INTERSPEECH 2009 BRIGHTON

Speaker Segmentation and Clustering for Simultaneously Presented Speech

Lingyun Gu[1] *and Richard M. Stern*[1,2]

[1]Language Technologies Institute
[1,2]Department of Electrical and Computer Engineering
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA, 15213, U.S.A.
{lgu, rms}@cs.cmu.edu

Abstract

This paper proposes a new scheme used to segment and cluster speech segments on an unsupervised basis in cases where multiple speakers are presented simultaneously at different SNRs. The new elements in our work are in the development of new feature for segmenting and clustering simultaneously-presented speech, the procedure for identifying a candidate set of possible speaker-change points, and the use of pair-wise cross-segment distance distributions to cluster segments by speaker. The proposed system is evaluated in terms of the F measure that is obtained. The system is compared to a baseline system that uses MFCC for acoustic features, the Bayesian Information Criterion (BIC) for detecting speaker-change points, and the Kullback-Leibler distance for clustering the segments. Experimental indicate that the new system consistently provides better performance than the baseline system with very small computational cost.[1]

Index Terms: speech segmentation, speaker clustering, feature extraction

1. Introduction

Speaker change detection and clustering are very important in many applications such as adapting speaker models online, improved speaker identification, etc. Several steps are needed for satisfactory results, including good feature extraction, efficient distance measurement for change-point detection, sound ways to cluster different segments into groups and reasonable evaluation systems to assess algorithm performance.

The Broadcast News problem has received a great deal of attention for a number of years, and has been the object of many speech segmentation and clustering approaches. Log Likelihood Ratio (LLR) [1] and its improved implementation, Bayesian Information Criterion (BIC) [2] have been widely used to find speaker switch boundaries, and the procedures developed in this paper will be compared to these approaches. Equations 1, 2, and 3 show how these three indicators are calculated.

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

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

[1]This research was supported by NSF Grant IIS-0420866.

$$d_{BIC} = L_1 - L_0 - \frac{\lambda}{2}\Delta K \log N \quad (3)$$

In the equations above N_x and N_y are the number of frames in two consecutive segments. θ_z is the distribution parameter for two consecutive segments merged together, while θ_x and θ_y are distribution parameters for each of the individual segments. ΔK is the difference between the number of parameters from Eq. 2 and the number of parameters from Eq. 1. While the parameter λ is commonly set to 1, the robustness of BIC in a variety of acoustical environments is largely dependent on the specific value that is chosen for λ. Other researchers [3] used Gaussian mixture models to replace the single Gaussian distribution in Eqs. 1 and 2 and achieved better results. In searching for speaker-change points, sliding windows are commonly used to obtain means and covariances from the above equations. But if the utterances are brief there will be only a small number of points that can be used.

Some algorithms (*e.g.*) [4] made an assumption that speaker-change points are very likely to occur in silence regions, which may not be the case when speech is highly overlapping. Another system [5] utilized the Generalized Likelihood Ratio (GLR) with good results. However, this system requires at least one second of clean speech from a target speaker before processing begins.

Clustering is usually accomplished by applying one of several distance measures to calculate how far every segment is away from others. Segments having smaller distance will be grouped together. The Kullback-Leibler distance [6] and GLR [5] have been utilized, along with other measures.

While these systems achieve a level of good performance, they have drawbacks. In general, they can perform reasonably well when each individual speaker speaks for a relatively long duration, but in many other cases obtaining a long segment of speech from one speaker may not be possible. Many of these results were obtained with relatively clean speech, and performance will degrade as the acoustical environment becomes less favorable.

This paper describes a new algorithm that accomplishes speech separation. With any system based on computational auditory scene analysis (CASA) there are two stages, segregation and grouping. Segregation refers to the categorization of frequency components into different groups corresponding to different speakers using combinations of intrinsic acoustic cues. Regrouping refers to the clustering of these segregated frequency components into different streams from which speech is reconstructed. This is usually accomplished using speaker

identification (SID) on speakers who are known to the system *a priori*, and efficient and accurate algorithm that segment and cluster speech for unknown speakers are generally not available. This paper considers the problem of separating segments of speech that are only 2-3 seconds long, with two speakers are almost completely overlapping.

2. System overview

Figure 1 shows a system block diagram of the proposed system. The combined speech is first decomposed into a two-dimensional time-frequency representation by applying short-time Fourier analysis (STFA). All time-frequency cells are then sorted according to power, retaining only the cells within each utterance that contained the upper 20^{th} percentile of power. (This threshold was empirically determined to be best for this purpose.) The 20% of the time-frequency cells that are retained constitute a new spectrographic representation. Frame-based features representing local power are derived from this representation and subjected to median smoothing to reduce fluctuations. A set of possible speaker-change points is determined by searching for local minima of power in time and frequency as will be discussed below.

