AlgBICMap–Voiced: An 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 the AlgBICMap algorithm, was proposed in [1]. The AlgBICMap–Voiced is a modification of that algorithm and enables us to decrease the number of false alarms. The algorithm allows to create a map of BIC (Bayesian Information Criterion). The map enables us to detect efficiently fields of speech of individual speakers. In comparison with a typical metric-based approach, the advantage of the proposed algorithm is its robustness because it uses more information, not only information provided by adjacent windows.

1. Introduction

In this paper, we have improved the present metric-based methods in order to get more information about the changes. When a distance is computed only between two adjacent windows, a lot of pieces of information will be lost. It is as if we wear blinkers and our eyes can look straight ahead only; then we have no information neither from the left side nor the right side. For that reason we have decided to compute not only a graph of distances (which is created by the distances computed between two adjacent windows shifted along the whole speech signal), but the whole map of distances. In such a case, the distance is computed between any two windows in the speech signal and the distances are depicted. We are interested in “reading from the map” in order to obtain as much information as possible.

The AlgBICMap algorithm [1] shows good results of the speaker change detection of TV and radio news. Our aim in this paper is a reduction of false alarms. For this purpose, the modification of the AlgBICMap, called the AlgBICMap–Voiced algorithm, uses mainly voiced sounds.

The organization of the paper is as follows: First, in Section 2, the AlgBICMap–Voiced algorithm is explained. In Section 3 experiments are described and achieved results are presented. Finally, in Section 4, some conclusions are given.

2. AlgBICMap–Voiced algorithm

The AlgBICMap–Voiced algorithm is based on the AlgBICMap method [1]. The algorithm detects speaker changes in a speech record. It uses the Bayesian Information Criterion (BIC) [3] in order to form a BIC-map.

2.1. Preprocessing

The first phase of our algorithm is preprocessing. The Schmidt trigger algorithm [2] was used for the detection of voiced sounds. The algorithm was optimized in such a way that it works properly also for records of several sentences. Only one threshold was used for thresholding. The threshold was computed by a median of the Schmidt function.

The example of the Schmidt function is showed in Fig. 1. The Schmidt function in the bottom view has higher values in the voiced segments. The voiced segments are chosen where the Schmidt function is higher than the threshold.

![Schmidt trigger algorithm](image1)

Figure 1: An Example of Schmidt trigger algorithm.

The vectors of 12 mel-frequency cepstral coefficients (MFCCs) are computed from the whole speech record. Then vectors of MFCCs, that correspond only to the voiced segments, are extracted. We call them as Voiced MFCCs. Voiced MFCCs will be used in the step called BIC-map computation.

2.2. Bayesian Information Criterion (BIC)

In speaker change detection algorithms, the $\Delta$BIC value (1) is commonly computed for two adjacent windows which are shifted along the whole speech signal. A speaker change is detected if the $\Delta$BIC value for the actual adjacent windows is positive [3]. The $\Delta$BIC value is given by

$$\Delta\text{BIC} = R - \lambda P,$$

where

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

and

$$P = \frac{1}{2} \left( d + \frac{1}{2} d (d + 1) \right) \log N.$$
and \( \Sigma_1 \) are respectively the number and the covariance matrix of the feature vectors in the window \( W_1 \). Similarly, \( N_2 \) and \( \Sigma_2 \) are the number and the covariance matrix of the feature vectors, respectively, in the window \( W_2 \). Further, \( N = N_1 + N_2 \), \( \Sigma \) is the covariance matrix of the feature vectors of both windows, and \( d \) is the dimension of the feature vectors. The value of the empirical factor \( \lambda \) has to be tuned in order to reduce the number of false alarms without increasing the number of missed detections.

### 2.3. BIC-map computation

The \( \Delta \text{BIC} \) value (1) is computed between the actual window (starting from the beginning of the speech stream) and all the following windows (the map is symmetric, that is why we do not need to compute the distance between the actual window and the previous windows). Then we move on to the next window and the same process is repeated until the end of the speech stream is reached. The BIC distances are stored in a matrix called the BIC-map (Fig. 2).

![Figure 2: An example of the BIC-map.](image)

As we can see from Fig. 3 and 4, the binary BIC-map computed only from the voiced segments (Fig. 3) is more compact than the binary BIC-map computed from the segments that contain both voiced and unvoiced parts (Fig. 4). Thanks to this fact, it is possible to reduce the number of false alarms.

![Figure 3: AlgBICMap–Voiced: An example of the binary BIC-map.](image)

![Figure 4: AlgBICMap–Voiced: An example of the binary BIC-map.](image)

### 2.4. Binary BIC-map

When we use a zero threshold, we can transform the BIC-map naturally into a binary BIC-map with the values 0 and 1. If a value of the distance in the BIC-map is positive, the corresponding value in the binary BIC-map is set to 1; otherwise, it is set to 0. An example of the binary BIC-map is presented in Fig. 3. The black color represents the value 1 and the white color represents the value 0.

