

(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:

$$\begin{aligned} |\max - \min_l| &> \alpha\sigma \\ \text{and} \\ |\max - \min_r| &> \alpha\sigma, \end{aligned} \quad (2)$$

where α is a real number, σ is the standard deviation of the distances along the plot, and \min_l and \min_r are the left and the right minima, respectively, around the peak \max .

2.2. Second step: BIC refinement

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

$$\Delta\text{BIC} = -R + \lambda P, \quad (3)$$

where

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

λ 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,

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

N_1 and Σ_1 are the number and the covariance matrix of the feature vectors in the window W_1 , respectively, N_2 and Σ_2 are the number and the covariance matrix of the feature vectors in the window W_2 , respectively, $N = N_1 + N_2$, Σ 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 ΔBIC 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.

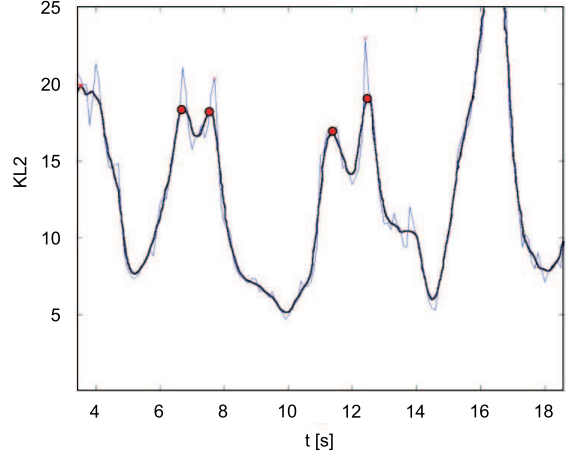


Figure 1: True speaker change points causing troubles (marked with a circle) in the condition (2).

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

$$\begin{aligned} |\max - \min_l| &> \alpha\sigma \\ \text{or} \\ |\max - \min_r| &> \alpha\sigma, \end{aligned} \quad (6)$$

where α , σ , \min_l , \min_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 ΔBIC value (3) to discard or validate the points similarly