Other features are extracted in parallel and subjected to feature extraction (analyzing attributes such as pitch, kurtosis, and zero crossings) on a frame-by-frame basis. Euclidean distance between consecutive frames is calculated using these features, which provides a parallel mechanism for identifying potential speaker-change points. These feature streams are merged at a later stage. As a result of these analyses the speech is (ideally) separated into segments belonging to different dominant speakers. In this paper we consider only the case of two simultaneously-presented speech sources.

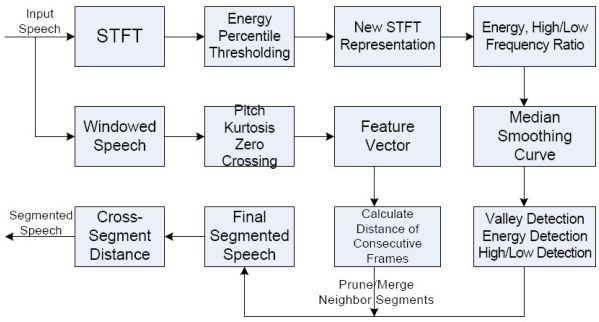

Figure 1: System block diagram.

In the next section, we discuss the development of new features that can process highly-overlapped speech sources, along with more conventional features. Section 4 describes a two-stage pruning segmentation procedure that produces a pool of speaker-change candidates by combining the best attributes of the new features and the conventional features. Clustering is discussed in Sec. 5. Section 6 describes the evaluation criteria along with experimental results, and Sec.7 summarizes our conclusions.

3. Feature extraction: new features based on energy percentiles

Many conventional segmentation systems use features that are already in use in speech recognition systems such as MFCC, PLP, LPC or LSP coefficients. Unfortunately, eatures of this type are not effective for segregating highly-overlapped or simultaneously-presented speech sources. In this section a new feature extraction scheme is proposed that addresses this challenge.

Our feature extraction technique begins with a conventional short-time Fourier transform (STFT) of the incoming speech. We focus on the strongest segments of the combined speech by selecting only those time-frequency cells that are among the top 20% in local power. Figure 2 shows such a STFT representation after this thresholding. It is clear to see that the filtered STFA preserves the clear structure of low-frequency harmonics combined with high-frequency unvoiced energy.

Let $M[n,k]$ refer to the binary mask that reflects whether or not a particular STFT coefficient $X[n,k]$ is among the top 20% in power. Specifically, $M[n,k] = 1$ if a particular frame n and frequency k is in the top 20% in power, and $M[n,k] = 0$ otherwise.

Two features are derived from this representation as follows:

$$S(n) = \sum_{k=1}^{K} M[n,k] \tag{4}$$

$$Ratio(n) = \frac{\sum_{k=\frac{K}{2}+1}^{K} M[n,k]}{\sum_{k=1}^{\frac{K}{2}} M[n,k]} \tag{5}$$

Equation 4 calculates how many high-energy time-frequency cells there are in a particular frame, while Eq. 5 calculates the ratio of the number of retained cells in the high-frequency versus low-frequency regions. K is half the size of the FFT.

In addition to these two features derived from modified time-frequency representation, several other features are derived directly from the original speech in the time domain, frame by frame, such as pitch, kurtosis and zero crossing rate (ZCR). Pitch is calculated from autocorrelation in the time domain, and it is extremely useful for separating two speakers of different genders. Kurtosis measures the peakedness for any given pdf, and it decreases as the number of simultaneous speakers increases. ZCR is a good way to detect unvoiced segments because they usually have high ZCR values. Kurtosis and ZCR are calculated as follows:

$$Kur = \frac{\mu_4}{\sigma^4} - 3 \tag{6}$$

$$ZCR = \frac{1}{T} \sum_{t=1}^{T} \frac{|sgn(x(t)) - sgn(x(t+1))|}{2} \tag{7}$$

where in Eq. 6, μ_4 is the 4th central moment and σ is the standard deviation of the speech signal in a given frame. In Eq. 7, T is the total number of speech samples in a given frame, and sgn is the signum function. As noted above, these features were selected because they provided better performance for this task than other similar features.

4. Segmentation: two-stage filtering

Many other algorithms in this field utilize the Log Likelihood Ratio (LLR), Bayesian Information Criterion (BIC), Kullback-

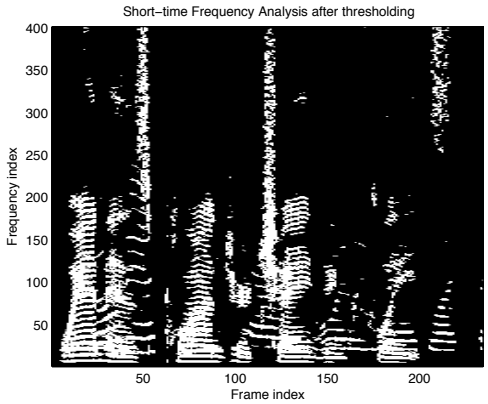

Figure 2: Short-time Fourier Analysis after 20% energy percentile thresholding.

