ADVANCES IN UNSUPERVISED AUDIO SEGMENTATION FOR THE BROADCAST NEWS AND NGSW CORPORA

Rongqing Huang, John H.L. Hansen

Robust Speech Processing Group, Center for Spoken Language Research
University of Colorado, Boulder, CO, USA
{huangr,jhlh}@cslr.colorado.edu

ABSTRACT

The problem of unsupervised audio segmentation continues to be a challenging research problem which significantly impacts Automatic Speech Recognition (ASR) and Spoken Document Retrieval (SDR) performance. This paper addresses novel advances in audio segmentation for unsupervised multi-speaker change detection. First, we investigate new features which are intended to be more appropriate for segmentation that include:PMVDR (Perceptual Minimum Variance Distortionless Response), SZCR (Smoothed Zero Crossing Rate), and FBLC (FilterBank Log Coefficients); next we consider a new distance metric, T^2-mean which is intended to improve segmentation for short segments (<5s). A novel false alarm compensation procedure is also developed and used after the segmentation phase. We establish a more effective evaluation procedure for segmentation versus the more traditional EER and Frame Accuracy approaches. Employing these advances within our new scheme, results in more than a 30% improvement in segmentation performance using the 3-hour Hub4 Broadcast news 1997 evaluation data. Evaluations are also presented for audio from the NGSW[13] corpus.

1. INTRODUCTION

The goals of effective audio/speaker segmentation are different than those for ASR, and therefore features, processing methods and modeling concepts successful for ASR may not necessarily be appropriate for segmentation. Features used for speech recognition attempt to minimize the differences across speakers and acoustic environments (i.e., *Speaker Variance*), and maximize the differences across phoneme space (i.e., *Phoneme Variance*). However, in speaker segmentation for audio streams, we want to maximize speaker traits to produce segments that contain a single acoustic event or speaker. The traditional MFCC features used for ASR may therefore not be as effective for speaker segmentation. Other studies have considered alternative features. For example, Adami, *et al.* [1] considered LSP features, Lu, *et al.* [9] used a multi-feature set that consisted of the MFCC, LSP, pitch features to detect change points, and then applied the Bayesian fusion model to combine segmentation results. Such approaches can be successful, however a method that employs multi-feature processing may be difficult to use since the *a priori* probability for each feature contribution must be set. Also, the Bayesian fusion model can only be used to accept or reject candidate change points, and cannot reduce the mismatch between experimental and actual change points. In the present study, we consider several novel features (e.g., PMVDR[11], SZCR, FBLC) and combine them instead of fusing.

If speaker segments are longer than 5 seconds, the BIC (Bayesian Information Criterion)[3] and many distance measure based approaches can achieve reliable segmentation performance[6]. However, these methods suffer from insufficient model estimation traits when segment turns are short (i.e., less than 5 seconds). We propose to use a new distance metric, the T^2-mean, to address this problem. A novel false alarm compensation routine is also developed in our segmentation scheme which can compensate the false alarm rate significantly with little cost to changes in the miss rate. Our algorithm is a Compound Segmentation method, so we call it *CompSeg*. Finally, in our experiments, we determine that the traditional segmentation evaluation criteria of EER(Equal Error Rate) and frame accuracy[5] are not appropriate and complete. We therefore propose a new evaluation criterion: Fused Error Score (FES), which encodes information of the average mismatch of the break points for a better overall performance criterion.

