Speaker Identification for Audio Indexing Applications

Sofia Tsekeridou and Ioannis Pitas

Dept. of Informatics, Aristotle Univ. of Thessaloniki,
Box 451, Thessaloniki 54006, GREECE
Tel: +30 31 996304, Fax: +30 31 996304, e-mail: pitas@zeus.csd.auth.gr

ABSTRACT

A method for identifying different speakers from an audio source of continuous speech is described in this paper aiming at extracting the speaker sequence, timing information and speaker identity label versus time diagram used successively for audio indexing. This approach enables further speaker specific content-based audio retrieval or browsing. Speaker classification is achieved by a variant of an LVQ3 classifier and speaker recognition is attempted in entire speech segments by evaluating a Mahalanobis distance measure. Speaker transitions are, in most cases, correctly detected and speaker recognition errors are limited in the range 2%-5% considering speech frames only.

I. INTRODUCTION

Content-based audio/video indexing and retrieval have recently grown to active research topics due to multimedia requirements for flexible audio/video manipulation [6, 7, 8, 9]. Since audio and/or video sources contain rich information, both tasks are complicated and usually focus on specific applications, e.g topic discrimination [7], news transcription [8], keyword spotting [7], speaker recognition [6], keyframe extraction [9]. Content-based audio/video indexing precedes retrieval and uses either low-level audio and/or visual features or semantic features based on objects and events. Retrieval exploits audio/video indexing to answer queries imposed by the end-user based on audio/video content.

A content-based audio indexing scheme is described in this paper that aims at temporally indexing an audio source of continuous speech according to the actual speaker. Prior to speaker recognition, discrimination between silence and voice frames is performed so that decisions on speaker identities are made to voice frames only. Each speaker is labelled by a number $1, .., k$. An LVQ3 classification algorithm is used to create speaker models based on apriori known training periods of continuous speech where a certain speaker speaks. After testing performed in entire speech segments, a speaker identity label versus time diagram is extracted. Such a task enables flexible audio retrieval in cases where queries according to speaker identities are imposed. If such a task was performed manually, it would be a very complex and time consuming task.

An overview of the voice-silence discrimination process as well as preprocessing techniques to the initial continuous speech source are described in Section II. Speaker codebooks are generated by a modified LVQ3 classifier in the way presented in Section III. Speaker recognition is accomplished by evaluating Mahalanobis distance measures from all speaker classes in entire speech segments and deciding on the one corresponding to the minimum distance (Section IV). Section V summarizes the simulation results obtained after evaluating the performance of the proposed scheme to the audio source of various news sequences with various speakers speaking at different time periods. In some cases, background noise or the nature of continuous speech are two aspects that limit the success of speaker recognition based on audio information. Furthermore, errors in silence frame estimations lead to further recognition errors. Recognition results exhibit the satisfactory performance of the audio indexing scheme. In many cases, it is also observed that speaker transitions are correctly detected, despite the fact that no effort towards this goal has been made. Finally, in Section VI conclusions are drawn on the performance of the speaker dependent audio indexing scheme and future work is discussed.

II. CONTINUOUS SPEECH PROCESSING

The continuous speech processing and feature extraction module prior to speaker modelling and recognition is illustrated in Figure 1. A more detailed de-

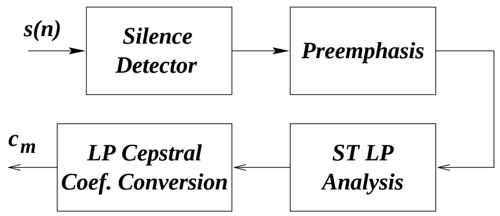

Figure 1: Audio processing and audio feature extraction module

scription about each sub-module is given in the successive sub-sections. In the following, the term *frame* denotes the minimum duration of speech used by the feature extraction module, whereas the term *segment* refers to a variable number of successive speech frames grouped together due to having a similar attribute (e.g. all being voice frames, thus creating a speech segment).

A. Silence Detection and Preemphasis

Before extracting audio feature vectors to be used by the speaker recognition classifier, *silence detection* is performed to discard silence periods from the initial audio source of continuous speech containing silence intervals. Silence frames are audio frames of background noise with a low energy level with respect to voice segments. The signal energy in each speech frame is evaluated by:

$$E_i = \sqrt{\sum_{k=1}^{M} x_i(k)^2}, \qquad i = 1,..,N \tag{1}$$

where M is the number of samples in a speech frame and N is the total number of speech frames. Thresholding is successively performed to detect silence frames with a threshold level equal to $E_{min}+0.1(E_{max}-E_{min})$. E_{min} and E_{max} are the lowest and greatest values of the N segment energy values respectively. The silence detector performs satisfactorily achieving to also approximately detect the boundaries between phonemes, words or phrases.

