Two Step Speaker Segmentation Method Using Bayesian Information Criterion and Adapted Gaussian Mixtures Models

Matej Grašič, Marko Kos, Andrej Žgank, Zdravko Kačič

Faculty of Electrical Engineering and Computer Science
University of Maribor, Smetanova ul. 17, SI-2000 Maribor, Slovenia
{matej.grasic, marko.kosl, andrej.zgank, kacic}@uni-mb.si

Abstract

This paper addresses the topic of online unsupervised speaker segmentation in a complex audio environment as it is present in the Broadcast News databases. A new two stage speaker change detection algorithm is proposed, which combines the Bayesian Information Criterion with an ABLS-SCD statistical framework where adapted Gaussian mixture models are used to achieve higher accuracy. To enhance the performance of the proposed method a sub-window dependent threshold selection strategy for the ABLS-SCD is introduced. Also an additional window selection strategy for the proposed method is presented. Experimental design and test evaluation were carried out on the Slovenian BNSI Broadcast News database.

Index Terms: speaker segmentation, Bayesian Information Criterion, statistical speaker modeling, Universal Background Model

1. Introduction

Nowadays complex audio sources are widely involved in various processing systems. Speaker turn detection (SCD) modules are applied in systems like speaker tracking, audio indexing and other multimedia applications where this information is used to detect scene or topic changes. SCD modules also benefit annotators who manually annotate speech databases. The main aim of speaker turn detection is to select speaker boundary change points within a conversational audio stream.

The most extensive speaker segmentation technique used reformulates segmentation as a model selection task between two competing models. The Bayesian Information Criterion (BIC) is often applied for the task of model selection, as it is very effective and straightforward [1], [2], [3]. Another advantage of this method is that no prior knowledge about the acoustic conditions is required and also no prior model training is needed, which makes online operation possible. In addition to BIC many other alternative approaches have been introduced, for example speaker segmentation using the distortion measure [4], speaker segmentation based on speaker pitch information [5] and speaker segmentation using a cross probabilities measure (XBC) [6].

Selection of a speaker boundary candidate using statistical speaker segmentation methods involves comparison of audio speech data in two segments of an analyses window divided by a hypothesized speaker change point. Statistical segmentation methods normally require a fixed size analysis window, with equal sub-windows. In a domain like Broadcast News speech recognition task one can expect speech segments of short length (less than 3s). To detect such short speaker turns while not mixing speech of different speakers, the sub-window size has to be limited to the minimum length, of the expected speaker segment. The drawback of the mentioned restriction is noticeable when processing longer speaker segments, where a short analysis window limits the amount of data, which is needed for high-quality speaker model representation. In order to generate a reliable speaker model, where phonetic variations are averaged out, at least 7s of individual speaker’s speech are needed [7].

Recently, a speaker segmentation method was introduced with the aim to provide high-quality speaker segmentation in adverse acoustic conditions with many short speaker turns (e.g. Broadcast News task). The method uses Universal Background Model Gaussian Mixture Models (UBM GMM) in a statistical framework called Adapted Model-based Bilateral Scoring-based Speaker Change Detection (ABLS-SCD) [7].

With the motivation to improve speaker segmentation in adverse acoustic conditions with many short speaker segments a new speaker segmentation method is proposed in this paper. The method proposed is based on the BIC and ABLS-SCD criterion in a combined two step system. Speaker boundary turns are selected using BIC and the candidates verified (accepted or rejected) using the ABLS-SCD criterion in a combined two step system.

The remainder of this document is organized as follows. Section 2 describes the motivation behind the presented work. Section 3 describes the BIC speaker segmentation method. Section 4 presents the statistical ABLS-SCD framework. Section 5 introduces the proposed segmentation method. Section 6 presents the experimental design. In section 7 we provide performance evaluation. The conclusion is given in Section 8.

