# Modified DISTBIC algorithm for speaker change detection

*Petra Zochová, Vlasta Radová*

Department of Cybernetics
University of West Bohemia, Pilsen, Czech Republic
pzochova@kky.zcu.cz, radova@kky.zcu.cz

## Abstract

The paper deals with the problem of automatic speaker change detection. A metric-based algorithm, called MDISTBIC, which means Modified DISTBIC, is proposed in this paper. The algorithm originates from the DISTBIC algorithm and modifies it in order to reach a higher efficiency. Both the DISTBIC and the MDISTBIC methods are tested in a number of experiments. As the results show, the MDISTBIC algorithm is more efficient than the DISTBIC algorithm in a majority of tests.

## 1. Introduction

The aim of automatic speaker change detection is to extract homogeneous segments containing the longest possible utterances produced by a single speaker. Many efforts have been devoted to this problem in the last years, mainly due to the large number of possible applications, e.g.:

- a reliable speaker change detection algorithm could be very helpful for annotators who manually annotate a large amount of speech data in order to organize an archive of audio documents (e.g. recordings of various sessions or meetings, broadcast news, etc.);
- performance of a speech recognition system that recognizes conversations or news broadcasts could be improved, when a speaker change detection system would detect the change of the speaker and parameters of the speech recognition system would be adapted accordingly;
- in speaker recognition systems it is supposed that the input speech belongs only to one speaker. It may cause problems in many real situations (e.g. in conversations or broadcast news), where the speech stream is continuous and there is no information about the beginning and ending of the speech segment of one speaker. A speaker change detection system could help to eliminate such problems.

There are three main speaker change detection approaches [1], [2]. In the *metric-based approach* the speaker changes are determined as the moments in which a distance measure computed between two adjacent windows shifted along the speech signal reaches a local maximum. In the *model-based approach* it is assumed that a model of each speaker the voice of which is contained in an utterance has been trained before the speaker change detection algorithm starts. The speaker changes are then detected as the instants when it is necessary to change the speaker model in order it match the speech signal. The last approach is the *decoder-guided* approach. Here, the speaker changes are determined according to information provided by a speech recognition system which decodes the spoken audio stream at first (e.g. possible speaker changes are at every silence location).

In this paper, we are interested in the metric-based approach, because it does not require any other information or things except the speech signal itself (i.e. neither speaker model, nor speech recognizer). We focus on the DISTBIC algorithm introduced in [3]. This algorithm is efficient in detecting speaker changes that are relatively close one another, however at the price that a lot of false speaker changes is detected. We will try to modify the algorithm in this paper and thereby to improve its efficiency.

The paper is organized as follows: In Section 2 the DISTBIC algorithm is briefly described. Next, in Section 3 the modifications of the DISTBIC algorithm are introduced. Experiments are described and their results are presented in Section 4. Finally, a conclusion is given in Section 5.

## 2. DISTBIC algorithm

The DISTBIC algorithm is based on a two-step analysis [3]: the first pass uses a distance computation to determine the speaker changes candidates and the second pass uses Bayesian Information Criterion (BIC) to validate or discard these candidates.

### 2.1. First step: detection of speaker change candidate points

The first step relies on a distance-based segmentation defined from the likelihoods of adjacent windows. In each window, the data are assumed to result from a single multi-dimensional Gaussian process. The question is, whether the data from the two adjacent windows together fit better with a single multi-dimensional Gaussian or whether a two-window representation justifies the data better. In order to answer this question, the Kullback-Leibler distance can be used for example.

A symmetric Kullback-Leibler distance KL2 between a vector $X$ coming from the multi-dimensional Gaussian process $N(\mu_X, \Sigma_X)$ and a vector $Y$ resulting from the multi-dimensional Gaussian process $N(\mu_Y, \Sigma_Y)$ can be computed as

$$\begin{aligned}\mathrm{KL2}(X,Y) =& \frac{1}{2}(\mu_Y - \mu_X)^{\mathrm{T}}(\Sigma_X^{-1} + \Sigma_Y^{-1})(\mu_Y - \mu_X) + \\ &+ \frac{1}{2}\mathrm{tr}\left((\Sigma_X^{\frac{1}{2}}\Sigma_Y^{-\frac{1}{2}})(\Sigma_X^{\frac{1}{2}}\Sigma_Y^{-\frac{1}{2}})^{\mathrm{T}}\right) + \\ &+ \frac{1}{2}\mathrm{tr}\left((\Sigma_X^{-\frac{1}{2}}\Sigma_Y^{\frac{1}{2}})(\Sigma_X^{-\frac{1}{2}}\Sigma_Y^{\frac{1}{2}})^{\mathrm{T}}\right) - d,\end{aligned} \tag{1}$$

where tr denotes the trace of a matrix, and $d$ is the dimension of the vectors $X$ and $Y$.

