



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

$$\text{Accuracy} = 100 \times \frac{\text{number of true speaker changes} - \text{number of MD}}{\text{number of true speaker changes}} [\%], \quad (7)$$

and the FAR is determined according to the formula

$$\text{FAR} = 100 \times \frac{\text{number of FA}}{\text{number of true speaker changes} + \text{number of FA}} [\%]. \quad (8)$$

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

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

where τ was set to 5, was used for the smoothing of the distance graph, and

- λ and α 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