# A TWO-STAGE MULTI-FEATURE INTEGRATION APPROACH TO UNSUPERVISED SPEAKER CHANGE DETECTION IN REAL-TIME NEWS BROADCASTING

*Lei Xie and Guangsen Wang*

Audio, Speech and Language Processing Group, School of Computer Science,
Northwestern Polytechnical University, Xi'an
lxie@nwpu.edu.cn

**ABSTRACT**

This paper presents a two-stage multi-feature integration approach for unsupervised speaker change detection in real-time news broadcasting. We integrate MFCC and LSP features (i.e. a perceptual feature plus a articulatory feature) in the metric-based potential speaker change detection stage to collect speaker boundary candidates as many as possible. We adopt a weighted Bayesian information criterion (BIC) to integrate boundary decisions from MFCC and LSP features in the speaker boundary confirmation stage. This multi-feature integration strategy makes use of the complementarity between perceptual features and articulatory features to achieve a performance gain. Speaker change detection experiments show that the multi-feature integration approach significantly outperforms the individual features with relative improvements of 26% over the LSP-only approach and 6% over the MFCC-only approach.

***Index Terms***— speaker change detection, speaker segmentation, audio segmentation, audio content analysis

## 1. INTRODUCTION

Speaker change detection or speaker segmentation aims at finding shift points between two successive speakers in an audio stream. The topic has drawn a great deal of interest in recent years [1]-[8] since detecting speaker changes is an important *preprocessing* step for various subsequent tasks such as speaker recognition, tracking and diarization, speaker normalization or adaptation for speech recognition, topic segmentation, multimedia indexing and retrieval. In broadcast news (BN), a major media channel delivered in continuous audio/video stream, the number and identities of speakers are often not known since such programs usually contain diverse speech from anchors, reporters, interviewees, spokesmen and other speakers. It is desirable to carry out unsupervised speaker change detection that finds out speaker shifts without prior information on the number, identities and acoustic information of speakers.

Approaches in unsupervised speaker change detection can be categorized into metric-based, model-based, decoder-guided, model-selection-based and hybrid approaches [1]. Metric-based methods simply measure the difference between two consecutive audio clips that are shifted along the audio signal, and speaker changes are identified at the maxima of the dissimilarity in terms of some distance metric, e.g. vector quantization distortion (VQD), K-L distance and divergence shape distance (DSD) [2]. Model-based approaches are based on recognizing specific speakers via Gaussian mixture models (GMM) [3] or hidden Markov Models (HMM). Recently, Sung [4] proposed a model-based approach that employs support vector machines (SVM). Hain *et al.* [5] proposed a decoder-guided approach that segments a speech stream into male and female clips via a gender-dependent phone recognizer. In model-selection-based methods, the segmentation problem is switched to a model selection problem between two nested competing models. Bayesian information criterion (BIC) is often adopted as the model selection criterion since it has some nice properties such as robustness, threshold-free and optimality [4][6][7].

Recently, much effort has been devoted to hybrid methods that combine merits from above different approaches to achieve better performance over single approaches [1][8]. Lu *et. al.* [8] proposed a two-stage approach for speaker segmentation in real-time news broadcasting. The potential change detection stage proposes potential speaker change points via a metric-based method (DSD of Linear Spectral Pairs between consecutive audio clips) and the speaker boundary refinement stage removes boundary false positives via a model-selection-based method (BIC). In their approach, GMM is used in speaker modeling and an incremental speaker model updating algorithm is proposed to guarantee real-time processing.

This paper extends Lu's work via a multi-feature integration strategy. We present a real-time speaker change detection system that integrates multiple features (LSP and MFCC) in both the potential change detection stage and the speaker boundary refinement stage. We explore the complementarity between articulatory features (LSP) and perceptual features (MFCC). We perform experiments to demonstrate the superiority of multi-feature integration in speaker change detection.

The remainder of this paper is organized as follow. Section 2 describes the two-stage speaker change detection approach. Section 3 presents our multi-feature integration strategy. Experiments are reported in Section 4. We summarize our work in Section 5.

