View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by City Research Online

Kotti, M., Benetos, E. & Kotropoulos, C. (2008). Computationally Efficient and Robust BIC-Based Speaker Segmentation. IEEE Transactions on Audio, Speech & Language Processing, 16(5), 920 - 933. doi: 10.1109/TASL.2008.925152 <http://dx.doi.org/10.1109/TASL.2008.925152>

CITY UNIVERSITY LONDON

EST 1894

City Research Online

Original citation: Kotti, M., Benetos, E. & Kotropoulos, C. (2008). Computationally Efficient and Robust BIC-Based Speaker Segmentation. IEEE Transactions on Audio, Speech & Language Processing, 16(5), 920 - 933. doi: 10.1109/TASL.2008.925152 <http://dx.doi.org/10.1109/TASL.2008.925152>

Permanent City Research Online URL: http://openaccess.city.ac.uk/2045/

Copyright & reuse

City University London has developed City Research Online so that its users may access the research outputs of City University London's staff. Copyright © and Moral Rights for this paper are retained by the individual author(s) and/ or other copyright holders. All material in City Research Online is checked for eligibility for copyright before being made available in the live archive. URLs from City Research Online may be freely distributed and linked to from other web pages.

Versions of research

The version in City Research Online may differ from the final published version. Users are advised to check the Permanent City Research Online URL above for the status of the paper.

Enquiries

If you have any enquiries about any aspect of City Research Online, or if you wish to make contact with the author(s) of this paper, please email the team at publications@city.ac.uk.

(c) 2008 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/ republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.
Published version: IEEE Transactions on Audio, Speech, and Language Processing, 16(5):920-933, Jul. 2008. doi: 10.1109/TASL.2008.925152

Computationally efficient and robust BIC-based speaker segmentation

Margarita Kotti, Emmanouil Benetos, and Constantine Kotropoulos*, *Senior Member, IEEE*

Dept. of Informatics, Aristotle University of Thessaloniki, Box 451, Thessaloniki 54124, Greece, Tel: +30-2310-996361, Fax: +30-2310-998453

{mkotti,empeneto,costas}@aiia.csd.auth.gr

Abstract

An algorithm for automatic speaker segmentation based on the Bayesian Information Criterion (BIC) is presented. BIC tests are not performed for every window shift (e.g. every milliseconds), as previously, but when a speaker change is most probable to occur. This is done by estimating the next probable change point thanks to a model of utterance durations. It is found that the inverse Gaussian fits best the distribution of utterance durations. As a result, less BIC tests are needed, making the proposed system less computationally demanding in time and memory, and considerably more efficient with respect to missed speaker change points. A feature selection algorithm based on branch and bound search strategy is applied in order to identify the most efficient features for speaker segmentation. Furthermore, a new theoretical formulation of BIC is derived by applying centering and simultaneous diagonalization. This formulation is considerably more computationally efficient than the standard BIC, when the covariance matrices are estimated by other estimators than the usual maximum likelihood ones. Two commonly used pairs of figures of merit are employed and their relationship is established. Computational efficiency is achieved through the speaker utterance modeling, whereas robustness is achieved by feature selection and application of BIC tests at appropriately selected time instants. Experimental results indicate that the proposed modifications yield a superior performance compared to existing approaches.

Index Terms

* Corresponding author

Automatic speaker segmentation, Bayesian Information Criterion, Speaker utterance duration distribution, Inverse Gaussian distribution, Simultaneous diagonalization, Speech analysis.

I. Introduction

Nowadays, a vast rise in multimedia archives has occurred, partially due to the increasing number of broadcast programs, the decreasing cost of mass storage devices, the advances in compression techniques, and the wide prevalence of personal computers. The functionality of such archives would be in doubt, unless data management is employed. Data management is necessary for organizing, navigating, and browsing the multimedia content, as is manifested by the MPEG-7 standard. Here, we focus on speech. Speaker segmentation is an efficient tool for multimedia archive management. It aims to find the speaker change points in an audio recording. Speaker segmentation finds numerous applications, since it is a prerequisite for audio indexing, speaker identification/verification/tracking, automatic transcription, and dialogue detection in movies. MPEG-7 audio low-level descriptors (e.g. AudioSpectrumProjection, AudioSpectrumEnvelope) can be used to describe efficiently a speech recording [1]. In addition, MPEG-7 high-level tools, (e.g. SpokenContent) exploit speakers' word usage or prosodic features that are also useful for speaker segmentation. A large number of groups and research centers compete for improved speaker segmentation. An example is the segmentation task administrated by the *National Institute of Standards and Technology* (NIST) [2]. NIST has also been working towards rich transcription evaluation (NIST/RT). Rich transcription includes speaker segmentation as a part of its diarization task [3].

A. Related Work

Extensive work in speaker segmentation has been carried out for more than two decades. Three major categories of speaker segmentation algorithms can be found: *model-based*, *metric-based*, and *hybrid* ones.

In *model-based segmentation*, a set of models is trained for different speaker classes and the incoming speech recording is classified using the trained models. Various methods have been used in order to create generic models. Starting from the less complex case, a universal background model (UBM) is utilized to separate speech from non-speech [4]. The UBM is trained by using a large volume of speech data off-line. The algorithm can be used in real-time, because the models have been pre-calculated. Second, instead of using just one generic model, two universal gender models (UGM), that discriminate between male and female speakers can be used [5]. Third, the so-called sample speaker model (SSM) can be adopted

[5]. This is a predetermined, generic, speaker-independent model, which is progressively adapted into a specific speaker-dependent one. Alternatively, an anchor model can be utilized, where a speaker utterance is projected onto a subspace of reference speakers [6]. Finally, more sophisticated models can be created with the help of hidden Markov models (HMMs) [7], [8], [9] or support vector machines (SVMs) [10].

Metric-based techniques detect the local extrema of a proper distance between neighboring windows in order to segment an input recording. Various distances have been employed. For example, a weighted squared Euclidean distance has been used, where the weights are updated by Fisher linear discriminant analysis [11]. Another criterion is the generalized likelihood ratio (GLR) test [5], [6], [9], [12], [23]. The Kullback-Leibler divergence is also commonly employed. It is used either in conjunction with the Bayesian Information Criterion (BIC) [12], [13] or independently [14]. Alternatively, second-order statistics could be used [12]. Another closely related measure is the Hotelling T^2 statistic, which is combined with BIC to achieve a higher accuracy than the standard BIC for turns of short duration [15], [16]. However, the most popular criterion is BIC [7], [12], [15], [17], [18], [19], [20], [21]. Commonly used features in speaker segmentation are the mel-cepstrum coefficients (MFCCs), applied in conjunction with BIC [1], [4], [5], [7], [8], [9], [11], [12], [13], [15], [18], [20], [22]. A milestone variant of BIC-based algorithms is DISTBIC, that utilizes distance-based pre-segmentation before applying BIC [12]. Most recently, BIC is compared to agglomerative clustering, minimum description length-based Gaussian modeling, and exhaustive search. It is found that applying audio classification into speech, noise, and music prior to speaker segmentation improves speaker segmentation accuracy [22]. An approach to segmentation and identification of mixed-language speech with BIC has been recently proposed [20]. In particular, BIC is employed to segment an input utterance into a sequence of language-dependent segments, each of which is used then as processing model for mixed language identification.

Many researchers have experimented with *hybrid algorithms*, where first metric-based segmentation creates an initial set of speaker models and model-based techniques refine the segmentation next. In [8], HMMs are combined with BIC. In [23], another hybrid system is proposed, where the audio stream is recursively divided into two subsegments and speaker segmentation is applied to both of them separately. Another interesting hybrid system is described in [9], where two systems are coupled, namely the LIA system and the CLIPS system. The LIA system is based on HMMs, while the CLIPS system is based on BIC speaker segmentation followed by hierarchical clustering. The aforementioned systems are combined

using different strategies to further improve performance.

B. Proposed Approach

In this paper, an unsupervised, BIC-based system for speaker segmentation is proposed. The first contribution of the paper is in modeling the distribution of the duration of speaker utterances. More specifically, the next probable change point is estimated by employing the utterance duration model. In this way, several advantages are gained, because the search is no longer "blind" and exhaustive, as is the common case in speaker segmentation algorithms. Consequently, a considerably less demanding algorithm in time and memory is developed. Several distributions have been tested as hypotheses for the distribution speaker utterance durations and their parameters have been estimated by maximum likelihood estimation (MLE). Both the log-likelihood criterion and the Kolmogorov-Smirnov criterion yield the inverse Gaussian (IG) distribution as the best fit. More specifically, distribution fitting in three datasets having substantially different nature verifies that IG models more accurately the distribution of speaker utterance duration. The first dataset, contains recorded speech of concatenated short utterances, the second dataset contains dialogues between actors that follow specific film grammar rules, while the last dataset contains spontaneous speech.

The second contribution is in feature selection applied prior to segmentation aiming to determine which MFCCs are most discriminative for the speaker segmentation task. The branch and bound search strategy using depth-first search and backtracking is employed, since its performance is near optimal [24]. In the search strategy, the performance measure resorts to the ratio of the inter-class dispersion over the intra-class one. That is, the trace of the product of the inverse within-class scatter matrix and the between-class scatter matrix is employed.

The third contribution is of theoretical nature. An alternative formulation of BIC for multivariate Gaussians is derived. The new formulation is obtained by applying centering and simultaneous diagonalization. A detailed proof can be found in Appendix I. It is shown that the new formulation is significantly less computationally demanding than the standard BIC, when covariance matrix estimators other than the sample dispersion matrices are employed, such as the robust estimators [25], [26] or the regularized MLE [27]. In particular, simultaneous diagonalization replaces matrix inversions and simplifies the quadratic forms to be computed. A detailed analysis of the computational cost can be found in Appendix II. The block diagram of the proposed approach is illustrated in Figure 1.

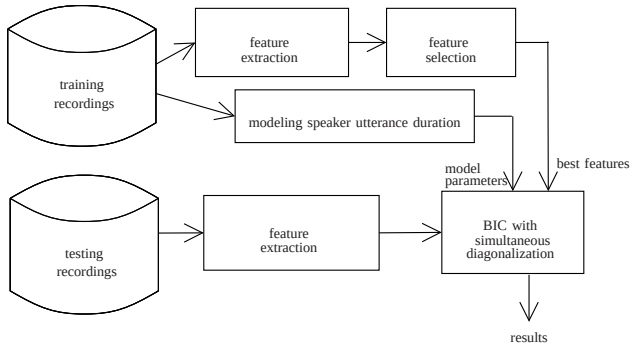

Fig. 1. The block diagram of the proposed approach.

Experimental results are reported with respect to the two commonly used sets of figures of merit, namely: (i) the precision rate (PRC), the recall rate (RCL), and the associated F_1 measure; (ii) the false alarm rate (FAR) and the miss detection rate (MDR). By utilizing both sets of figures of merit, a straightforward comparison with other experimental results reported in related works is enabled. Furthermore, relationships between the aforementioned figures of merit are derived. The proposed approach yields a significant improvement in efficiency compared to previous approaches. Experiments were carried out on two datasets. The first dataset has been created by concatenating speakers from the TIMIT database [28]. This dataset will be referred to as the conTIMIT test dataset. The second dataset has been derived from RT-03 MDE Training Data Speech [44]. In essence, the HUB-4 1997 English Broadcast News Speech part has been utilized. The greatest improvement is achieved for missed speaker turn points. Compared with other approaches, the number of missed speaker change points is smaller as explained in Section V. This is attributed to the fact that BIC tests take place when a speaker change is most probable to occur.