The speech signal is successively preemphasized by a first-order FIR filter with transfer function $H(z) = 1 - 0.95z^{-1}$. The main purpose of preemphasis is to spectrally flatten the speech signal, that is to increase the relative energy of its high-frequency spectrum [1]. Furthermore, preemphasis ensures prevention of numerical instability in case that Linear Predictive (LP) analysis follows.

B. Audio Feature Extraction

Audio feature extraction is performed only in voice frames [2]. The preemphasized audio signal is processed in order to obtain the LP cepstral coefficients in every speech frame. Speech frames of 30msec are extracted with an overlap of 20msec from each other. Each frame is windowed by a Hamming window of size M in order to minimize the signal discontinuities at the beginning and end of it. Short Term (ST) LP analysis succeeds and the LP parameter set is later on converted to LP cepstral coefficients. These coefficients have proven to be a more robust parameter set for speaker recognition than the LP coefficients themselves. Given the LP parameter set $\alpha_m, 1 \leq m \leq p$ (p: order of LP model) of an audio frame, the LP cepstral coefficients are defined by:

$$c_m = \alpha_m + \sum_{k=1}^{m-1} (\frac{k}{m}) c_k a_{m-k}, \qquad 1 \leq m \leq p \tag{2}$$

III. SPEAKER CLASSIFICATION

After the generation of a sequence of p-dimensional audio feature vectors $\mathbf{c}(i), i = 1,..,N$ from the entire audio source, *speaker modelling* is attempted by an LVQ3 classifier [3, 4]. Vector Quantization based speaker recognition is a conventional and successful method even with a limited number of training vectors. All speaker codebooks are trained in parallel using training vectors from apriori known time periods in which a certain speaker speaks and misclassification cases for the training vectors are compensated for. Initialization of the codebooks, each having 32 codevectors, is performed by the LBG algorithm. Classification of a training vector $\mathbf{c}$ to a speaker codebook $\mathbf{m}_i$ is performed judging from Mahalanobis codevector distance values $d(\mathbf{m}_{ik}, \mathbf{c})$, $k = 1 \cdots 32$:

$$d(\mathbf{m}_{ik}, \mathbf{c}) = (\mathbf{m}_{ik} - \mathbf{c})^T \mathbf{R}_i^{-1} (\mathbf{m}_{ik} - \mathbf{c}) \tag{3}$$

where $\mathbf{m}_{ik}$ represents the k-th codevector of the speaker codebook $\mathbf{m}_i$ and $\mathbf{R}_i$ denotes the global covariance matrix derived from the training vectors of the speaker codebook $\mathbf{m}_i$. The codevectors of every class that were assigned the minimum codevector distances are selected. Finally, the training vector $\mathbf{c}$ is classified to the speaker codebook $\mathbf{m}_i$ if the respective codevector distance is the minimum of all others.

In succession, correct classification and misclassification cases are appropriately used for speaker codebook updating and total distortion minimization. If a training vector $\mathbf{c}$, known to belong to the speaker codebook $\mathbf{m}_j$, is classified at iteration t to the speaker codebook $\mathbf{m}_i$ and the training vector $\mathbf{c}$ lies in a predefined "window", then updating of the codebooks of both speaker classes at iteration $t+1$ is performed in the following way:

$$\begin{aligned} \mathbf{m}_i(t+1) &= \mathbf{m}_i(t) - \alpha(t)[\mathbf{c} - \mathbf{m}_i(t)] \\ \mathbf{m}_j(t+1) &= \mathbf{m}_j(t) + \alpha(t)[\mathbf{c} - \mathbf{m}_j(t)] \end{aligned} \tag{4}$$

A training vector $\mathbf{c}$ is assumed to lie in the "window" if $\frac{s_j - s_i}{s_j} > w$ [5], where w is a constant set equal to 0.25 and s_j and s_i are the Mahalanobis distances of vector $\mathbf{c}$ from the true codevector of the codebook $\mathbf{m}_j$ that it belongs to and from the codevector of the codebook $\mathbf{m}_i$ that it has been classified to respectively. In the case that the training vector $\mathbf{c}$ is classified correctly to the codebook $\mathbf{m}_j$ then updating at iteration $t+1$ is performed by:

$$\mathbf{m}_j(t+1) = \mathbf{m}_j(t) + \epsilon\alpha(t)[\mathbf{c} - \mathbf{m}_j(t)] \tag{5}$$

It is noted that updating is performed only to the true codebook $\mathbf{m}_j$. $\alpha(t)$ is a linearly decreasing function depending on the number of iterations which controls the learning rate. The constant ϵ is set usually to a small value (0.25 in our case). Training is performed iteratively until the overall average distortion falls below a small value or the maximum number of iterations has been reached.