## 2. TWO-STAGE SPEAKER CHANGE DETECTION

In Lu's approach [8], the front-end processing module first segments the input speech signal into short clips with overlaping and the short clips are further divided into speech frames. Silence frames are not considered in speaker modeling and they are removed by a threshold on short time energy (STE). LSP features are extracted and speaker changes are detected on the sequence of LSP feature vectors.

### 2.1. Potential Speaker Change Detection

Potential speaker change detection stage proposes speaker change candidates by measuring the LSP divergence distance of two consecutive short audio clips along the audio signal [8]. Suppose the

This work was supported in part by the National Natural Science Foundation of China (60802085), the Research Fund for the Doctoral Program of Higher Education in China (20070699015), the Natural Science Basic Research Plan of Shaanxi Province (2007F15), the NPU Aoxiang Star Plan (07XE0150), the NPU Foundation for Fundamental Research (W018103).

LSP vectors are Gaussian, the divergence shape distance (DSD) between two consecutive clips $i$ and $i+1$ is defined by

$$\mathcal{D}_{i,i+1} = \frac{1}{2} tr[(\mathbf{C}_i - \mathbf{C}_{i+1})(\mathbf{C}_{i+1}^{-1} - \mathbf{C}_i^{-1})], \quad (1)$$

where $\mathbf{C}$ is the estimated LSP covariance matrix.

A speaker change candidate is proposed if a local DSD peak is detected, i.e, matching the following conditions:

$$\begin{aligned} \mathcal{D}_{i,i+1} &> \mathcal{D}_{i+1,i+1}, \\ \mathcal{D}_{i,i+1} &> \mathcal{D}_{i-1,i}, \\ \mathcal{D}_{i,i+1} &> \tau_i. \end{aligned} \quad (2)$$

The last condition prevents very low peaks to be selected. The threshold $\tau_i$ is set dynamically by the weighted moving average of the previous $N$ successive distances, i.e.,

$$\tau_i = \alpha \frac{1}{N} \sum_{n=0}^{N} \mathcal{D}(i-n-1, i-n), \quad (3)$$

where $\alpha$ is an empirical amplifier.

### 2.2. Incremental Speaker Model Updating

In Lu's approach, GMM is used as the speaker model. Since the EM algorithm does not match the need of real-time processing due to its recursive model parameter estimation progress, an incremental speaker model updating algorithm is adopted [8].

Suppose the current speaker model $\mathcal{N}(\mu, \mathbf{C})$ is estimated from the previous $K-1$ clips and there is no potential speaker change is detected between $K$ and $K-1$ clips according the the potential speaker change detection stage. We update the speaker model $\mathcal{N}(\mu, \mathbf{C})$ by the speaker model of the $K$th clip $\mathcal{N}(\mu_K, \mathbf{C}_K)$:

$$\mathbf{C}' = \frac{N}{N+N_K}\mathbf{C} + \frac{N_K}{N+N_K}\mathbf{C}_K \quad (4)$$

where $\mathbf{C}_K$ is the covariance matrix of the speaker model $\mathcal{N}(\mu_K, \mathbf{C}_K)$, $N$ and $N_K$ are the number of frames used for modeling $\mathcal{N}(\mu, \mathbf{C})$ and $\mathcal{N}(\mu_K, \mathbf{C}_K)$, respectively. The means $\mu$ and $\mu_K$ are not considered in the speaker modeling process because they are easily biased by different acoustic conditions. Broadcast news programs usually have diverse acoustic conditions such as studio, street, factory, meeting room and stadium, etc. This procedure is repeated until the difference between $\mathbf{C}$ and $\mathbf{C}_K$ is lower than a pre-set threshold or a potential speaker change point is detected. When the update is terminated, a new speaker model is initiated.

According to the above model updating algorithm, one potential speaker may have several Gaussian models if there are enough speech data for the speaker. Segmental clustering is then adopted to form a quasi-GMM model for the speaker. The quasi-GMM is formed by combining all the Gaussian models of the speaker:

$$\mathbf{C} = \frac{N_j}{N}\mathbf{C}_j \quad (5)$$

where $N_j$ is the number of frames used in the estimation of Gaussian model $j$, and $N = \sum_{j=1}^{J} N$ is the total number of frames. The number of Gaussian models for a potential speaker is limited to 32. If 32 is reached, the updating of the quasi-GMM model is terminated.