The remainder of the paper is organized as follows. In Section II, various distributions are tested for modeling the duration of speaker utterance and the IG distribution is demonstrated to be the best fit. The feature selection algorithm is sketched in Section III. In Section IV, the standard BIC is presented and its equivalent transformed BIC is derived. In Section V, the evaluation of the proposed approach is undertaken. Finally, conclusions are drawn in Section VI. The derivation of the transformed BIC can be found in Appendix I and the computational cost analysis is detailed in Appendix II.

II. Modeling the duration of speaker utterances

A. Distribution of speaker utterance durations

The first contribution of this paper is in modeling the distribution of the duration of speaker utterances. Let us argue why such a modeling is advantageous. By estimating the duration of a speaker's utterance, the search is no longer "blind". After modeling, it is safe to claim that the next speaker change point most probably occurs after as many seconds dictated by a statistic of speaker utterance durations. In this context, several distributions have been tested for a goodness-of-fit to the empirical distribution of speaker utterance durations. A question that arises is why fitting a theoretical distribution of speaker utterance durations to the empirical one is necessary. The answer is that by doing so, distribution parameters take into account the structure of the data. Moreover, finding such a theoretical distribution is interesting per se and this result may find additional applications, e.g. in speech synthesis.

The following distributions have been considered: Birnbaum-Saunders, Exponential, Extreme value, Gamma, IG, Log-logistic, Logistic, Lognormal, Nakagami, Normal, Rayleigh, Rician, t-location scale, and Weibull. MLE has been used to calculate the best fitting parameters of each distribution. Here, the parameters under consideration are the mean and the variance. In order to evaluate the goodness-of-fit, the log-likelihood and the Kolmogorov-Smirnov criteria have been computed.

The TIMIT database was used first in order to model the duration of speaker utterances. The TIMIT database includes 6300 sentences uttered by 630 speakers, both male and female ones, who speak various U.S. English dialects. The recordings are mono-channel, the sampling frequency is 16 KHz, and the audio PCM samples are quantized in 16 bits [28]. In total, 55 recordings of artificially created dialogues along with the ground-truth associated with speaker changes comprise the conTIMIT dataset.[1] The recordings have a total duration of about 1 hour. Since a transition between speech and silence is not similar to a transition between two speakers, the inter-speaker silences have been reduced so that conversations sound like real [12]. Thus, each segment of a speaker is followed by a segment of another speaker. This is equivalent to silence removal, which is a common pre-processing step [4], [8], [13], [15]. 935 speaker change points occur in the conTIMIT dataset. Throughout the conTIMIT dataset, the minimum duration of an utterance is 1.139 s and the maximum one is 11.751 s, while the mean duration is 3.286 s with a standard deviation equal to 1.503 s. 10 out of the 55 recordings of the conTIMIT dataset, randomly

[1] conTIMIT dataset is available at http://poseidon.csd.auth.gr/LAB_RESEARCH/Latest/data/conTIMITdataset.zip

chosen, were used to create the conTIMIT training-1 dataset, employed in modeling the speaker utterance duration.

IG distribution has been found to be the best fit with respect to both log-likelihood and Kolmogorov-Smirnov test. The values of mean utterance duration and standard deviation for the conTIMIT training-1 dataset under the IG model equal 3.286 s and 1.388 s, respectively. An illustration of the best fit for the aforementioned distributions to the empirical distribution of speaker utterance durations can be seen in Figure 2 with respect to the probability-probability (P-P) plots. Let us denote the mean and the standard deviation of durations by μ and σ. Let $F_k(\cdot)$ be the kth normalized theoretical cumulative density function (cdf) tested. To make a P-P plot, uniform quantiles $q_i = \frac{i}{N+1}$, $i = 1, 2, \ldots, N$, are defined in the horizontal axis, where N is the number of the samples whose distribution is estimated. The vertical axis represents the value admitted by the theoretical cdf at $\frac{t_{(i)}-\mu}{\sigma}$, i.e. $F_k(\frac{t_{(i)}-\mu}{\sigma})$, where $t_{(i)}$ is the ith order statistic of utterance durations. That is, the durations are arranged in an increasing order and the ith sorted value is chosen. The better the theoretical cdf approximates the empirical one, the closer the points $(q_i, F_k(\frac{t_{(i)}-\mu}{\sigma}))$ are to the diagonal [29].

To validate that IG distribution generally fits best the empirical speaker utterance duration distribution, another dataset has been employed, to be referred to as the movie dataset [30]. In this dataset, 25 audio recordings are included that have been extracted from six movies of different genres, namely: Analyze That, Cold Mountain, Jackie Brown, Lord of the Rings I, Platoon, and Secret Window. Indeed, Analyze That is a comedy, Platoon is an action, and Cold Mountain is a drama. Thus, dialogues of different natures are included in the movie dataset. Speaker segmentation has been performed by human agents. Having the ground-truth speaker change points, the distribution of speaker utterance duration for the movie dataset is modeled by applying the same procedure as for the conTIMIT training-1 dataset. The best fit is found to be the IG distribution, once again, with respect to both log-likelihood and Kolmogorov-Smirnov test. This outcome is of great importance, since the dialogues are not recorded in a clean environment. Longer pauses or overlaps between the actor utterances exist and background music or noise occurs. However, as expected, different parameters from those estimated in the conTIMIT training-1 dataset are obtained. In the movie dataset, the mean duration equals 5.333 s, while the standard deviation is 6.189 s. Accordingly, modeling the duration of speaker utterances by an IG distribution helps to predict the next speaker change point.

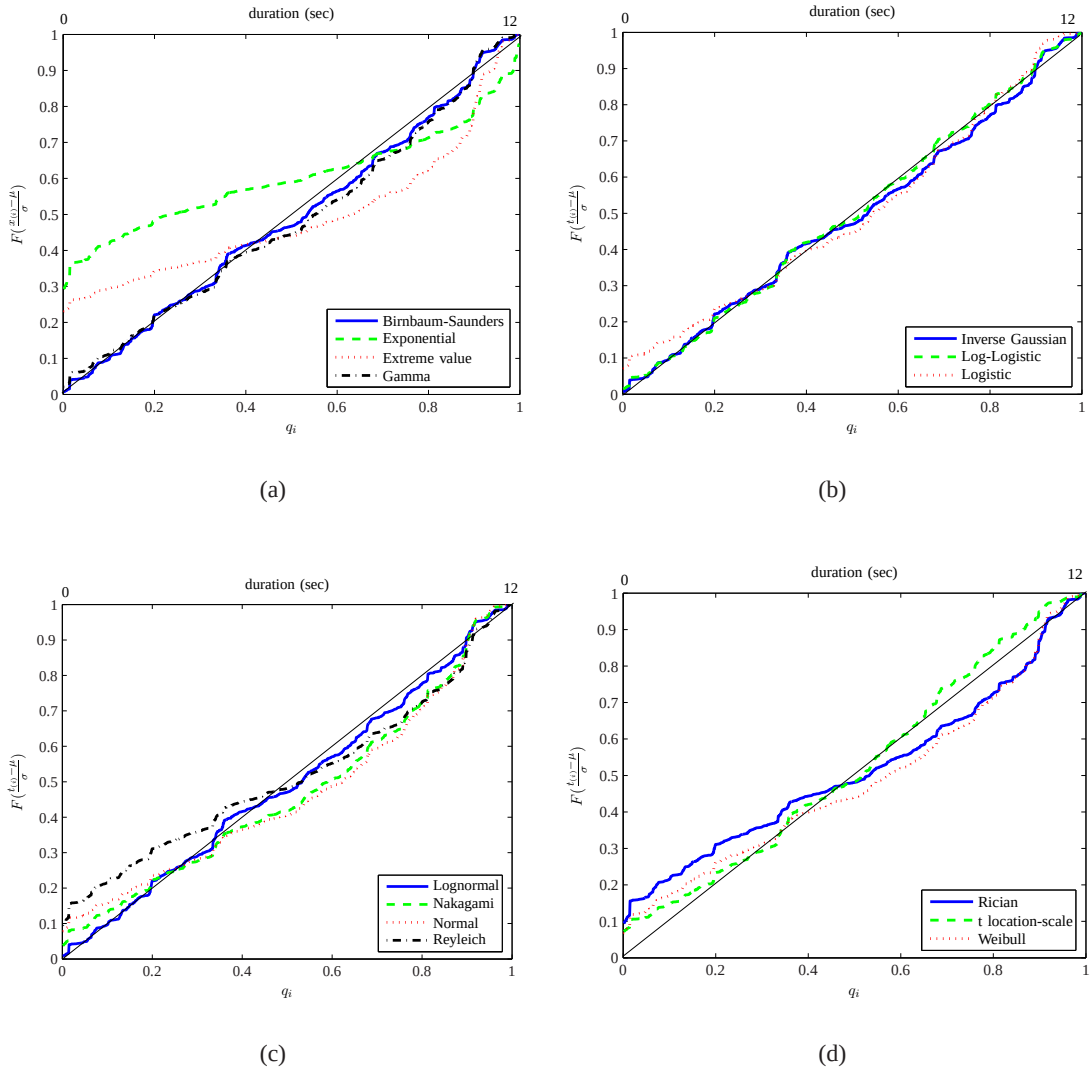

Fig. 2. (a) The P-P plots for distributions Birnbaum-Saunders, Exponential, Extreme value, Gamma for the conTIMIT training-1 dataset. (b) The P-P plots for distributions IG, Log-Logistic, Logistic for the conTIMIT training-1 dataset. (c) The P-P plots for distributions Lognormal, Nakagami, Normal, Reyleigh for the conTIMIT training-1 dataset. (d) The P-P plots for distributions Rician, t-location scale, Weibull for the conTIMIT training-1 dataset.

B. Mathematical properties of the IG distribution and its application

Some of the main mathematical properties of the IG distribution are briefly discussed next. The IG distribution, or Wald distribution, has probability density function (pdf) with parameters μ and λ_{IG} [31], [32]:

$$f(t) = \sqrt{\frac{\lambda_{IG}}{2\pi t^3}} \exp\left(\frac{-\lambda_{IG}(t-\mu)^2}{2\mu^2 t}\right) \qquad t \in (0, \infty) \tag{1}$$

where $\lambda_{IG}, \mu > 0$. The IG probability density is always positively skewed. Kurtosis is positive as well. For N realizations, t_1,t_2,...,t_N of an IG random variable the MLEs of σ and μ are: $\widehat{\sigma} = \sqrt{\frac{\overline{t}^3}{\widehat{\lambda_{IG}}}}$, where $\widehat{\lambda_{IG}} = \frac{N}{\sum_{i=1}^{N}(\frac{1}{t_i} - \frac{1}{\overline{t}})}$ and $\overline{t} = \widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} t_i$. Obviously, $\widehat{\sigma}$ does not coincide with the sample dispersion. One of the most important properties of the IG distribution is its infinite divisibility, which implies that the IG distribution generates a class of increasing Lèvi processes [33]. Lèvi processes contain both "small jumps" and "big jumps". Such jumps are Poisson point processes that are commonly used to model the arrival time of an event. In the case under consideration, the arrival time refers to a speaker change point. This property makes Lèvi processes candidates for modeling the speaker utterance duration. "Small jumps" occur if there are lively exchanges (stichomythia) and "big jumps" occur for monologues.

In conclusion, modeling the speaker utterance duration enables us to perform BIC tests when a speaker change point is most probable to occur. The simplest approach is to assume that a probable speaker change point occurs every r seconds, where r is a submultiple of the expected duration of speaker utterances. r should be chosen at the same order of magnitude as the sample dispersion. This technical solution does not exclude other alternatives, such as setting r to a submultiple of the mode of (1) that is given by $\widehat{\mu}\big[\big(1 + \frac{9\widehat{\mu}^2}{4\widehat{\lambda_{IG}}^2}\big)^{\frac{1}{2}} - \frac{3\widehat{\mu}}{2\widehat{\lambda_{IG}}}\big]$ [34].