2. A NEW EVALUATION CRITERION

The goal of reliable segmentation in audio streams requires that we measure the mismatch between hand/human segmentation and automatic segmentation. Hain, *et al.*[5] applied frame accuracy as the evaluation criterion for speaker segmentation. While this criterion is useful, we believe it

This work was supported by NSF under Cooperative Agreement No.IIS-9817485.

can be misleading. To explain why, consider Fig.1. Here, the experimental break points in (b) are inaccurate, even though the frame accuracy is still high. This is because with pre-classification, you are more likely to get small duration segments inserted, resulting in a "toggling" action between potential speakers. This causes problems for other applications such as model adaptation for ASR, which requires long duration homogeneous segments. So, we feel frame accuracy may not be the best criterion for audio/speaker segmentation. EER (Equal Error Rate) is another popular evaluation criterion. However, in many circumstances, the miss rate is more important than the false alarm rate. Also, the average mismatch between experimental and actual break points is an important norm which reflects break points accuracy for the features and data. We propose a new combined evaluation criterion, similar in principle to WER and accuracy in ASR, that fuses these three terms into an overall score as follows:

$$Fused\ Error\ Score = (False\ Alarm\ Rate_{\%} + 2 * Miss\ Rate_{\%}) * Average\ Mismatch_{ms} \quad (1)$$

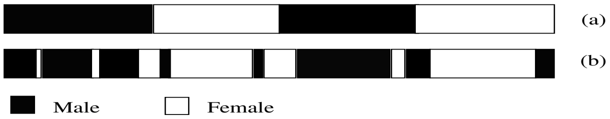

Fig. 1. *An Example of Break Points Detecting of an Audio File. (a) actual break points; (b) experimental break points*

3. THREE NOVEL FEATURES

Having developed a new integrated evaluation criterion, we now turn to features for segmentation. We consider three new features here, and compare them to traditional MFCCs in subsequent evaluations. All features use a 20 ms analysis window with a 10 ms skip frame rate between windows.

3.1. PMVDR

High order MVDR(Minimum Variance Distortionless Response) models provide better upper envelope representations of the short-term speech spectrum than MFCCs[4, 11]. Furthermore, it has been shown that the MVDR spectrum can be simply obtained from a non-iterative computation of LP coefficients[4]. Another trait is that PMVDRs do not require an explicit filterbank analysis of the speech signal. For the application of speaker segmentation, we increase the order of the LP model to reflect more speaker dependent information in the features. We also apply a detailed Bark frequency warping for better results.

3.2. SZCR

A High Zero Crossing Rate Ratio has also been proposed for speaker classification[8]. In our experiments, we find that a smoothed ZCR is more efficient, which is computed as follows: *(i)* we compute 5 sets of ZCR evenly spaced across the analysis window with no intermediate overlap; *(ii)* next we use the mean of the 5 sets as the feature of this frame, which reduces the feature variance and thereby increase the class separability[4].

3.3. FBLC

Although in [11], it was suggested that direct warping of the FFT power spectrum without filterbank processing can preserve almost all the information in the short-term speech spectrum, we find that filterbank processing is more sensitive than other features in detecting speaker change (i.e., the mismatch between the experimental break points and the actual break points is very small). As such, the FBLC are simply the 20 Mel frequency FilterBank Log energies Coefficients.

4. A NEW SEGMENTATION DISTANCE METRIC

If the segments are more than 5 seconds long, BIC and other distance metric based methods perform segmentation very well[3, 6]. However in real audio data from Broadcast News or two-way conversations, many segments are very short (i.e., less than 5 seconds). Since BIC and most distance metric based methods need the second statistics (i.e., the covariance), they often suffer in estimation error due to insufficient data.

The Kullback Leibler distance (KL2) is a popular distance metric in speaker segmentation[10]. Fig.2(a) shows the KL2 distance of a 35 seconds audio stream which has only one real break point at 19 seconds. From the figure, we find that the KL2 distance measure of the first and final 5 seconds are not correct. This occurs because of insufficient data in the estimation of the covariance when the segment is shorter than 5 seconds. In contrast, we see in Fig.2(b) that the T^2 distance measure detects the break point accurately with no initial or trailing edge effects.