The KL2 distance is computed for two adjacent windows $W_1$ and $W_2$ of the same size (2 s) shifted by a fixed step

(100 ms) along the whole speech signal. This process results in a graph of distances with respect to time. The graph is smoothed by a low-pass filtering operation, and then all the significant local maxima are searched because they represent potential speaker change points. A local maximum is regarded as significant when the differences between its value and those of the minima surrounding it are above a certain threshold, and when there is no higher local maximum in its vicinity. Thus, the local maximum has to fulfill the following condition to be significant:

$$\begin{array}{l} |\max - \min_{\mathrm{l}}| > \alpha\sigma \\ \text{and} \\ |\max - \min_{\mathrm{r}}| > \alpha\sigma, \end{array} \tag{2}$$

where $\alpha$ is a real number, $\sigma$ is the standard deviation of the distances along the plot, and $\min_{\mathrm{l}}$ and $\min_{\mathrm{r}}$ are the left and the right minima, respectively, around the peak $\max$.

### 2.2. Second step: BIC refinement

A $\Delta$BIC value is computed for each potential speaker change point detected in the first step to validate or discard this point. The $\Delta$BIC value is given by [3]

$$\Delta\mathrm{BIC} = -R + \lambda P, \tag{3}$$

where

$$R = \frac{N}{2}\log|\Sigma| - \frac{N_1}{2}\log|\Sigma_1| - \frac{N_2}{2}\log|\Sigma_2|, \tag{4}$$

$\lambda$ is a penalty factor which has to be experimentally tuned in order to reduce the number of false alarms without increasing the number of missed detections,

$$P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\log N, \tag{5}$$

$N_1$ and $\Sigma_1$ are the number and the covariance matrix of the feature vectors in the window $W_1$, respectively, $N_2$ and $\Sigma_2$ are the number and the covariance matrix of the feature vectors in the window $W_2$, respectively, $N = N_1 + N_2$, $\Sigma$ is the covariance matrix of the feature vectors of both windows together, and $d$ is the dimension of the feature vectors.

A potential speaker change point is regarded as a true speaker change if the $\Delta$BIC value for this point is negative.

## 3. Modified DISTBIC algorithm

The DISTBIC algorithm allows to obtain good speaker change detection results, nonetheless it has some weak points. We have focused on these points and suggest some improvements of the algorithm in order to obtain even better results. The improvements are specified in next subsections.

### 3.1. Silence and breathing elimination

Silence and breathing may cause a lot of false alarms in speaker change detection tasks. Therefore we used a simple but efficient silence detector before the speaker change detection process. The speech signal was divided into segments the length of which was 10 ms. Short-time energy and the number of zero crossings [4] were computed for each segment. If both the short-time energy and the number of zero crossings were lower than experimentally derived thresholds, the segment was regarded as containing silence and was temporarily eliminated from the utterance.

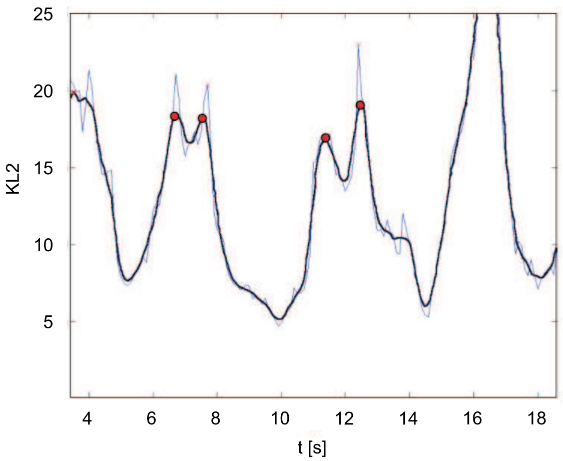

Figure 1: *True speaker change points causing troubles (marked with a circle) in the condition (2).*

Sometimes short sections with a high energy can occur in silent parts of the utterance. Such sections can be caused for example by the speaker breathing. The high energy impels the silence detector to regard these sections as speech. In order to overcome this problem, we implemented a clustering algorithm: if a part of a speech signal shorter than 645 ms was surrounded by silent segments, this part was also regarded as silence. On the contrary, if there was a silent segment shorter than 250 ms between two segments containing speech, this segment was regarded as containing speech.