If the total length of the audio recording is L_a, then $\lfloor \frac{L_a}{r} \rfloor$ BIC tests take place. In straightforward implementation of BIC-based speaker segmentation, $\lfloor \frac{L_a}{u} \rfloor$ BIC tests are performed, where u is the window shift of several ms. Thus, $\frac{r-u}{r}$% less BIC tests are performed by taking into account the duration of utterances, when compared to the straightforward implementation of BIC-based speaker segmentation. If a probable change point is not confirmed through the BIC test, the information contained in the last r seconds updates the existing speaker model. The use of a submultiple of the expected speaker utterance duration enables us to reduce the probability of missed speaker change points, as explained in Section V. Previous experiment demonstrates that under-segmentation, caused by a high number of miss detections, is more cumbersome to remedy than over-segmentation caused by a high number of false alarms [12], [13], [15], [16], [23], [40]. For example, over-segmentation could be alleviated by clustering and/or merging. The use of r within the context of BIC is described in Section IV.

III. Feature Extraction and Selection

Different features yield a varying performance level in speaker segmentation applications [13]. This fact motivated the authors to invest in feature selection for speaker segmentation, which is the second

contribution of the paper.

MFCCs, sometimes with their first-order (delta) and/or second-order differences (delta-delta) are the most commonly used features in speaker segmentation. Furthermore, MFCCs were used in various techniques besides BIC such as HMMs [9] or SVMs [10]. Still, not all the researchers employ the same MFCC order in BIC-based approaches. For example, 24 MFCCs are employed in [7], while 12 MFCCs and their first differences are utilized in [12] and [20]. 32 MFCCs are used in [11]. In [4], 16 MFCCs are applied, while 12 MFCCs along with their delta and delta-delta coefficients are employed in [9]. 23 MFCCs are utilized in [1], [8] and 24 MFCCs are applied in [5], [15], [18]. A more detailed study is presented in [22], where a comparative study between 12 MFCCs, 13 MFCCs and their delta coefficients, and 13 MFCCs, their delta, and delta-delta coefficients is performed.

A different approach is investigated here. Instead of trying to reveal the MFCC order that yields the most accurate speaker turn point detection results, an effort is made to find out an MFCC subset that is more suitable for detecting a speaker change. The MFCCs are calculated every 10 ms with 62.5% overlap by the algorithm described in [35]. An initial set consisting of 36 MFCCs is formed and the goal is to derive the subset, which contains the 24 more suitable MFCCs for speaker segmentation, since utilizing 24 coefficients is commonplace in [5], [15], [18].

Let us test the hypothesis there is a speaker change point against the hypothesis there is no speaker change point. Speakers change once under the first hypothesis, while a monologue is observed under the second one. For training purposes, an additional dataset of 50 recordings was created from speaker utterances derived in the TIMIT database, referred to as the conTIMIT training-2 dataset, that is disjoint to the conTIMIT dataset[2]. 25 out of the 50 recordings contain a speaker change point and the remaining 25 recordings do not. In this way, two classes are presumed: the first class represents a speaker change and includes 25 recordings with one speaker change and the second class corresponds to no speaker changes and includes 25 recordings with monologues. We assume that the mean feature vectors in the two different classes are different in order to enable discrimination [24]. The goal of feature selection is to find a feature subset $F_i(D)$ of dimension D. In our case, $D = 24$. Let J denote the performance measure. Feature selection finds $F_i(D)$ such that $J(F_i(D)) \geq J(F_j(D))$, where $j \in \{1, \ldots, q(D)\}$ and

[2]conTIMIT training-2 dataset is available at http://poseidon.csd.auth.gr/LAB_RESEARCH/Latest/data/conTIMITtraining-2dataset.zip

$q(D)$ is the number of distinguishable subsets containing D elements. If 24 out of the 36 coefficients are to be selected, then $q(24)=\binom{36}{24}$ is enormous. As a result, a more efficient search strategy than exhaustive search is required. Such an alternative is branch and bound, which attains an almost optimal performance [24]. The search process is accomplished systematically by means of a tree structure consisting of 36-24+1=13 levels. A level is composed of a number of nodes and each node corresponds to a coefficient subset. At the highest level, there is only one node corresponding to the full set of coefficients. At the lowest level, there are nodes containing 24 coefficients. The search process starts from the highest level by systematically traversing all levels until the lowest level is reached. The traversing algorithm uses depth-first search with a backtracking mechanism. This means that if J_1 is the best performance found so far, then branches whose performance is worse than J_1 are skipped [24].

The selection criterion J can be defined in terms of scatter matrices. A scatter matrix gives information about the dispersion of samples around their mean. The within-class scatter matrix, $\boldsymbol{S}_w$, describes the within-class dispersion. The between-class scatter matrix, $\boldsymbol{S}_b$, describes the dispersion of the class-dependent sample means around the gross mean. Mathematically, J is defined by

$$J = \mathrm{tr}(\mathbf{S}_w^{-1}\mathbf{S}_b) \tag{2}$$

where $\mathrm{tr}(\cdot)$ stands for the matrix trace operator. J, as defined in (2), is a monotonically increasing function of the distance between the mean vectors and a monotonically decreasing function of the scattering around the mean vectors. Moreover, (2) is invariant to reversible linear transformations. In addition, it is ideal for Gaussian distributed feature vectors. To a first approximation, MFCCs are assumed to follow the Gaussian distribution. Under this assumption, (2) guarantees the best performance [24].

From each recording, 36-order MFCCs are extracted. The selected 24 MFCCs for the conTIMIT training-2 dataset can be seen in Table I. Although the selection of MFCCs, as in Table I, might depend on the dataset used for feature selection, the aforementioned feature selection can be applied to any dataset. MFCCs shown in Table I are used in conjunction with their delta and delta-delta coefficients, in order to capture their temporal evolution that carries additional useful information. In general, the temporal evolution is found to increase efficiency [9], [15], [20], [22]. However, in [12], it is reported that using delta and delta-delta coefficients impairs efficiency.

Alternatively, one could replace (2) with BIC (5) itself. Then feature selection is performed by a

TABLE I
THE SELECTED 24 MFCCS FOR THE CONTIMIT TRAINING-2 DATASET.

#	1	2	3	4	5	6	7	8	9	10	11	12
MFCC	1st	3rd	4th	5th	6th	7th	8th	9th	10th	11th	13th	16th
#	13	14	15	16	17	18	19	20	21	22	23	24
MFCC	22th	23th	24th	25th	26th	27th	28th	29th	31th	33th	35th	36th

wrapper instead of a filter [36], as (2) implies. In such case, the selected MFCCs are: : 1st-18th, 22nd, 23rd, 27th, 28th, 31st, and 35th. The computation time required by (2) is less by 187.06% than that required by BIC, when a PC with a 3 GHz Athlon processor and 1 GB of RAM is used. In Section V, we comment on the accuracy of the latter feature selection method.

IV. BIC-BASED SPEAKER SEGMENTATION

In this section, the BIC criterion is detailed and an equivalent BIC criterion is derived, that is considerably less computationally demanding. It is also explained how the contributions of Sections II and III are utilized in conjunction with BIC.

BIC is a maximum likelihood, asymptotically optimal, Bayesian model selection criterion penalized by the model complexity. For speaker turn detection, two different models are employed. Assume that there are two neighboring chunks X and Y around time t_j. The problem is to decide whether or not a speaker change point exists at t_j. Let $Z = X \cup Y$ and N_X, N_Y, N_Z be the numbers of samples in chunks X, Y, and Z, respectively. Obviously, $N_Y = N_Z - N_X$. The problem is formulated as a two hypothesis testing problem.

Under H_0 there is no speaker change point at time t_j. MLE is used to compute the parameters of a Gaussian distribution that models the data samples in Z. Let us denote by $\boldsymbol{\theta}_Z$ the parameters of the Gaussian distribution, i.e. the mean vector $\boldsymbol{\mu}_Z$ and the full covariance matrix $\boldsymbol{\Sigma}_Z$. The log-likelihood L_0 under H_0 is

$$L_0 = \sum_{i=1}^{N_X} \ln p(\mathbf{z}_i|\boldsymbol{\theta}_Z) + \sum_{i=N_X+1}^{N_Z} \ln p(\mathbf{z}_i|\boldsymbol{\theta}_Z) \tag{3}$$

where $\mathbf{z}_i \in \mathbf{R}^d,\ i = 1, 2, \ldots, N_Z$ which are assumed to be independent. $\mathbf{z}_i$ consists of the 24 selected MFCCs with their delta and delta-delta coefficients, i.e. $d = 72$. Under H_1 there is a speaker change point at time t_j. The chunks X and Y are modeled by distinct multivariate Gaussian densities, whose

parameters are denoted by $\boldsymbol{\theta}_X$ and $\boldsymbol{\theta}_Y$, respectively. Their definition is similar to $\boldsymbol{\theta}_Z$. The log-likelihood L_1 under H_1 is given by:

$$L_1 = \sum_{i=1}^{N_X} \ln p(\mathbf{z}_i|\boldsymbol{\theta}_X) + \sum_{i=N_X+1}^{N_Z} \ln p(\mathbf{z}_i|\boldsymbol{\theta}_Y). \tag{4}$$

The BIC is defined as

$$\delta = L_1 - L_0 - \frac{\lambda}{2} \underbrace{\left(d + \frac{d(d+1)}{2}\right)}_{\text{model parameters}} \ln N_Z \gtrless 0 \tag{5}$$

where λ is a data-dependent penalty factor (ideally 1.0). If $\delta > 0$, then time t_j is considered to be a speaker change point. Otherwise, there is no speaker change point at time t_j. The standard BIC formulation for multivariate Gaussian densities $p(\mathbf{z}_i|\boldsymbol{\theta}_X)$, $p(\mathbf{z}_i|\boldsymbol{\theta}_Y)$, $p(\mathbf{z}_i|\boldsymbol{\theta}_Z)$ can be analytically written as

$$-\sum_{i=1}^{N_Z}(\boldsymbol{z}_i-\boldsymbol{\mu}_Z)^T\boldsymbol{\Sigma}_Z^{-1}(\boldsymbol{z}_i-\boldsymbol{\mu}_Z)+\sum_{i=1}^{N_X}(\boldsymbol{z}_i-\boldsymbol{\mu}_X)^T\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{z}_i-\boldsymbol{\mu}_X)+\sum_{i=N_X+1}^{N_Z}(\boldsymbol{z}_i-\boldsymbol{\mu}_Y)^T\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{z}_i-\boldsymbol{\mu}_Y) \gtrless \gamma_{BIC}, \tag{6}$$

where γ_{BIC} is defined as

$$\gamma_{BIC} = N_Z \ln|\boldsymbol{\Sigma}_Z| - N_X \ln|\boldsymbol{\Sigma}_X| - N_Y \ln|\boldsymbol{\Sigma}_Y| + \lambda \left(d + \frac{d(d+1)}{2}\right) \ln N_Z. \tag{7}$$

In the light of the discussion made in Section II, BIC tests are performed every r seconds, where r is a submultiple of the expected duration of speaker utterances. The window size is also set equal to r taking into consideration as many data as possible. When more data are available, more accurate Gaussian models are built, since BIC behaves better for large windows, whereas short changes are not easily detectable by BIC [12], [16]. Moreover, it was shown in [22], that the bigger the window size, the better the performance.