Leibler distance (KL) as well as the Generalized Likelihood Ratio (GLR) to calculate distance in order to find speaker change points (*e.g.* [7]). Nevertheless, all the distance measures mentioned above need relatively long segments of incoming speech to perform well (sometimes as much as several seconds with more than 100 ms of overlap). In our application, the utterances are simply not long enough to accommodate this requirement.

A new two-stage filtering procedure is introduced to address the problems mentioned above. Equation 4 is used to generate a frame-based "energy" curve for the entire processed utterance. The word "energy" has been put quotes here to differentiate it from the real power curve that was discussed in Sec. 3. A reasonable assumption is that most possible speaker-change points should occur in either silent or low-power regions or valleys in the "energy" curve. By using thresholding, certain low-power regions can be highlighted for better detection. Therefore, silence and valley detection have been applied to the "energy" curve to identify these potential change points. In addition, a dramatic change in the high/low frequency ratio (Eq. 5) also suggests a possible speaker switch point. A pool of potential speaker-change points is obtained by combining all change candidates obtained by all of the above methods.

In the second pass of processing, all three features (pitch, kurtosis and ZCR) are combined into a single vector. Unlike the traditional cepstral and spectral-based features mentioned above, kurtosis and ZCR continue to perform well in speech-on-speech situations. The conventional Euclidean distance is calculated for every pair of consecutive frames. Because this statistic used by itself will fluctuate as the local power of speech from the dominant speaker fluctuates, false alarms produced for this reason are eliminated when they correlate too much with local power fluctuations.

By performing this two-stage processing, the combined speech can be segmented into several speech segments either with extremely low energy or each segment only contains one dominant speaker. It is worth noting that even though a dominant speaker may be present in a specific segment, there is a high chance that a simultaneously-presented weaker speaker could be dominant during a particular segment, due to fluctuations in power in the signals.

5. Clustering: cross-segment distance calculation

We used the procedures described in Sec. 4 to apply a distance measure to calculate the distance among all segments. We also use a distance measure that is commonly used to calculate distance among Gaussian distributions. Specifically, consider the multi-variate Gaussian pdf

$$p(X) = \frac{1}{(2\pi)^{d/2}|C|^{1/2}} \exp -\frac{1}{2}(X-\mu)^T C^{-1}(X-\mu) \quad (8)$$

where X is the feature vector, C is the covariance matrix, μ is the mean vector and d is the length of feature vector X. The distance measure can be expressed by the equation:

$$D(i,j) = \int_X |p_i(X) - p_j(X)| \ln \frac{p_i(X)}{p_j(X)} dX \quad (9)$$

To further simplify the above equation, the final distance can be given by

$$D(i,j) = \frac{1}{2} tr[(C_i - C_j)(C_j^{-1} - C_i^{-1})] + \frac{1}{2} tr[(C_j^{-1} + C_i^{-1})(\mu_i - \mu_j)(\mu_i - \mu_j)^T] \quad (10)$$

where i and j represent the Gaussian distributions corresponding to each segment.

To determine how many segments are believed to belong to the same speaker, every pair-wise cross-segment distance is calculated. Those segments whose distances are close to each other are categorized as belonging to the same speaker, while other segments are associated with the other speaker according to the *a priori* assumption of only two speakers.

6. Evaluation criterion and experimental results

6.1. Evaluation criterion

The F value, which combines precision and recall, is used to determine system performance. We first compare the proposed system with a baseline system that uses MFCC as its acoustic features, BIC to detect segments and speaker change points, and the Kullback Leibler (KL) distance [7] to cluster segments. The standard precision, recall, F value, and KL distance formulae are

$$precision = \frac{\#\ of\ correctly\ labeled\ items}{\#\ of\ total\ labeled\ items} \quad (11)$$

$$recall = \frac{\#\ of\ correctly\ labeled\ items}{\#\ of\ total\ correct\ items} \quad (12)$$

$$F = \frac{2 \times precision \times recall}{precision + recall} \quad (13)$$

$$KL2(X,Y) = \frac{1}{2}(\mu_Y - \mu_X)^T (C_X^{-1} + C_Y^{-1})(\mu_Y - \mu_X) + \frac{1}{2} tr((C_X^{1/2} C_Y^{-1/2})(C_X^{1/2} C_Y^{-1/2})^T) + \frac{1}{2} tr((C_X^{-1/2} C_Y^{1/2})(C_X^{-1/2} C_Y^{1/2})^T) - d \quad (14)$$

where μ and C are the mean vector and covariance matrix from each distribution and d is the vector length.

