IDIAP RESEARCH REPORT

Robust Speaker Change Detection

Jitendra Ajmera, Iain McCowan, Hervé Bourlard

IDIAP–RR 02-39

September 2003

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP Research Report 02-39

Robust Speaker Change Detection

Jitendra Ajmera, Iain McCowan, Hervé Bourlard

September 2003

Abstract. Most commonly used criteria for speaker change detection like log likelihood ratio (LLR) and Bayesian information criterion (BIC) have an adjustable threshold/penalty parameter to make speaker change decisions. These parameters are not always robust to different acoustic conditions and have to be tuned. In this letter, we present a criterion which can be used to identify speaker changes in an audio stream without such tuning. The criterion consists of calculating the LLR of two models with the same number of parameters. Results on the Hub4 1997 evaluation set indicate that we achieve a performance comparable to using BIC with optimal penalty term.

1 Introduction

Most of the metric based approaches toward speaker change detection formulate the problem like this: to decide if a speaker change point exists at time t or not, two neighboring windows of relatively small size are considered as shown in Figure 1. The content of these windows are usually sequences of feature vectors extracted from the audio signal. In Figure 1, these sequences are denoted as $X = \{x_1, x_2, ..., x_{N_x}\}$ and $Y = \{y_1, y_2, ..., y_{N_y}\}$, where N_x and N_y are the number of data points in the two windows respectively. Let Z denote the union of the contents of the two windows having $N = N_x + N_y$ data points. The contents of these two windows are compared using a dissimilarity function. Local optima of this dissimilarity function compared to a threshold, are considered to be speaker change points.

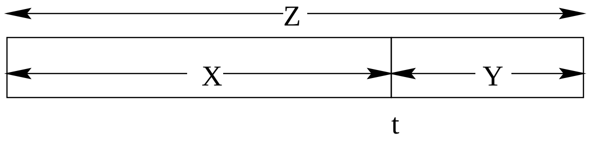

Figure 1: Two neighboring windows with sequences X and Y around time t, when we want to decide if a change point exists or not.

Various metric based algorithms differ in the kind of dissimilarity function they employ, the size of the two windows, the time increments of the shifting of the two windows, and the way the resulting dissimilarity values are evaluated and thresholded. Generally, the threshold parameters are calculated on the basis of experiments and are not robust to different acoustic and environmental conditions. As a result, these parameters have to be tuned before applying them to unseen conditions either manually or with the help of some development data [5].

The main focus of this paper is to introduce a criterion which does not need tuning of any parameter to make decisions about speaker changes and hence proves to be robust to different acoustic conditions. In most commonly used criteria like log likelihood ratio (LLR) and Bayesian information criterion (BIC), a threshold parameter (or a penalty term, as explained later) is required because likelihoods of two different models with different number of parameters are compared. Moreover, these threshold parameters are highly sensitive to different data. In our approach, we force the number of parameters to be the same in the two models used in LLR, to make the likelihoods directly comparable. By using models that are directly comparable, we obtain a natural decision threshold around zero which is robust across a variety of data.

The remainder of this letter is organized as follows: Section 2 summarizes LLR and BIC and their requirements of a tunable threshold to make decisions. Section 3 presents the proposed criterion. Finally, section 4 explains the experimental setup and results for the proposed criterion.

2 Common criteria

LLR [1, 2] and BIC [4, 3, 5] are among the most commonly used criteria for the purpose of speaker change detection. In these cases, the problem is formulated as a hypothesis test, comparing the following two hypotheses:

H_0: The null hypothesis is that there is no speaker change at time t. The maximum likelihood (ML) estimates of the parameters of the single Gaussian density of the complete data (Z) are calculated and denoted as θ_z. Assuming that the data points in X and Y are independent and identically distributed (i.i.d.), the log likelihood of the complete dataset Z under this hypothesis L_0, is calculated as:

$$L_0 = \sum_{i=1}^{N_x} \log p(x_i|\theta_z) + \sum_{i=1}^{N_y} \log p(y_i|\theta_z) \tag{1}$$

where $p(x|\theta)$ is the likelihood of data point x given θ.
H_1: A speaker change point is hypothesized at time t. The ML estimates of the parameters of the Gaussian densities of data set X and Y are individually estimated and denoted as θ_x and θ_y, respectively. The log likelihood, L_1, in this case is calculated as:

$$L_1 = \sum_{i=1}^{N_x} \log p(x_i|\theta_x) + \sum_{i=1}^{N_y} \log p(y_i|\theta_y) \tag{2}$$

The resulting LLR dissimilarity between two windows, d_{llr}, is then calculated as:

$$d_{llr} = L_1 - L_0 \tag{3}$$

Since, the 'maximum' likelihood parameters for subsets X and Y in hypothesis H_1 are estimated individually, each of the two terms on right hand side of (2) are greater than the corresponding term in 1. Hence, we always have $L_1 \geq L_0$ or $d_{llr} \geq 0$. Thus, a threshold is needed to make a decision about a speaker change point using this dissimilarity measure.