Next, a novel formulation of the BIC is theoretically derived. It is assumed that $\boldsymbol{\Sigma}_X$, $\boldsymbol{\Sigma}_Y$, and $\boldsymbol{\Sigma}_Z$ are full covariance matrices. Moreover, the covariance matrix estimators are not limited to sample dispersion matrices for which BIC defined in (6)-(7) obtains the simplified form (21), as explained in Appendix I. For example, one may employ the robust estimators of covariance matrices [25], [26] or the regularized MLEs [27]. To obtain the novel formulation, we apply first centering and then simultaneous diagonalization for the pairs of $\boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Z$ and $\boldsymbol{\Sigma}_Y, \boldsymbol{\Sigma}_Z$. Let us define the mean vector in Z chunk as $\boldsymbol{\mu}_Z$. The centering

transformation is $\widetilde{\boldsymbol{z}}_i = \boldsymbol{z}_i - \boldsymbol{\mu}_Z$. Next, simultaneous diagonalization of $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Z$ is applied. Let $\boldsymbol{\Lambda}_Z$ be the diagonal matrix of the eigenvalues of $\boldsymbol{\Sigma}_Z$ and $\boldsymbol{\Phi}$ be the corresponding modal matrix. Let us define $\mathbf{K} = \boldsymbol{\Lambda}_Z^{-\frac{1}{2}}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_X\boldsymbol{\Phi}\boldsymbol{\Lambda}_Z^{-\frac{1}{2}} = \boldsymbol{\Psi}\boldsymbol{\Lambda}_K\boldsymbol{\Psi}^T$, where $\boldsymbol{\Lambda}_K$ is the diagonal matrix of eigenvalues of $\mathbf{K}$ and $\boldsymbol{\Psi}$ is the corresponding modal matrix. The simultaneous diagonalization transformation yields for $\boldsymbol{z}_i \in Z \cap X = X$

$$\widetilde{\boldsymbol{w}}_i = \boldsymbol{\Psi}^T\boldsymbol{\Lambda}_Z^{-\frac{1}{2}}\boldsymbol{\Phi}^T\widetilde{\boldsymbol{z}}_i. \tag{8}$$

Let $\mathbf{H} = {\boldsymbol{\Lambda}_Z}^{-\frac{1}{2}}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_Y\boldsymbol{\Phi}{\boldsymbol{\Lambda}_Z}^{-\frac{1}{2}}$ and $\mathbf{H} = \boldsymbol{\Xi}\boldsymbol{\Lambda}_H\boldsymbol{\Xi}^T$. Following the same strategy, we obtain for $\boldsymbol{z}_i \in Z \cap Y = Y$

$$\widetilde{\boldsymbol{v}}_i = \boldsymbol{\Xi}^T\boldsymbol{\Lambda}_Z^{-\frac{1}{2}}\boldsymbol{\Phi}^T\widetilde{\boldsymbol{z}}_i. \tag{9}$$

In Appendix I, it is shown that (6) is equivalently rewritten as

$$\sum_{i=1}^{N_X}\widetilde{\boldsymbol{w}}_i^T({\boldsymbol{\Lambda}_K}^{-1} - \mathbf{I})\widetilde{\boldsymbol{w}}_i + \sum_{i=N_X+1}^{N_Z}\widetilde{\boldsymbol{v}}_i^T({\boldsymbol{\Lambda}_H}^{-1} - \mathbf{I})\widetilde{\boldsymbol{v}}_i \gtrless \gamma' \tag{10}$$

where γ' is an appropriate threshold derived analytically in (26).

Concerning the computational cost, simultaneous diagonalization replaces matrix inversions and simplifies the quadratic forms to be computed. This leads to a substantially less computational costly transformed BIC, as opposed to the standard BIC. As it is detailed in Appendix II, the computational cost of the standard BIC in flops, excluding the cost of γ_{BIC}, is

$$3d^3 + 6N_Zd^2 + (8N_Z + 3)d + 2, \tag{11}$$

whereas the computational cost of the transformed BIC, excluding the cost of γ_{BIC}, equals

$$30d^3 + (4N_Z + 4)d^2 + (7N_Z + 9)d + 5. \tag{12}$$

Since $d \ll N_Z$, by comparing (11) and (12), it can be seen that the standard BIC is more computationally costly than its transformed alternative.

To sum up, the algorithm can be roughly sketched as follows:

1) Initialize the interval $[a, b]$ to $[0, 2r]$ and let $v = \frac{a+b}{2}$.
2) Until the audio recording end, use BIC with the selected MFCCs to evaluate if there is a change point in $[a, b]$.

3) If there is no speaker change point in $[a, b]$, then $b = b + r$. Go to step 2).
4) If there is a speaker change point in $[a, b]$, then $a = v$, $b = v + r$. Go to step 2).

It is reminded that r is a submultiple of the mean utterance duration, which is obtained by the analysis in Section II and the term selected MFCCs refers to the MFCCs chosen by the feature selection algorithm, described in Section III.

V. Experimental Results and Discussion

A. Figures of Merit

To judge the efficiency of speaker turn point detection algorithms, two sets of figures of merit are commonly used. On the one hand, one may use the false alarm rate (FAR) and the miss detection rate (MDR) defined as

$$FAR = \frac{FA}{GT + FA}, \quad MDR = \frac{MD}{GT} \tag{13}$$

where FA denotes the number of false alarms, MD the number of miss detections, and GT stands for the number of actual speaker turns, i.e. the ground truth. A false alarm occurs when a speaker turn is detected although it does not exist, while a miss detection occurs, when the process does not detect an existing speaker turn [12], [23]. On the other hand, one may employ the precision (PRC), recall (RCL) and F_1 rates given by

$$PRC = \frac{CFC}{DET} = \frac{CFC}{CFC+FA}, \quad RCL = \frac{CFC}{GT} = \frac{CFC}{CFC+MD}, \quad F_1 = 2\frac{PRC\ RCL}{PRC+RCL} \tag{14}$$

where CFC denotes the number of correctly found changes and $DET = CFC + FA$ is the number of detected speaker changes. F_1 admits a value between 0 and 1. The higher its value is, the better performance is obtained [8], [18]. Between the pairs (FAR, MDR) and (PRC, RCL), the following relationships hold:

$$MDR = 1 - RCL, \quad FAR = \frac{RCL\ FA}{DET\ PRC+RCL\ FA}. \tag{15}$$

In order to facilitate a comparative assessment of our results with others reported in the literature, the evaluation of the proposed approach is carried out using all the aforementioned figures of merit.

B. Evaluation

1) Comparative performance evaluation on the conTIMIT test dataset: The total number of recordings in the conTIMIT dataset is 55. The evaluation is performed upon 49 randomly selected recordings of the conTIMIT dataset, forming the conTIMIT test dataset. The remaining 6 were used to determine the value of the BIC penalty-factor λ and to create the conTIMIT training-3 dataset. Although the conTIMIT dataset is an artificially created dataset and includes no real conversations, performance assessment over the conTIMIT test dataset is still informative. It is also reminded that 10 randomly chosen recordings out of the 55 ones of the conTIMIT dataset, are employed to model the speaker utterance duration (conTIMIT training-1 dataset). Consequently, there is a partial overlap between the conTIMIT test dataset and the conTIMIT training-1 dataset. It has been reported that BIC performance is likely to reach a limit [20]. There are 4 reasons for that: (a) estimates of the BIC model parameters are used, (b) the penalty-factor λ may not be tuned properly, (c) the data are assumed to be jointly normal, but this is an assumption, frequently not validated, for example when voiced speech is embedded into noise [37], and (d) researchers have found that BIC faces problems for small sample sets [37], [38]. Researchers tend to agree that BIC performance deteriorates when a speaker utterance does not have sufficient duration, which should be more than about 2 s. For the conTIMIT test dataset, the tolerance equals 1 s. That is, 0.5 s before and after the actual speaker change point.

For evaluation purposes, 4 systems are assessed, namely: (a) The BIC system without speaker utterance duration estimation and feature selection. This is the baseline system (system 1). (b) The BIC system with speaker utterance duration estimation (system 2). (c) The BIC system with feature selection (system 3). (d) The proposed system, that is the BIC system with speaker utterance duration estimation and feature selection (system 4). The window shift r is set equal to the half of the average speaker utterance duration for systems 2 and 4, whereas r is equal to 0.2 s for systems 1 and 3.

BIC performance on the conTIMIT test dataset without modeling the speaker utterance duration and feature selection is depicted in Table II. The performance of BIC with modeling the distribution of the speaker utterance is exhibited in Table III, while its performance when feature selection is applied only is summarized in Table IV. The overall performance of the proposed system (system 4) is summarized in Table V. For all systems, the figures of merit are computed for each audio recording and then their corresponding mean value and standard deviation are reported [43]. Concerning system 4, if (2) is replaced

TABLE II
PERFORMANCE OF BIC ON THE CONTIMIT TEST DATASET WITHOUT MODELING THE SPEAKER UTTERANCE DURATION NOR APPLYING FEATURE SELECTION.

	PRC	RCL	F_1	FAR	MDR
mean	0.446	0.647	0.516	0.352	0.353
standard deviation	0.094	0.137	0.081	0.157	0.137

TABLE III
PERFORMANCE OF BIC ON THE CONTIMIT TEST DATASET WITH MODELING THE SPEAKER UTTERANCE DURATION.

	PRC	RCL	F_1	FAR	MDR
mean	0.613	0.895	0.723	0.311	0.105
standard deviation	0.079	0.116	0.077	0.114	0.116

TABLE IV
PERFORMANCE OF BIC ON THE CONTIMIT TEST DATASET WITH FEATURE SELECTION.

	PRC	RCL	F_1	FAR	MDR
mean	0.527	0.654	0.567	0.295	0.335
standard deviation	0.159	0.137	0.110	0.177	0.150

TABLE V
PERFORMANCE OF THE PROPOSED SYSTEM ON THE CONTIMIT TEST DATASET SYSTEM (WITH MODELING THE SPEAKER UTTERANCE DURATION AND FEATURE SELECTION).

	PRC	RCL	F_1	FAR	MDR
mean	0.670	0.949	0.777	0.289	0.051
standard deviation	0.106	0.056	0.069	0.139	0.056

by BIC, it is found that PRC=0.685, RCL= 0.951, F_1=0.974, FAR= 0.303, and MDR=0.049. However, it is reminded that the improvement is achieved at the cost of constraining the generalization ability.

Our aim is to validate that each system differentiates significantly from the others concerning their mean figures of merit. First, one-way analysis of variance (one-way ANOVA) is applied. The null hypothesis tested is that the 4 system mean figures of merit are equal. The alternative hypothesis states that the differences among the figures of merit are not due to random errors, but due to variation among unequal mean figures of merit. That is, the null hypothesis declares that the systems do not differentiate significantly from one another, while the alternative hypothesis suggests that at least one of the systems differs from the remaining. The F-statistic value and its p-value for all five efficiency measures are indicated in Table VI. From Table VI, it is evident that the 4 systems are statistically different, with

TABLE VI
F-STATISTIC VALUES AND P-VALUES FOR PRC, RCL, F_1, FAR, AND MDR OF THE 4 SYSTEMS TESTED ON THE CONTIMIT TEST DATASET.

	PRC	RCL	F_1	FAR	MDR
F-statistic	36.322	90.295	103.931	1.794	81.576
p-value	$1.945\ 10^{-6}$	$2.743\ 10^{-8}$	$1.009\ 10^{-9}$	0.150	$9.324\ 10^{-7}$

respect to PRC, RCL, F_1, and MDR, whereas there appears to be no significant difference with respect to the FAR at 95% confidence level.

