Automatic Speaker Change Detection with the Bayesian Information Criterion using MPEG-7 Features and a Fusion Scheme

Margarita Kotti, Emmanouil Benetos, Constantine Kotropoulos
Artificial Intelligence and Information Analysis Laboratory
Department of Informatics, Aristotle Univ. of Thessaloniki
Box 451, Thessaloniki 541 24, Greece
E-mail: {mkotti, empeneto, costas}@zeus.csd.auth.gr

***Abstract*— This paper addresses unsupervised speaker change detection, a necessary step for several indexing tasks. We assume that there is no prior knowledge either on the number of speakers or their identities. Features included in the MPEG-7 Audio Prototype are investigated such as the AudioWaveformEnvelope and the AudioSpecrtumCentroid. The model selection criterion is the Bayesian Information Criterion (BIC). A multiple pass algorithm is proposed. It uses a dynamic thresholding for scalar features and a fusion scheme so as to refine the segmentation results. It also models every speaker by a multivariate Gaussian probability density function and whenever new information is available, the respective model is updated. The experiments are carried out on a dataset created by concatenating speakers from the TIMIT database, that is referred to as the TIMIT data set. It is and demonstrated that the performance of the proposed multiple pass algorithm is better than that of other approaches.**

I. Introduction

Speaker segmentation aims at finding the speaker change points in an audio stream. This task is a necessary preprocessing task for audio indexing, speaker identification - verification - tracking, automatic transcription, etc. Massive research has been carried out during the last decade in this area. Tritschler and Gopinnath proposed the use of the Bayesian Information Criterion (BIC) over mel-cepstrum coefficients (MFCCs) [4]. Delacourt and Wellekens proposed a new two-pass segmentation technique called DISTBIC and improved performance by utilizing distance-based segmentation before applying the BIC [2]. Ajmera et al introduced an alternative of the BIC which does not need tuning, and some heuristics [3]. Meanwhile, novel features like the smoothed zero crossing rate (SZCR), the perceptual minimum variance distortionless response (PMVDR), and the filterbank log coefficients (FBLC) were introduced by Huang and Hansen [10]. Another method is the so-called METRIC-SEQDAC [9]. Finally, a hybrid algorithm was proposed, which combines metric-based segmentation with the BIC criterion and model-based segmentation with Hidden Markov Models (HMMs) [7].

In this work, we employ an algorithm that improves the performance of BIC-based segmentation by introducing a number of novelties. First of all, two new features are utilized: the AudioSpectrumCentroid and the AudioWaveformEnvelope, both derived from the MPEG-7 Audio Standard [1], and an adaptive dynamic thresholding is introduced. However, the most significant proposal we make is that of a fusion scheme, which combines the partial results so as to achieve better scores than those obtained by the same algorithm without fusion. Every speaker is modelled by a Gaussian probability density function and whenever more information is available the speaker model is updated [8]. The evaluation criterion used is the BIC-type criterion proposed by Ajmera et al [3]. A multiple pass algorithm employing a distinct feature at each pass is utilized. Each pass is executed independently from others because this has the advantage that if time efficiency is of greater importance than performance, we can prune the last passes at the expense of performance deterioration.

The rest of this paper is organized as follows. In Section 2, the criterion applied for speaker change detection is described. In Section 3, the selected features are presented. The proposed algorithm is introduced in Section 4. In Section 5, our experiments are described, and in Section 6 conclusions and perspectives of future work are presented.

II. Speaker Change Detection via the BIC criterion

A BIC-type criterion is applied [2] [3] [4] [7] [9] and the BIC variant proposed in [3] is used. Speaker change detection is formulated as a hypothesis testing problem. We assume that there are two neighboring chunks X and Y around time c_j and the problem is to decide whether or not a speaker change point exists on c_j. Let $Z = X \cup Y$.

Under H_0 there is no speaker change point at time c_j. The maximum likelihood (ML) principle is used to estimate the parameters of the chunk Z that is modelled by a GMM of two components. Let us denote the GMM parameters estimated using the expectation-maximization (EM) algorithm as θ_z. The log likelihood L_0 is calculated as:

$$L_0 = \sum_{i=1}^{N_x} \log p(x_i|\theta_z) + \sum_{i=1}^{N_y} \log p(y_i|\theta_z) \qquad (1)$$

where N_x and N_y are the total numbers of samples in chunks X and Y, respectively.

