Investigation on Effectiveness of Mid-level Feature Representation for Semantic Boundary Detection in News Video

R. Radharkishnan, Z. iong, A. Divakaran, B. Raj

TR2003-119 September 2003

Abstract

In our past work, we have attempted to use a mid-level feature namely the state population histogram obtained from the Hidden Markov Model (HMM) of a general sound class, for speaker change detection so as to extract semantic boundaries in broadcast news. In this paper, we compare the performance of our previous approach with another approach based on video shot detection and speaker change detection using the Bayesian Information Criterion (BIC). Our experiments show that the latter approach performs significantly better than the former. This motivated us to examine the mid-level feature closely. We found that the component population histogram enabled discovery of broad phonetic categories such as vowels, nasals, fricatives etc, regardless of the number of distinct speakers in the test utterance. In order for it to be useful for speaker change detection, the individual components should model the phonetic sounds of each speaker separately. From our experiments, we conclude that state/component population histograms can only be useful for further clustering or semantic class discovery if the features are chosen carefully so that the individual states represent the semantic categories of interest.

SPIE Internet Multimedia Management Systems IV

Publication History:

1. First printing, TR-2003-119, September 2003

Investigation on Effectiveness of Mid-level Feature Representation for Semantic Boundary Detection in News Video

Regunathan Radhakrishan, Ziyou Xiong, Ajay Divakaran, Bhiksha Raj
Mitsubishi Electric Research Labs, Cambridge, MA, USA.
{regu, zxiong, ajayd, bhiksha}@merl.com

ABSTRACT

In our past work, we have attempted to use a mid-level feature namely the state population histogram obtained from the Hidden Markov Model (HMM) of a general sound class, for speaker change detection so as to extract semantic boundaries in broadcast news. In this paper, we compare the performance of our previous approach with another approach based on video shot detection and speaker change detection using the Bayesian Information Criterion (BIC). Our experiments show that the latter approach performs significantly better than the former. This motivated us to examine the mid-level feature closely. We found that the component population histogram enabled discovery of broad phonetic categories such as vowels, nasals, fricatives etc, regardless of the number of distinct speakers in the test utterance. In order for it to be useful for speaker change detection, the individual components should model the phonetic sounds of each speaker separately. From our experiments, we conclude that state/component population histograms can only be useful for further clustering or semantic class discovery if the features are chosen carefully so that the individual states represent the semantic categories of interest.

1. INTRODUCTION

Knowledge of the semantic boundaries in news video helps in quick topic based browsing of the content.[1] Past work on semantic boundary detection in news video can be broadly categorized into two approaches. The first approach achieves topic segmentation using closed caption information, embedded captions and text obtained through speech recognition, by themselves or in combination with each other.[2,3] The second approach relies on the detection of news-anchors for topic segmentation. Since news video is typically arranged topic-wise and the news-anchor introduces each topic at the beginning, semantic segmentation of news video can also be achieved through detection of the principal cast. Principal cast detection can be carried out using color, motion and audio features. Wang et al carry out a speaker separation on the audio track and then use the visual or video track to locate the faces of the most frequent speakers.[4] The speaker separation is carried out by first identifying homogenous speech segments. Each homogenous speech segment is then modelled using a GMM. This is followed by comparison of GMM parameters for different speech segments to identify and merge segments belonging to the same speaker.

In this paper, we propose and evaluate two competing approaches for semantic boundary detection that relies on speaker change detection. Approach 1 uses low-level MFCC features at video shot boundaries to detect the points of speaker change using the BIC. Approach 2 is based on a mid-level feature namely the state/component population histogram obtained from the (HMM)/(GMM) of a general sound class such as male or female speech.

The rest of the paper is organized as follows. In section 2, we present a review on prior art for speaker segmentation and our proposed architectures for semantic segmentation of news video. In section 3, we present our experiments and discussion on the results and we conclude in section **??**.