TABLE VII
95% CONFIDENCE INTERVALS FOR ALL PAIRWISE COMPARISONS OF THE 4 SYSTEMS FOR PRC.

systems compared	95% confidence interval
1st - 2nd	[-0.224,-0.107]
1st - 3rd	[-0.140,-0.022]
1st - 4th	[-0.282,-0.164]
2nd - 3rd	[0.027,0.145]
2nd - 4th	[-0.116,-0.002]
3rd - 4th	[-0.201,-0.083]

TABLE VIII
95% CONFIDENCE INTERVALS FOR ALL PAIRWISE COMPARISONS OF THE 4 SYSTEMS FOR RCL.

systems compared	95% confidence interval
1st - 2nd	[-0.309,-0.188]
1st - 3rd	[-0.067,-0.054]
1st - 4th	[-0.363,-0.241]
2nd - 3rd	[0.181,0.302]
2nd - 4th	[-0.114,-0.007]
3rd - 4th	[-0.356,-0.235]

One-way ANOVA assures us that at least one system is different from the others. However no information is provided about the pairs of systems that differentiate. Tukey's method or honestly significant difference method is applied to find the pairs of systems that differentiate [39]. Tukey's method is designed to make all pairwise comparisons of means, while maintaining the confidence level at a pre-defined level. Moreover, it is optimal for balanced one-way ANOVA, which is our scenario. For k systems, there are $k\left(\frac{k-1}{2}\right)$ possible combinations (e.g. 6 possible combinations are examined for $k = 4$). Tukey's method for the same number of measurements is applied, i.e. 49. The critical test statistic is obtained from the Studentized range statistic.

Since one-way ANOVA has validated that FAR differences are not significant, Tukey's method is applied to the remaining figures of merit i.e. PRC, RCL, F_1, and MDR for the same confidence level 95%. The corresponding confidence intervals for all pairwise comparisons among the 4 systems for the aforementioned figures of merit can be seen in Tables VII - X, respectively. If the confidence interval includes zero, the difference is not significant. It is clear from Tables VII - X that zero is not included in any interval. Thus, for any pairwise system comparison and any figure of merit from PCR, RCL, F_1, and MDR the difference is significant.

Accordingly, there is statistical evidence that both speaker utterance modeling and feature selection improve performance significantly either individually or combined for PRC, RCL, F_1 and MDR. This is not the case for FAR.

2) MDR histogram of the conTIMIT test dataset: We focus on the results of the proposed system for the conTIMIT test dataset depicted in Table V. The MDR histogram is plotted in Figure 3. A clear peak exists near to 0, which is the ideal case. The latter is a direct outcome of the fact that two-hypothesis BIC tests are carried out at times, where a speaker change point is most probable to occur.

TABLE IX
95% CONFIDENCE INTERVALS FOR ALL PAIRWISE COMPARISONS OF THE 4 SYSTEMS FOR F_1.

systems compared	95% confidence interval
1st - 2nd	[-0.251,-0.162]
1st - 3rd	[-0.095,-0.006]
1st - 4th	[-0.307,-0.218]
2nd - 3rd	[0.112,0.201]
2nd - 4th	[-0.101,-0.012]
3rd - 4th	[-0.257,-0.168]

TABLE X
95% CONFIDENCE INTERVALS FOR ALL PAIRWISE COMPARISONS OF THE 4 SYSTEMS FOR MDR.

systems compared	95% confidence interval
1st - 2nd	[0.186,0.311]
1st - 3rd	[-0.044,-0.081]
1st - 4th	[0.240,0.364]
2nd - 3rd	[-0.292,-0.168]
2nd - 4th	[-0.009,-0.116]
3rd - 4th	[0.221,0.346]

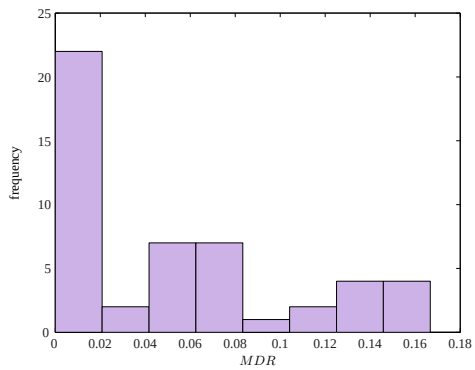

Fig. 3. The histogram of MDR in the conTIMIT test dataset.

3) Correlation among figures of merit on the conTIMIT test dataset: The correlation coefficient between the figures of merit for the conTIMIT test dataset can be seen in Table XI. The correlation coefficient between RCL and MDR is -1, as a consequence of (15). The pairs: (i) (PRC, RCL) and (PRC, MDR) (ii) (F_1, RCL) and (F_1, MDR) (iii) (FAR, RCL) and (FAR, MDR) have opposite signs. That is, when the first quantity increases, the second decreases and vice versa. The degree of linear dependence is indicated by the absolute value of the correlation index. It is seen that (PRC,F_1) and (PRC,FAR) exhibit the strongest correlation.

TABLE XI
THE CORRELATION COEFFICIENT BETWEEN THE PAIRS OF FIGURES OF MERIT FOR THE CONTIMIT TEST DATASET.

	PRC	RCL	F_1	FAR	MDR
PRC	1	-0.344	0.939	-0.945	0.344
RCL		1	-0.014	0.628	-1
F_1			1	-0.778	0.014
FAR				1	-0.628
MDR					1

4) Performance evaluation on the HUB-4 1997 English Broadcast News Speech dataset: Aiming to verify the efficiency of the proposed contributions, i.e. modeling the speaker utterance duration and feature selection on real data, RT-03 MDE Training Data Speech is utilized [44]. To facilitate performance comparisons between the proposed system and other systems, we confine ourselves to broadcast news audio recordings, that is the HUB-4 1997 English Broadcast News Speech dataset [41]. The recordings are mono-channel, the sampling frequency is 16 KHz, and the audio PCM samples are quantized in 16 bits. The selected audio recordings have a duration of approximately 1 hour.

20% of the selected audio recordings are used for estimating the speaker utterance duration distribution. For the third time, IG distribution is verified to be the best fit for modeling speaker utterance duration by both the log-likelihood and the Kolmogorov-Smirnov criteria. In this case, the mean duration equals 23.541 s and the standard deviation equals 24.210 s. Since the standard deviation value is considerably large, r is set equal to one eighth of the mean speaker utterance duration. This rather small submultiple aims to reduce the probability of missing a speaker change for the reasons explained in Section II-B.

To assess the robustness of the MFCCs shown in Table I, the same coefficients along with their corresponding delta and delta-delta coefficients have been used in the HUB-4 1997 English Broadcast News Speech dataset. The proposed algorithm is tested on the remaining 80% of the selected audio recordings. For the HUB-4 1997 English Broadcast News Speech dataset, since the dialogues are real, the tolerance should be greater, as is explained in [12]. Motivated by [23], that also employs broadcasts, the tolerance is equal to 2 s. The achieved figures of merit are: $PRC = 0.634$, $RCL = 0.922$, $F_1 = 0.738$, $FAR = 0.309$, and $MDR = 0.078$.

5) Performance discussion: Before discussing the performance of the proposed system with respect to other systems, let us argue why it is generally a good choice to minimize MD even if FA is high [12]. FA can be more easily removed [13], [15], [16], [40], for example through clustering. PRC and FAR are associated with FA, while RCL and MDR depend on MD. This means that PRC and FAR are less cumbersome to remedy than RCL and MDR.

The proposed system is evaluated on the conTIMIT test dataset first. It outperforms three other systems tested on a similar dataset, created by concatenating speakers from the TIMIT database, as described in [21]. Although the dataset in [21] is substantially smaller than the conTIMIT test dataset, the nature of the audio recordings is the same enabling us to conduct fair comparisons. The performance achieved

by the previous approaches is summarized in Table XII. Missing entries are due to the fact that not all researchers use the same set of figures of merit, which creates further implications in direct comparisons. PRC and FAR of the proposed system on the conTIMIT test dataset are slightly deteriorated than those obtained by the multiple pass system with a fusion scheme when speakers are modeled by quasi-GMMs system. However, RCL and MDR are substantially improved. RCL and MDR are also improved with respect to the two remaining systems. Finally, the superiority of the proposed system against the three systems developed in [21] is demonstrated by the fact that its F_1 value is relatively improved by 7.917%, 6.438%, and 28.007%, respectively. In [12], the used dataset was created by concatenating

TABLE XII
AVERAGE FIGURES OF MERIT IN [12] AND [21] ON A SIMILAR DATASET CREATED BY CONCATENATING SPEAKERS FROM THE TIMIT DATABASE.

system	**Database used**	PRC	RCL	F_1	FAR	MDR
Proposed system	concatenated utterances from speakers of the TIMIT database (not the same concatenation as in [12])	0.670	0.949	0.777	0.289	**0.051**
Multiple pass system with a fusion scheme [21]	concatenated utterances from speakers of the TIMIT database	0.780	0.700	0.720	0.218	0.305
Speakers modeled by quasi-GMMs system [21]	concatenated utterances from speakers of the TIMIT database	0.680	0.800	0.730	0.280	0.200
Auxiliary second-order and T^2 Hotelling statistic system [21]	concatenated utterances from speakers of the TIMIT database	0.490	0.812	0.607	0.455	0.188
Delacourt and Wellekens [12]	concatenated utterances from speakers of the TIMIT database				0.282	0.156

speaker utterances from the TIMIT database, too. However, this concatenation is not the same to the one employed here. Although FAR is slightly better that ours, the reported MDR for the proposed system is considerably lower. The relative MDR improvement equals 67.308%.

The proposed system is also assessed on the HUB-4 1997 English Broadcast News Speech dataset. As is demonstrated in Table XIII, the same dataset is utilized in [13]. The system presented in [13] is a two-step system. The first step is a ”coarse to refine” step, whereas the second step is a refinement one that aims at reducing FAR. Both our algorithm and the one in [13] apply incremental speaker model updating to deal with the problem of insufficient data in estimating the speaker model. However, the updating strategy is not the same. In [13], quasi-GMMs are utilized. Both algorithms consider FAs less cumbersome than MDs. In [13], down-sampling takes place from 16 KHz to 8 KHz and an adaptive

background noise level detection algorithm is applied. This is not required here. Another interesting point is that the tolerance in our approach is 2 s, whereas in [13] the tolerance is 3 s. Improved FAR in [13] may be partially attributed to the increased tolerance. In summary, the proposed system, when compared to that in [13], yields a relatively improved RCL by 3.600% at the expense of doubling FAR. The latter is more easily manageable than RCL.

A dataset of the same nature with the HUB-4 1997 English Broadcast News Speech is employed in [18]. The greatest PRC relative improvement of the system proposed in [18] when compared to the proposed system is 7.256%. The corresponding relative RCL deterioration is 29.501%.

TABLE XIII
THE EFFICIENCY AND THE DATASET USED BY THE PROPOSED SYSTEM AND OTHER BENCHMARK SYSTEMS.

System	**Database used**	PRC	RCL	F_1	FAR	MDR
Proposed system	HUB-4 1997 English Broadcast News Speech [41]	0.634	0.922	0.738	0.309	**0.078**
Lu and Zhang [13]	HUB-4 1997 English Broadcast News Speech [41]		0.89		0.15	
Ajmera et al. [18]	HUB-4 English Evaluation Speech and Transcripts [42]	0.68	0.65	0.67		
Cheng and Wang [23]	MATBN-2002 [23]				0.289	0.100
Kim et al. [8]	audio track from television talk show program [8]	0.754	0.864	0.805		

It should be noted that the efficiency of a speaker segmentation algorithm depends highly on the nature of the data it is designed for. There are experimental results available for different datasets, such as the MATBN-2002 database. However, a direct comparison is not feasible. For the remaining of the section, a rough discussion between systems tested on different datasets is attempted.

