Audio Segmentation and Speaker Localization in Meeting Videos

Himanshu Vajaria¹, Tanmoy Islam², Sudeep Sarkar¹, Ravi Sankar², Ranga Kasturi¹

¹Dept. of Computer Science & Engineering, ²Dept. of Electrical Engineering

University of South Florida, Tampa, FL, 33620, USA

{hvajaria,sarkar,r1k}@cse.usf.edu, {tislam,sankar}@eng.usf.edu

Abstract

Segmenting different individuals in a group meeting and their speech is an important first step for various tasks such as meeting transcription, automatic camera panning, multimedia retrieval and monologue detection. In this effort, given a meeting room video, we attempt to segment individual person's speech and localize them in the video, based on data from a single audio and video source. The segmentation method is driven by audio and enhanced by video cues. We used Bayesian Information Criterion (BIC) to segment the feature vector streams and graph spectral partitioning to cluster them. We compare our results with audio based segmentation method and our localization technique with the commonly used mutual information.

1. Introduction

In this effort, we analyze meeting room video clips where multiple people are present, conversing with each other. Our goal is to locate individual speakers in the image sequence (source localization) and segment their speech (speaker diarization). The objectives are similar to those of the NIST evaluation [4]. However, instead of multiple cameras and microphones used in the NIST evaluation, our recording setup is simple, consisting of a single camera and a single microphone. The benefit of this simple setup is that it is easily portable - in essence any laptop with a microphone and camera can be used to record the meeting.

The single camera, single microphone scenario has been investigated in [3, 5, 6, 7, 8]. Here, the problem is posed as one of detecting synchronous audio and video events. Usually in these scenarios, speakers explicitly address the camera. Thus a speaker's face and lips are clearly visible so that when speech is heard, there is a corresponding source in the video. Mutual information (MI) based approaches have been successfully demonstrated in these works.

In the meeting room scenario, the camera usually focusses on the torso or even the entire body of the participants. Participants address each other rather than the camera. Also, having only a single view puts severe constraints on feature extraction. Non-frontal faces make face detec-

Figure 1. System Flowchart: Audio and Video features are fused. Clustering is performed after change point detection and the speaker is localized within the clustered frames

tion a non-trivial task and detecting lip movements is even harder - if at all possible. Thus a system based solely on features from face and lip movements, will not be robust. Secondly, the “synchronization” hypothesis does not hold as speech may be detected without any corresponding video motion when the speaker is not facing the camera. For this reason, as we show later, the mutual information (MI) based approach does not perform well on our dataset.

Since speech and gestures are co-expressive [9], we assume that on an average a speaker has more movement than a listener. Thus in this work, we use video cues from not only lip and face movements but also from body movements to segment the continuous speech signal. In general, robust audio-visual person segmentation is difficult when multiple persons are present in a scene and not all faces are frontal or even in profile view. Here we present a framework where audio and video information is fused for robust segmentation of individuals in both - the audio and video sense.

The flowchart in Figure 1 describes our technique. The audio and video features are combined to form a fused feature vector. The BIC is used to detect change points in the

feature vector streams (audio, video and fused). Graph spectral partitioning is then used to cluster the segments. We also use the changepoints detected in the fused vector stream to cluster the audio vector stream.

2. Features

The video clips were captured using a Canon Optura videocamera. The video frame rate is 30 Hz at 720 x 480 resolution. The audio signal is captured at 44 kHz and was processed to extract 23 Mel-Frequency Cepstral Coefficients (MFCC) at a 30 Hz rate. This was done to bring the temporal dimensionality of the audio to the same rate as the video. The MFCC are then projected into principal component analysis (PCA) space to obtain a 23 dimensional feature vector corresponding to each video frame.