2. PROPOSED APPROACHES

2.1. Prior Art for speaker segmentation

Speaker segmentation involves identification of time stamps which mark the beginning and the ending of the utterance of each speaker in an audio record. With prior knowledge of the speakers in the audio record, segmentation can be achieved relatively easy by using supervised learning models for each of the speaker. Unsupervised or

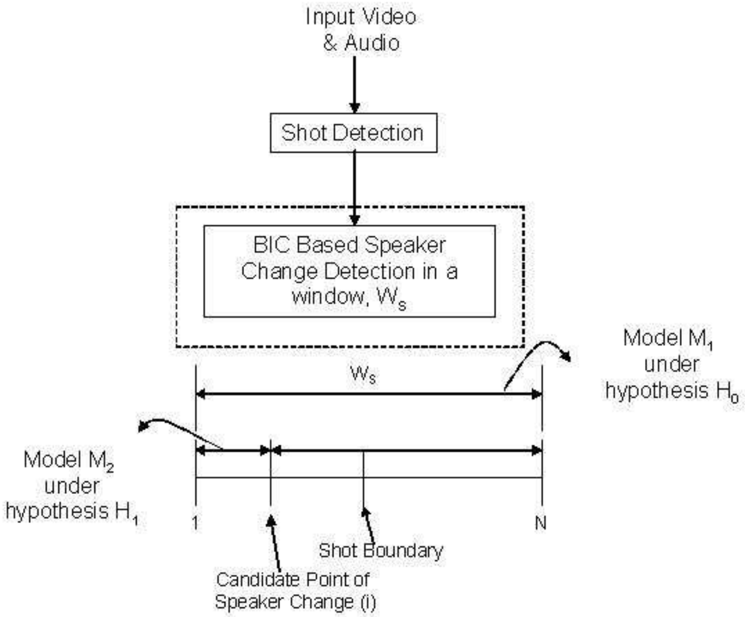

Input Video
& Audio
Shot Detection
BIC Based Speaker Change Detection in a window, W$_s$
Model M$_1$ under hypothesis H$_0$
W$_s$
Model M$_2$ under hypothesis H$_1$
1
N
Shot Boundary
Candidate Point of Speaker Change (i)

Figure 1. Speaker Change Detection Using Low-level MFCC Features at Video Shot-Boundaries

blind speaker segmentation which assumes no apriori knowledge of the speakers is a more difficult and challenging problem.

Prior work on unsupervised speaker segmentation can be broadly categorized into two approaches. The first approach detects signal breaks to extract homogenous audio segments and performs a clustering operation to merge and label similar homogenous segments.[5,4] Wang et al fit a GMM to each homogenous segment and merge two homogenous segments based on similarity between two GMMs.

The second approach performs segmentation and classification jointly.[6,7] It is often based on an iterative method in which statistical models are built for each speaker during each iteration and hypothesis testing is performed to evaluate the likelihood of models at each iteration. The iteration stops when N speaker model at iteration ‘k’ has higher a likelihood than N+1 speaker model at iteration ‘k+1’. Evolutive Hidden Markov Model (E-HMM) and the BIC based speaker segmentation are examples of this iterative approach and are computationally more expensive than the first approach.

2.2. Semantic Segmentation Using Low-Level MFCC Features

Our proposed approach for semantic segmentation of news video using low-level MFCC features is shown in figure 1. The proposed approach is based on the observation that semantically related video shots typically have the same speaker talking without a break. Points of video shot changes are assumed to be candidate points of topic change. Around each point of shot change, low-level MFCC features are extracted for a window of size W_s and BIC is used to determine if there is a speaker change. If there is no speaker change, the video shot is assumed to be belonging the same topic.

