# UNSUPERVISED SPEAKER CHANGE DETECTION FOR MOBILE DEVICE RECORDED SPEECH

*Olli Vuorinen, Johannes Peltola, Satu-Marja Mäkelä*

VTT Technical Research Centre of Finland, Oulu, Finland
email: firstname.surname@vtt.fi

**ABSTRACT**

In this paper we propose an unsupervised Speaker Change Detection (SCD) system developed for mobile device applications. We use Bayesian Information Criterion (BIC) to find initial speaker changes, which are then verified or discarded in the second phase by utilizing modified BIC and silence detector information. Silence information usage after initial BIC in decision making is useful to separate real changes from noise peaks. Enhanced Peak detector adjusts BIC penalty parameter automatically, which improve the robustness and feasibility. Improved BIC based False Alarm Compensation (FAC) merges effectively consecutive segments belonging to same speaker. Our experiments have shown the robustness of the algorithm and it produces very satisfactory results for difficult mobile phone recorded speech data.

***Index Terms***— speaker segmentation, speaker change detection, mobile audio segmentation, multimedia database, metadata

## 1. INTRODUCTION

Mobile devices and their novel applications are handling increasing amounts of multimedia content such as video, audio, images, messages and music. Automatic metadata extraction from video and audio recordings enables the development of sophisticated multimedia content management applications which can help users to manage their personal recordings.

Speaker segmentation is a necessary step for several indexing tasks and speaker segmentation research have been applied on audio material from news archives, digital libraries and TV program/movies. In these indexing tasks audio material is professionally created and edited with high quality.

The personal recordings created with camcorders or camera phones do not have as good recording properties as professional equipments. Currently used sampling rate in mobile video capture tools is as low as 8 kHz. Speech signal is also affected with speech coding, such as Advanced Multi Rate (AMR), and speech enhancement algorithms. Amateur users also will typically not create a clear structure for their videos. Video clips are short and audio quality is typically worse than in professionally created camcorder videos [1]. Above mentioned properties of mobile audio recordings make speaker segmentation task even more challenging compared with earlier mentioned indexing tasks using professionally created high quality audio.

Useful preceding task for speaker segmentation is audio analysis, which classifies audio recording to speech and non speech segments [1, 2]. For applications like Automatic Speaker Recognition (ASR), Spoken Document Retrieving (SDR) or speaker emotion recognition, speech segmentation is crucial, because the performance of these heavily relied on it.

Speaker change detection approaches can roughly be divided into three classes: energy –based, metric –based and model –based methods. Metric based methods basically measure the difference between two consecutive frames that are shifted along the audio signal. Several distance measures have been investigated as Kullback-Leibler, Bhattacharyya, and Mahalanobis [3]. Parametric models corrected for finite samples using the Bayesian Information Criterion are widely used [4]. Model-based methods are based on recognizing specific known audio objects e.g. speakers, and classify the audio stream accordingly.

To be feasible for mobile devices Speaker Change Detection should be free from the tuning of data dependent parameters and thresholds forehand. In above mentioned basic speaker change detection approaches there is typically a need to define data dependent thresholds or parameters forehand. To optimize performance BIC penalty term is often defined differently for each data set. Lack of robustness decrease the feasibility of this kind of approaches in mobile device applications.

Used methods in speaker segmentation algorithm should also be chosen in the way that computational and memory costs are not too demanding for mobile device's limited resources.

In this paper we present a novel mobile device targeted speaker segmentation algorithm including an adaptive threshold, corresponding to BIC penalty term, integrated silence - peak detector, which distinguishes effectively real change points from noise peaks, and enhanced robust False Alarm Compensation method using BIC profiles instead of one BIC value.

This article is organized as follows. In Section 2, the proposed speaker segmentation algorithm is explained in details. Simulation results are presented in Section 3 and a short summary is presented in Section 4.