### 3.2. Speaker change candidate detection

Equally as in Section 2.1, the potential speaker change points were detected on the smoothed graph of the symmetric Kullback-Leibler distance. However, the condition (2) necessary to detect the potential speaker change points was changed. The reason for the change was the fact, that the condition (2) did not allow to detect some local maxima of the graph as the potential speaker change points. The problems were caused mainly by the maxima that were rather near each other, so that the minimum between them was too high to satisfy the condition (2). An example of such a kind of peaks is shown in Figure 1. For that reason the conjunction in (2) was substituted with the disjunction, i.e. a local maximum was regarded as a potential speaker change point if it satisfied the condition

$$\begin{array}{l} |\max - \min_{\mathrm{l}}| > \alpha\sigma \\ \text{or} \\ |\max - \min_{\mathrm{r}}| > \alpha\sigma, \end{array} \tag{6}$$

where $\alpha$, $\sigma$, $\min_{\mathrm{l}}$, $\min_{\mathrm{r}}$, and $\max$ have the same meaning as before.

In order to avoid the situation that two different maxima belonging in fact to one true speaker change would be detected as two potential speaker change points, we required a minimal distance between two maxima: if two maxima were closer than 0.5 s, the lowest one was discarded. This condition protects the algorithm against false alarms.

### 3.3. Speaker change position location

Having detected the potential speaker change points, we used the $\Delta$BIC value (3) to discard or validate the points similarly

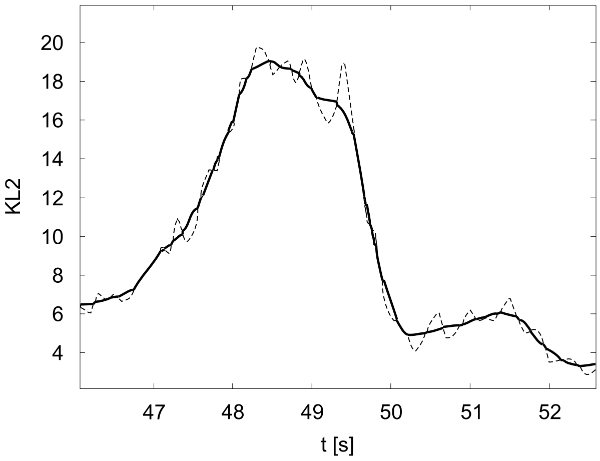

Figure 2: *Illustration of the correspondence between the peak on the smoothed distance graph (solid line) and the peaks on the unsmoothed distance graph (dashed line).*

like in Section 2.2. However, more than one local peak on the unsmoothed distance graph may correspond to the speaker change point detected on the smoothed graph (see Figure 2 for illustration). For that reason, we computed the $\Delta$BIC value for all local peaks on the unsmoothed graph that correspond to the peak validated as the speaker change on the smoothed graph, and the local peak with the lowest $\Delta$BIC value was chosen as the true speaker change.

### 3.4. Time alignment

Silent parts temporarily eliminated from the speech signal in Section 3.1 have to be inserted back into the utterance now, and the points of the speaker changes have to be aligned accordingly. In addition, if a detected speaker change was close (less than 0.2 s) to a silent part, it was moved into the centre of the silent part.

## 4. Experiments and results

The purpose of the experiments was to compare the performance of the DISTBIC algorithm and the Modified DISTBIC (MDISTBIC) algorithm. Both of these algorithms were tested in several experiments, where audio records of TV news, radio news and radio discussions were automatically segmented with respect to the speaker changes.