During the segmentation evaluation procedure, all the speaker-change points identified by the system are compared with the ground truth, which is determined automatically on a frame-by-frame basis using oracle information about the local SNR. (This is done without any manual labeling of ground truth.) In our evaluations we omit frames for which the local frame-based SNR lies between -2 dB and $+2$ dB because it is not meaningful to draw an inference about ground truth in frames for which the two speakers have almost the same energy. We also omit during frames of near silence for similar reasons.

6.2. Experimental results

While standard DARPA/NIST databases such as Resource Management or Broadcast News are widely used to test segmentation and clustering performance due to the availability of hand-labeled segmentation information, these databases do not contain many segments of simultaneously-presented speech. For this reason we evaluated our algorithm on a database in which 1600 pairs of standard DARPA Resource Management (RM) sentences were digitally added at SNRs of 5 dB, 10 dB, and 15 dB. For each SNR, four different combination are generated: male-male, female-female, male-female and female-male, where the first gender in each pair is the globally-dominant speaker (who is not necessarily the dominant speaker in every frame). The vocabulary size of the RM database is nominally 1000 words.

	5 dB	10 dB	15 dB
F value(new system)	0.72	0.76	0.85
F value(conventional system)	0.69	0.73	0.80

Table 1: Segmentation performance comparison at different SNRs

	5 dB	10 dB	15 dB
F value(new system)	0.82	0.86	0.80
F value(conventional system)	0.70	0.76	0.79

Table 2: Clustering performance comparison at different SNRs

Segmentation performance is given in Table 1. The table shows the new system demonstrates better performance than the conventional system. We noted a greater number of insertion than deletion errors in the new system, while conventional system does not exhibit this bias. It is possible that the newer features are sensitive to any possible changes. While capturing more true-change points, the system also pays the price of introducing more false alarms.

Table 2 shows that the clustering performance of the new system also consistently outperforms the baseline system at all SNR levels. But it is interesting to note that the performance of the new system does not increase linearly with SNR, while the conventional system does. A possible reason is that the new features do not match the characteristics of the clean environment very well, perhaps because the the energy-percentile thresholding may throw away some useful information. While it is widely believed that segmentation error can adversely affect clustering performance, this does not always happen in our observations. This possibly happens because segmentation errors frequently occur during frames in which target an masker strength are very close, which are excluded from analysis in this study.

Table 3 breaks out these results according to speaker gender and indicates (unsurprisingly) that separation is more accurate when the two speakers are of different genders, which produce different pitch contours and less overlap of harmonic structures, among other things. All of the differences in results described above are statistically significant at the $p = .05$ level except for the 15-dB case in Table 2.

	5 dB	10 dB	15 dB
F value (same gender)	0.79	0.83	0.78
F value (different gender)	0.86	0.88	0.83

Table 3: Clustering performance comparison at gender-dependent groups by using new system

7. Conclusions

We describe a new segmentation and clustering scheme for speech that is based on new features implemented in a two-stage procedure that detects and clusters speaker-change points. The system was tested in using simultaneously-presented speech samples at various SNRs and evaluated using the F measure. Using databases derived from the DARPA Resource Management corpus, good performance was demonstrated compared to a baseline system that used MFCC features, BIC, and a KL distance metric. The new system shows consistently better performance over several SNR levels and confirms the expectation that speech segmentation and clustering is more challenging when the two simultaneous speakers are of the same gender.

8. References

[1] H. Gish, M.H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *Proceedings of ICASSP 1991*, 1991, pp. 873–876.

[2] A. Tritschler and R. Gopinath, "Improved spaker segmentation and segments clustering using the bayesian information criterion," in *Proceedings of Eurospeech 1999*, 1999, pp. 679–682.

[3] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *IEEE Signal Processing Letters*, vol. 11, pp. 649–651, 2004.

[4] D.A. Reynolds, R.B. Dunn, and J.J. McLaughlin, "The lincoln speaker recognition system: Nist eval2000," in *Proceeding of ICSLP 2000*, 2000, pp. 470–473.

[5] A. G. Adami, S.S. Kajarekar, and H. Hermansky, "A robust speaker change detection method for two-speaker segmentation," in *Proceeding of ICASSP 2002*, 2002, pp. 3908–3911.

[6] M.A. Siegler, U. Jain, B. Raj, and R.M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *DARPA Speech Recognition Workshop 1997*, 1997, pp. 97–99.

[7] P. Delacourt and C.J. Wellekens, "Distbic: A speaker-based segmentation for audio data indexing," *speech communication*, vol. 32, pp. 111–126, 2000.