### 2.3. Speaker Boundary Refinement

The potential speaker change detection stage raises possible boundaries that may contain false alarms. The speaker boundary refinement stage is thus used to make final boundary confirmation, and this is performed by a model-selection-based approach, i.e. BIC [8].

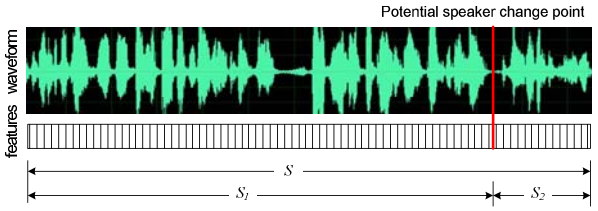


**Fig. 1**. A potential speaker change point splits two neighboring speech segments with feature sequence $S_1$ and $S_2$.

BIC is a penalized maximum likelihood model selection criterion that has been widely used in statistical data processing. As shown in Fig. 1, we wish to check if there is a real boundary at the potential speaker change point. Suppose two Gaussian models estimated from $S_1$ and $S_2$ are $\mathcal{N}(\mu_1, \mathbf{C}_1)$ and $\mathcal{N}(\mu_2, \mathbf{C}_2)$, respectively, and the number of data used to estimate the two models are $N_1$ and $N_2$. $\mathcal{N}(\mu, \mathbf{C})$ is the Gaussian model estimated from $S$ with the number of data $N$ ($N = N_1 + N_2$). The BIC difference between the two models can be defined as

$$\begin{aligned} BIC(\mathbf{C}_1, \mathbf{C}_2) &= \frac{1}{2}(N \log|\mathbf{C}| - N_1 \log|\mathbf{C}_1| - N_2 \log|\mathbf{C}_2|) \\ &\quad - \frac{1}{2}\lambda(d + \frac{1}{2}d(d+1)) \log N \end{aligned} \quad (6)$$

where $\lambda$ is a penalty factor and $d$ is the feature dimension. If $BIC(\mathbf{C}_1, \mathbf{C}_2)$ takes a positive value, the two speech segments are likely originate from different speakers, and a speaker change is confirmed. Otherwise, no speaker change is declared.

Since we compare the current speech clip modelled by $\mathcal{N}(\mu_2, \mathbf{C}_2)$, with the current quasi-GMM speaker model having S Gaussian densities denoted by $\mathcal{N}(\mu_{1j}, \mathbf{C}_{1j})$, $j = 1, 2, \cdots S$ and $S \leq 32$, over $N_{1j}$ feature vectors, the BIC difference can be roughly estimated as [8]

$$D = \sum_{j-1}^{S} w_{1j} BIC(\mathbf{C}_{1j}, \mathbf{C}_2), \quad (7)$$

where $w_{1j} = N_{1j}/N_1$ and $N_1 = \sum_{j=1}^{S} N_{1j}$. When $D > 0$, the potential speaker change previously reported by the metric approach is confirmed as a real speaker boundary by the BIC-based refinement procedure.

## 3. MULTI-FEATURE INTEGRATION

We extend the above two-stage approach by a multi-feature integration strategy to further improve the speaker segmentation performance. Since different kind of speech features may complement each other, we integrate different speech features in both the potential speaker change detection stage and the speaker boundary refinement stage.

### 3.1. Feature Selection

MFCC and LSP are chosen as the speech features due to their popularity in speech and speaker recognition literature. The LSP feature originates from linear prediction (LP) analysis that simulates human speech production process, i.e., an articulatory feature. Articulatory analysis of speech extracts parametric representations of human articulatory organs (i.e. vocal tract) and their actions. The MFCC feature is an another frequently used perceptual feature that is based on the facts of human speech perception. Different from articulatory features witch stems from mechanisms of speech production, speech perceptual analysis relies on how people understand speech. We expect the combination of articulatory and perceptual features

**Table 1**. The audio corpus used in the experiments.

| | |
|---|---|
| Nature | Mandarin BN audio recordings from CCTV-1 19:00-19:30 |
| No. of recordings | 8 (Development set:3,test set: 5) |
| Audio format | 22.05KHz, 16bit, mono |
| Audio duration | ∼240 mins (30mins/recording) |
| No. of speaker change points | 685 (Development set:253, test set:432) |

can improve the discriminative ability among different speakers and thus lead to a better speaker segmentation performance than individual features. The feature orders of LSP and MFCC was 10 and 12 respectively in our system. Only two kinds of features are involved in the integration because more features cannot match the need of real-time processing and will induce intolerable delay.