A lot of interesting things can be seen from Fig. 3. We can see for example, that the main diagonal of the binary BIC-map is scattered with square-shape elements. We call them diagonal squares. One diagonal square represents speech of one speaker. The place, where a diagonal square ends and another one starts, represents a speaker change.

From the BIC-map we can also trace all the segments of the speech stream where a particular speaker speaks. When we chose a diagonal square in the binary BIC-map and find all black rectangle-shape elements placed above and below that square in the vertical direction, the segments on the vertical axis corresponding to the rectangle-shape elements represent the segments where the same speaker speaks.

![Figure 3: AlgBICMap–Voiced: An example of the binary BIC-map.](image)

![Figure 4: AlgBICMap: An example of the binary BIC-map.](image)

### 2.5. Detection of the diagonal squares

The algorithm of speaker change detection consists of two phases, that are performed only on voiced parts of the utterance. The aim of these phases is to detect the bottom-right and upper-left corners of the diagonal squares, because the corners represent potential speaker changes.

From now on it will be supposed that the BIC-map can be expressed as a matrix

\[
M(i, j), \ i = 1 \ldots m, \ j = 1 \ldots m, \quad (4)
\]

where \( m \) represents the length of the speech stream, \( i \) denotes a row of the BIC-map, \( j \) is a column of the BIC-map, \( M(i, j) = 1 \) for the black points of the BIC-map, and \( M(i, j) = 0 \) for the white points.

**First phase: Vertical examination**

The aim of the first phase is to find the bottom-right corners of the diagonal squares. We have used the following steps for this purpose:
Step 1: Initialization – start at the upper-left corner of the BIC-map (i.e. set $i = 1$ and $j = 1$) and set the counter of the potential speaker changes to 1 (i.e. $c = 1$).

Step 2: Compute the ratio
\[
\text{ratio}_v = \frac{\sum_{k=i}^j M(k,j)}{j-i+1} \times 100\%, \quad (5)
\]
which is the percentage of the black points (i.e. the points with the value 1) in the column $j$ between the rows $i$ and $j$.

Step 3: If $\text{ratio}_v$ is lower than a percentage threshold $t_{\text{vertical}}$, there is a potential speaker change in time $j - 1$ (in samples). Then set $T_v(c) = j - 1$, $i = j$, and $c = c + 1$.

Step 4: Set $j = j + 1$.

Step 5: Repeat Steps 2, 3, and 4 until the end of the BIC-map is reached (i.e. until $i = m$ and $j = m$).

### Second phase: Horizontal Validation

The aim of the second phase is to find the upper-left corners of the diagonal squares. In the ideal case the upper-left corner of a diagonal square is equal to the bottom-right corner of the previous diagonal square. The first upper-left corner is supposed to be at the beginning (upper-left corner) of the BIC-map. In this phase only several rows around the potential speaker changes found in the first phase are examined.

Step 1: Initialization – start with the first potential speaker change found in the first phase (i.e. set $c = 1$) and set $i = T_v(c) - n$, where $n$ is an experimentally determined constant.

Step 2: Compute the ratio
\[
\text{ratio}_h = \frac{\sum_{k=i} M(i,k)}{T_v(c) - i + 1} \quad (6)
\]
which is the percentage of the black points (i.e. the points with the value 1) in the row $i$ between the columns $i$ and $T_v(c)$.

Step 3: If $\text{ratio}_h$ is higher than a percentage threshold $t_{\text{horizontal}}$, there is a potential speaker change in time $i$ (in samples). Then set $T_h(c) = i$, $c = c + 1$, and $i = T_v(c) - n$. Go to Step 5.

Step 4: Set $i = i + 1$.

Step 5: Repeat Steps 2, 3, and 4 until the end of the BIC-map is reached (i.e. until $i = m$ and $j = m$).

### 2.6. Accurate position of the speaker change

Using only the voiced sounds, the found speaker changes need not agree with the true speaker change.

The found change occurs always in some voiced sounds but the true speaker change can happen in the adjacent unvoiced parts. For that reason, we need an interval for each found change where we will search a corresponding true change. The interval will be determined by the help of horizontal and vertical changes.

We have 2 values for each potential speaker change: one from the first phase, and the other from the second phase (see Table 1 for illustration). The aim of this step is to find the accurate positions of the speaker changes.

Suppose that the accurate position of the $c$-th speaker change is investigated and denote the position $T(c)$.

