Unsupervised Speaker Change Detection Using Probabilistic Pattern Matching

A. Malegaonkar, A. Ariyaeeinia, P. Sivakumaran, and J. Fortuna

***Abstract*—This letter presents an investigation into the use of a probabilistic pattern matching approach for detecting speaker changes in audio streams. The experiments are conducted using clean speech as well as broadcast news material. It is shown that, in the proposed approach, the use of bilateral scoring is considerably more effective than unilateral scoring. Appropriate score normalization methods are considered in the study. It is observed that in all the cases, the bilateral scoring approach outperforms the currently popular method of Bayesian information criterion (BIC) for speaker change detection. This letter discusses the principles of the proposed approach and details the experimental investigations.**

***Index Terms*—Bilateral scoring, probabilistic pattern matching, score normalization, speaker change detection.**

I. Introduction

THE PROBLEM of detecting speaker changes in a given audio stream, without prior acoustic information on the speakers, has received a great deal of interest in recent years [1]–[6]. This is mainly due to its repeated occurrence in various applications of speaker recognition, such as speaker tracking and speaker diarization, improving the accuracy of speech recognition systems (via speaker normalization, adaptation), indexing audio recordings, and providing cues for scene, topic, and program changes in multimedia applications.

Addressing this problem involves two phases: the extraction of the feature parameters that uniquely represent an individual and the utilization of the available feature parameters in the best possible way to detect the speaker changes. This letter is concerned with the latter phase. The pioneering work in this phase of operation [1] involves using a sliding window through the audio stream and measuring the similarity between the adjacent subsets of the data within each window positioning. If the level of similarity falls below a threshold, then a speaker change is registered. In that work, the generalized log-likelihood ratio is used as the similarity measure. Since then, various other measures have been investigated. These include the Kullback–Leibler symmetrical measure (KL-2) [2], Bhattacharyya measure [3], divergence measure [2], and distances derived from second-order statistics [3].

Manuscript received December 6, 2005; revised February 3, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Steve Renals.

A. Malegaonkar, A. Ariyaeeinia, and J. Fortuna are with the School of Electronic, Communication and Electrical Engineering, University of Hertfordshire, Herts AL10 9AB, U.K. (e-mail: A.Malegaonkar@herts.ac.uk; A.M.Ariyaeeinia@herts.ac.uk; J.Fortuna@herts.ac.uk).

P. Sivakumaran is with Canon Research Centre Europe Ltd., Berks RG12 2XH, U.K. (e-mail: siva@cre.canon.co.uk).

Digital Object Identifier 10.1109/LSP.2006.873656

An alternative to the above approach is the method based on Bayesian information criterion (BIC) [4]. This is based on a model selection approach and involves statistical hypothesis testing. This has been the most dominant approach for speaker change detection in recent years. Its popularity is mainly due to its superior ability to detect various acoustic changes, including speaker changes [4], [5].

In this letter, a probabilistic pattern matching approach is proposed for speaker change detection. This is based on quantifying the likelihood of the speaker model built on each side of the hypothesized speaker change point generating the data on the other side. The quantified score is then enhanced using a set of background models and is used to confirm or reject the speaker change in question. This letter details the proposed method and experimentally examines its effectiveness in relation to BIC. The remainder of this letter is structured as follows. Section II introduces the proposed approach for unsupervised speaker change detection. Section III describes the experimental investigation, and the overall conclusions are presented in Section IV.

II. Proposed Approach

In this method, a fixed-size analysis window is slid through the given audio stream at a predetermined rate. At each instance, a speaker change is hypothesized at the midpoint of the window. This results in the following two hypothesized speaker segments:

$$\mathbf{O}^i_{\mathrm{LHS}} = \left\{\mathbf{o}^i_1, \mathbf{o}^i_2, \ldots, \mathbf{o}^i_{N/2}\right\} \tag{1}$$

$$\mathbf{O}^i_{\mathrm{RHS}} = \left\{\mathbf{o}^i_{1+N/2}, \mathbf{o}^i_{2+N/2}, \ldots, \mathbf{o}^i_{N}\right\} \tag{2}$$

