

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

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

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