### 3.2. Multi-feature Integration in Potential Speaker Change Detection

The potential speaker change detection stage aims to recall speaker change points as many as possible. We hope to recall more potential speaker change points by multi-feature integration. We build individual speaker models by MFCC and LSP as two independent change detection agents. We measure MFCC and LSP divergence shape distances separately through Eq. (1). As a result, each agent proposes a set of potential speaker change points via formula (2). We combine the two sets (union operation) as the final set of potential speaker change points that is subject to the speaker boundary refinement procedure. The complementarity between the two features will decrease the speaker boundary missing rate and let as many as possible real speaker change points go through the final boundary confirmation stage.

### 3.3. Multi-feature Integration in Speaker Boundary Refinement

Similar to the potential speaker change detection stage, we calculate the BIC distances for MFCC and LSP. The final speaker boundary decision is made by the following weighted BIC integration:

$$BIC = \omega_{lsp} BIC_{mfcc} + \omega_{lsp} BIC_{lsp} \quad (8)$$

where $\omega$ is the integration weight, and $\omega_{lsp} + \omega_{mfcc} = 1$. The integration weights are tuned through a development data set. Whenever $BIC > 0$, a speaker change point is declared. We intend to remove speaker boundary false alarms as many as possible via the complementarity between different speech features.

## 4. EXPERIMENTS

### 4.1. Corpus

We experiment with audio recordings from CCTV news programs, shown in Table 1. We manually remove the non-speech content (e.g. music) from the audio recordings and manually annotate the speaker change points as the evaluation references. The speech content involves not only pure speech but also noisy speech with background music or with environment sound. The length of speaker segments range from 1.02s to 407.7s with a mean of 18.7s. Fig. 2 shows the histogram of the speaker segment lengths in the corpus. The corpus is randomly separated into a development set and a test set, where the former is for parameter tuning and the latter is for testing.

### 4.2. Experiment Setup

We built a real-time speaker change detection system according to the proposed approach, which was implemented by the Marsyas toolkit[1]. The system integrates three different configurations: MFCC-only, LSP-only and multi-feature integration. The system can swiftly report speaker changes while the BN audio was playing.

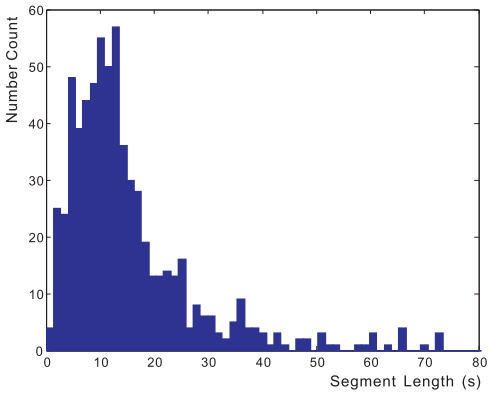


**Fig. 2**. Histogram of the speaker segment length in the corpus.

In the experiments, recall, precision and F1 measure were used to evaluate the speaker change detection performance. A detected speaker change point was considered correct if it lies within a 2s tolerance window on each side of a hand-annotated reference. Empirical parameter tuning was performed on the development set that selects parameters achieving the best F1 measure of speaker segmentation. In the front-end processing, the speech stream was divided into short clips of 3.4s with 3.1s overlapping. The short clips were further segmented into non-overlapping windows of 15ms for speech feature extraction. We set the amplifier of the dynamic threshold $\alpha = 0.6$, the BIC penalty factor $\lambda = 0.8$, and feature integration weights $\omega_{lsp} = 0.15, \omega_{mfcc} = 0.85$.