Under H_1 there is a speaker change point at time c_j. The chunks X and Y are modelled by a distinct single multivariate Gaussian densities whose parameters are denoted by θ_x and θ_y. Then, the log likelihood L_1 is given by:

$$L_1 = \sum_{i=1}^{N_x} \log p(x_i|\theta_x) + \sum_{i=1}^{N_y} \log p(y_i|\theta_y). \quad (2)$$

The dissimilarity is estimated by:

$$d = L_1 - L_0 - \frac{\lambda}{2} \cdot \Delta K \cdot \log N_Z \quad (3)$$

where N_Z is the total number of samples in chunk Z, λ is the penalty factor (ideally 1.0), tuned according to data, and ΔK is the number of the model parameters (i.e. $\Delta K = 13$ in case we use the first 13 MFCCs) [2] [3]. If $d > 0$ then a local maximum is found, and time c_j is considered to be a speaker change point. In the case of $d < 0$, there is no change point at time c_j.

III. Feature Extraction

The selection of the appropriate features is vital for the accurate description of the audio signal. In general, the chunks X and Y are sets of feature vectors. In this paper, a couple of new features, namely the AudioWaveformEnvelope and the AudioSpecrtumCentroid where chosen after investigating several MPEG-7 features. Commonly used features were utilized as well. In the following, the features exploited are listed.

A. Commonly Used Features

- *Mel Cepstrum Coefficients (MFCCs)*.
- *The maximum magnitude of the DFT coefficients in a speech frame*: This feature is more efficient, when the two speakers are of different genders.
- *Short Time Energy (STE)*: In general, around a speaker change point no energy is present. Accordingly a great variation of energy is a clue of a potential speaker change point.

B. MPEG-7 Features

The MPEG-7 Audio standard [1] includes Description Schemes, Descriptors, Datatypes, a Description Definition Language, and a number of System Tools. A subcategory of Descriptors are those having a low level complexity that are called Low-Level-Descriptors (LLD). Among them are the AudioWaveformEnvelope and the AudioSpecrtumCentroid.

- *AudioWaveformEnvelope*: It is a descriptor that describes the audio waveform envelope using a small set of values that represent the extrema (minimum and maximum) of the speech waveform. In this implementation, we retain only the maximum value.
- *AudioSpecrtumCentroid*: It indicates whether the power spectrum is dominated by low or high frequencies. It is an economical descriptor of the shape of the power spectrum. It describes the center of gravity of the log-frequency power spectrum and is defined as the power weighted log-frequency centroid.

IV. Multiple Pass Speaker Change Point Detection

The proposed algorithm employees multiple passes, where different features may be used. The technique of using the same feature in multiple passes was first proposed by Tritschler [4] and Delacourt [2] and aims to increase the segmentation efficiency. The reason why we decided to have multiple passes is that after each pass, the number of chunks is decreased, because specific potential change points are discarded, since there are found to be false. So the length of chunks is becoming larger. Several researchers have aimed to the conclusion that the larger the chunks are, the better the performance is, because there is enough data for satisfactory parameter estimation of the speaker model [2] [4] [5] [6] [8] [10].

The algorithm starts with chunks of 1 sec duration in the first pass and checks the hypothesis that two adjacent chunks belong to different speakers. Every speaker is represented with a multivariate Gaussian probability density function (pdf) with mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The pdf parameters are automatically updated when more data are available. Utilizing the fact that the chunks are becoming larger, we employ a constant updating of the speaker models [5] [6] [10].