Two models are built so as to detect a speaker change within the window W_s of N frames. Model M_1 under hypothesis H_0 assumes there is no speaker change within W_s and approximates the distribution of MFCC features using a single multivariate gaussian with mean μ and covariance Σ. Model M_2 under hypothesis H_1 assumes there is a speaker change at i^{th} frame and approximates the distributions of the features of the corresponding segments with multivariate gaussians with means μ_1 and μ_2 and covariances Σ_1 and Σ_2. Then, ΔBIC (derived from likelihood ratio test for these two hypotheses) is computed for every second within W_s to locate speaker change. ΔBIC is defined as below:

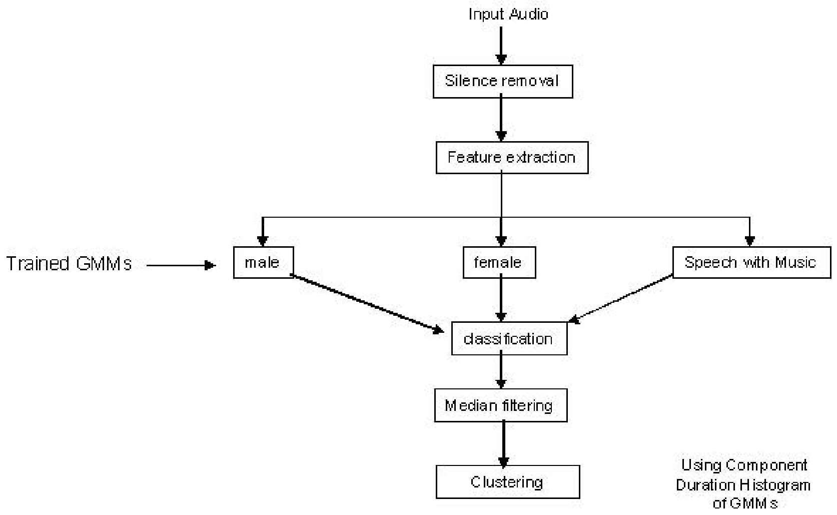


Figure 2. Proposed Approach for Speaker Change Detection Using Mid-Level Features

$$\Delta BIC = -\frac{N}{2}\log|\Sigma| + \frac{i}{2}\log|\Sigma_1| + \frac{N-i}{2}\log|\Sigma_2| + \frac{1}{2}(d + \frac{d(d+1)}{2})\log N \tag{1}$$

where d is the dimension of the MFCC features.

2.3. Semantic Segmentation Using Mid-Level Features

Our approach for semantic segmentation which uses a mid-level feature consists of the following three modules as shown in Fig. 2

- Silence detection and removal
- Extraction of MFCC features and classification into one of the following sound classes: male, female and speech with music.
- Median filtering and Unsupervised clustering.

Our motivation for this architecture stems from the need for a common platform for audio analysis for sports highlights and news video browsing. We have shown that this architecture helps in identifying “interesting” segments in sports video based on audience reaction.[8]

The input audio is broken down into sub-clips of smaller duration. The energy of each sub-clip is calculated and a threshold is used to detect and remove silent sub-clips. MFCC features are extracted from non-silent sub-clips and are classified into one of the three sound classes namely male, female and speech with music.

At this point, most of the male-female speaker transitions are detected. Median filtering is performed on the classification labels to eliminate abrupt changes in speaker genders. In order to identify speaker changes within a contiguous section of male and female sound class, a unsupervised clustering step is performed based on the state population histogram or component population histogram descriptor.

Each classified sub-clip is associated with a state population histogram or component population histogram descriptor depending on whether a HMM or a GMM was used as a classifier in the previous stage. Each bin of the state/component population histogram represents the probability of occurrence of a particular state/component. Since each speaker has a characteristic spectral content, we are motivated to use this descriptor to identify clusters belonging to each speaker in each sound class.

The clustering approach adopted was bottom-up agglomerative dendrogram construction based. In this approach, a distance matrix is first obtained by computing pairwise distance between all utterances to be clustered and then a dendrogram is constructed by merging two closest clusters according to the distance matrix until there is only one cluster. Then, the dendrogram is pruned to obtain the clusters of individual speakers.