#### Step 1:
Form two adjacent windows: The first window starts at the time $T(c-1)$,
\[
T(c-1) = \max\{T_v(c-1), T_h(c-1)\}, \quad (7)
\]
and ends at a time $T$; the second window starts at the time $T$ and ends at the time $T(c+1)$.
\[
T(c+1) = \min\{T_v(c+1), T_h(c+1)\}. \quad (8)
\]

#### Step 2:
Let $T_{\min}(c) = \min\{T_v(c), T_h(c)\}$ and $T_{\max}(c) = \max\{T_v(c), T_h(c)\}$. For $T$ from $T_{\min}(c)$ to $T_{\max}(c)$ compute the $\Delta$BIC value (1) between the windows formed in Step 1. The $T$ for which the $\Delta$BIC value is the lowest is the wanted position $T(c)$.

In the AlgBICMap–Voiced algorithm we have to think of big enough hysteresis around limit values $T(c-1)$ and $T(c+1)$ because the true change may be in big time interval around the found change. Experiments showed that the suitable interval is $\pm 1.5$ s. Localization of accurate speaker change positions is performed by the BIC method. A maximum of the BIC function (1) characterizes the true speaker change.

We obtain all speaker changes after passing the whole record. We have to emphasize that localization of accurate positions is done in the original MFCCs (MFCCs of the whole record).

#### 2.7. Diagonal belt

Computation of the whole BIC-map is very time demanding. For a speaker change detection task, we do not need to compute the whole map. It is sufficient to compute only a belt around the diagonal of the map (see Fig. 5). However, the algorithm of the diagonal squares detection has to be adapted accordingly.

---

**Table 1: Table of potential speaker changes (PSC).**

<table>
<thead>
<tr>
<th>PSC order</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSC (1st phase)</td>
<td>$T_v(1)$</td>
<td>$T_v(2)$</td>
<td>$T_v(3)$</td>
<td>...</td>
</tr>
<tr>
<td>PSC (2nd phase)</td>
<td>$T_h(1)$</td>
<td>$T_h(2)$</td>
<td>$T_h(3)$</td>
<td>...</td>
</tr>
</tbody>
</table>

---

**Figure 5:** Diagonal belt for the BIC-map in Fig. 3.
3. Experiments and results

In our experiments, the length of the windows, for which the $\Delta$BIC value (1) is computed, was 2 s, and the window shift was 0.25 s. The penalty factor $\lambda$ was set to a value from 1.5 to 1.8. The threshold $t_{vertical}$ was set to 80 %, and the threshold $t_{horizontal}$ was set to 70 %. The constant $n$ in the second phase of the algorithm for the detection of the diagonal squares was equal to 7. Width of the diagonal belt was 10 seconds.

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 were achieved for the record using a segmentation algorithm. The accuracy is defined as

\[
\text{Accuracy} = 100 \times \frac{\text{number of true speaker changes} - \text{number of MD}}{\text{number of true speaker changes}} \%.
\]

and the FAR is determined according to the formula

\[
\text{FAR} = 100 \times \frac{\text{number of FA}}{\text{number of true speaker changes} + \text{number of FA}} \%.
\]

The purpose of the experiments was to compare the performance of the AlgBICMap algorithm and the AlgBICMap–Voiced algorithm. Both of these algorithms were tested in several experiments, where audio records of TV news and radio news 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 3 records of newscasts of different Czech TV channels. The length of the records ranged from 15 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 accuracy and the false alarm rates achieved for records (TV news and Prague radio news) from the above mentioned test sets using both the AlgBICMap and the AlgBICMap–Voiced algorithms are given in Fig. 6–9 in detail. The average values of the accuracy and false alarm rates for each of the test

\[
\text{TV News accuracy}
\]

\[
\text{Prague Radio News accuracy}
\]

\[
\text{TV News - False Alarm Ratio (FAR)}
\]

\[
\text{Prague Radio News - FAR}
\]
sets are also presented.

It can be found out after an inspection of the results, that the AlgBICMap–Voiced algorithm gives a little bit lower accuracy, but much less false alarms than the AlgBICMap algorithm. A summary of the results are presented in table 2.

4. Conclusion

Conclusion is not easy. Results depend on the type of record and on the requirements that the user has. We can compare both algorithms (the AlgBICMap and the AlgBICMap–Voiced) by its positive and negative aspects.

The BIC-map can be used for speaker tracking tasks and therefore it enables us to detect the segments where the same speaker speaks in the record. Speaker tracking can also be useful for false alarm elimination.

<table>
<thead>
<tr>
<th>AlgBICMap</th>
<th>AlgBICMap–Voiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>fast computation</td>
<td>time-consuming computation</td>
</tr>
<tr>
<td>higher accuracy (about 5%)</td>
<td>lower accuracy</td>
</tr>
<tr>
<td>more false alarms</td>
<td>FA reduction (70% decrease)</td>
</tr>
</tbody>
</table>

Table 2: Table of the comparison of algorithms.

5. Acknowledgements

The work was supported by the Grant Agency of the Czech Republic, project no. 102/05/0278.

6. References