Every pass utilizes a specific feature either scalar or vector that is averaged over the frames creating the chunk under consideration. For example, the MFCCs and the AudioSpectrumCentroid computed on frame basis result in feature vectors. The remaining features computed over the chunk are scalar: the maximum magnitude of DFT, the STE, and the maximum of AudioWavformEnvelope. In the first four passes, we use the MFCCs; the fifth pass uses the maximum of DFT magnitude; in the sixth pass the STE is exploited; in the seventh pass we re-use the MFCCs; in the eighth pass we rely on the AudioSpectrumCentroid; in the ninth pass the maximum DFT magnitude is used and finally in the tenth pass the maximum value of the AudioWaveformEnvelope is exploited. It is worth mentioning that AudioSpectrumCentroid and AudioWaveformEnvelope are used for speaker segmentation for first time. The BIC criterion is used only for the vectors, while for scalar features the city-block distance is utilized for time efficiency.

The dynamic thresholding refers only to scalar features. Its need is justified by the fact that the nature of every recording is unique. We start with an ad hoc threshold ϑ that is determined after a considerable number of experiments. In these experiments we compute the F_1 measure, as defined in (9), for several threshold values and then retain the value which maximizes the F_1 measure. For example, in Figure 1, for the fifth pass, the selected ad hoc threshold is ϑ value is 1.8, since this value maximizes the F_1 measure.

By adjusting ϑ we manage to enhance the scores. Let us consider a recording that has I chunks which means that it also has $I - 1$ possible speaker change points. The value of I is determined at the previous pass. We test the possible speaker change point c_j which lays between chunks k and $k + 1$. If $f(k)$ is the current feature value computed at chunk k, we estimate $f(k)$ and $f(k + 1)$. Then, we calculate the value of

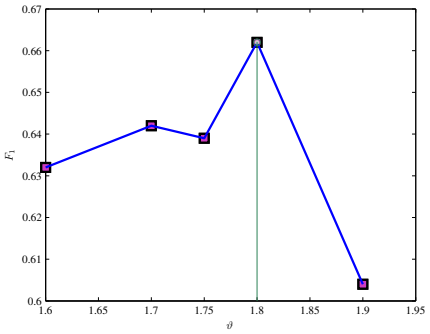

Fig. 1. The diagram of F_1 vs. the ad hoc threshold for the fifth pass.

the absolute difference between these values, which is in fact the city-block distance and is denoted by ϵ:

$$\epsilon = |f(k+1) - f(k)|. \tag{4}$$

Let $\bar{\epsilon}$ be the mean value of ϵ over all chunks of a recording:

$$\bar{\epsilon} = \frac{\sum_{l=1}^{I-1} |f(l+1) - f(l)|}{I-1}. \tag{5}$$

Then ϵ is compared to ϑ, whose value is adjusted by adding or reducing 0.5% of $\bar{\epsilon}$'s. The new adjusted threshold ϑ' is:

$$\vartheta' = \begin{cases} \vartheta + 0.005\bar{\epsilon} & \text{when } \vartheta < \bar{\epsilon} \\ \vartheta - 0.005\bar{\epsilon} & \text{when } \vartheta > \bar{\epsilon}. \end{cases} \tag{6}$$

Whenever a feature vector is employed the BIC is applied. In order to estimate the GMM needed in (1), the EM algorithm is used, which may converge at local minima. Although, there is no guarantee that a local minimum coincides with the global minimum or that there is only one local minimum. That issue, combined with the fact that the BIC is a weak classifier lead us to propose a fusion scheme so as to improve performance since it is possible for the same input set to obtain different output sets. Thus, we could theoretically reduce the error introduced by the EM algorithm by repeating the experiment multiple times, say R times and applying majority voting in each pass.

To be more specific, for each repetition we obtain a set of possible speaker turn points. Let us denote it by $\mathbf{C}_i = \{c_1, c_2, ..., c_j\}$, where i is the running number of the experiment and $c_1, c_2, ..., c_j$ are the potential speaker change points. The final set of change points $\mathbf{C}_f$ for the pass under consideration consists of those potential speaker change points c_j that appear at least S times. Both R and S are determined heuristically and their values depend mainly on the nature of the respective features and the position in the chain of passes. Typical values for R and S are 5 and 4 respectively. The algorithm is summarized as:

1) Initialize R, S, $\mathbf{C}_f$=$\emptyset$
2) For $i = 1 : R$ find $\mathbf{C}_i = \{c_1, c_2, ..., c_j\}$
3) $\forall$ distinct c_j
   if total count of $c_j > S$ then $\mathbf{C}_f$=$\mathbf{C}_f \cup \{c_j\}$.