## 2. SPEAKER SEGMENTATION ALGORITHMS

A block diagram of the speaker segmentation system is shown in Fig. 1. Initial speaker change detection consist of an Energy based Voice Activity Detector (VAD), audio feature extraction using Mel Frequency Cepstral Coefficients (MFCC), dissimilarity measurement using BIC calculated using single Gaussian Mixture Models (GMM) and decision logic, where information of detected silences is integrated with peak detection. After the initial speaker

changes are detected they are validated or discarded using BIC based False Alarm Compensation (FAC). Finally metadata is extracted from SCD. The modules are described in more details in the following subsections.

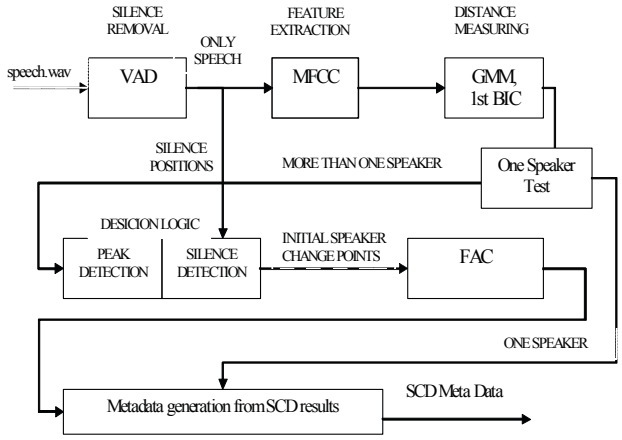


Fig. 1. Block diagram of Speaker Segmentation system.

In the first step Energy based Voice Activity Detector (VAD) classifies frames to speech/silence comparing frame energy against classification threshold which is defined experimentally as 10% of the median of power energy of audio frame. Then 20 Mel Frequency Cepstral Coefficients (MFCC) are computed every 10 ms from a 30 ms analysis window. Two consecutive analysis frames overlap each other over 20 ms [3], [6]. The first MFCC coefficient is discarded. Single Gaussian Mixture Model (GMM) is calculated using the MFCC feature vectors. We use diagonal GMM covariance matrix, which is a good compromise between quality and model size. One Speaker Test (OST) is done to detect if recording contains only one speaker. If all BIC values are above experimentally set threshold, about 50% of maximum BIC value, metadata is generated without the execution of speaker change detection. If OST fails, wrongly detected speaker segments are merged in false alarm compensation.

### 2.1. Bayesian Information Criteria

BIC is one of the most commonly used methods for the purpose of speaker change detection. It was first proposed by Chen & Gopalakrishnan [4]. The BIC is a maximum likelihood criterion penalized by the complexity of model parameters. A one data segment has two hypotheses, it either consist speech of one speaker when there exists a single Gaussian model or it consists speech of two speakers with two multidimensional Gaussian models. The maximum likelihood ratio between the two hypotheses is then formulated as

$$R(i)=\frac{N_x}{2}\log|\Sigma_x|-\frac{N_{x1}}{2}\log|\Sigma_{x1}|-\frac{N_{x2}}{2}\log|\Sigma_{x2}|, \qquad (1)$$

Where $\Sigma$ is the corresponding covariance matrix and $N$ is the number of acoustic vectors in the complete sequence. The variations between one speaker (one Gaussian) and two speakers (two different Gaussians) is given by

$$\Delta BIC(i) = -R(i) + \lambda P, \qquad (2)$$

where $P$ is the penalty term $P = \frac{1}{2}(p + \frac{1}{2}p(p+1))$ x log $N_X$ and $p$ is the dimension of the acoustic space and $\lambda$ is the penalty factor. The negative value of BIC denotes the speaker turn change in the sequence.

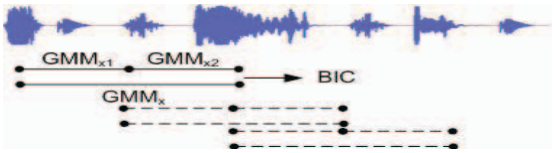


Fig. 2. Example of window sliding for BIC

The BIC is achieved by comparing GMMs calculated for two adjacent windows and to window including the both smaller windows, see Fig. 2.