Concerning the performance of the proposed system on 2 different datasets, namely the conTIMIT dataset and the subset of the HUB-4 1997 English Broadcast News Speech dataset, all five figures of merit are slightly deteriorated for the HUB-4 1997 English Broadcast News Speech dataset, when compared to those measured on the conTIMIT. This is expected, due to the nature of the conTIMIT.

Metric-SEQDAC is another approach introduced by Cheng and Wang [23]. The dataset employed is the MATBN-2002 Mandarin Chinese broadcast news corpus. A ROC-diagram is provided in [23] to demonstrate the efficiency of metric-SEQDAC algorithm. From the diagram it can be deducted that for FAR=0.289 (equal to the FAR of the proposed system on the conTIMIT test dataset), the reported

MDR is roughly equal to 0.100. Once again, a great relative MDR improvement equal to 28.205% is achieved at the expense of a 6.472% relative FAR deterioration.

Kim et al. [8] presented a hybrid speaker-based segmentation, which combines metric-based and model-based techniques. Audio track from a television talk show program is used to evaluate the performance. Comparing the proposed system to the one in [8], PRC is relatively deteriorated by 15.915%, while RCL is relatively improved by 6.713%.

To sum up, the proposed system demonstrates a very low MDR compared to state-of-the-art systems.

VI. Conclusions

A novel efficient and robust approach for automatic BIC-based speaker segmentation is proposed. Computational efficiency is achieved through the speaker utterance modeling and the transformed BIC formulation, whereas robustness is attained by feature selection and application of BIC tests at appropriately selected time stamps. The first contribution of the paper is in modeling the duration of speaker utterances. As a result, computational needs are reduced in terms of time and memory. The IG distribution is found to be the best fit for the empirical distribution of speaker utterance duration. The second contribution of the paper is in MFCC selection. The third contribution is in the new theoretical formulation of BIC after centering and simultaneous diagonalization, whose computational complexity is less than that of the standard BIC, when covariance matrix estimators other than the sample dispersion matrices are used.

In order to attest that speaker utterance duration modeling and feature selection yield more robust systems, 4 systems are tested on the conTIMIT test dataset: the first utilizes the standard BIC approach, the second applies speaker utterance duration estimation, the third employs feature selection, and the fourth is the proposed system that combines all proposals made in this paper. One-way ANOVA and a posteriori Tukey's method confirm that the 4 systems, when compared pairwise, are significantly different from one another for PRC, RCL, F_1, and MDR. Accordingly, the proposed contributions either individually or in combination improve performance. Moreover, to overcome the restrictions posed by the artificial dialogues in the conTIMIT dataset, first experimental results on the HUB-4 1997 English Broadcast News Speech dataset have verified the robustness of the proposed system.

ACKNOWLEDGEMENT

M. Kotti was supported by the "Propondis" Public Welfare Foundation and E. Benetos by the "Alexander S. Onassis" Public Benefit Foundation through scholarships.

APPENDIX I

A detailed proof of the new formulation of BIC follows. Assuming that the chunks X, Y, and Z are modeled by Gaussian density functions, we define

$$\begin{aligned} A \triangleq & -N_X(\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_Z|) - \frac{1}{2}\sum_{i=1}^{N_X}(\boldsymbol{z}_i - \boldsymbol{\mu}_Z)^T\boldsymbol{\Sigma}_Z^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_Z) \\ & +N_X(\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_X|) + \frac{1}{2}\sum_{i=1}^{N_X}(\boldsymbol{z}_i - \boldsymbol{\mu}_X)^T\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_X), \end{aligned} \tag{16}$$

$$\begin{aligned} B \triangleq & -N_Y(\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_Z|) - \frac{1}{2}\sum_{i=N_X+1}^{N_Z}(\boldsymbol{z}_i - \boldsymbol{\mu}_Z)^T\boldsymbol{\Sigma}_Z^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_Z) \\ & +N_Y(\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_Y|) + \frac{1}{2}\sum_{i=N_X+1}^{N_Z}(\boldsymbol{z}_i - \boldsymbol{\mu}_Y)^T\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_Y). \end{aligned} \tag{17}$$

Under these assumptions, (5) equals to

$$\delta = A + B - \frac{\lambda}{2}\left(d + \frac{d(d+1)}{2}\right)\ \ln N_Z \gtrless 0. \tag{18}$$

If $\boldsymbol{\Sigma}_X$, $\boldsymbol{\Sigma}_Y$, and $\boldsymbol{\Sigma}_Z$ are estimated by sample dispersion matrices, it is true that

$$\sum_{i=1}^{N_X}(\boldsymbol{z}_i - \boldsymbol{\mu}_Z)^T\boldsymbol{\Sigma}_Z^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_Z) = \mathrm{tr}\{\boldsymbol{\Sigma}_Z^{-1}\sum_{i=1}^{N_X}(\boldsymbol{z}_i - \boldsymbol{\mu}_Z)(\boldsymbol{z}_i - \boldsymbol{\mu}_Z)^T\}, \tag{19}$$

$$\sum_{i=1}^{N_X}(\boldsymbol{z}_i - \boldsymbol{\mu}_X)^T\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_X) = \mathrm{tr}\{\boldsymbol{\Sigma}_X^{-1}\sum_{i=1}^{N_X}(\boldsymbol{z}_i - \boldsymbol{\mu}_X)(\boldsymbol{z}_i - \boldsymbol{\mu}_X)^T\} = dN_X. \tag{20}$$

So, (16) can be written as: $A = -N_X(\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_Z|) - \frac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}_Z^{-1}\sum_{i=1}^{N_X}(\boldsymbol{z}_i - \boldsymbol{\mu}_Z)(\boldsymbol{z}_i - \boldsymbol{\mu}_Z)^T\}$ $+N_X(\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_X|) + \frac{d}{2}N_X$. Applying the same estimation for B allows us to rewrite (18) as

$$-\frac{N_Z}{2}\ln|\boldsymbol{\Sigma}_Z| + \frac{N_X}{2}\ln|\boldsymbol{\Sigma}_X| + \frac{N_Y}{2}\ln|\boldsymbol{\Sigma}_Y| - \frac{\lambda}{2}\left(d + \frac{d(d+1)}{2}\right)\ \ln N_Z \gtrless 0, \tag{21}$$

which according to (7) corresponds to $\gamma_{BIC} \lesseqgtr 0$.

If we apply simultaneous diagonalization to $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Z$, then $\mathbf{\Sigma}_Z = \mathbf{\Phi}\mathbf{\Lambda}_Z\mathbf{\Phi}^T$, where $\mathbf{\Lambda}_Z$ and $\mathbf{\Phi}$ are the diagonal matrix of eigenvalues and modal matrix of $\mathbf{\Sigma}_Z$, respectively. Moreover, let $\mathbf{K}$=$\mathbf{\Lambda}_Z{}^{-\frac{1}{2}}\ \mathbf{\Phi}^T\mathbf{\Sigma}_X$ $\mathbf{\Phi}\mathbf{\Lambda}_Z{}^{-\frac{1}{2}}$. $\mathbf{\Lambda}_K$ is the diagonal matrix of eigenvalues of $\mathbf{K}$ and $\mathbf{\Psi}$ is the corresponding modal matrix, i.e. $\mathbf{\Lambda}_K = \mathbf{\Psi}^T\mathbf{K}\mathbf{\Psi}$. $\mathbf{W}$ is defined as $\mathbf{W} \triangleq \mathbf{\Phi}\mathbf{\Lambda}_Z{}^{-\frac{1}{2}}\mathbf{\Psi}$. It is straightforward to prove that $\mathbf{W}^T\mathbf{\Sigma}_Z\mathbf{W} = \mathbf{I}$ and $\mathbf{W}^T\mathbf{\Sigma}_X\mathbf{W} = \mathbf{\Lambda}_K$. The same procedure for simultaneous diagonalization of $\mathbf{\Sigma}_Y$ and $\mathbf{\Sigma}_Z$ takes place. Let $\mathbf{H} = \mathbf{\Lambda}_Z{}^{-\frac{1}{2}}\mathbf{\Phi}^T\mathbf{\Sigma}_Y\mathbf{\Phi}\mathbf{\Lambda}_Z{}^{-\frac{1}{2}}$. Additionally, $\mathbf{\Lambda}_H$ is the diagonal matrix of eigenvalues of $\mathbf{H}$ and $\mathbf{\Xi}$ is the corresponding modal matrix i.e. $\mathbf{\Lambda}_H = \mathbf{\Xi}^T\mathbf{H}\mathbf{\Xi}$. If $\mathbf{\Omega}$ is defined as $\mathbf{\Omega} \triangleq \mathbf{\Phi}\mathbf{\Lambda}_Z{}^{-\frac{1}{2}}\mathbf{\Xi}$, it is straightforward to prove that $\mathbf{\Omega}^T\mathbf{\Sigma}_Z\mathbf{\Omega} = \mathbf{I}$ and $\mathbf{\Omega}^T\mathbf{\Sigma}_Y\mathbf{\Omega} = \mathbf{\Lambda}_H$. The transformed (21) is $\frac{N_X}{2}\ln\frac{|\mathbf{W}\mathbf{\Lambda}_K\mathbf{W}^T|}{|\mathbf{W}\mathbf{W}^T|}$ $+\frac{N_Y}{2}\ln\frac{|\mathbf{\Omega}\mathbf{\Lambda}_H\mathbf{\Omega}^T|}{|\mathbf{\Omega}\mathbf{\Omega}^T|}$-$\frac{\lambda}{2}\left(d+\frac{d(d+1)}{2}\right)\ \ln N_Z \gtrless 0$ or equivalently,

$$\frac{N_X}{2}\sum_{i=1}^{d}\ln\lambda_i(\mathbf{\Lambda}_K) + \frac{N_Y}{2}\sum_{i=1}^{d}\ln\lambda_i(\mathbf{\Lambda}_H) - \frac{\lambda}{2}\left(d+\frac{d(d+1)}{2}\right)\ \ln N_Z \gtrless 0, \tag{22}$$

where $\lambda_i(\mathbf{\Lambda}_K)$ stands for the ith eigenvalue of $\mathbf{\Lambda}_K$ and $\lambda_i(\mathbf{\Lambda}_H)$ stands for the ith eigenvalue of $\mathbf{\Lambda}_H$.

However, sample dispersion matrices are not the only estimators for $\mathbf{\Sigma}_X$, $\mathbf{\Sigma}_Y$, and $\mathbf{\Sigma}_Z$. Besides the sample dispersion matrix, there exist other estimators of the covariance matrix, such as the robust estimators [25], [26] or the regularized MLEs [27]. For that reason, in the remaining of Appendix I, (19) and (20) are not required. Accordingly, the transformation holds for any covariance matrix estimators.