In addition a Bayesian network with a serial (or tandem) architecture is employed to forward evidence (data) between nodes. Since serial Bayesian networks have independent connections, every pass in the proposed algorithm is independent. Diagrammatically, this is a directed graph which represents a causal network depicted in Figure 2, were the data can be transmitted only forward. Apparently, pass1 affects pass2-pass10 and so on. Finally, f, f', f'' are the features utilized in each pass.

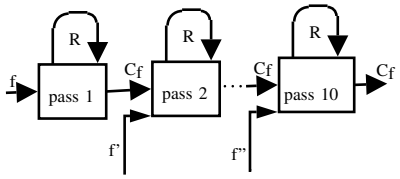

Fig. 2. The flow of the proposed algorithm.

V. Experiments

In order to assess the performance of the aforementioned algorithm the TIMIT dataset was created by concatenating speakers from the TIMIT database. TIMIT is an acoustic-phonetic database including 6300 sentences and 630 speakers who speak English. The audio format is PCM, and the audio samples are quantized in 16 bit. The recordings are single-channel with mean duration of 3.28 sec and standard deviation (st. dev.) of 1.52 sec. The parameter λ is fine-tuned using the complete TIMIT dataset (43 speech files), not including any of the ten files used on the evaluation.

In a change detection system there are two types of errors. The first type takes place when a true change is not spotted and is called precision (PRC) while the second type happens when the system detects a change that does not actually exists and is called recall (RCL). They are defined as:

$$PRC = \frac{number\ of\ correctly\ found\ changes}{total\ number\ of\ changes\ found}. \tag{7}$$

$$RCL = \frac{number\ of\ correctly\ found\ changes}{total\ number\ \ of\ correct\ changes}. \tag{8}$$

There is a third measure of the algorithm effectiveness, which is called F_1 measure and is defined as:

$$F_1 = \frac{2\ \ PRC\ \ RCL}{PRC + RCL}. \tag{9}$$

F_1 measure admits a value between 0 and 1 and the higher its value is, the better performance is obtained.

Table I demonstrates the performance for 10 randomly selected test recordings extracted from TIMIT database not included in the training procedure. The efficiency has been presumed dropping whenever the speaker's utterance has a duration of less than 1-2 sec, as it was expected [2] [4] [5] [6] [8] [10]. It is observed that the maximum value of F_1 measure is 0.903 while the minimum value is 0.588. However, it should be noted that the performance of the algorithm is rather stable, since the standard deviation of the F_1 measure is 0.109. For comparison, the PRC, RCL, and F_1 measure values achieved by Ajmera [3] were 0.68, 0.65, and 0.67, respectively.

TABLE I

EFFICIENCY OVER 10 AUDIO RECORDINGS.

Index	PRC	RCL	F_1 measure	duration (sec)	#change points
1	0.889	0.889	0.889	27	9
2	0.800	0.667	0.727	26	12
3	0.857	0.545	0.667	39	11
4	0.800	0.727	0.762	31	11
5	0.533	0.800	0.640	37	10
6	0.643	0.563	0.600	49	16
7	0.834	0.714	0.769	49	14
8	0.588	0.588	0.588	37	17
9	1.000	0.824	**0.903**	45	17
10	0.917	0.579	0.710	45	19
mean	0.786	0.690	0.726	38.5	13.6
st. dev.	0.151	0.121	0.109	8.523	3.47

Additional experiments were carried out on a second dataset, called the INESC dataset [13]. This dataset was created by using recordings from the MPEG-7 test set CD1 and broadcast news. In this dataset, the audio format is PCM, the audio samples are quantized in 16 bit, and the recordings are single-channel. In the INESC dataset, where long dialogues are included the first system achieved an F_1 measure of 0.256, because the mean duration of a speaker's utterance in INESC dataset (19.81sec) is much longer than the mean duration in TIMIT dataset (3.28sec) that the system was designed for. This leads to over-segmentation as can be seen from the large FAR. A possible enlargement of the windows length will probably lead to better results Further information and results can be found in [13].