### 4.3. Results and Analysis

Experimental results on the test set for systems using LSP-only, MFCC-only and multi-feature integration (LSP+MFCC) are summarized in Table 2, Table 3 and Table 4, respectively. We observe that the MFCC-only system outperforms the LSP-only system and the multi-feature integration system achieves the best performance. Multi-feature integration achieves a high F1 value of 0.80 with relative improvements of 26% over the LSP-only system (Lu's approach [8]) and 6% over the MFCC-only system.

In broadcast news, sometimes speaker may shift very swiftly, for example chit-chat between anchors and quick interviews. This kind of quick speaker shifts holds about 8% of the total speaker changes in our corpus. In these cases, the small amount of speaker data may not build a robust speaker model. Thus we specifically analyzed the short speaker segments ($< 5s$) from the test set, and results are summarized in Table 5. As a comparison, we also list the results on brief news in which speaker segments are also not long ( but commonly longer than 10s). As can be seen, the detection performance on quick speaker shifts is much worse than the general performance on all speaker changes. However, multi-feature integration still achieves the best performance. Detection performance on brief news is much better with highest F1 of 0.906 provided by multi-feature integration. The superior performance on brief news is because of the enough speaker data and clean anchor speech in noisy-free studio.

Fig. 3 illustrates the LSP-DSD time curve, the MFCC-DSD time curve and the BIC difference time curve (calculated using multi-feature integration) for a 170s-long brief news clip extracted from the test set. The BN clip contains real speaker changes at about 47s,

[1]http://marsyas.sness.net/

**Table 2**. Experimental results on LSP feature only (Lu's approach)

| Recording | Original | Detected | Miss | False | Recall | Precision | F1 |
|---|---|---|---|---|---|---|---|
| 1 | 103 | 116 | 23 | 36 | 0.727 | 0.690 | 0.732 |
| 2 | 90 | 80 | 39 | 29 | 0.567 | 0.638 | 0.600 |
| 3 | 74 | 56 | 32 | 14 | 0.568 | 0.750 | 0.646 |
| 4 | 73 | 102 | 21 | 50 | 0.721 | 0.510 | 0.594 |
| 5 | 91 | 59 | 49 | 17 | 0.462 | 0.712 | 0.560 |
| All | 432 | 413 | 164 | 146 | 0.620 | 0.647 | **0.633** |

**Table 3**. Experimental results on MFCC feature only

| Recording | Original | Detected | Miss | False | Recall | Precision | F1 |
|---|---|---|---|---|---|---|---|
| 1 | 103 | 94 | 28 | 19 | 0.728 | 0.798 | 0.761 |
| 2 | 90 | 94 | 20 | 24 | 0.778 | 0.745 | 0.761 |
| 3 | 74 | 85 | 16 | 27 | 0.784 | 0.682 | 0.730 |
| 4 | 73 | 75 | 17 | 19 | 0.767 | 0.747 | 0.757 |
| 5 | 91 | 88 | 23 | 20 | 0.747 | 0.773 | 0.760 |
| All | 432 | 436 | 104 | 109 | 0.759 | 0.750 | **0.755** |

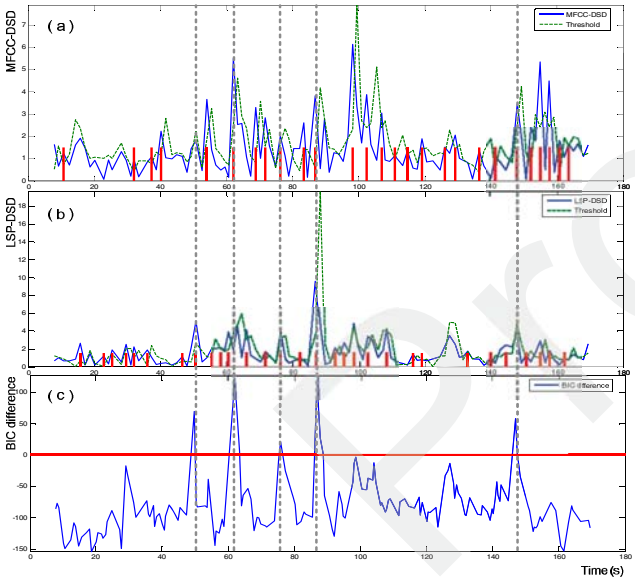


**Fig. 3**. MFCC-DSD (a), LSP-DSD (b) and BIC difference (c) calculated on a short segment of brief news, where the gray dotted lines denote the real speaker change points. Red and gray vertical lines denote potential and real speaker boundaries, respectively.