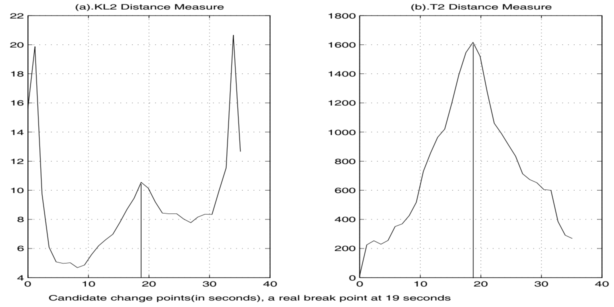

Fig. 2. *KL2 and T^2 distance of an audio window, which has a real break point at 19 seconds*

The idea of using the Hotelling T^2-Statistic[2, 12] for speaker segmentation is that: for two audio segments, if they can be modeled by multivariate Gaussian distributions: $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$, we assume their covariances are equal but unknown, then the only difference between them is the mean values reflected in the T^2 distance as:

$$T^2 = \frac{ab}{a+b}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) \qquad (2)$$

where a, b are the number of frames within each of the audio segments respectively. Under the equal covariance assumption, we can use more data to estimate the covariance and reduce the impact of insufficient data in the estimation. That is why the T^2 distance measure can detect the break point accurately in Fig.2(b). If the processing audio window is shorter than 2 seconds, even a global covariance will suffer from insufficient estimation. We can then further assume the global covariance to be an identity matrix, in which case we call this the weighted Mean Distance. Fig.3 clearly shows that if there is a break point in the processing window, the distance measure has *one and only one prominent peak*. Therefore, the T^2-Mean can be used to detect the break point in the short processing window (<5s) efficiently. As the window grows in duration, the covariance can be estimated more accurately, and we can then apply BIC to detect the break points directly as in [3].

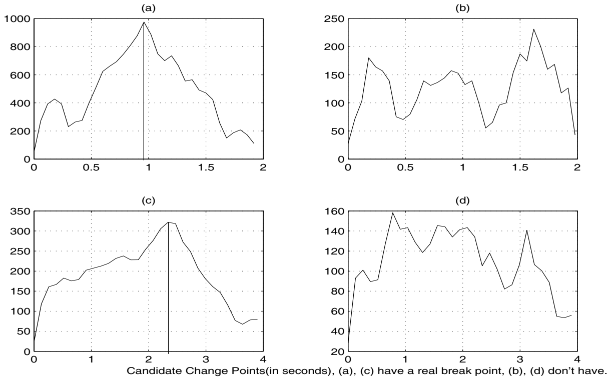

Fig. 3. *T^2-Mean distance of processing audio windows, x-axis is the window length (in seconds), y-axis is the distance.(a), (b) is the weighted Mean Distance; (c), (d) is the T^2 distance. (a),(c) have one real break point at 0.9 second, 2.4 second respectively, (b),(d) have no true break points*

5. FALSE ALARM COMPENSATION

5.1. Audio Clustering

It is common that speech from the same speaker might appear multiple times in an audio stream. In general, it would be useful to pool the homogeneous data from the same speaker for subsequent processing (e.g., speaker adaptation, speaker identification, etc.). The application of traditional BIC for a hierarchical clustering is straightforward[3]. Our clustering is implemented in a bottom-up framework, where each segment is a node, where we calculate the distance between all nodes, and apply BIC to examine node pairs with the nearest distance if they can be merged. If they are homogeneous, we merge them to a single new node and the distance matrix is re-calculated. If they are not homogeneous, then consider the second nearest node pair. This procedure continues until all nodes have been examined.

5.2. False Alarm Compensation

If the two mergable nodes are adjacent segments in the audio clustering routine, it means we can compensate the false alarm rate. However, a false alarm compensation method based on clustering is not powerful, because it cannot compensate the false alarm caused by short segments due to reduced data size. Conceptually, the false alarm compensation is similar to classification. Here, we calculate the distance between two adjacent segments, and if the distance is below a threshold, then they belong to the same class, (i.e., we find a false alarm break point), otherwise they are from different classes. We apply the T^2-mean as the distance metric for short segments and regular T^2 with covariance matrix estimation for long segments. This scheme can compensate the false alarm rate significantly with little cost to the miss rate. The feature used in false alarm compensation can be different than that used in segmentation. Here, we use the first half of the PCA(Principal Components Analysis[7]) projections of the combined feature: PMVDR+FBLC. The PCA projection has more discriminative power than the original feature, so it is suitable in this classification-like task.