### 2.2. Peak Detector for BIC using adaptive threshold

Based on the BIC theory the decision threshold is set to zero, and penalty factor $\lambda$ to one [4]. The need for manual tuning for optimal results has been noticed in the literature and the role of $\lambda$ is seen as a definition of a threshold [5]. Also methods for eliminating the effect of predefined penalty factor have been studied. The number of parameters was then forced to be same in the two models used in log likelihood criterion (LLR) in [6].

In our experiments speaker changes are often located in places, where BIC does not give negative values, instead a drop down in BIC values is typically notified, see Fig 4. In our approach we tune automatically threshold level based on the properties of the BIC curve. Threshold is initially set as median value of BIC multiplied by the threshold factor. Threshold factor is then changed step by step in a limited interval and variance of masked peaks is calculated. The threshold factor value corresponding to the maximum value of variance is selected. In Fig. 3 is presented an example of the correlation between the maximum of variance and high $F$-score values. Variance values are normalized to one for the comparison.

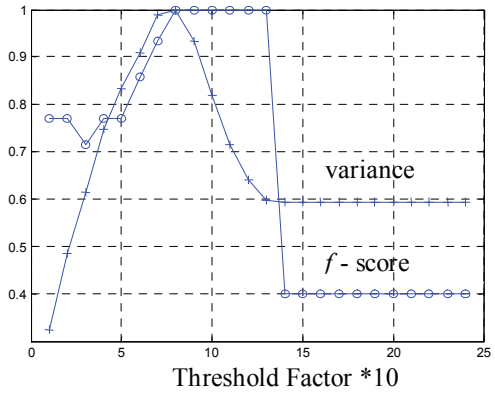


Fig. 3. Variance of masked BIC curve and $F$-score

In peak detector those parts of BIC curve that fall below adapted threshold are masked out to be analyzed further in peak detector. After threshold is adapted adjacent BIC distance values are summed in sliding window and normalized to eliminate too close

local peaks. Visually it was noticed that noise peaks in BIC curve were often sharper than peaks pointing real speaker changes. Peak is scaled by multiplying it by peak breath and then normalized. Peak detecting phases are visualized in Fig. 4.

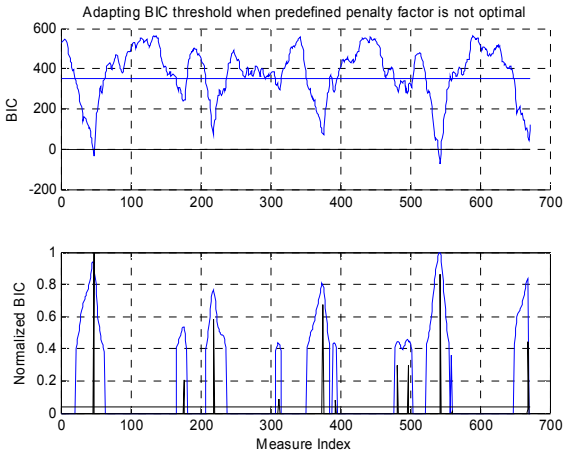


Fig. 4. Example of BIC peak detecting from BIC distance curve

If there are significant side lobes (depth is greater than 10% of main lobe height), those lobes are also kept for decision making. Acceptance threshold is used to do peak detecting decision, see lower picture in Fig 4. Acceptance threshold is for fine tuning, when masking threshold affects more heavily on results.

### 2.3. Integrating Silence Information to peak detection

Silence usage in speaker change detection has been discussed earlier e.g. in [8]. In our approach we use integrated peak - silence detector before FAC, which improve the final performance. Threshold value 0.3 s has been used to classify silences into two categories: silences within speech (< 0.3 s) and potential silence between two speakers (>0.3 s). If there is an inter-speaker silence within acceptance window from initial speaker change candidate, this candidate is accepted as true changing point. Initial candidates are quantized to nearest silent changing point.

There is not always an inter-speaker silence in true speaker changing point. In these cases if BIC distance indicates clear changing point, candidate is accepted as true changing point although there is no inter-speaker silence.