61s, 73s, 85s and 146s (gray dotted vertical lines in Fig. 3). This figure clearly demonstrates the complementarity between different features. Although the individual features propose a large number of speaker boundary candidates (much more than the number of real boundaries), there are still several real speaker changes missing. However, the missing speaker changes at 47s and 146s by the LSP feature were recovered by the MFCC feature. The feature combination helps the final BIC decision stage successfully detect all the speaker changes.

We also analyzed the time delay between the detect speaker boundaries and the references for our multi-feature integration approach. Results show that about 70% detected points have a time delay lower than 1s, and 97% lower than 1.8s. For computation complexity, to perform speaker change detection on a 30min-long audio

**Table 4**. Experimental results on multi-feature integration

| Recording | Original | Detected | Miss | False | Recall | Precision | F1 |
|---|---|---|---|---|---|---|---|
| 1 | 103 | 103 | 21 | 21 | 0.796 | 0.796 | 0.796 |
| 2 | 90 | 94 | 15 | 19 | 0.833 | 0.798 | 0.815 |
| 3 | 74 | 82 | 15 | 23 | 0.797 | 0.720 | 0.757 |
| 4 | 73 | 74 | 11 | 12 | 0.849 | 0.838 | 0.844 |
| 5 | 91 | 84 | 22 | 15 | 0.758 | 0.821 | 0.788 |
| All | 432 | 436 | 84 | 90 | 0.806 | 0.794 | **0.800** |

**Table 5**. Speaker change detection results on short speaker segments($< 5s$) and brief news.

| | **Short speaker segments** | | | **Brief news** | | |
|---|---|---|---|---|---|---|
| | Recall | Precision | F1 | Recall | Precision | F1 |
| LSP | 0.182 | 1 | 0.308 | 0.679 | 0.844 | 0.753 |
| MFCC | 0.485 | 0.818 | 0.609 | 0.732 | 0.804 | 0.766 |
| Multi-feature | 0.485 | 0.941 | 0.640 | 0.946 | 0.869 | 0.906 |

recording (without playing), the MFCC approach, LSP approach and the multi-feature integration approach spend 13s, 179s and 186s, respectively, on a Celeron 1.5GHz PC with 768M memory.

## 5. SUMMARY

This paper has presented a multi-feature integration strategy for unsupervised speaker change detection in real-time news broadcasting, as an extension of Lu's approach [8]. We have integrated MFCC (a perceptual feature) and LSP (an articulatory feature) in both the potential speaker change detection stage and the speaker boundary refinement stage. Experimental results have shown the complementarity between different speech features, and the integration between MFCC and LSP significantly outperforms the individual features.

In the experiments, we have found that a large number of speaker boundary false alarms are raised by background noise and various audio channels. We plan to test with more robust speech features to further improve the speaker segmentation performance.

## 6. REFERENCES


[1] S.L. Zhang, S.W. Zhang, and B. Xu, "A two-level method for unsupervised speaker-based audio segmentation," in *ICPR*, 2006, pp. 298–301.

[2] K. Jørgensen, L. Mølgaard, and L.K. Hansen, "Unsupervised speaker change detection for broadcast news segmentation," in *EUSIPCO*, 2006.

[3] M.H. Sui, G. Yu, and H. Gish, "An unsupervised, sequential learning algorithm for the segmentation of speech waveform with multiple speakers," in *ICASSP*, 1992, pp. 189–192.

[4] H.-C. Sung, "A novel approach for speaker change detection based on support vector machine," M.S. thesis, National Cheng Kung University, Taiwan, 2005.

[5] T. Hain, S.E. Hahnson, A. Tuerk, and Young S. J. Woodland, P.C., "Segment generation and clustering in the htk broadcast news transcriptin system," in *DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 133–137.

[6] S. Chen and P.S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[7] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," in *Eurospeech*, 1999, vol. 2.

[8] L. Lu and H.-J. Zhang, "Real-time unsupervised speaker change detection," in *ICPR*, 2002, vol. 2.