- The radio news test set consisted of 8 records containing news broadcasted by the Czech radio station Český rozhlas 2 – Praha. The length of each record was about 10 minutes, each record contained about 23 speaker changes on average. Speakers in the records did not speak simultaneously and the interval between two consecutive speaker changes was quite long.
- The TV news test set contained 7 records of newscasts of different Czech TV channels. The length of the records ranged from 11 to 20 minutes, each record contained about 94 speaker changes on average. Similarly as in the radio news, the speakers did not speak simultaneously. However, unlike the radio news, about 9% of speaker changes were quite close (less than 2 s).
- The radio discussions test set contained 9 records of the programme Radiofórum broadcasted by the Czech radio station Český rozhlas 1 – Radiožurnál. The length of each record was approximately 30 minutes, each record contained about 105 speaker changes on average. Approximately one third of the changes occurred very soon after the previous change (i.e. the changes were closer than 2 seconds). In addition, the speakers spoke often simultaneously.

Two types of errors could happen during the segmentation. A false alarm (FA) occurred when a speaker change was detected, although it did not exist. On the contrary, a missed detection (MD) occurred when the algorithm did not detect an existing speaker change. If we know the number of FA and MD for a record, we can determine the accuracy and the false alarm rate (FAR) that ware achieved for the record using a segmentation algorithm. The accuracy is defined as

$$\begin{aligned}\text{Accuracy} &= 100\times \\ &\times \frac{\text{number of true speaker changes} - \text{number of MD}}{\text{number of true speaker changes}}\ [\%],\end{aligned} \tag{7}$$

and the FAR is determined according to the formula

$$\begin{aligned}\text{FAR} &= 100\times \\ &\times \frac{\text{number of FA}}{\text{number of true speaker changes} + \text{number of FA}}\ [\%].\end{aligned} \tag{8}$$

The accuracy and the false alarm rates achieved for records from the above mentioned test sets using both the DISTBIC and the MDISTBIC algorithms are given in Tables 1, 2, and 3 in detail. It can be found out after an inspection of the results, that the MDISTBIC algorithm gives better results (i.e. a higher accuracy and a lower FAR) in a majority of the tests. It outperforms the DISTBIC algorithm both for the radio news where there were long intervals between the speaker changes and for the radio discussions where the speaker changes were relatively very close one another and the speakers often spoke simultaneously. This can be seen also from the Figures 3, 4, and 5, where the average values of the accuracy and false alarm rates for each of the 3 test sets are lucidly presented.

In order to provide all information about the experiments we should also say that

- the sample frequency was 8 kHz and 12 mel-frequency cepstral coefficients were used as feature vectors for the representation of the speech signal,
- a Gaussian function

  $$h(t) = \exp(-\frac{t^2}{2\tau^2}), \tag{9}$$

  where $\tau$ was set to 5, was used for the smoothing of the distance graph, and
- $\lambda$ and $\alpha$ in (3) and (6), respectively, were tuned separately for each algorithm and test set so that none of the methods was privileged to the other.

## 5. Conclusion

The Modified DISTBIC (MDISTBIC) algorithm for automatic speaker change detection in audio records has been introduced in this paper. The algorithm has been tested in a number of tests, and the results have been compared with the results achieved using the original DISTBIC algorithm. It follows from the results

Table 1: *Speaker change detection results achieved for records of radio news.*

| | DISTBIC algorithm | | MDISTBIC algorithm | |
|---|---|---|---|---|
| record | accuracy | FAR | accuracy | FAR |
| 1 | 100.00% | 50.00% | 100.00% | 50.00% |
| 2 | 73.91% | 46.51% | 86.96% | 30.30% |
| 3 | 85.19% | 49.06% | 100.00% | 46.00% |
| 4 | 96.55% | 45.28% | 100.00% | 40.82% |
| 5 | 84.21% | 45.71% | 100.00% | 40.63% |
| 6 | 89.66% | 47.27% | 100.00% | 42.00% |
| 7 | 83.33% | 66.04% | 100.00% | 53.85% |
| 8 | 96.67% | 45.46% | 93.33% | 28.57% |