### 2.4. Enhanced False Alarm Compensation

BIC has been used also for audio clustering and false alarm compensation [7]. BIC distance values were then calculated between two adjacent segments, and if the distance is below a threshold, then they belong to the same class and were merged. Drawbacks were then threshold tuning and short segments due to reduced size of data.

Our approach is based on the assumption that, if consecutive sequences belong to the same speaker, their BIC distance relations, which we call BIC profiles, to all other segments are mostly similar. BIC distance matrix is calculated between all detected speech segments. Each row represents a BIC profile, one segment BIC distances against other segments.

$$BIC_{Matrix} = \begin{bmatrix} BIC(S_{1,1}) & BIC(S_{1,2}) & \cdots & BIC(S_{1,i-1}) & BIC(S_{1,i}) \\ BIC(S_{2,1}) & BIC(S_{2,2}) & \cdots & & BIC(S_{2,i}) \\ \vdots & & \ddots & & \vdots \\ BIC(S_{j-1,1}) & & & & BIC(S_{j-1,i}) \\ BIC(S_{j,1}) & BIC(S_{j,2}) & \cdots & BIC(S_{j,i-1}) & BIC(S_{j,i}) \end{bmatrix}$$

In Fig. 5 are visualized normalized BIC profiles calculated from BIC matrix containing 7 speech segments. Visually it can be noted that there are two kinds of profiles (segments of two different speakers).

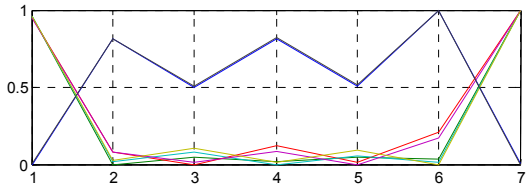

Fig. 5. Example of normalized BIC profiles.

Merging decision is made by comparing the similarity of normalized BIC profiles of two adjacent segments by subtracting them from each other. Rms difference, the variance of remaining difference vector, is calculated. Decision is made based on the rms difference between consecutive segments. Sequence pairs are merged, if rms difference between their BIC profiles goes under merging threshold. Merging threshold gets values (0 - 0.5), where higher value allows les false alarms.

## 3. EXPERIMENTS

### 3.1. Database

For evaluation of the algorithm we used database collected with mobile phone Nokia 6630 in AMR format. Speech sequence durations are presented in Fig. 6. Total number of speaker changes is 321, including 167 short segments (< 5 s). Percentage value of short segments is about 52%, which lay out that natural conversation tend to have short few word sentences.

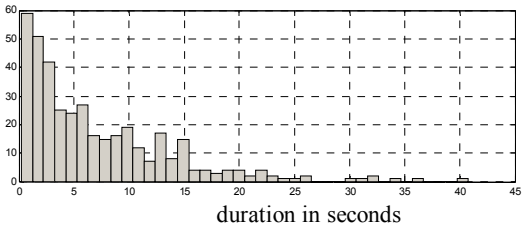


Fig. 6. Speech sequence durations for test database

Some recordings contain also overlapped noise like faucet. Number of recordings is 70 and the duration of test files in total is about 40 minutes.

### 3.2. Evaluation methods

For comparing target and hypothesized changes, we adopted from literature the precision *PRC*, recall *RCL* and *F*-score measures.

$$PRC = \frac{\text{number of correctly found changes}}{\text{total number of changes found}} \quad (3a)$$

$$RCL = \frac{\text{number of correctly found changes}}{\text{total number of correct changes}}. \quad (3b)$$

The evaluation of the segmentation quality is made in terms of *F*-score, a combined measure of *PRC* and *RCL* of change detection. *F*-score is often defined as

$$F - score = \frac{2.0 * \text{PRC} * \text{RCL}}{\text{PRC} + \text{RCL}}. \quad (4)$$

The *F*-score values vary from 0 to 1, with a higher *F*-score indicating better performance.

### 3.3 Simulations

Simulations were done using manually annotated test database presented in section 3.1. In table1 are presented the results using adaptive and manually tuned BIC masking thresholds. Also proposed false alarm compensation method is compared against baseline FAC. The effect of silence information integration to peak detecting is shown in intermediate results Test 2. Summary of executed tests:

*Test 1*: Results are calculated after initial peak detector output, before silence information is used.