where $\mathbf{o}^i_n$ is the $(i+n)$th feature vector of the audio stream, and N is the size of the analysis window. These segments can then be used to build two speaker models, $\boldsymbol{\lambda}^i_{\mathrm{LHS}}$ and $\boldsymbol{\lambda}^i_{\mathrm{RHS}}$, respectively, as shown in Fig. 1. One possible method to detect the speaker change point is to quantify the likelihood of $\boldsymbol{\lambda}^i_{\mathrm{LHS}}$ being the data generator, given $\mathbf{O}^i_{\mathrm{RHS}}$, i.e., determining the conditional probability $p(\boldsymbol{\lambda}^i_{\mathrm{LHS}} \mid \mathbf{O}^i_{\mathrm{RHS}})$. This can be estimated using the Bayes' theorem as

$$L\left(\boldsymbol{\lambda}^i_{\mathrm{LHS}} \mid \mathbf{O}^i_{\mathrm{RHS}}\right) = L\left(\mathbf{O}^i_{\mathrm{RHS}} \mid \boldsymbol{\lambda}^i_{\mathrm{LHS}}\right) - L\left(\mathbf{O}^i_{\mathrm{RHS}}\right) \tag{3}$$

where $L(\cdot) = \log p(\cdot)$. In this equation, the prior probability of the speaker model, $p(\boldsymbol{\lambda}^i_{\mathrm{LHS}})$, is not included, as it can be considered equal for all the instances of the analysis window. It should be noted that the above formulation could equally be used for quantifying the likelihood of $p(\boldsymbol{\lambda}^i_{\mathrm{RHS}})$ being the data generator,

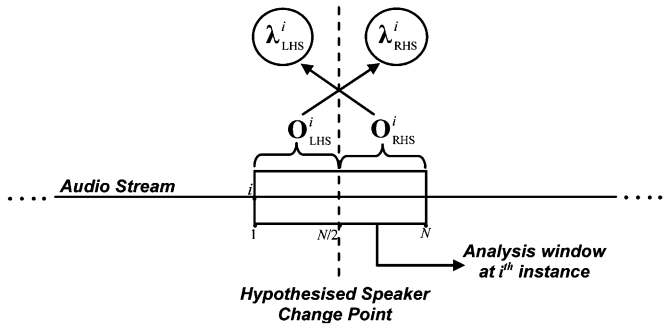

Fig. 1. Probabilistic approach to speaker change detection.

given $\mathbf{O}^i_{\mathrm{LHS}}$, i.e., by estimating $p(\boldsymbol{\lambda}^i_{\mathrm{RHS}} \mid \mathbf{O}^i_{\mathrm{LHS}})$. The presumed speaker change is confirmed by comparing the estimated value of $p(\boldsymbol{\lambda}^i_{\mathrm{LHS}} \mid \mathbf{O}^i_{\mathrm{RHS}})$ with a preset threshold. In theory, setting this threshold is relatively easy since $L(\boldsymbol{\lambda}^i_{\mathrm{LHS}} \mid \mathbf{O}^i_{\mathrm{RHS}})$ is less affected by the variations in the speech originated from the same speaker. It should be noted that $L(\mathbf{O}^i_{\mathrm{RHS}} \mid \boldsymbol{\lambda}^i_{\mathrm{LHS}})$ is indeed affected by such speech variations and (3) relies on $L(\mathbf{O}^i_{\mathrm{RHS}})$ to compensate for these effects. In practice, therefore, the estimation of $L(\mathbf{O}^i_{\mathrm{RHS}})$ is a critical factor to the success of the above approach. Section II-A provides further information on the methods used in this letter for this estimation. For the purpose of this letter, the above approach is referred to as *unilateral scoring-based speaker change detection* (ULS-SCD).