2. Motivation

BIC is an efficient segmentation method with reasonable computational complexity and only one parameter to set. Its advantage is also the capability of segmenting data segments of different size, as opposed to XBIC [6] and ABLS-SCD [7] where a fixed-window is required. In our previous work [15] we presented the performance advantage of BIC achieved when using variable window selection (where segments of different size are selected) as opposed to using fixed window selection. In a variable window selection scheme more data can be employed for speaker modelling, which significantly improves the speaker segmentation performance.

Although good performance can be achieved, BIC still has some performance disadvantages. Particularly segmentation of short speaker segments (shorter than 3 seconds) can be problematic using the BIC as it was pointed out in [2]. The ABLS-SCD statistical framework has
advantages regarding this aspect, as it applies a priori speech model and score normalization to compensate for speaker and noise variations. In our previous analysis [15] where BIC and ABLS-SCD were compared using the same fixed windows selection system, ABLS-SCD achieved a higher performance as BIC. Nevertheless, BIC with a variable window selection still outperformed ABLS-SCD.

This paper proposes a new speaker segmentation method based on BIC and ABLS-SCD, with the motivation of combining the advantages of both methods.

3. Speaker Segmentation Using BIC

The Bayesian Information Criterion is an asymptotically optimal likelihood criterion penalized by the model complexity, introduced by Schwarz in 1971 [8]. BIC is often applied as a model-selection criterion used to decide, which of parametric models best represents given data samples [9]. The BIC likelihood expression is defined as:

$$BIC_i = \log P(X | M_i) - \frac{1}{2} \beta_i \log N.$$  \hspace{1cm} (1)

Where $\log P(X | M_i)$ is the log likelihood of training data $X$ for model $M_i$ and $\beta_i$ is the weight of the second term. The item $\beta_i$ represents the number of parameters in the model $M_i$ [10].

3.1. Speaker turn segmentation

To find a speaker segment boundary at frame $i$ a model selection test is done, where model $M_2$ over $M_1$ is chosen if the expression $\Delta BIC = BIC_i - BIC_2$ is negative. $M_1$ is defined using data $X$ drawn from a single full-covariance Gaussian whereas $M_2$ is defined using data $X=[X_1, X_2]$ from two full covariance Gaussians.

Full covariance matrices are employed to define model complexity. Actual speaker boundary points are selected according to:

$$\Delta BIC_i = \frac{N}{2} \log |\Sigma_0| + \frac{i}{2} \log |\Sigma_1|$$

$$+ \frac{N \cdot i}{2} \log |\Sigma_2| + \frac{1}{2} \beta_i \frac{d(d+1)}{2} \log N,$$  \hspace{1cm} (2)

at frame $i$ when $\Delta BIC_i$ is a negative value. Here $|\Sigma_0|$ is the determinant of the covariance of the tested window (model $M_1$), $|\Sigma_1|$ is the determinant of the covariance of the first sub-window (corresponding to model $M_2$: first Gaussian) and $|\Sigma_2|$ is the determinant of the covariance of the second sub-window (corresponding to model $M_2$: second Gaussian).

4. Segmentation using adapted GMMs

Model based speaker segmentation approaches where prior speech data is given and statistically modeled (e.g. with GMM) usually achieves good segmentation results. If enough statistically relevant speech data is available, reliable acoustic models can be generated, which are also a fine representation of short speaker segments. This directly influences speaker segmentation performance. When online speaker segmentation is needed, no prior information can be given and therefore such an approach can’t be used directly.

4.1. ABLS-SCD statistical framework

The ABLS-SCD method is based on a statistical framework which involves a cross log likelihood measure and score normalization to measure speaker differences. A none-adjusting fixed-size window is utilized for speaker analysis. Potential speaker boundary points are hypothesized in the mid point of the window as shown on Figure 1. The resulting sub-windows define two speech sub-segments as:

$$X_a^i = \{x_1^i, x_2^i, ..., x_{i+1}^i\},$$

$$X_b^i = \{x_{i+1}^n, x_{i+2}^n, ..., x_n^i\},$$  \hspace{1cm} (3)

where $x^i_{i+n}$ is the $(i+n)th$ feature vector of the conversational audio stream, and $N$ is the size of the analysis window.