In the following section, we present some of the experimental results comparing our architecture with the one that uses low-level audio features for speaker segmentation.

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1. Data Set

Since broadcast news contains mainly three sound classes viz, male speech, female speech and speech with music, we collected training examples for each of the sound classes from three and a half hours of news video from four different TV channels by hand labeling. The audio signals are all mono-channel, 16 bits per sample with a sampling rate of 16 KHz. The database for training GMMs is partitioned into 90%/10% training/testing set for cross-validation. The test sequences for speaker change detection were two audio tracks from TV broadcast news: "news 1" with duration 34 minutes and "news 2" with duration 59 minutes respectively.

3.2. Feature Extraction

The input audio signal is cut into segments of length 1 second and silent segments are removed. For each non-silent 1s segment, MFCC features are extracted as follows. Each segment is divided into overlapping frames of duration 16ms with 8ms overlapping for consecutive frames. Each frame is then multiplied a hamming window function. After performing FFT on each windowed frame, energy in each of 40 the Mel frequency subbands are computed and DCT is performed on the resulting vector for dimensionality reduction.

3.3. Results On Using Low-level MFCC features

At each point of video shot change, low-level features were extracted within a window W_s of size 20s. ΔBIC was computed for every second within this window to locate a speaker change. The BIC based speaker change detection algorithm described in the previous section assumes that there is atmost one speaker change within W_s. However, there can be more than one speaker change within a time window of 20s around a video shot change. This results in multiple valleys in the contour of ΔBIC values for the 20s segment. Therefore, we only look for valleys that are close enough to the video shot boundary, i.e, valleys that are within a 2s window of the shot boundary. The Precision-Recall curve was generated by varying the threshold for detecting the valley as shown in Fig. 3.

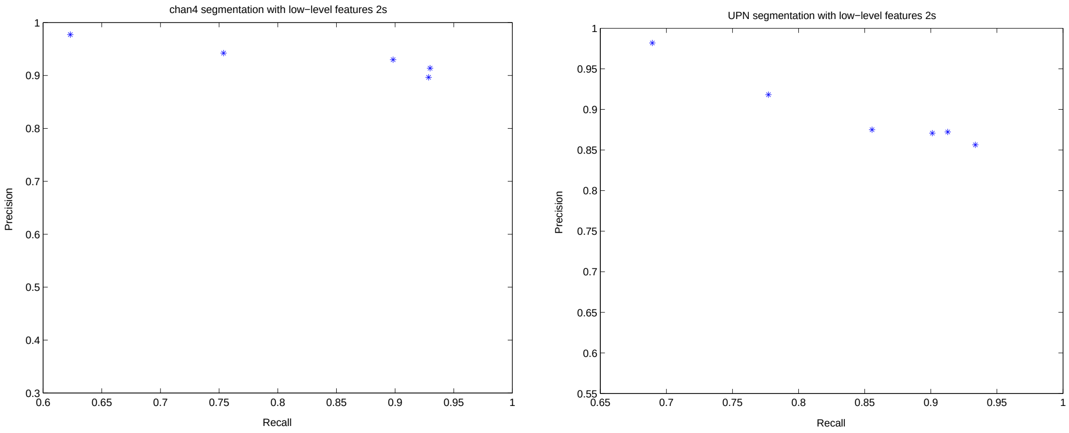
3.4. Results On Using Mid-Level features

A GMM with 10 mixture components is chosen to model a single sound class. Therefore, each 1s segment after classification is associated with component population histogram. Male and female speaker changes are detected after the classification step. In order to detect speaker changes within a contiguous set of male or female segments, the component population histogram was used in the one of the following ways.

