UNIVERSITY OF WEST BOHEMIA IN PILSEN,
DEPARTMENT OF CYBERNETICS

# A Method for Speaker-Based Segmentation of Audio Signals

**Petra Zochová, Vlasta Radová**

E-mail: zpetruska@seznam.cz, radova@kky.zcu.cz

## 1. INTRODUCTION

The paper deals with the problem of speaker-based segmentation. The goal of this task is to extract homogeneous segments containing the longest possible utterances produced by a single speaker. In the method presented here, no assumption is made about prior knowledge of the speaker or speech signal characteristics (there is no speaker model, no speech model, even the number of speakers in the recording is not known).

## 2. PRINCIPLE OF THE METHOD

The main idea of the method is to use a dissimilarity measure between two consecutive parts of the parameterised speech signal for the detection of the most likely speaker turn points. Next, the Bayesian Information Criterion (BIC) [1] is employed to validate or discard the detected speaker turn candidates. In fact, the method consists of 6 steps.

**Step 1: Silence elimination.** The aim of this step is the elimination of silent segments. We used the approach described in [2]. The speech signal is divided into segments the length of which is 10 ms. The short-time energy and the number of zero crossings are computed for each segment. If both the short-time energy and the number of zero crossings are lower than experimentally derived thresholds, the segment is regarded as containing silence. If silent segments surround a part of the speech signal shorter than 645 ms this part is also regarded as silence. If there is a silent segment shorter than 250 ms between two parts containing speech this segment is regarded as containing speech.

**Step 2: Distance computation.** A distance measure is computed for two adjacent windows $W_1$ and $W_2$ of the same size (2 s), the windows are shifted by a fixed step (100 ms) along the whole speech signal. In our method, several distances were tested. The best results were reached by the use of the symmetrical Kullback-Leibler distance

$$\begin{aligned} KL2(W_1,W_2) = {} & \tfrac{1}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}}(\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \\ & + \tfrac{1}{2}\,\mathrm{tr}\left((\boldsymbol{\Sigma}_1^{\frac{1}{2}}\boldsymbol{\Sigma}_2^{-\frac{1}{2}})(\boldsymbol{\Sigma}_1^{-\frac{1}{2}}\boldsymbol{\Sigma}_2^{-\frac{1}{2}})^{\mathrm{T}}\right) + \\ & + \tfrac{1}{2}\,\mathrm{tr}\left((\boldsymbol{\Sigma}_1^{-\frac{1}{2}}\boldsymbol{\Sigma}_2^{\frac{1}{2}})(\boldsymbol{\Sigma}_1^{-\frac{1}{2}}\boldsymbol{\Sigma}_2^{\frac{1}{2}})^{\mathrm{T}}\right) - d, \end{aligned} \tag{1}$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$ are the mean and the covariance matrix, respectively, of the feature vectors of the window $W_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$ are the mean and the covariance matrix, respectively, of the feature vectors of the window $W_2$, tr denotes the trace of a matrix, and $d$ is the dimension of the feature vectors (we used feature vectors of 12 mel-cepstral coefficients).

**Step 3: Detection of speaker turn candidate points.** The distance is depicted in a graph with respect to time and smoothed by a low-pass filtering operation. A graph of the distance (1) for a part of an utterance is presented in Figure 1. The solid line represents the distance *KL*2 given by (1), the dashed line represents the distance after the filtering operation. We have used a Gaussian for filtering in our experiments.

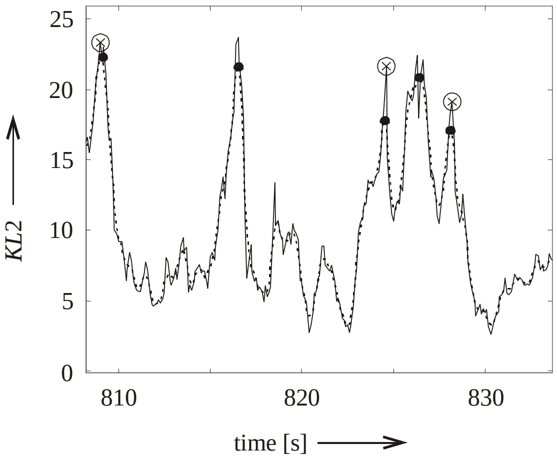

**Fig. 1: Graph of the distance**

All the local maxima of the smoothed graph are searched in order to detect the speaker turn candidate points. A local maximum is regarded as a speaker turn candidate point when the differences between its value and those of the minima surrounding it are above a certain threshold defined as a fraction of the standard deviation of the amplitude of the distance signal. It means if at least one of the conditions

$$|\max - \min_l| > \alpha\sigma \tag{2}$$

or

$$|\max - \min_r| > \alpha\sigma, \tag{3}$$

is true. $\alpha$ is a real number, $\sigma$ is the standard deviation, and $\min_l$ and $\min_r$ are the left and right minima, respectively, around the peak max.

In addition, we require a minimal duration between two maxima: if two maxima are too close (< 0.5s), the lowest one is discarded. The speaker turn candidate points are marked with a dot in Figure 1.