*Test 2*: Results are calculated after silence information is integrated to peak detecting decision making.

*Test 3*: Results are calculated after proposed FAC, which use BIC profile instead of one BIC value.

*Test 4*: Results are calculated using only one BIC distance in false alarm compensation. Test 4 results are compared with Test 3 results.

*Test 5*: Results are calculated using manually tuned global BIC threshold (=430). Test 5 results are compared with Test 3 results.

We assume that a speaker change point is true if the bias from the hand-labelled break point is less than 1 s [8]. Window length used in BIC calculations was 1.8 s and window shift 0.1 s.

TABLE 1
SIMULATION RESULTS FOR SPEAKER CHANGE DETECTOR

| Tests | *F*-score | *RCL* | *PRC* |
|---|---|---|---|
| Results of Proposed methods | | | |
| Test 1 | 0.53 | 0.81 | 0.40 |
| Test 2 | 0.644 | 0.804 | 0.537 |
| Test 3 | 0.720 | 0.717 | 0.724 |
| Baseline Results | | | |
| Test 4 | 0.684 | 0.627 | 0.75 |
| Test 5 | 0.714 | 0.655 | 0.784 |

## 4. CONCLUSION

We have implemented speaker segmentation algorithms that can be applied in mobile phone applications. Algorithms were tested with audio data collected with mobile phones. Comparing the *F*-score results between adaptive and manually tuned thresholds, it can be noted that results are near the same level, but adaptive threshold is more robust because no beforehand tuning is needed. Silence information integration to peak detection reduces effectively false alarms, which improves also FAC performance. Developed false alarm compensation using BIC profiles instead of one BIC distance performs better comparing with baseline FAC. Advantages of proposed FAC method are enhanced robustness in the decision making and in the setting of merging threshold.

Despite the challenging mobile device recorded test database presented in (3.1), our speaker change detection products very satisfactory results being comparable with results presented in literature with high quality data [6], [7], [8]. The processing performance in Intel Pentium M processor at 1.6 GHz is about 12 times faster than the real-time performance, which indicates actual real-time processing capabilities for lower capacity mobile terminal targeted processors.

## 5. ACKNOWLEDGEMENTS

This work is partly funded by Nokia Research Center, Finland and partly by National Technology Agency of Finland (TEKES).

## 6. REFERENCES

[1] Mäkelä S.-M., Peltola J., Myllyniemi M. (2006) Mobile Video Capture Targeted Narrowband Audio Content Classification, ICASSP 2006, Toulouse, France.

[2] Iain McCowan, Jitendra Ajmera, Darren Morre, “An Online System for Automatic Annotation of Audio Documents”, IDIAP-RR 03-39, 2003.

[3] O. Pietquin, L. Couvreur, P. Couvreur, "Applied Clustering for Automatic Speaker-Based Segmentation of Audio Materials", *Belgian Journal of Operations Research, Statistics and Computer Science (JORBEL), Special Issue Operations Research and Statistics in the Universities of Mons*, volume 41, n° 1-2, 2001

[4] Scott Shaobing Chen, P.S. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the bayesian information criterion”, 1998 DARPA Broadcast News Transcription & Understanding Workshop.

[5] Perrine Delacourt, Christian J. Wellekens, “DISTBIC: A Speaker-Based Segmentation for Audio Data Indexing”, Speech Communication 32 (2000) 111 -126

[6] Jitendra Ajmera, Iain MCCowan and Hervé Bourlard, “Robust Speaker Change Detection”, IEEE Signal Processing Letters, Vol. 11, No.8, August 2004

[7] R. Huang, J.H.L. Hansen, ”Unsupervised Audio Segmentation and Classification for Robust Spoken Document Retrieval”, IEEE ICASSP-2004, volume 1, pp. 741-744, May 2004.

[8] Bowen Zhou, J.H.L. Hansen, “Efficient Audio Stream Segmentation via the combined T2 Statistic and Bayesian Information Criterion”, IEEE Transactions on Speech and Audio Processing. vol. 13, NO. 4, July 2005