In the general case, the first transformation that takes place is centering for $\boldsymbol{z}_i \in X \cap Z = X$

$$\widetilde{\boldsymbol{z}}_i = \boldsymbol{z}_i - \boldsymbol{\mu}_Z, \quad \boldsymbol{\mu'}_X = \tfrac{1}{N_X}\textstyle\sum_{i=1}^{N_X}\boldsymbol{z}_i - \boldsymbol{\mu}_Z = \tfrac{1}{N_X}\textstyle\sum_{i=1}^{N_X}\widetilde{\boldsymbol{z}}_i, \tag{23}$$

the centered A is re-written as A' = - $\frac{N_X}{2}\ln\frac{|\mathbf{\Sigma}_Z|}{|\mathbf{\Sigma}_X|}$-$\frac{1}{2}\sum_{i=1}^{N_X}\widetilde{\boldsymbol{z}}_i^T\mathbf{\Sigma}_Z^{-1}\widetilde{\boldsymbol{z}}_i$+$\frac{1}{2}\sum_{i=1}^{N_X}\widetilde{\boldsymbol{z}}_i^T\mathbf{\Sigma}_X^{-1}\widetilde{\boldsymbol{z}}_i$ $-\frac{N_X}{2}\boldsymbol{\mu'}_X^T\mathbf{\Sigma}_X^{-1}\boldsymbol{\mu'}_X$. B is transformed to B' by an exactly similar procedure. For $\boldsymbol{z}_i \in Y \cap Z = Y$, it holds

$$\widetilde{\boldsymbol{z}}_i = \boldsymbol{z}_i - \boldsymbol{\mu}_Z, \quad \boldsymbol{\mu'}_Y = \tfrac{1}{N_Y}\textstyle\sum_{i=N_{X+1}}^{N_Z}\boldsymbol{z}_i - \boldsymbol{\mu}_Z = \tfrac{1}{N_Y}\textstyle\sum_{i=N_{X+1}}^{N_Z}\widetilde{\boldsymbol{z}}_i. \tag{24}$$

By doing so (6) can be written as:

$$-\sum_{i=1}^{N_X}\widetilde{\boldsymbol{z}}_i^T\mathbf{\Sigma}_Z^{-1}\widetilde{\boldsymbol{z}}_i + \sum_{i=1}^{N_X}\widetilde{\boldsymbol{z}}_i^T\mathbf{\Sigma}_X^{-1}\widetilde{\boldsymbol{z}}_i - \sum_{i=N_X+1}^{N_Z}\widetilde{\boldsymbol{z}}_i^T\mathbf{\Sigma}_Z^{-1}\widetilde{\boldsymbol{z}}_i + \sum_{i=N_X+1}^{N_Z}\widetilde{\boldsymbol{z}}_i^T\mathbf{\Sigma}_Y^{-1}\widetilde{\boldsymbol{z}}_i \gtrless \gamma' \tag{25}$$

where

$$\gamma' = \gamma_{BIC} + N_X\boldsymbol{\mu'}_X^T\mathbf{\Sigma}_X^{-1}\boldsymbol{\mu'}_X + N_Y\boldsymbol{\mu'}_Y^T\mathbf{\Sigma}_Y^{-1}\boldsymbol{\mu'}_Y \tag{26}$$

with γ_{BIC} defined in (7). Let us define the following auxiliary variables A'' and B'' as: $A'' \triangleq$ - $\sum_{i=1}^{N_X} \widetilde{\boldsymbol{z}}_i^T \boldsymbol{\Sigma}_Z^{-1} \widetilde{\boldsymbol{z}}_i + \sum_{i=1}^{N_X} \widetilde{\boldsymbol{z}}_i^T \boldsymbol{\Sigma}_X^{-1} \widetilde{\boldsymbol{z}}_i$, $B'' \triangleq -\sum_{i=N_X+1}^{N_Z} \widetilde{\boldsymbol{z}}_i^T \boldsymbol{\Sigma}_Z^{-1} \widetilde{\boldsymbol{z}}_i + \sum_{i=N_X+1}^{N_Z} \widetilde{\boldsymbol{z}}_i^T \boldsymbol{\Sigma}_Y^{-1} \widetilde{\boldsymbol{z}}_i$. The second transformation is the simultaneous diagonalization of $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Z$. For that case, $\boldsymbol{z}_i \in X \cap Z = X$ are transformed to $\widetilde{\boldsymbol{w}}_i = \boldsymbol{\Psi}^T {\boldsymbol{\Lambda}_Z}^{-\frac{1}{2}} \boldsymbol{\Phi}^T \widetilde{\boldsymbol{z}}_i = \mathbf{W}^T \widetilde{\boldsymbol{z}}_i$. Therefore, A'' is equal to

$$\begin{aligned} A'' &= -\sum_{i=1}^{N_X} \widetilde{\boldsymbol{w}}_i^T \boldsymbol{\Psi}^T {\boldsymbol{\Lambda}_Z}^{\frac{1}{2}} \boldsymbol{\Phi}^T \boldsymbol{\Sigma}_Z^{-1} \boldsymbol{\Phi} {\boldsymbol{\Lambda}_Z}^{\frac{1}{2}} \boldsymbol{\Psi} \widetilde{\boldsymbol{w}}_i + \sum_{i=1}^{N_X} \widetilde{\boldsymbol{w}}_i^T \boldsymbol{\Psi}^T {\boldsymbol{\Lambda}_Z}^{\frac{1}{2}} \boldsymbol{\Phi}^T \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Phi} {\boldsymbol{\Lambda}_Z}^{\frac{1}{2}} \boldsymbol{\Psi} \widetilde{\boldsymbol{w}}_i \\ &= -\sum_{i=1}^{N_X} \widetilde{\boldsymbol{w}}_i^T \widetilde{\boldsymbol{w}}_i + \sum_{i=1}^{N_X} \widetilde{\boldsymbol{w}}_i^T {\boldsymbol{\Lambda}_K}^{-1} \widetilde{\boldsymbol{w}}_i. \end{aligned} \tag{27}$$

The same procedure for simultaneous diagonalization of $\boldsymbol{\Sigma}_Y$ and $\boldsymbol{\Sigma}_Z$ is applied. Then, $\boldsymbol{z}_i \in Y \cap Z = Y$ are transformed to $\widetilde{\boldsymbol{v}}_i = \boldsymbol{\Xi}^T {\boldsymbol{\Lambda}_Z}^{-\frac{1}{2}} \boldsymbol{\Phi}^T \widetilde{\boldsymbol{z}}_i = \boldsymbol{\Omega}^T \widetilde{\boldsymbol{z}}_i$. Accordingly, we obtain

$$B'' = -\sum_{i=N_X+1}^{N_Z} \widetilde{\boldsymbol{v}}_i^T \widetilde{\boldsymbol{v}}_i + \sum_{i=N_X+1}^{N_Z} \widetilde{\boldsymbol{v}}_i^T {\boldsymbol{\Lambda}_H}^{-1} \widetilde{\boldsymbol{v}}_i. \tag{28}$$

By using (27) and (28), (25) is rewritten as:

$$\sum_{i=1}^{N_X} \widetilde{\boldsymbol{w}}_i^T ({\boldsymbol{\Lambda}_K}^{-1} - \mathbf{I}) \widetilde{\boldsymbol{w}}_i + \sum_{i=N_X+1}^{N_Z} \widetilde{\boldsymbol{v}}_i^T ({\boldsymbol{\Lambda}_H}^{-1} - \mathbf{I}) \widetilde{\boldsymbol{v}}_i \gtrless \gamma' \tag{29}$$

whose left side is a weighted sum of squares.

APPENDIX II

The computational cost of the left part of standard BIC, as appears in (6), is calculated here approximately. By flop we denote a single floating point operation, i.e. a floating point addition or a floating point multiplication [45]. This is a crude method for estimating the computational cost. Robust statistics [25], [26] are assumed for the computation of $\boldsymbol{\Sigma}_X$, $\boldsymbol{\Sigma}_Y$, $\boldsymbol{\Sigma}_Z$. The standard BIC left part computational cost is detailed in Table XIV.

Adding all the above computational costs plus 2 flops for the additions among the terms (13)-(15) of Table XIV, the final cost is

$$3d^3 + 6N_Z d^2 + (8N_Z + 3)d + 2. \tag{30}$$

For the left part of the transformed BIC, the calculation is summarized in Table XV. It includes the cost for all the transformations, as described in Appendix I, and for the computation of the left part of transformed BIC, as appears in (29). It should be noted that the computational cost for the derivation

TABLE XIV
STANDARD BIC LEFT PART COMPUTATIONAL COST.

Term Index	Evaluated Term	Computational Cost
1	$\mathbf{\Sigma}_X$	$N_X d^2$
2	$\mathbf{\Sigma}_Y$	$N_Y d^2$
3	$\mathbf{\Sigma}_Z$	$N_Z d^2$
4	$\boldsymbol{\mu}_Z$	$N_Z d + d$
5	$\boldsymbol{\mu}_X$	$N_X d + d$
6	$\boldsymbol{\mu}_Y$	$N_Y d + d$
7	$\boldsymbol{z}_i - \boldsymbol{\mu}_Z$, $i = 1, \ldots, N_Z$	$N_Z d$
8	$\boldsymbol{z}_i - \boldsymbol{\mu}_X$, $i = 1, \ldots, N_X$	$N_X d$
9	$\boldsymbol{z}_i - \boldsymbol{\mu}_Y$, $i = 1, \ldots, N_Y$	$N_Y d$
10	${\mathbf{\Sigma}_Z}^{-1}$	d^3
11	${\mathbf{\Sigma}_X}^{-1}$	d^3
12	${\mathbf{\Sigma}_Y}^{-1}$	d^3
13	$\sum_{i=1}^{N_Z} (\widetilde{\boldsymbol{z}}_i - \boldsymbol{\mu}_Z)^T {\mathbf{\Sigma}_Z}^{-1} (\widetilde{\boldsymbol{z}}_i - \boldsymbol{\mu}_Z)$	$N_Z(2d^2 + 2d)$
14	$\sum_{i=1}^{N_X} (\widetilde{\boldsymbol{z}}_i - \boldsymbol{\mu}_X)^T {\mathbf{\Sigma}_X}^{-1} (\widetilde{\boldsymbol{z}}_i - \boldsymbol{\mu}_X)$	$N_X(2d^2 + 2d)$
15	$\sum_{i=N_{X+1}}^{N_Z} (\widetilde{\boldsymbol{z}}_i - \boldsymbol{\mu}_Y)^T {\mathbf{\Sigma}_Y}^{-1} (\widetilde{\boldsymbol{z}}_i - \boldsymbol{\mu}_Y)$	$N_Y(2d^2 + 2d)$

TABLE XV
TRANSFORMED BIC LEFT PART COMPUTATIONAL COST.

Term Index	Evaluated Term	Computational Cost
1	$\mathbf{\Sigma}_X$	$N_X d^2$
2	$\mathbf{\Sigma}_Y$	$N_Y d^2$
3	$\mathbf{\Sigma}_Z$	$N_Z d^2$
4	$\boldsymbol{\mu}_Z$	$N_Z d + d$
5	$\boldsymbol{z}_i - \boldsymbol{\mu}_Z$, for $X \cap Z$ in (23)	$N_X d$
6	$\boldsymbol{\mu'}_X$ in (23)	$N_X d + d$
7	$\boldsymbol{z}_i - \boldsymbol{\mu}_Z$, for $Y \cap Z$ in (24)	$N_Y d$
8	$\boldsymbol{\mu'}_Y$ in (24)	$N_Y d + d$
9	$\mathbf{W}$	$14d^3$[45]
10	$\mathbf{\Omega}$	$14d^3$ [45]
11	$\widetilde{\boldsymbol{w}}_i = \mathbf{W}^T \widetilde{\boldsymbol{z}}_i$, for $\widetilde{\boldsymbol{w}}_i \in X$	$2N_X d^2$
12	$\widetilde{\boldsymbol{v}}_i = \mathbf{\Omega}^T \widetilde{\boldsymbol{z}}_i$, for $\widetilde{\boldsymbol{v}}_i \in Y$	$2N_Y d^2$
13	$\mathbf{\Lambda}_K^{-1}$	d
14	$\mathbf{\Lambda}_H^{-1}$	d
15	$\sum_{i=1}^{N_X} \widetilde{\boldsymbol{w}}_i^T ({\mathbf{\Lambda}_K}^{-1} - \mathbf{I}) \widetilde{\boldsymbol{w}}_i$	$4N_X d$
16	$\sum_{i=N_X+1}^{N_Z} \widetilde{\boldsymbol{v}}_i ({\mathbf{\Lambda}_H}^{-1} - \mathbf{I}) \widetilde{\boldsymbol{v}}_i$	$4N_Y d$

of $\mathbf{W}$ that simultaneously diagonalizes $\mathbf{\Sigma}_Z$, and $\mathbf{\Sigma}_X$, such that $\mathbf{W}^T \mathbf{\Sigma}_Z \mathbf{W} = \mathbf{I}$ and $\mathbf{W}^T \mathbf{\Sigma}_X \mathbf{W} = \mathbf{\Lambda}_K$ is included [45, pp. 463-464]. This is also true for matrix $\mathbf{\Omega}$ that simultaneously diagonalizes $\mathbf{\Sigma}_Z$, and $\mathbf{\Sigma}_Y$. The total cost of the transformations and the left part of the transformed BIC equals the sum of the terms that appear in Table XV plus 1 for the addition between the terms (15) and (16). This cost is $28d^3 + 4N_Z d^2 + (7N_Z + 5)d + 1$. Moreover, as can be seen in (26), there is an additional differential cost with respect to BIC for the right part of the transformed BIC. This cost is analyzed in Table XVI. The total differential cost for the right part of the transformed BIC, is the sum of the terms (1)-(4) that

TABLE XVI
DIFFERENTIAL COMPUTATIONAL COST FOR THE RIGHT PART OF THE TRANSFORMED BIC.