IV. SPEAKER RECOGNITION

During the recognition process, the distortion $s_i(\mathbf{c}_t)$ of an input test vector $\mathbf{c}_t$ when it is quantized using the speaker codebook $\mathbf{m}_i$ is calculated by:

$$s_i(\mathbf{c}_t) = \min_{\mathbf{m}_{ik} \in \mathbf{m}_i} [d(\mathbf{m}_{ik}, \mathbf{c}_t)], k = 1, .., 32 \quad (6)$$

In the above equation $\mathbf{m}_{ik}$ denotes the k-th codevector of the $\mathbf{m}_i$ codebook and the index i spans the number of speaker classes. $d(\mathbf{m}_{ik}, \mathbf{c}_t)$ is evaluated by (3). The global covariance matrix $\mathbf{R}_i$ has been evaluated during the training process.

In our approach, instead of assigning each input test vector separately to a speaker class, we use an entire speech segment consisting of a variable number of speech frames. A speech segment is determined as a successive sequence of speech frames, that is, cepstral vectors, lying between two silence frames. A coarse detection of words or phrases is performed in this way. Such an approach is adopted based on the fact that between two silence frames only one speaker is bound to be talking. Thus, the average distortion of the whole speech segment is evaluated by:

$$S_i = \frac{1}{T} \sum_{t=1}^{T} s_i(\mathbf{c}_t) \quad (7)$$

where T is the number of speech frames consisting the current speech segment and $s_i(\mathbf{c}_t)$ are given by (6). The unknown speaker is identified as the reference speaker whose model gives the smallest quantization error:

$$SpeakerID = \arg\min_{1 \leq i \leq N_{Speakers}} S_i \quad (8)$$

Finally, the time interval that corresponds to the speech segment determined beforehand is assigned to the estimated speaker.

Two more recognition methods have been implemented to be compared with the already described recognition scheme (referred to as *Segments (1)* from now on). Firstly, distances from speaker classes are evaluated in speech segments, as already presented, with the difference that the distortion $s_i(\mathbf{c}_t)$ of an input test vector $\mathbf{c}_t$ when it is quantized using the speaker codebook $\mathbf{m}_i$ is estimated by:

$$s_i(\mathbf{c}_t) = \sum_{k=1}^{32} [d(\mathbf{m}_{ik}, \mathbf{c}_t)] \quad (9)$$

and the unknown speaker is identified using (8). S_i is defined by (7). This recognition scheme will be referred to as *Segments (2)*. Finally, recognition has been performed by evaluating distances from speaker classes in speech frames independently and not entire speech segments. This method is labelled as *Frames*.

After testing of all test input vectors has been performed by the three approaches, *the speaker label time diagram* is produced. Detected silence periods are also included. Speaker interchangeability has been also tested (i.e. the same speaker speaking in different parts of the video sequence). The training algorithm is computationally intensive, due to the parallel training of all speaker models, the misclassification compensation and the evaluation of the global covariance matrices of each speaker class as well as their inverse for Mahalanobis distance calculations. However, training and even testing can be performed in non-real time since the only medium used afterwards for audio indexing and retrieval is the output speaker label time diagram.

V. SIMULATION RESULTS

We show the application of the speaker dependent audio indexing method to an audio source of continuous speech extracted from a news sequence of approximately 6min duration. 5 different speakers exist two of whom appear in different time periods. The audio signal has been sampled at 22050Hz and each sample is a 16bit signed integer. Speech frames of 30msec with an overlap of 20msec have been used during the audio feature extraction.

The silence detector exhibits a detection error of 0.15% in all frames and of 0.6% in silence frames only. It fails to detect a silence segment of high background noise. In general, however, its performance is satisfactory.

In order to assess the performance of the audio indexing scheme, a speaker recognition error criterion is used, which is defined as the percentage of the number of audio frames in which a wrong speaker has been assigned divided by the total number of them. The estimation of the recognition error has been performed either in audio frames solely or in both silence and audio frames.