- Method 1: Agglomerative Clustering.
- Method 2: Break Detection

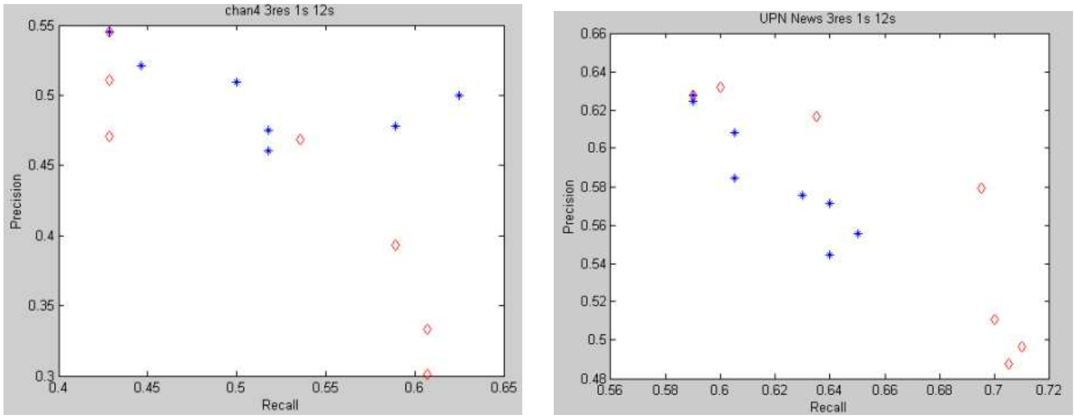
Method 2 attempts to detect speaker changes by comparing a local statistic (component population histogram for the current 1s segment) with a global statistic (average of component population histograms) in its context. At the boundary of a speaker change one would expect, the local statistic for a new speaker to be different from the global statistic in its context. Method 2 can be thought of as a clustering algorithm which deals with only two clusters at a time but incorporates time continuity more naturally than Method 1.

Fig. 4 shows the precision and recall curves for 'news1' and 'news2'. The precision-recall curve for method 1 was generated by varying the height at which dendrogram was pruned to generate clusters. The precision-recall curve for method 2 was generated by varying the break detection threshold.



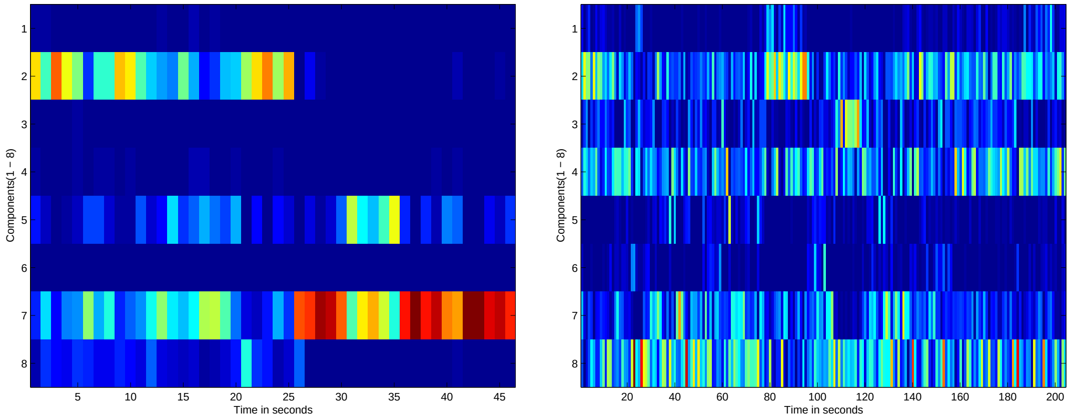
'news1' 'news2'

Figure 3. Speaker Change Detection Performance '*': Using BIC criterion on low-level MFCC features



'news1' 'news2'

Figure 4. Speaker Change Detection Performance '*':Agglomerative Clustering, '◇':Break Detection



Music Clip

Speech Clip

Figure 5. Component Population Histograms for a 8 component mixture

3.5. Discussion

From the experiments in the previous section, it can be observed that the speaker change detection using low-level MFCC features outperforms the approaches based on Mid-level features. Our initial motivation to use the mid-level feature for speaker change detection stems from the fact that these mid-level features were shown to be effective in retrieving similar music clips.[9] This led us to further investigate the meaning of this mid level feature (component population histogram) in the context of music and speech.