Term Index	Evaluated Term	Computational Cost
1	$\mathbf{\Sigma}_X^{-1}$	d^3
2	$\mathbf{\Sigma}_Y^{-1}$	d^3
3	$\boldsymbol{\mu'}_X^T \mathbf{\Sigma}_X^{-1} \boldsymbol{\mu'}_X$	$2d^2 + 2d$
4	$\boldsymbol{\mu'}_Y^T \mathbf{\Sigma}_Y^{-1} \boldsymbol{\mu'}_Y$	$2d^2 + 2d$

appear in Table XVI, plus 2 multiplications and 2 additions, i.e. $2d^3 + 4d^2 + 4d + 4$. Accordingly, the total computational cost for the transformed BIC, excluding the cost of γ_{BIC} is

$$30d^3 + (4N_Z + 4)d^2 + (7N_Z + 9)d + 5. \tag{31}$$

Since $N_Z \gg d$, it is obvious that N_Z bears the main computational cost. In particular, typical values are $d = 72$, $N_Z = 25,000$. By comparing (30) and (31), it is clear that transformed BIC has a

significantly reduced computational cost. The computational gain in flops is defined as the subtraction of the standard BIC computational cost minus the transformed BIC computational cost. The aforementioned computational gain, with respect to various N_Z and d values, can be seen in Figure 4. The total

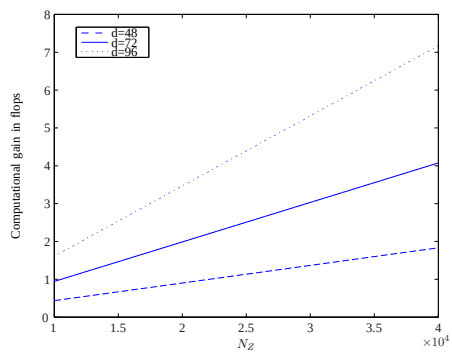

Fig. 4. The computational gain in flops for several N_Z and d values.

computational cost gain if the speaker utterance duration estimation is used in conjunction with the transformed BIC rather than the standard BIC with no speaker utterance duration estimation, equals $1 - \frac{u}{r}\frac{30d^3+(4N_Z+4)d^2+(7N_Z+9)d+5}{3d^3+6N_Zd^2+(8N_Z+3)d+2}\%$.

REFERENCES

[1] 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. *2004 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 5, pp. 925-928, Montreal, Canada, May 2004.

[2] The Segmentation Task: Find the Story Boundaries, http://www.nist.gov/speech/tests/tdt/tdt99/presentations/NIST_segmentation/index.htm

[3] NIST Rich Transcription Evaluation, http://www.nist.gov/speech/tests/rt/

[4] T. Wu, L. Lu, K. Chen, and H. Zhang, "UBM-based real-time speaker segmentation for broadcasting news", in Proc. *2003 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, pp. 193-196, Hong Kong, April, 2003.

[5] S. Know and S. Narayanan, "Unsupervised speaker indexing using generic models", *IEEE Trans. Audio, Speech, and Language Processing*, vol. 13, no. 5, pp. 1004-1013, September 2005.

[6] M. Collet, D. Charlet, and F. Bimbot, "A correlation metric for speaker tracking using anchor models", in Proc. *2005 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, pp. 713-716, Philadelphia, USA, March 2005.

[7] 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, Budapest, Hungary, September 1999.

[8] H. Kim, D. Elter, and T. Sikora, “Hybrid speaker-based segmentation system using model-level clustering,” in Proc. *2005 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. I, pp. 745-748, Philadelphia, USA, March 2005.

[9] S. Meignier, D. Moraru, C. Fredouille, J. F. Bonastre, and L. Besacier, “Step-by-step and integrated approaches in broadcast news speaker diarization”, *Computer Speech and Language*, vol. 20, no. 2-3, pp. 303-330, April-July 2006.

[10] J. A. Arias, J. Pinquier, and R. André-Obrecht, “Evaluation of classification techniques for audio indexing”, in Proc. *13th European Signal Processing. Conf.*, Antalya, Turkey, September 2005.

[11] S. Know and S. Narayanan, “Speaker change detection using a new weighted distance measure,” in Proc. *Int. Conf. Spoken Language*, vol. 4, pp. 2537-2540, Colorado, USA, September 2002.

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

[13] L. Lu and H. Zhang, “Unsupervised speaker segmentation and tracking in real-time audio content analysis”, *Multimedia Systems*, vol. 10, no. 4, pp. 332-343, April 2005.

[14] H. Harb and L. Chen, “Audio-based description and structuring of videos”, *Int. J. Digital Libraries*, vol. 6, no. 1, pp. 70-81, February 2006.

[15] B. Zhou and J. H. L. Hansen, “Efficient audio stream segmentation via the combined T^2 statistic and the Bayesian information criterion”, *IEEE Trans. Audio, Speech, and Language Processing*, vol. 13, no. 4, pp. 467-474, July 2005.

[16] J. H. L. Hansen, R. Huang, B. Zhou, M. Seadle, J. R. Deller, A. R. Gurijala, M. Kurimo, and P. Angkititrakul, “SpeechFind: Advances in spoken document retrieval for a national gallery of the spoken word”, *IEEE Trans. Audio, Speech, and Language Processing*, vol. 13, no 5, pp. 712- 730, September 2005.

[17] M. Cettolo and M. Vescovi, “Efficient audio segmentation algorithms based on the BIC”, in Proc. *2003 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 6, pp. 537-540, Hong Kong, April 2003.

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

[19] D. A. Reynolds and P. Torres-Carrasquillo, “Approaches and applications of audio diarization”, in Proc. *2005 IEEE Int. Conf. Acoustics, Speech, and signal Processing*, vol. 5, pp. 953-956, Philadelphia, USA, March 2005.

[20] C. H. Wu, Y. H. Chiu, C. J. Shia, and C. Y. Lin, “Automatic segmentation and identification of mixed-language speech using delta-BIC and LSA-based GMMs”, *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 266-276, January 2006.

[21] M. Kotti, L. G. P. M. Martins, E. Benetos, J. S. Cardoso, and C. Kotropoulos, “Automatic speaker segmentation using multiple features and distance measures: A comparison of three approaches”, in Proc. *2006 IEEE Int. Conf. Multimedia and Expo*, pp. 1101-1104, Toronto, Canada, July 2006.

[22] C. H. Wu and C. H. Hsieh, “Multiple change-point audio segmentation and classification using an MDL-based Gaussian model”, *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 647- 657, March 2006.

[23] S. Cheng and H. Wang, “Metric SEQDAC: A hybrid approach for audio segmentation”, in Proc. *8th Int. Conf. Spoken Language Processing*, Jeju, Korea, October 2004.

[24] F. Van der Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*, London, UK: Wiley, 2004.

[25] N. A. Campbell, “Robust procedures in multivariate analysis I: Robust covariance estimation”, *Applied Statistics*, vol. 29, no. 3, pp. 231-237, 1980.

[26] G. A. F. Seber, *Multivariate Observations*, N.Y.: John Wiley and Sons, 1994.

[27] S. Tadjudin and D. A. Landgrebe, “Covariance estimation with limited training samples,”, *IEEE Trans. Geoscience and Remote Sensing*, pp. 2113-2118, vol. 37, no. 4, July 1999.

[28] J. S. Garofolo, “DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus”, Linguistic Data Consortium, Philadelphia, 1993.

[29] R. D. Reiss and M. Thomas, *Statistical Analysis of Extreme Values*, Basel : Birkhäuser Verlag, 1997.

[30] M. Kotti, E. Benetos, C. Kotropoulos, and I. Pitas, “A neural network approach to audio-assisted movie dialogue detection”, *Neurocomputing, Special Issue: Advances in Neural Networks for Speech and Audio Processing*, vol. 71, no. 1-3, pp. 157-166, December 2007.

[31] J. L. Folks and R. S. Chhikara, “The inverse Gaussian distribution and its statistical application - A review”, *J. R. Statist. Soc. B*, vol. 40, pp. 263-289, 1978.

[32] N. L. Johnson, S. Kotz, and S. Balakrishnan, *Continuous Univariate Distributions, Volume 1*, N.Y.: Wiley, 1994.

[33] S. I. Boyarchenko and S. Z. Levendorskiĭ, “Perpetual american processes under Lèvi processes”, *SIAM J. Control Optim.*, vol. 40, no. 6, pp. 1514-1516, June 2001.

[34] M. C. K. Tweedie, “Statistical properties of inverse Gaussian distributions I”, *Annals of Mathematical Statistics*, vol. 28, no. 2, pp. 362-377, June 1957.

[35] X. D. Huang, A. Acero, and H. -S. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Upper River Saddle: Pearson Education - Prentice Hall, 2001.

[36] R. Kohavi and G. H. John, “Wrappers for feature subset selection”, *Artificial Intelligence*, pp. 273-324, vol. 97, no. 1-2, December 1997.

[37] G. Almpanidis and C. Kotropoulos, “Phonemic segmentation using the generalised Gamma distribution and small sample Bayesian information criterion,” *Speech Communication*, vol. 50, no. 1, pp. 38-55, January 2008.

[38] M. Kotti, V. Moschou, and C. Kotropoulos, “Speaker segmentation and clustering”, *Signal Processing*, vol. 88, no. 5, pp. 1091-1124, May 2008.

[39] J. J. Barnette and J. E. McLean, “The Tukey honestly significant difference procedure and its control of the type I error-rate”, in Proc. *Annual meeting Mid-South Educational Research Association*, New Orleans, LA, 1998.

[40] C. Barras, X. Zhu, S. Meignier, and J. L. Gauvain, “Multistage speaker diarization of broadcast news”, *IEEE Trans. Audio, Speech, and Language Processing*, pp. 1505-1512, vol. 14, no. 5, September 2006.

[41] J. Fiscus, “1997 English Broadcast News Speech (HUB4)”, Linguistic Data Consortium, Philadelphia, 1998.

[42] D. Graff, J. Fiscus, and J. Garofolo, “1997 HUB4 English Evaluation Speech and Transcripts”, Linguistic Data Consortium, Philadelphia, 2002.

[43] R. R. Korfhage, *Information Storage and Retrieval*, USA: Wiley and Sons, 1997.

[44] S. Strassel, C. Walker, and H. Lee, “RT-03 MDE Training Data Speech”, Linguistic Data Consortium, Philadelphia, 2004.

[45] G. H. Golub and C. F. Van Loan, *Matrix Computations*. 3rd Ed., Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1996.