During the speaker modelling, training periods of 20, 30 and 40 secs of continuous speech for every speaker have been used. Table 1 summarizes the results obtained using different training periods in solely audio or both audio and silence frames. The recognition method *Segments (1)* has been used during testing. The observation of Table 1 leads to the conclusion that an increase in the training periods leads to reduction in the total speaker recognition error. Furthermore, since the silence detector attains a small detection error, evaluation of the speaker

Table 1: Speaker recognition errors obtained for different training periods.

Frames	Training Periods	Rec. Error (%)
Audio	20secs	3.26%
	30secs	2.63%
	40secs	2.38%
Audio & Silence	20secs	2.45%
	30secs	1.98%
	40secs	1.79%

Table 2: Speaker recognition error percentages obtained by the three different recognition methods.

Frames	Rec. Method	Rec. Error (%)
Audio	*Segments (1)*	2.38%
	Segments (2)	4.67%
	Frames	22.39%
Audio & Silence	*Segments (1)*	1.79%
	Segments (2)	3.52%
	Frames	16.86%

recognition errors in both audio and silence frames leads to better results.

Training periods of 40secs have been selected for further evaluation of the audio indexing method performance. Speaker recognition errors obtained by the three recognition methods of Section IV are illustrated in Table 2 for either audio frames or both audio and silence frames. Table 2 illustrates the satisfactory performance of the recognition approach *Segments (1)*. The final speaker identity label versus time diagrams achieved by all three recognition methods as well as the true one can be seen in Figure 2, when silence frames are included in the evaluation process, and in Figure 3, when only audio frames have been used during the evaluation process. The latter figure is included in order to enable better assessment of the results. It is easily observed that the indexing diagram achieved by *Segments (1)* converges to the true one and attains the minimum percentage of wrong recognitions.

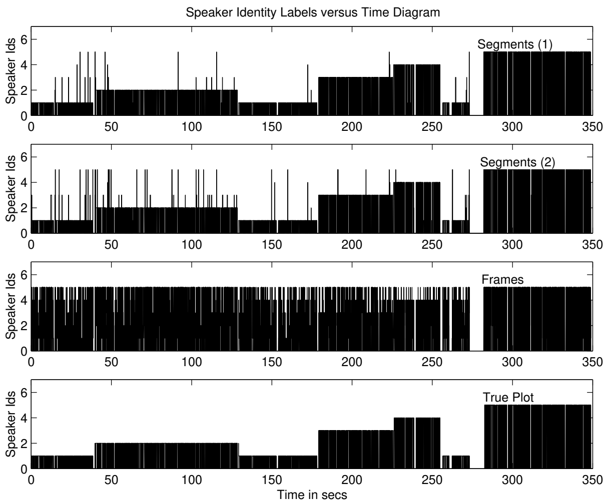

Figure 2: Speaker label time diagrams derived by all three recognition approaches and the correct one. Silence segments are included in the evaluation.

A second aspect considered in the evaluation of the audio indexing scheme performance is the correct estimation of speaker transitions when one speaker succeeds the other, that is the correct estimation of the exact time that the transition occurs. Although no special attempt has been made to improve the performance of the proposed method, it seems that, in some cases, transitions are correctly estimated. Figure 4 shows three examples of speaker transitions at different time intervals. The vertically aligned plots correspond to the same time interval. The top plots represent the true transitions, the middle plots the transitions achieved by *Segments (1)* and the bottom plots the transitions achieved by *Segments (2)*. The left aligned plots (time interval 20-50secs) illustrate a case of wrong speaker transition estimation by both recognition approaches. The center aligned plots (time interval 160-200secs) present a case of correct speaker transition estimation, whereas the right aligned plots correspond to a wrong speaker transition estimation due to error in silence detection at the beginning of the audio indexing algorithm (silence segment: 274-283secs). This is the only case of wrong silence segment estimation due to high background noise at the specific time period. The linearly increasing part in all transition plots indicate silence segment existence.

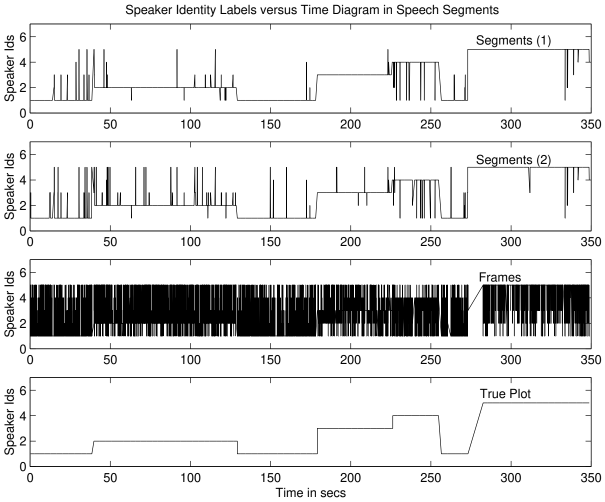

Figure 3: Speaker label time diagrams derived by all three recognition approaches and the correct one. Silence segments have been removed from all diagrams.