It is reported in speaker recognition that two different speakers are usually not reciprocal. That is, when the models built using speech from a speaker (speaker A) are matched against speech from another speaker (speaker B), they may not return high likelihoods, while speech from speaker A matched against the models built using speech from speaker B giving high likelihoods [6]. This implies that if there is a speaker change at the hypothesized point and $L(\boldsymbol{\lambda}^i_{\mathrm{LHS}} \mid \mathbf{O}^i_{\mathrm{RHS}})$ is high, $L(\boldsymbol{\lambda}^i_{\mathrm{RHS}} | \mathbf{O}^i_{\mathrm{LHS}})$ is not necessarily high. This leads to an alternative method that, as described below, is considered to be superior to the ULS-SCD approach in reducing the misdetection rate, that is, the rate of missing correct speaker changes. In this method, the score used to decide the hypothesized speaker change is computed as follows:

$$S^i_{\mathrm{SCD}} = p\left(\boldsymbol{\lambda}^i_{\mathrm{LHS}} \mid \mathbf{O}^i_{\mathrm{RHS}}\right) \times p\left(\boldsymbol{\lambda}^i_{\mathrm{RHS}} \mid \mathbf{O}^i_{\mathrm{LHS}}\right). \quad (4)$$

In the log likelihood domain, the score is expressed as

$$\left\{L\left(\mathbf{O}^i_{\mathrm{RHS}} \mid \boldsymbol{\lambda}^i_{\mathrm{LHS}}\right) - L\left(\mathbf{O}^i_{\mathrm{RHS}}\right)\right\} + \left\{L\left(\mathbf{O}^i_{\mathrm{LHS}} \mid \boldsymbol{\lambda}^i_{\mathrm{RHS}}\right) - L\left(\mathbf{O}^i_{\mathrm{LHS}}\right)\right\}. \quad (5)$$

There are two assumptions behind the above formulation: $p(\boldsymbol{\lambda}^i_{\mathrm{LHS}} \mid \mathbf{O}^i_{\mathrm{RHS}})$ and $p(\boldsymbol{\lambda}^i_{\mathrm{RHS}} \mid \mathbf{O}^i_{\mathrm{LHS}})$ are statistically independent, and the samples of speech from the same speaker always match well, irrespective of which is used to build the model. For the purpose of this letter, the above method is referred to as *bilateral scoring-based speaker change detection* (BLS-SCD).

The basis for the view about the superiority of BLS-SCD over ULS-SCD is the earlier work in speaker verification [6]. There, it has been established that the use of bilateral scoring is advantageous over unilateral scoring due to its capability in reducing false-alarm errors. The fact that a false alarm in speaker verification corresponds to a missed detection in speaker change detection suggests that BLS-SCD should be more effective than ULS-SCD.

It should be noted that a measure known as XBIC, which is formulated using a comparison between BIC and a distance measure for hidden Markov models (HMMs), has recently been proposed for speaker change detection [7]. This measure can also be formed from (5) by excluding the terms that offer robustness against variations in speech (from the same speaker). Therefore, it appears that BLS-SCD, which incorporates a mechanism for tackling the effects of such variations, has a clear advantage over XBIC.

A. Score Normalization Methods

As it can be understood from the above discussions, the estimation of the normalization terms, $L(\mathbf{O}^i_{\mathrm{LHS}})$ and $L(\mathbf{O}^i_{\mathrm{RHS}})$, in (5) is a critical factor to the effectiveness of the proposed method. For this purpose, various techniques can be adopted from the field of speaker recognition [8]–[11]. These techniques can be categorized into two groups. The first group is based on the Bayes' theorem, and it involves approximating the normalization terms in (5) with likelihoods that may be derived by using various forms of "anti-speaker" modeling [8]. The main approaches in this group are world model normalization (WMN), cohort normalization (CN), and unconstrained cohort normalization (UCN) [8]. WMN approximates the anti-speaker model by using a single background model known as world model (or universal background model), whereas both CN and UCN provide an estimate of the anti-speaker model by using a set of background speaker models. The main difference between CN and UCN is in the way the background speaker models are chosen. In the context of this letter, the background speaker models for CN are chosen based on their closeness to the considered hypothesized speaker models, (e.g., $\boldsymbol{\lambda}^i_{\mathrm{LHS}}$ and $\boldsymbol{\lambda}^i_{\mathrm{RHS}}$). On the other hand, in UCN, this choice is based on the closeness of background speaker models to the speech segments under test, i.e., $\mathbf{O}^i_{\mathrm{LHS}}$ and $\mathbf{O}^i_{\mathrm{RHS}}$, in the analysis window. In BLS-SCD, UCN is computationally less costly than CN. This is because the latter involves a pair-wise background scoring procedure [10]. Hence, only UCN is considered in this letter.