Table 2: *Speaker change detection results achieved for records of TV news.*

| | DISTBIC algorithm | | MDISTBIC algorithm | |
|---|---|---|---|---|
| record | accuracy | FAR | accuracy | FAR |
| 1 | 78.18% | 40.22% | 85.46% | 36.05% |
| 2 | 75.97% | 30.00% | 75.16% | 18.18% |
| 3 | 88.89% | 33.68% | 84.13% | 19.23% |
| 4 | 84.48% | 40.82% | 87.93% | 32.56% |
| 5 | 87.23% | 42.68% | 91.49% | 28.24% |
| 6 | 73.98% | 30.90% | 86.18% | 25.46% |
| 7 | 83.02% | 28.05% | 83.90% | 21.85% |

Table 3: *Speaker change detection results achieved for records of radio discussions.*

| | DISTBIC algorithm | | MDISTBIC algorithm | |
|---|---|---|---|---|
| record | accuracy | FAR | accuracy | FAR |
| 1 | 52.42% | 51.56% | 52.85% | 45.33% |
| 2 | 52.34% | 61.51% | 56.60% | 52.89% |
| 3 | 48.65% | 54.88% | 53.64% | 45.55% |
| 4 | 60.00% | 61.54% | 52.24% | 52.24% |
| 5 | 67.24% | 74.45% | 71.93% | 70.31% |
| 6 | 53.13% | 57.19% | 52.76% | 53.65% |
| 7 | 67.74% | 52.79% | 70.65% | 49.73% |
| 8 | 69.61% | 53.64% | 68.32% | 50.25% |
| 9 | 58.42% | 59.44% | 62.25% | 55.86% |

that the MDISTBIC algorithm can reach a higher accuracy and a lower number of false alarms in the speaker change detection task than the DISTBIC algorithm.

## 6. Acknowledgements

The work was supported by the Ministry of Education of the Czech Republic, project no. MSM 235200004, and by the Grant Agency of the Czech Republic, project no. 102/05/0278.

## 7. References

[1] Couvreur L., Boite J.M.: ”Speaker tracking in Broadcast audio material in the framework of the THISL Project”, Proc. of ESCA ETRW Workshop on Accessing Information in Spoken Audio, Cambridge (UK), 1999.

[2] Kwon S., Narayanan S.: ”Speaker change detection using a new weighted distance”, Proc. of ICSLP 2002, pp. 2537–2540, Denver, Colorado, USA, 2002.

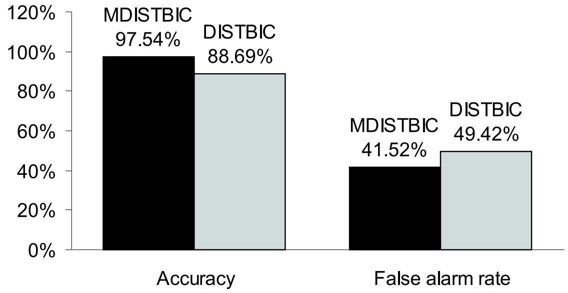

Figure 3: *Average speaker change detection results achieved for radio news.*

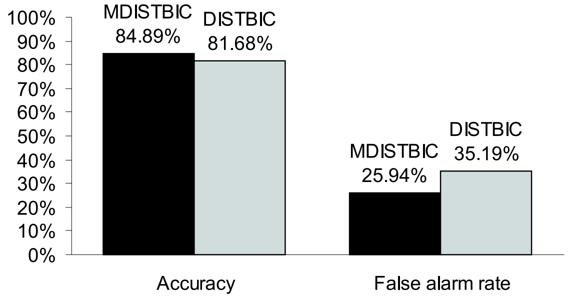

Figure 4: *Average speaker change detection results achieved for TV news.*

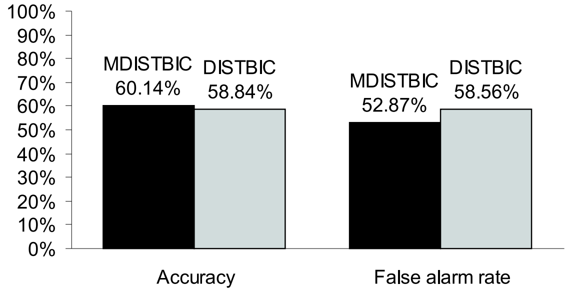

Figure 5: *Average speaker change detection results achieved for radio discussions.*

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

[4] Lu L., Jiang H. and Zhang H. J.: ”A Robust Audio Classification and Segmentation Method”, Proc. of 9th ACM Multimedia, pp. 203–211, Ottawa, Canada, 2001.