**Step 4: Validation of the candidates.** The validation is based on a ΔBIC value computed for the adjacent windows that caused the corresponding local peak on the distance graph [1]. The ΔBIC value is given by

$$\Delta\,\mathrm{BIC} = -R + \lambda P, \tag{4}$$

where

$$R = \tfrac{N}{2}\log\left|\mathbf{\Sigma}\right| - \tfrac{N_1}{2}\log\left|\mathbf{\Sigma}_1\right| - \tfrac{N_2}{2}\log\left|\mathbf{\Sigma}_2\right|, \tag{5}$$

$\lambda$ is a penalty factor,

$$P = \tfrac{1}{2}\left(d + \tfrac{1}{2}d(d+1)\right)\log N, \tag{6}$$

$N_1$ and $\mathbf{\Sigma}_1$ are respectively the number and the covariance matrix of the feature vectors in the window $W_1$, $N_2$ and $\mathbf{\Sigma}_2$ are the number and the covariance matrix of the feature vectors, respectively, in the window $W_2$, $N = N_1 + N_2$, $\mathbf{\Sigma}$ is the covariance matrix of the feature vectors of both windows, and $d$ is the dimension of the feature vectors. A speaker turn candidate can be regarded as a true speaker turn if the ΔBIC value for this candidate is negative.

The ΔBIC value (4) is computed for all local peaks on the unsmoothed distance graph that correspond to a speaker turn candidate point on the smoothed distance graph. The local peak with the lowest ΔBIC value is regarded as the true speaker turn. The true speaker turns are marked with a cross in a circle in Figure 1.

**Step 5: Time alignment.** Since the silent parts were eliminated from the speech signal in Step 1, the points of the speaker turns have to be aligned now accordingly. In addition, if a speaker turn is close (less than 0.2 s) to a silent part, the speaker turn is moved into the centre of the silent part.

## 3. EXPERIMENTAL RESULTS

The method described in Section 2 was used for the segmentation of several audio recordings. Results achieved for a recording of radio broadcast news are presented in Table 1. In the odd columns from the left there are the true speaker turns, in the even columns there are the detected speaker turns. The algorithm has detected all the true speaker turns, however, about 10% of the detected turns are false alarms. It seems to be quite a high number, however, after an inspection of the recording, we have found out, that some false alarms occur in the moments when a background noise started, the speaker changed the loudness of his voice, etc.

Comparing the time instant of the true speaker turns with the detected speaker turns one could think that the method is not accurate so much. However, the speaker turns occur mainly in a silence between two utterances, and the displacement of the turn by a fraction of a second is not important.

**Tab 1: Time of the true and detected speaker turns**

| true speaker turn [s] | detected speaker turn [s] | true speaker turn [s] | detected speaker turn [s] | true speaker turn [s] | detected speaker turn [s] | true speaker turn [s] | detected speaker turn [s] |
|---|---|---|---|---|---|---|---|
| 3.40 | 3.40 | 223.00 | 222.93 | 543.00 | 542.62 | 754.30 | 753.79 |
| 7.66 | 7.70 | 275.60 | 275.33 | 562.30 | 561.83 | 795.20 | 794.95 |
| 12.85 | 12.82 | 291.90 | 291.74 | 569.70 | 569.59 | 815.70 | 815.19 |
| 17.30 | 17.29 | 334.60 | 334.41 | 613.30 | 613.07 | 870.30 | 870.12 |
| 42.67 | 42.60 | 350.50 | 350.25 | 620.00 | 619.82 | 876.90 | 877.60 |
| 44.60 | 44.70 | 406.70 | 406.61 |  | 627.85 | 880.30 | 883.94 |
| 51.10 | 50.98 | 410.70 | 410.67 | 648.50 | 648.38 | 884.10 | 886.26 |
| 84.90 | 84.71 | 425.20 | 425.63 | 661.10 | 660.57 | 903 | 902.74 |
| 139.80 | 139.51 | 459.60 | 459.24 |  | 669.35 |  | 906.34 |
| 147.10 | 146.67 |  | 463.08 | 706.30 | 706.07 | 914.20 | 914.04 |
| 172.60 | 172.87 | 481.30 | 481.28 | 714.50 | 714.47 |  | 915.47 |
| 190.50 | 190.15 |  | 493.28 | 729.30 | 729.20 | 959.60 | 959.54 |
| 203.60 | 204.41 | 511.20 | 511.08 | 731.30 | 731.29 | 967.70 | 967.56 |
| 206.70 | 206.51 | 513.70 | 513.78 | 743.60 | 743.51 | 1019.30 | 1018.97 |

## ACKNOWLEDGEMENT

The work was supported by the Ministry of Education of the Czech Republic, projects no. F2286/03/G1 and MSM 235200004.

## REFERENCES

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

[2] M. Greenwood, A. Kinghorn, *SUVing: Automatic Silence/Unvoiced/Voiced Classification of Speech*. Undergraduate Coursework, Department of Computer Science, The University of Sheffield, UK, 1999.