The second group is based on the standardization of score distributions, and it mainly includes two methods: T-norm and Z-norm [11]. The latter involves a fixed set of utterances, whereas the former involves a fixed set of background speaker models [8]. It should be noted that, in speaker recognition, Z-norm is applied in conjunction with one of the methods in the first group or T-norm [8]. The score normalization method that is used in conjunction with Z-norm is referred to as Z*-norm in this letter. The reason for such a combination of normalization methods can be described as follows. Z-norm aims to tackle the problem of misalignment amongst the registered speaker models due to variations in training conditions. In order for this to be effective, the variations in the test conditions have to be pre-normalized by using Z*-norm [8]. BLS-SCD, however, represents a unique situation in which the speaker models are

built in the test phase, and the cause of model misalignment is the variation in the test conditions. As a result, in BLS-SCD, it is likely that any model misalignment is dealt with by Z*-norm effectively, before getting to the Z-norm stage. Therefore, it is believed that the deployment of Z-norm is redundant in BLS-SCD, and it is not considered for the purpose of this letter.

Amongst WMN, CN, UCN, and T-norm, WMN would be the preferred choice because it involves a single background model for scoring. This can reduce the computational cost and hence increase the processing speed. The score for BLS-SCD with WMN is given as

$$L\left(\mathbf{O}^i_{\mathrm{RHS}} \mid \boldsymbol{\lambda}^i_{\mathrm{LHS}}\right) - L\left(\mathbf{O}^i_{\mathrm{RHS}} \mid \boldsymbol{\lambda}_{\mathrm{WM}}\right) + L\left(\mathbf{O}^i_{\mathrm{LHS}} \mid \boldsymbol{\lambda}^i_{\mathrm{RHS}}\right) - L\left(\mathbf{O}^i_{\mathrm{LHS}} \mid \boldsymbol{\lambda}_{\mathrm{WM}}\right) \quad (6)$$

where $\boldsymbol{\lambda}_{\mathrm{WM}}$ is a world model trained using utterances from a large number of male and female speakers. It should be noted that an equation similar to (6), which is referred to as the cross-likelihood ratio (CLR), is extensively used for clustering purposes in the task of speaker diarization [12]–[15]. Moreover, equations similar to that given in (6) can be derived for the cases of T-norm and UCN by applying appropriate modifications to (5) [8].

III. Experimental Investigation

A. Speech Data

The experiments in this letter are conducted using speech data obtained in clean audio conditions as well as in broadcast news audio conditions. The data with clean audio conditions is an artificial recording created using a subset of the TIMIT database. This recording is similar to that adopted in [2] and has 1000 speaker turns. This constitutes test speech data of 4.5 h in length, where the average duration of a speaker-specific segment is 7 s. The experiments with broadcast news audio are conducted using the speech data from five candidate recordings from the "CNN prime news" show of the HUB-4 database. The number of speaker turns in this data is 500, and the length of the test speech data in this case is around 2.5 h. The average duration of a speaker-specific segment in this data is about 25 s.

B. Feature Representation

For the purpose of this letter, the tth frame of the input speech data is represented as $\mathbf{c}_t = \{c_t(1), c_t(2), \ldots, c_t(20)\}$, where $c(i)$ is the ith, linear predictive coding-derived cepstral (LPCC) parameter. The extraction of LPCC parameters is based on first pre-emphasising the input speech data using a first-order digital filter and then segmenting it into 20 ms frames at intervals of 10 ms using a Hamming window.