BIC is similar to LLR except the fact that the likelihoods now are penalized by the number of parameters used in the model. This is also equivalent to saying that the threshold in this case is a function of the difference of the number of parameters (ΔK) in the two hypotheses. BIC based dissimilarity, d_{bic}, is calculated as:

$$d_{bic} = L_1 - L_0 - \frac{\lambda}{2} \cdot \Delta K \cdot \log N \tag{4}$$

where λ is the penalty factor which should ideally be 1.0. BIC is supposed to have the advantage of not having any thresholding. Ideally, local maxima of d_{bic} greater than 0 should be speaker change points. However, this is only true if $\lambda = 1$ or if there a systematic way to find the optimal value of λ. In absence of this, λ is an implicit threshold embedded into the penalty term. This fact has been mentioned in previous work and was also noticed during our experiments as discussed later. In [1], it was mentioned that the threshold found using BIC principle (with $\lambda = 1$) yielded significantly worse results compared to the best possible threshold selection. In [4], the value of λ used was different than 1.0. In [5] a development dataset was used to find the optimal value of this parameter.

During our experiments we noticed that higher values of λ result in a higher threshold, and thus ignore many genuine speaker changes. A lower value, on the other hand, results in many false alarms.

3 Proposed Criterion

For our proposed criterion, we also formulate the problem as that of hypothesis testing like in previous section. The only difference lies in the fact that in hypothesis H_0 (that there is no speaker change), we model the data Z with a Gaussian mixture model (GMM) with two Gaussian components instead of a single Gaussian density. The ML estimates of the parameters of this GMM, θ'_z, are calculated using expectation-maximization (EM) algorithm. The log likelihood, L'_0 in this case is calculated as:

$$L'_0 = \sum_{i=1}^{N_x} \log p(x_i|\theta'_z) + \sum_{i=1}^{N_y} \log p(y_i|\theta'_z) \tag{5}$$

It should be noted at this point that we always have,

$$L'_0 \geq L_0 \tag{6}$$

since a GMM can always fit the data better (or equally well) compared to a single Gaussian density.

The proposed criterion d_{llrc} is then simply the log likelihood ratio (with constant number of parameters):

$$d_{llrc} = L_1 - L'_0 \tag{7}$$

All local maxima of d_{llrc} greater than 0 are considered to be speaker change points. It can be seen that, in contrast to the standard LLR and BIC techniques, all terms in this criterion are derived directly from the data, and thus the criterion can be expected to be robust to changing data conditions. To get a further insight into the working of this criterion, let us consider two extreme theoretical cases.
Case 1: Let us assume that the subsets X and Y come from two very different speakers and hence have very distinct probability density functions (PDF) such that:

$$p(x_i|\theta_y) \ll p(x_i|\theta_x) \qquad \forall x_i \in X \tag{8a}$$

$$p(y_i|\theta_x) \ll p(y_i|\theta_y) \qquad \forall y_i \in Y \tag{8b}$$

In this case the parameters of the two Gaussian components of the GMM (estimated by EM algorithm) will approximately converge to θ_x and θ_y, and their weights (w_1 and w_2) will approximately converge to $\frac{N_x}{N}$ and $\frac{N_y}{N}$, respectively. Based on this, L_0' can be written as:

$$\begin{aligned} L_0' \quad \approx \quad & N_x \log w_1 + \sum_{i=1}^{N_x} \log p(x_i|\theta_x) \\ & + N_y \log w_2 + \sum_{i=1}^{N_y} \log p(y_i|\theta_y) \end{aligned} \tag{9}$$

Using (9) and (2) in (7), d_{llrc} can be written as:

$$d_{llrc} \approx -N_x \log \frac{N_x}{N} - N_y \log \frac{N_y}{N} \tag{10}$$

It is then easy to see that:

$$0.0 < d_{llrc} \leq N \log 2.0 \tag{11}$$

Since $d_{llrc} > 0$, the criterion will favor the hypothesis that there exists a change point at time t.
Case2: Let us assume that X and Y come from the same speaker, and have very similar PDFs. In an extreme case, we can assume that $\theta_x \approx \theta_y \approx \theta_z$ (θ_z was defined in section 2). This implies that $L_1 \approx L_0$. This together with 6 also implies that $d_{llrc} \leq 0$ and hence the proposed criterion will favor the hypothesis that there is no speaker change. The equality sign in this case is only of theoretical interest and holds in a limiting case when the *real* PDFs of X and Y are mono-Gaussians, which will not occur in practice.

With this, it can be seen that more negative values of this criterion provide a strong confidence toward no speaker change and on the other hand, more positive values provide high confidence toward potential speaker change points.

Of course, it may still be advantageous to tune (whenever possible) an explicit threshold term to further improve performance on a given dataset. However, the principal advantage of this approach is that it provides robust results even without tuning any parameter.