The ABLS-SCD framework proposes a statistical criterion for speaker boundary detection as follows:

$$\rho_{scd} = \left\{ \left( L(X_a^i | M_{i-1}^i) - L(X_b^i | M_{i-1}^i) \right) \right\}$$

$$+ \left\{ \left( L(X_a^i | M_{i-1}^i) - L(X_b^i | M_{i-1}^i) \right) \right\},$$  \hspace{1cm} (4)

where the notation $M_{UBM}$ represents an UBM GMM model, $M_i$ represents the model adapted with data from the left sub-window and $M_b$ represents the model adapted with data from the right sub-window. In (4) $L$ is represented as $L(\cdot) = \log(\cdot)$. $L_{i-1}^a$, speaker vector is considered as speaker boundary candidate if the criterion (4) is greater than DTH, which represents the decision threshold. DTH is defined during the development phase using an empirical approach.

5. The two step segmentation method

The proposed segmentation method based on the Bayesian Information Criterion and Adapted Model-based Bilateral Scoring-based Speaker Change Detection (BIC-ABLS-SCD) specifies a two step segmentation scheme where speaker turn candidates are hypothesized in the first step and accepted or rejected during the second step. BIC with a lower $\lambda$ penalty threshold (low miss rate) and variable window selection is included in the first step. The second step is based on the ABLS-SCD framework with variable window and equal sub-window size, which detects the final speaker change point candidates. In the proposed approach BIC is applied to pre-select speaker candidates and to select sub-window boundaries in which only speech of one speaker has to be found. Our previous analysis [15] indicated that ABLS-SCD is a more robust segmentation criterion than BIC. To take advantage of this ability, longer equal size sub-windows are extracted from the BIC window selection information. This allows more data to be used with ABLS-SCD and a performance advantage should be noticed compared with the baseline approach where a fixed size window is used.

5.1. Window selection strategy

Proper window selection strategy is crucial for high accuracy. Window selection has to be performed in a way that prevents a mix-up of more than one speaker vectors in one sub-segment. Additional care should be taken when selecting the size of sub-windows, as a small number of vectors results in an imperfect speaker model representation. At this point false insertions or deletions of boundaries can occur [3].

For the proposed segmentation method we use variable window selection for BIC [11] and variable equal size sub-
window selection for ABLS-SCD. Figure 1 presents the sub-window selection strategy as it is used with BIC. Figure 1A is showing sub-window selection in the initial window; BIC is not applied at the edges of the window as there are not enough data to build a proper speaker model. The process shown on Figure 1A is repeated after each window re-selection. Figures 1B and 1C show the window selection-reselection strategy, where the window size is incremented to $\text{MAXWINSIZE}$. $\text{MAXWINSIZE}$ was defined using prior information collected during the development phase. If the window is equal or greater than $\text{MAXWINSIZE}$ the window is reselected according to Figure 1C.

We propose a new variable sub-window selection strategy for the ABLS-SCD, where the size of the sub-windows is adapted to the size of the smaller sub-window used in BIC. Figures 1A and 1D show the adaptation process used for the ABLS-SCD sub-window reselection, where $T$ point represents the BIC hypothesized speaker turn candidate, at which the ABLS-SCD speaker turn test is applied using the adapted window and sub-window.

5.2. ABLS–SCD threshold selection

To compare the values of the ABLS-SCD criterion under various conditions (e.g. influence of sub-window size) the mean ABLS-SCD criterion value with different sub-window sizes was analyzed on the BNSI Broadcast News database [12]. The result of this analysis is presented on Figure 2 where the correlation of sub-window size and the mean ABLS-SCD criterion value is shown using a quadratic polynomial fit. This polynomial curve was later applied to compute the threshold in relationship to the sub-window size. Threshold adaptation was carried out as:

$$DTH = P_1 \cdot SWS^3 + P_2 \cdot SWS + P_3,$$

where $DTH$ represents the calculated threshold, $SWS$ the sub-window size and $(P_1, P_2, P_3)$ are polynomial coefficients. In (5) $P_3$ is used as a threshold tuning parameter to control the miss and insertion rate.