VI. CONCLUDING REMARKS

Content-based audio indexing systems offer a flexible and efficient tool for further audio and/or video retrieval and browsing which is made easier and user-friendly. A speaker dependent audio indexing system has been presented in this paper based on a variant of an LVQ3 classifier. VQ classification is a successful tool for speaker recognition tasks. Recognition errors are significantly reduced following the proposed recognition approach when compared to a frame by frame recognition method. Errors in silence detection affect speaker recognition errors. The performance of the audio indexing scheme is very satisfactory limiting recognition errors to 2%. The speaker dependent audio indexing diagram obtained by the proposed approach proves to converge to the true one. Transitions among speakers are, in

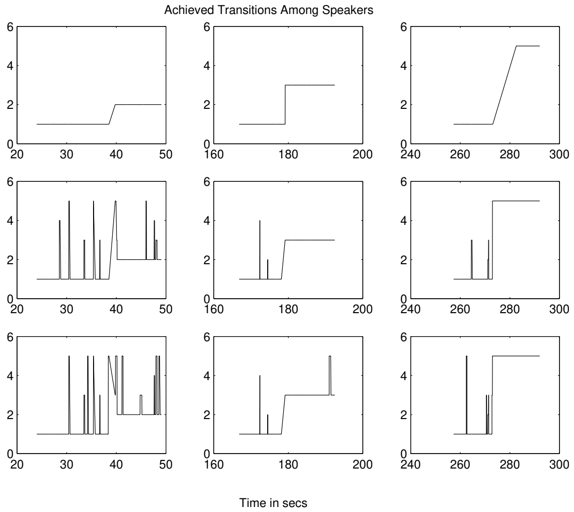

Figure 4: Three cases of achieved speaker transitions at different time intervals. Top plots are the actual transitions, middle plots show the ones achieved by *Segments (1)* and bottom plots represent the transitions achieved by *Segments (2)*.

many cases, correctly estimated. Enhancements to the proposed algorithm can be introduced so that more than one speakers can be detected and tracked when simultaneously speaking, or when background noise or music exist. Improvement in the estimation of word or phrase segments will lead to improvement in recognition rates.

REFERENCES

[1] J.R. Deller, J.G. Proakis and J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*, Upper Saddle River, N.J.: Prentice Hall, 1987.

[2] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, N.J.: Prentice Hall, 1993.

[3] T. Kohonen, "The Self-Organizing Map", *Proceedings of the IEEE*, vol. 78, no. 9, September 1990, pp. 1464-1480.

[4] I. Pitas, C. Kotropoulos, N. Nikolaidis, R. Yang and M. Gabbouj, "Order Statistics Learning Vector Quantizer", *IEEE Trans. on Image Processing*, vol. 5, no.6, June 1996, pp. 1048-1053.

[5] J. He, L. Liu and G. Palm, "A New Codebook Training Algorithm for VQ-Based Speaker Recognition", *in Proc. of ICASSP'97*, vol.2, Munich, Germany, 21-24 April 1997, pp. 1091-1094.

[6] D. Roy and C. Malamud, "Speaker Identification based Text to Audio Alignment for an Audio Retrieval System", *in Proc. of ICASSP'97*, vol. 2, Munich, Germany, 21-24 April 1997, pp. 1099-1102.

[7] S.J. Young, M.G. Brown, J.T. Fooye, G.J.F. Jones and K. Sparck Jones, "Acoustic Indexing for Multimedia Retrieval and Browsing", *in Proc. of ICASSP'97*, vol. 1, Munich, Germany, 21-24 April 1997, pp. 199 -202.

[8] F. Kubala, H. Jin, S. Matsoukas, L. Nguyen and R. Schwartz, "Broadcast News Transcription", *in Proc. of ICASSP'97*, vol. 1, Munich, Germany, 21-24 April 1997, pp. 727-730.

[9] A.G. Hauptmann and H.D. Wactlar, "Indexing and Search of Multimodal Information", *in Proc. of ICASSP'97*, vol. 1, Munich, Germany, 21-24 April 1997, pp. 195 -198.