C. Speaker Representation

The speaker representation in this letter is based on a single Gaussian model having a full covariance structure. Such modeling has already been shown [3] to be capable of reliably representing the speaker-related information in short data segments. Additionally, it can help suppress the effects of phonetic diversity present in small datasets. This is particularly beneficial when the test data segments are produced by the same speaker. The individual background models (in the case of UCN and T-Norm) as well as the world model all have the same topology as that of the speaker models. Albeit it is possible to use Gaussian models of higher orders for representing WM, the said choice helps faster processing of audio material. More importantly, the earlier studies in speaker recognition, conducted by the authors [16], have indicated that there are clear advantages (in terms of score normalization effectiveness) in using the same topology for speaker models and the background model. In the case of each of the two databases used, the dataset adopted for training the background models includes an equal number of male and female speakers. The world model is trained by pooling all the speech material from all the speakers in this dataset. In the case of TIMIT, this dataset consists of 90 speakers. This represents speech material with duration of 1 h. In the case of HUB-4, this dataset is based on three of the recordings in the database. The number of background speakers in this case is 40, and the associated speech material accounts for approximately 1.5 h in duration.

TABLE I
COMPARISON OF THE XBIC PERFORMANCE WITH THAT OF BLS-SCD (WITH THE CONSIDERED SCORE NORMALIZATION TECHNIQUES) IN TERMS OF EER (%)

	XBIC	BLS-SCD$_{\mathrm{WMN}}$	BLS-SCD$_{\mathrm{UCN}}$	BLS-SCD$_{\mathrm{T\text{-}norm}}$
TIMIT	4.28	2.78	2.71	3.42
HUB-4	21.90	17.22	16.97	18.90

D. Audio Scanning and Testing Procedure

The testing in this experiment is conducted by sliding a window of 4 s duration through the recording at a rate of 0.1 s between two successive instances of the window. The length of the sliding window and the sliding rate is decided *a priori* using pilot experiments. A detected speaker change point is declared to be correct if it is within a 0.5-s margin around the actual speaker change point. The error rates are calculated in terms of the percentage of correct speaker change points that are missed, i.e., missed detection rate (MDR), and the percentage of test points wrongly identified as speaker change points, i.e., the false-alarm rate (FAR). This error calculation is the same as that given in [17]. The equal error rate (EER %) is then calculated by adjusting the decision threshold such that MDR and FAR are equal. This is then used as the measure of performance in this letter.

E. Experimental Conditions, Results, and Discussions

The aim of the first experiment is to compare the relative effectiveness of XBIC and BLS-SCD for both the TIMIT and HUB-4 test data. Although WMN is envisaged as the preferred choice in Section II-A, the results of experiments with the other considered score normalization methods are also given in Table I for the sake of comparison.

It is observed from this table that BLS-SCD is more effective than XBIC. This is mainly due to the incorporation of score normalization in the measure. It is also observed that amongst the considered score normalization techniques, UCN and WMN achieve better results than T-Norm. This trend in

TABLE II
COMPARISON BETWEEN BLS-SCD $_{\mathrm{WMN}}$ AND ULS-SCD IN TERMS OF EER (%)

	BLS-SCD$_{\mathrm{WMN}}$	ULS-SCD$_{\mathrm{WMN}}$
TIMIT	2.78	7.50
HUB-4	17.22	23.80

TABLE III
COMPARISON BETWEEN BIC AND BLS-SCD $_{\mathrm{WMN}}$ IN TERMS OF EER (%)

	BLS-SCD$_{\mathrm{WMN}}$	BIC
TIMIT	2.78	4.24
HUB-4	17.22	21.66

relative performance is observed for both datasets (i.e., TIMIT and HUB 4). Hence, the number of background speakers deployed does not seem to significantly influence the relative performance of T-Norm. WMN exhibits similar performance to that of UCN but is computationally less intensive. Thus, as asserted in Section II-A, WMN can be considered a better choice in the proposed approach.

The next set of experiments compares the effectiveness of BLS-SCD with that of ULS-SCD for both sets of the test data (see Table II). In this case, only WMN is considered for score normalization.

It is quite clear from these results that BLS-SCD is considerably more effective than ULS-SCD. This proves the assertions made in Section II.

The final experiment in this letter investigates the relative performance of the currently popular technique of BIC and $\mathrm{BLS{-}SCD_{WMN}}$. The results are presented in Table III.