4 Experimental Setup

An experimental setup was designed to compare the performance of the proposed criterion with that of BIC. It was mentioned earlier that the speaker change detection framework also involves shifting and resizing of the two neighboring windows. We basically adopted the algorithm presented in [3] for this purpose while also implementing some useful tricks presented in [4]. The algorithm now runs as follows:

```
1. initialize the interval [a, b]
     a =0, b = MIN_WINDOW;
2. find the change point in [a, b]
```

```
   according to a criterion.
3. if(no change in [a, b])
      b = b+ MORE_FRAMES;
   else if(t is the changing point)
      a = t+1, b=a + MORE_FRAMES;
4. if(b-a > MAX_WINDOW)
      a=b-MAX_WINDOW;
5. go to (2)
```

Where, the values like `MORE_FRAMES, MAX_WINDOW` etc. are also optimized for good performance as well as keeping the computational complexity reasonable. In our experiments `MORE_FRAMES` and `MAX_WINDOW` were chosen to correspond to 1 second and 20 seconds of speech respectively.

For the comparison of different criteria, the basic framework was kept unchanged except the criterion in step 2 above.

The HUB-4 1997 evaluation set was used to test the performance of the proposed criterion. The HUB-4 database consists of nearly 3 hours of broadcast news data in different acoustic conditions, totaling 515 speaker changes from a large variety of speakers, and thus gives a good test of robustness.

Feature vectors used were 24-dimensional mel frequency cepstral coefficients (MFCC) [6] extracted every 10ms, with a window size of 30ms. Diagonal covariance matrices were used for all the experiments, allowing real-time implementation .

4.1 Evaluation Criterion

A change detection system has two possible types of error. Type-I errors occur if a true change is not spotted within a certain window (1 second in our case). Type-II errors occur when a detected change does not correspond to a true change in the reference (false alarm). Type I and II errors are also referred to as precision (PRC) and recall (RCL) respectively, which are defined as:

$$PRC = \frac{\text{number of correctly found changes}}{\text{total number of changes found}} \tag{12a}$$

$$RCL = \frac{\text{number of correctly found changes}}{\text{total number of correct changes}} \tag{12b}$$

In order to compare the performance of different systems, the F-measure is often used and is defined as:

$$F = \frac{2.0 * PRC * RCL}{PRC + RCL} \tag{13}$$

The F-measure varies from 0 to 1, with a higher F-measure indicating better performance.

4.2 Results

The results using the proposed criterion and the BIC are presented in Table 1. It is clear from the results that the performance of the BIC depends heavily on the value of λ (penalty factor). If this value is too high (greater than 7.0 in this case), the algorithm avoids many false alarms (higher PRC), but at the cost of deleting many genuine changes (lower RCL). On the other hand, if the λ is too low (less than 6.0 in this case), the algorithm generates too many false alarms (lower PRC), in addition to detecting most of the genuine changes (higher RCL). This demonstrates that λ in the BIC serves as an implicit data-dependent threshold.

Conversely, the proposed criterion is free of tuning of any such threshold or penalty factor. The performance using the proposed criterion is comparable to that of the BIC with optimal λ (= 6.0 or 7.0), and better when compared to other values of λ, such as the theoretically motivated case of $\lambda = 1$.

Criterion	RCL	PRC	F
Proposed	0.65	0.68	0.67
BIC ($\lambda = 1.0$)	0.81	0.22	0.35
BIC ($\lambda = 4.0$)	0.77	0.46	0.58
BIC ($\lambda = 5.0$)	0.74	0.57	0.64
BIC ($\lambda = 6.0$)	0.71	0.66	0.68
BIC ($\lambda = 7.0$)	0.66	0.71	0.68
BIC ($\lambda = 8.0$)	0.60	0.73	0.66

Table 1: Results of the proposed criterion and BIC (with different values of λ) on HUB-4 1997 evaluation data.

5 Conclusion

A novel criterion for speaker change detection was proposed in this letter. In contrast to other metric based approaches, the proposed criterion gives robust results without tuning any penalty/threshold term. This is achieved by making a hypothesis test where the number of parameters used to model the data in the two hypothesis are forced to be the same. Thus, the likelihoods in these two hypotheses are directly comparable. The usefulness and robustness of this measure was further illustrated with the help of experiments where the proposed criterion achieved comparable results to that of using BIC with optimal penalty term.

References

[1] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, "Strategies for automatic segmentation of audio data," *ICASSP*, vol. 3, pp. 1423–1426, 2000.

[2] H. Gish, M. H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," *ICASSP*, 1991.

[3] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," Tech. Rep., IBM T.J. Watson Research Center, 1998.

[4] Alain Tritschler and Ramesh Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian information criterion," *Eurospeech*, pp. 679–682, 1999.

[5] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker based segmentation for audio data indexing," *Speech Communication*, vol. 32, pp. 111–126, Sept. 2000.

[6] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken utterances," *IEEE Trans. ASSP*, vol. 28, pp.357–366, Aug. 1980