6. Experimental design

6.1. Slovenian BNSI Broadcast News speech database

The Slovenian BNSI Broadcast News speech database was designed in cooperation between University of Maribor and the Slovenian national broadcaster RTV Slovenia [12]. The BNSI database comprehends two different types of TV-news shows. The first type is evening news where general overview of daily events is given. The second types of show are late night news where major events of the day are analyzed. The speech corpus consists of 42 news shows, which account for 36 hours of speech material. This material is further grouped into three sets: training, development and evaluation, respectively. The size of the training set is 30 hours, whereas the size of the development and evaluation set is 3 hours each.

Two most frequent focus conditions in the BNSI database are F0 (read studio speech, 36.6%) and F4 (read or spontaneous speech with background other than music, 37.6%). 16.2% of speech in the database is spontaneous in studio environment (F1), while 6.0% is spoken in presence of background music (F3). Altogether 1565 different speakers are present in the BNSI database. The majority, 1069 of them, are male, while 477 are female. The gender of remaining 19 speakers was annotated as unknown.

Distribution of speaker segments in the BNSI database according to their length is shown on Figure 3. It can be seen that approx. 9% of speaker segments are shorter than 3 seconds. Such short segments are often incorrectly segmented when a segmentation method like BIC is applied.
speech/non-speech classification was done using the manual reference transcriptions. The UBM GMM for ABLS-SCD model was trained using 450 mixtures and 2 hours of speech data from the train set. The training data included only audio segments labeled as speech, which could also occur in adverse acoustic conditions (e.g. background music, overlapping speech). The non-speech segments were excluded from the training and testing phase. Only diagonal-covariance matrices were involved in training and adaptation of the UBM model.

The UBM GMM model training required for ABLS-SCD during the development phase was performed with the Expectation-maximization algorithm (EM) [13], where speech samples from a given speech database were used, whereas non-speech samples were discarded. According to (4), adaptation of the UBM GMM was performed using the maximum a posteriori (MAP) adaptation [14]. To reduce the computational complexity only means were adapted. Further on, MAP adaptation was carried out using a relevance factor of 14. BIC was employed with the advance window selection strategy as described in section 5.1, where the MAXWINSIZE was set to 20 seconds, with a 4 seconds initial window size.

7. Results

7.1. Performance evaluation

Three types of performance measures were used in the evaluation process: R - Recall (% of detected speaker boundary points), P - Precision (% of correct speaker boundary points found) and F - measure (defined as $2 \cdot R \cdot P / (R + P)$). At each detected speaker change a margin was set for evaluation. If the hypothesized speaker change was within the 1 second margin before or after the reference point, then the turn was counted as correct.

7.2. Results

Table 1 presents the speaker segmentation results for all three approaches involved. The proposed method achieved better overall performance, when compared with other two included in the experiment. The overall performance (F – measure) is increased for approx. 3% absolute. The results show that the proposed method tends to a lower false alarm rate (Precision) and a higher miss rate (Recall). A plausible assumption is that the ABLS-SCD step in the proposed method compensates the false alarm rate. It can be concluded from the results that the reduced false alarm rate represents the main benefit of the two step segmentation method.

During the initial experiments with the two step segmentation method a lower λ penalty factor was set for the BIC segmentation. The results showed that the same factor as for standalone BIC had to be used for optimal performance. We have noticed that when the λ penalty factor for BIC was decreased (which increases Recall and decreases Precision) speaker turns are set earlier. This doesn’t present a problem for standalone BIC segmentation, but in the proposed method an incorrectly positioned hypothesized speaker candidate often gets erroneously rejected in the second step of segmentation (ABLS-SCD).

8. Conclusions

A new two step speaker segmentation method (BIC-ABLS-SCD) was introduced that combines the well known BIC segmentation method and the ABLS-SCD statistical framework. The proposed method was tested on the Slovenian BNSI Broadcast News database where a performance increase was shown when compared with the results of the baseline methods. In future work the threshold selection strategy will be compared with an approach where various threshold values will be tuned manually for different sub-window sizes.

9. References