It is clear from the results that $\mathrm{BLS{-}SCD_{WMN}}$ outperforms BIC in detecting speaker change points. Comparing the BIC results given in Table III with those obtained with $\mathrm{BLS{-}SCD_{UCN}}$ and $\mathrm{BLS{-}SCD_{T{-}Norm}}$ (see Table I), it can be said that using any of the considered score normalization techniques with BLS-SCD results in better performance than that obtained with BIC. The only drawback of BLS-SCD when compared with BIC is the slower speed of operation as a result of the background scoring procedure.

IV. CONCLUSION

The probabilistic pattern matching approach is advocated for speaker change detection. In this approach, the technique of bilateral scoring is proved highly beneficial for speaker change detection. This technique is shown to be significantly more effective than the currently popular technique of BIC and the recently introduced technique of XBIC. This is mainly due to its inclusion of score normalization techniques in the procedure. In using score normalizations with bilateral scoring, WMN and UCN are observed to achieve similar level of performance that is better than that obtained with T-norm. The advantage of using WMN is the higher processing speed due to the involvement of a single background model. The operational speed of BLS-SCD obtained with WMN is still slower than that of BIC. Indeed, in some applications, this may be found acceptable considering the higher accuracy offered. While the effectiveness of normalized bilateral scoring for unsupervised speaker change detection is established in this letter, further work is being considered for further enhancement of the approach. In this regard, a projected study is that of investigating the performance of adapted speaker models. This is to benefit from the robustness offered by the model adaptation procedure against the problem of short data length [17], providing opportunities for dealing with unseen data [9], and optimizing the topology of speaker models for the length of the data available.

REFERENCES

[1] H. Gish *et al.*, "Segregation of speakers for speech recognition and speaker identification," in *Proc. ICASSP*, 1991, vol. 2, pp. 873–876.
[2] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Commun.*, vol. 32, no. 1–2, pp. 111–126, 2000.
[3] S. Johnson, "Speaker tracking," M.Phil. thesis, CUED, Univ. Cambridge, Cambridge, U.K., 1997.
[4] S. Chen and P. Gopalakrishnan, "Speaker, environment, and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Speech Recognition Workshop*, 1998, pp. 127–132.
[5] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian information criterion," in *Proc. Eurospeech*, 1999, vol. 2, pp. 679–682.
[6] E. Parris and M. Carey, "Multilateral techniques for speaker recognition," in *Proc. ICSLP*, 1998, pp. 1343–1346.
[7] X. Anguera, XBIC: Real-Time Cross Probabilities Measure for Speaker Segmentation Aug. 2005, ICSI—Berkeley Tech. Rep.
[8] J. Fortuna, P. Sivakumaran, A. Ariyaeeinia, and A. Malegaonkar, "Relative effectiveness of score normalization methods in open-set speaker identification," in *Proc. Speaker Odyssey*, 2004, pp. 369–376.
[9] D. A. Reynolds *et al.*, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.
[10] A. E. Rosenberg *et al.*, "The use of Cohort normalized scores for speaker verification," in *Proc. ICSLP*, 1992, vol. 1, pp. 599–602.
[11] R. Auckenthaler *et al.*, "Score normalization for text-independent speaker verification systems," *Digit. Signal Process.*, vol. 10, no. 1–3, pp. 42–54, 2000.
[12] D. Reynolds *et al.*, "Blind clustering of speech utterances based on speaker and language characteristics," in *Proc. ICSLP*, 1998, pp. 610–613.
[13] D. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. ICASSP*, 2005, vol. 5, pp. 953–956.
[14] R. Sinha *et al.*, "The Cambridge University march 2005 speaker diarization system," in *Proc. Interspeech*, 2005, pp. 2437–2440.
[15] X. Zhu *et al.*, "Combining speaker identification and BIC for speaker diarization," in *Proc. Interspeech*, 2005, pp. 2441–2444.
[16] J. Fortuna, A. Malegaonkar, A. Ariyaeeinia, and P. Sivakumaran, "On the use of decoupled and adapted Gaussian mixture models for open-set speaker identification," in *Proc. 3rd COST-275 Workshop Biometrics Internet*, 2005, pp. 41–44.
[17] M. Roch and Y. Cheng, "Speaker segmentation using MAP-adapted Bayesian information criterion," in *Proc. Speaker Odyssey*, 2004, pp. 349–354.