6. EXPERIMENTS

For our experiments, the evaluation data is drawn from broadcast news Hub4 1996 training data, Hub4 97 evaluation data and NGSW data[13]. We evaluate the advances in CompSeg one by one.

6.1. Feature Evaluation

Table 1 shows that PMVDR can outperform MFCC on all levels. FBLCs have very small average mismatch implying they are very sensitive to the changes between speakers and environments. Because PMVDR does not apply filterbank processing, we combine PMVDR and FBLC together. Also, the SZCR encodes information directly from the waveform which we combine as well. We did consider other prosodic features such as pitch, but the results did not improve. We believe this is because pitch only encodes information from voiced speech, and does not contain information from unvoiced speech and noise, making it less effective for segmentation. We select the 24 features from PMVDR, all 20 features from FBLC, and 1 SZCR(i.e., a 45-dimensional set). We normalize the features to zero mean and unit variance for improved discrimination ability.

6.2. T^2-Mean & False Alarm Compensation Evaluation

The result of the new segmentation T^2-Mean is shown in Table 2, where a 24-dimensional PMVDR feature set is used in both baseline and T^2-Mean segmentation. The baseline system in this study uses BIC only. With this advance, there is a 2.2% absolute improvement in the miss rate, with 2.0%

coming from the short segments. This suggests that the contribution of T^2-Mean is mainly on the short duration turn detection.

In order to apply the proposed false alarm compensation routine, the initial segmentation is set to find all possible break points regardless of the false alarm rate. Table 3 shows that the false alarm compensation scheme is very efficient. In the segmentation stage, the feature is a 24-dimensional MFCC set. In the false alarm compensation stage, the feature is the first half of the PCA projections of the combined feature: PMVDR+FBLC. The threshold is determined from one hour of broadcast news development data. The baseline system uses MFCCs with traditional BIC.

6.3. DARPA Hub4 Evaluation

The DARPA Hub4 1997 Evaluation Data was used for performance assessment. The set contains 3 hours of Broadcast News data, with 584 break points, including 178 short segments(<5s). CompSeg uses the PMVDR, SZCR, FBLC features, applies T^2-Mean measure for segments less than 5 seconds, and applies the novel False Alarm Compensation post-processing routine. The improvement using these advances is shown in Table 4. We see that for all metrics, performance improves significantly on the Hub4 data. The baseline system uses MFCCs and traditional BIC only.

6.4. NGSW Data Evaluation

We also evaluate the CompSeg algorithm with a portion of the NGSW corpus[13], using audio material from the 1960s. From Table 5, we see that CompSeg can detect not only the speaker changes, but also the music and long silence(>2s) segments.

Table 1. *Feature Evaluation. '( )' is the relative improvement, FA: False Alarm Rate(%); MIS: Miss Detection Rate(%); MMatch: Average Mismatch(msec); FES: Fused Error Score. Same as in Table 2,3,4*

Feature	FA	MIS	MMatch	FES
MFCC	29.6%	25.0%	298.47	237.58
FBLC	29.8% (-0.7%)	25.3% (-1.2%)	266.80 (10.6%)	214.51 (9.7%)
PMVDR	25.9% (12.5%)	24.9% (0.4%)	284.29 (4.8%)	215.21 (9.4%)
Combine 45-D	23.8% (19.6%)	24.3% (2.8%)	265.06 (11.2%)	191.99 (19.2%)

Table 2. *Evaluation of T^2-Mean Segmentation*

Scheme	FA	MIS	MMatch	FES
Baseline	27.6%	27.4%	277.50	228.56
T^2**-Mean**	23.5% (14.9%)	25.2% (8.0%)	281.21 (-1.3%)	207.53 (9.2%)

