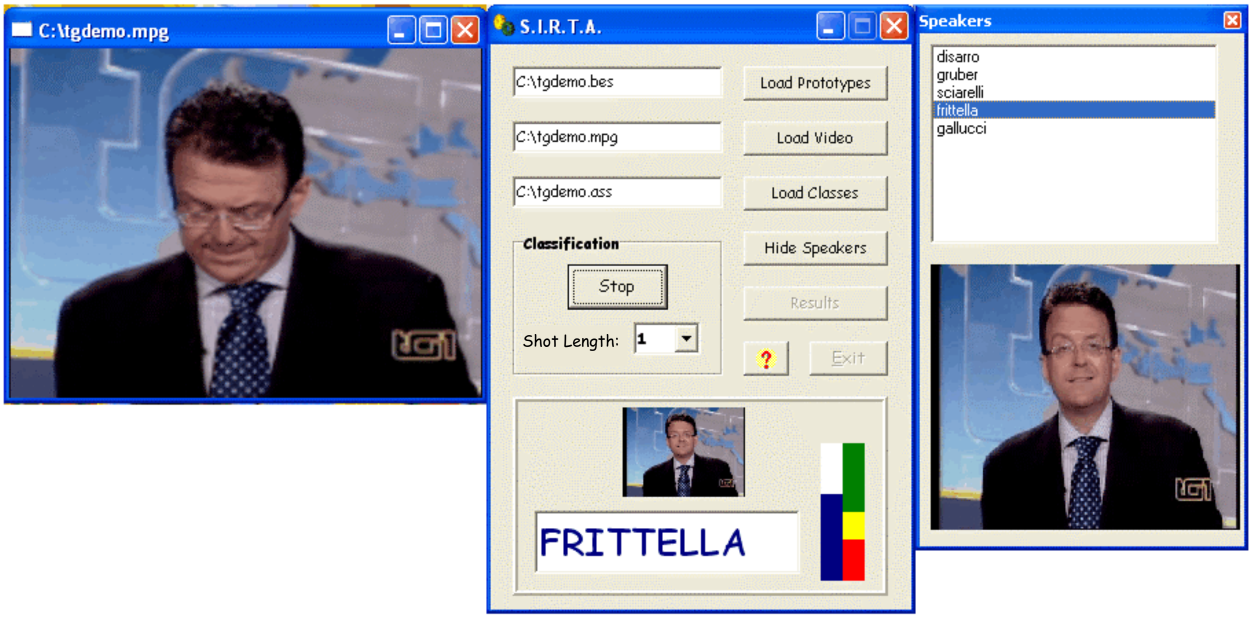
[13] L. Wang, K. Chen, H. Chi, "Capture interspeaker information with a neural network for speaker identification", *IEEE Transactions on Neural Networks*, vol 13, no. 2, 2002, pp. 436 -445.

[14] M.H. Hayers, *Statistical Digital Signal Processing and Modeling*, John Wiley & Sons Inc. 1996

[15] R. Mammone, X. Zhang, R. Ramachandran, *Robust speaker recognition*, IEEE Signal Processing Magazine, September 1996, pp. 58-71.

[16] L.P. Cordella, C. Sansone, F. Tortorella, M. Vento, C. De Stefano, "Neural Network Classification Reliability: Problems and Applications", in *Image Processing and Pattern Recognition*, Academic Press, San Diego, 1998, pp. 161-200.

[17] S.C. Ahalt, A.K. Krishnamurthy, P. Chen, D.E. Melton, "Competitive Learning Algorithms for Vector Quantization", *Neural Networks*, vol.3, 1990, pp. 277-290.



C:\tgdemo.bes

C:\tgdemo.mpg

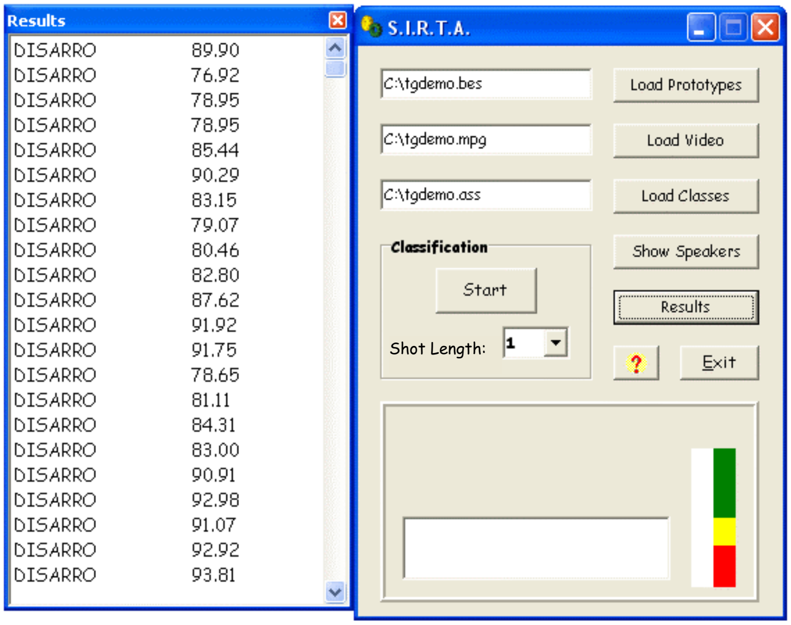
C:\tgdemo.ass

Shot Length: 1

disarro
gruber
sciarelli
frittella
gallucci

FRITTELLA

**Figure 5: A snapshot of the real-time speaker identification application. Five speakers have been considered (see right window), and a TV news made up of 2 min and 30 sec has been used in this experiment. The MPEG video is played in the left window. The shot length for classification was set to 1 sec, as shown in the center window. In the same window it can be noted that, if the bar on the left overcomes the level indicated by the yellow bar (on the right), as in this case, the speaker identification is considered reliable.**



```
DISARRO        89.90
DISARRO        76.92
DISARRO        78.95
DISARRO        78.95
DISARRO        85.44
DISARRO        90.29
DISARRO        83.15
DISARRO        79.07
DISARRO        80.46
DISARRO        82.80
DISARRO        87.62
DISARRO        91.92
DISARRO        91.75
DISARRO        78.65
DISARRO        81.11
DISARRO        84.31
DISARRO        83.00
DISARRO        90.91
DISARRO        92.98
DISARRO        91.07
DISARRO        92.92
DISARRO        93.81
```

C:\tgdemo.bes

C:\tgdemo.mpg

C:\tgdemo.ass

Shot Length: 1

**Figure 6: By clicking on the 'Results' button, the system allows us to examine the obtained classification results, together with their reliability. In this experiment, the recognition rate on the five speakers was 96.46%; but if we consider only reliable classifications (i.e. classifications with a reliability value greater or equal to 50.00), the recognition rate reached the 99.79%.**