A GMM was trained for a music clip containing two different kinds of music one with sounds of mainly percussion instruments and other with the sound of mainly piano. The frames belonging to each mixture component were appended and listened to. It was observed that one component contained mainly audio frames from the percussion instruments and another component contained mainly audio frames from piano. Therefore, the component population histogram distinguishes the two clips of music in feature space as shown in Fig. **??**. It is easy to observe that there is a change in the music characteristics at 25s.

A similar experiment was performed for a speech clip containing 5 different speakers and the audio frames corresponding to each mixture component were listened to. It was observed that the components represented broad phonetic categories corresponding to vowels, nasals, fricatives etc, regardless of the number of distinct speakers in the test utterance.[10] Increasing the number of mixture components to separate phonetic categories of individual speakers also failed. Therefore, a component population histogram would not capture a speaker change. It would only capture those speaker changes that are accompanied by change in audio characteristics in the background. For example, around 80s there is a speaker change from indoor to noisy outdoor in the news audio segment. This change can be observed by looking at the component population histogram for the speech clip in Fig. **??**

The component population histogram can only be useful for speaker change detection only when the individual components model the phonetic sounds of each speaker separately. When a GMM is trained for a particular speaker's identity, the components represent the phonetic sounds of that speaker and hence will give a lower likelihood score for the same utterance of another speaker.[11] However, when a single GMM is trained for all male speakers the meaning of components is not speaker specific anymore but phonetic sound specific for all the speakers.

4. CONCLUSION

We proposed and evaluated the performance of the following two competing approaches for semantic segmentation of news video. The first approach is based on low-level MFCC features and detects speaker changes using the BIC at shot boundaries. The second approach is based on a mid-level feature namely state/component population histogram from a HMM/GMM for a general sound class. The former approach outperformed the latter. In order to understand, the mid-level features better, we investigated the meaning of components/states and found them to be phonetic class specific instead of speaker specific. From our experiments, we conclude that state/component population histograms can only be useful for further clustering or semantic class discovery if the features are chosen carefully so that the individual states represent the semantic categories of interest.

REFERENCES

1. A. Divakaran, R. Radhakrishnan, Z. Xiong and M. Casey, "A procedure for audio-assisted browsing of news video using generalized sound recognition," *Proc. SPIE Conference on Storage and Retrieval for Media Databases* , 2003.
2. A. Hanjalic, G. Kakes, R.L. Lagendijk and J. Biemond, "Dancers: Delft advanced news retrieval system," *Storage and retrieval for media databases, San Jose* , 2001.
3. R. Jasinschi, N. Dimitrova and D. Li, "Integrated multimedia processing for topic segmentation and classification," *Proceedings of ICIP , Thessaloniki, Greece* , pp. 366–369, 2001.
4. Y. Wang,Z. Liu and J.C. Huang, "Multimedia content analysis," *IEEE Signal Processing Magazine, November* , 2000.
5. H. Gish, H-H Siu and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," *Proc. Of ICASSP* , 2001.
6. S. Meignier, J-F Bonastre , C. Fredouille and T. Merlin, "Evolutive hmm for multi-speaker tracking system," *Proc. Of ICASSP* , 2000.
7. S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," *Proc. Broadcast News Trans. and Under. Workshop* , 1998.
8. Z. Xiong, R. Radhakrishnan, A. Divakaran and T.S. Huang, "Audio-based highlights extraction from baseball, golf and soccer games in a unified framework," *Proc. of ICASSP* , 2003.
9. M. Casey, "Mpeg-7 sound-recognition tools," *IEEE Transactions on Circuits and Systems for Video Technology* , 2001.
10. A. Poritz, "Linear predictive hidden markov models and the speech signal," *Proc. of ICASSP* , 1982.
11. D.A. Reynolds and R.C. Rose, "Robust text independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing* , 1995.