Table 3. *Evaluation of False Alarm Compensation Scheme*

Scheme	FA	MIS	MMatch	FES
Baseline	44.2%	18.7%	307.83	250.90
FA-COMP	23.8% (23.5%)	21.3% (-13.9%)	292.28 (5.1%)	194.07 (22.7%)

Table 4. *Evaluation of CompSeg with Hub4-97 Evaluation Data*

Algorithm	FA	MIS	MMatch	FES
Baseline	26.7%	26.9%	293.02	235.82
CompSeg	21.1% **(21.0%)**	20.6% **(23.4%)**	262.99 **(10.2%)**	163.84 **(30.5%)**

Table 5. *Evaluation of the NGSW Data*

Speaker Change	Speaker MMatch	Music & Sil Change	Music & Sil MMatch	False Alarm
100%	129ms	100%	118ms	5.6%

7. CONCLUSION AND FUTURE WORK

We have shown that a systematic set of advances integrated into our new algorithm, *CompSeg*, can achieve efficient unsupervised audio segmentation, especially for short duration segments. We achieved more than a 30% improvement over a traditional BIC with MFCC based segmentation algorithm. In the future, we plan to develop more advanced features for audio segmentation which can maximize the speaker variance and minimize the phoneme variance simultaneously. We also plan to apply CompSeg in our Spoken Document Retrieval project[13].

8. REFERENCES

[1] A.Adami, S.Kajarekar, H.Hermansky, "A New Speaker Change Detection Method for Two-Speaker Segmentation", *ICASSP 2002*, Orlando, USA, 2002

[2] T.Anderson, "An Introduction to Multivariate Statistical Analysis", *John Wiley and Sons. Inc.*, New York, USA, 1958

[3] S. Chen, P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via The Bayesian Information Criterion", *Proc. Broadcast News Transcr. & Under. Workshop*, 1998

[4] S.Dharanipragada, B.Rao, "MVDR-Based Feature Extraction for Robust Speech Recognition", *ICASSP 2001*, Salt Lake City, Utah, USA, 2001

[5] T.Hain, S.Johnson, A.Tuerk, P.Woodland, S.Young, "Segment Generation and Clustering in the HTK: Broadcast News Transcription System", *DARPA Broadcast News Transcr. & Under. Workshop*, Lansdowne, VA, USA, 1998

[6] S.Johnson, "Speaker Tracking", *Master Thesis*, Engineering Department,Cambridge University, UK, 1997

[7] I. T. Jolliffe. "Principal Component Analysis", *Springer-Verlag*, New York, 1986

[8] L.Lu, H.Zhang, H.Jiang, "Content Analysis for Audio Classfication and Segmentation", *IEEE Trans. Speech & Audio Processing*, pp. 504-516, Vol.10, No.7, 2002

[9] L.Lu, H.Zhang, "Speaker Change Detection and Tracking in Real-Time News Broadcasting Analysis", *ACM Multimedia 2002*, France, Dec., 2002

[10] M.Siegler, U.Jain, B.Raj, R.M.Stern, "Automatic Segmentation, Classification and Clustering of Broadcast News Audio", *DARPA Speech Recognition Workshop*, Chantilly, Virginia, USA, pp. 97-99, 1997

[11] U.Yapanel, J.H.L.Hansen, "A New perspective on Feature Extraction for Robust In-Vehicle Speech Recognition", *EuroSpeech 2003*, Geneva, Sep., 2003

[12] B.Zhou, J.H.L.Hansen, "Unsupervised Audio Stream Segmentation and Clustering Via the Baysian Information Criterion", *ICSLP 2000*, vol.1, pp.714-717, Beijing, China, Oct., 2000

[13] National Gallery of the Spoken Word (NGSW), NSF DLI-2 Project. http://www.ngsw.org; http://speechfind.colorado.edu