In a second set of experiments, we investigated the way F_1 measure improves over the passes. For that purpose the experiments were terminated in the respective pass and the F_1 measure was recorded. We conducted several experiments. For all the created recordings and we recorded the behavior of the mean value of F_1 measure. In Figure 3 it is shown that the rate of improvement of the mean F_1 measure is greater in the first six passes and drops in the last four passes. Thus, if time efficiency is of greater importance than performance some of the last passes may be pruned.

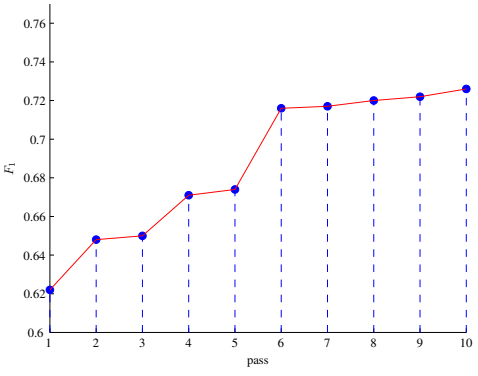

Fig. 3. The values of F_1 measure over the passes.

VI. CONCLUSIONS AND FUTURE WORK

In this work, we have proposed a new multiple pass algorithm for unsupervised speaker change detection. We investigated a considerable number of MPEG-7 features and concluded that the two most appropriate features for our objectives are the AudioWaveformEnvelope and the AudioSpecrtumCentroid. Moreover, we proposed a novel dynamic thresholding and keep updating constantly the speaker models. Furthermore, we developed a fusion scheme which improves the BIC performance.

In the future, we intend to enforce our fusion scheme by combining the results of every distinct pass. Additional features proposed in the MPEG-7 Audio standard could also be applied in the existing or additional passes.

ACKNOWLEDGEMENT

This work has been supported by the FP6 European Union Network of Excellence MUSCLE "Multimedia Understanding through Semantics, Computation and Learning" (FP6-507752).

REFERENCES

[1] ISO/IEC 15938-4:2001, "Multimedia Content Description Interface - Part 4: Audio", Version 1.0.

[2] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Communication*, vol. 32, pp. 111-126, September 2000.

[3] I. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 649-651, August 2004.

[4] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," in Proc. *6th European Conf. Speech Communication and Techology*, pp. 679-682, September 1999.

[5] L. Lu and H. Zhang, "Speaker change detection and tracking in real-time news broadcast analysis," in Proc. *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, pp. 741-744, June 2004.

[6] T. Wu, L. Lu, K. Chen, and H. Zhang, "UBM-Based Real-Time Segmentation for Broadcasting News," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, pp. 193-196, Hong Kong, April, 2003.

[7] H. Kim, D. Elter, and T. Sikora, "Hybrid Speaker-Based Segmentation System Using Model-Level Clustering," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, pp. 745-748, Philadelphia, March, 2005.

[8] L. Lu and H. Zhang, "Real-time unsupervised speaker change detection", in Proc. *16th Int. Conf. Pattern Recognition*, vol. 2, pp. 358-361, August 2002.

[9] S. Cheng and H. Wang , "Metric Seqdac: A hybrid approach for audio segmentation," in Proc. *6th Int. Conf. Spoken Language Processing*, Corea, October 2004.

[10] R. Huang and J. H. L. Hansen, "Advances in unsupervised audio segmentation for the broadcast news and ngsw corpora," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, pp. 741-744, May, 2004.

[11] H. G. Kim and T. Sikora "Comparison of Mpeg-7 audio spectrum projection features and MFCC applied to speaker recognition, sound classification and audio segmentation ," in Proc. *IEEE Int. Conf. Acoustics, Speech, and Signal Processing,* vol. 5, pp. 925-928, May 2004.

[12] J. A. Arias, J. Pinquier, and R. Andè-Obrecht, "Evaluation of Classification Techniques for Audio Indexing," in *13th European Signal Processing Conf.*, September 2005.

[13] M. Kotti, E. Benetos. C. Kotropoulos, L.G.P.M. Martins, "Speaker change detection alogithms using Bayesian Information Criterion," in Proc. *2006 IEEE Int. Symp. Circuits, Communications, and Signal Prossesing*, Marocco.