In previous work, dealing with audio-visual synchrony, various features extraction schemes have been used, such as those based on optic flow [3], image differences [7, 8], hand [9] and face detection [3]. Here, we use the grey-scale difference image as it is computationally less expensive and performs better than optic flow on our data where the inter-frame displacements are high [9]. We downsample the difference image by a factor of 20. The image dimensions are further reduced to 23 by PCA to correspond with the dimensionality of the audio features. Once a reduced representation of the audio and video features is obtained, the features are concatenated to obtain a fused feature vector F ,

$$F = [A^T V^T]^T \quad (1)$$

where A and V are the audio and video vectors respectively. Before concatenation, the individual vectors are scaled so that the sum of their eigenvalues is the same. The fused feature vectors are then projected into another PCA space.

3. Segmentation

Segmenting speakers from a continuous speech signal has been studied extensively in the last few years. Among the more promising approaches, is the use of the Bayesian Information Criterion (BIC) first suggested in [1]. It was shown in [2] that this method outperforms segmentation methods based on the KL2 and Generalized likelihood ratio based approaches. In [1], the authors demonstrated how the BIC could be used to detect audio changepoints. Various modifications were suggested to improve speed and robustness in [2, 11]. In these works, the goal is to segment an audio stream into individual clusters - one for each speaker.

The BIC is a model selection criterion; given a set of features, $X = (X_1, X_2 \dots X_N)$, it selects the model that maximizes the likelihood of the data. Since the likelihood increases with the number of model parameters, a penalty term proportional to the number of parameters 'd' is introduced to favor simpler models. The BIC for a model M with parameters θ is defined as

$$BIC(M) = \log pr(X|\theta) - \frac{1}{2} \lambda d \log N \quad (2)$$

where $\lambda=1$ and N is the number of feature vectors.

The task of changepoint detection is cast into the model selection framework. We want to determine whether the entire feature set X is better represented by a single model M_C or whether there exist two models M_1 and M_2 that can better represent feature sets $X_{W_1} = (X_1, X_2 \dots X_i)$ and $X_{W_2} = (X_{i+1} X_{i+2} \dots X_N)$ respectively.

Defining ΔBIC to be $BIC(M_1; M_2) - BIC(M_C)$, we have

$$\Delta BIC(i) = \frac{i}{2} \log |\Sigma_{X_{W_1}}| + \frac{N-i}{2} \log |\Sigma_{X_{W_2}}| - \frac{N}{2} \log |\Sigma_X| - \frac{i}{2} \lambda \left(d + \frac{d+1}{2} \right) \log N \quad (3)$$

Here, $|\Sigma_X|$ is the determinant of the covariance of the entire feature set and $|\Sigma_{X_{W_1}}|$, $|\Sigma_{X_{W_2}}|$ are the determinant of the covariance of the two subdivisions of the feature sets. A positive $\Delta BIC(i)$ indicates that the data is better represented by two models. The frame 'i' for which $\Delta BIC(i)$ is maximum represents the best split of the feature set.

We modified the technique in [1] to increase the speed. The sliding window is split into three equal parts and the BIC is computed only for the middle sub-window. The window is then slid by 1/3rd of its length and the process continues. This improves robustness of the BIC computation. The BIC curve was then low pass filtered and salient peaks were extracted. A peak is considered salient if it exceeds both its neighboring minima by $\alpha\sigma$ [2], where σ is the standard deviation of the BIC curve and α is a constant.

4. Clustering

Audio segments corresponding to individual speakers are clustered together using graph spectral clustering [10], which recursively bipartitions each cluster until a stopping criterion is met. Since partitioning is done in a top down manner, clusters at the upper levels are larger and have low purity while clusters at lower levels are smaller and have higher purity. To select the clusters, we first need to select the ideal level. This is done by computing a tightness criterion for each level using equation (4)

$$L_T = \frac{1}{N_C} \sum_{i=1}^{N_C} C_T(i) \quad (4)$$

where N_C is the number of clusters in the level and C_T is the tightness of the cluster computed using equation (5)

$$C_T = \frac{1}{N_S^2 - N_S} \sum BIC(S_i, S_j) \quad \forall S_i, S_j \subset C, i \neq j \quad (5)$$

where S_i, S_j are segments belonging to the cluster and N_S is the total number of segments in the cluster. We select the uppermost level which exceeds a tightness threshold and use its clusters.

5. Localization

Once clusters for individual speakers are obtained, the next step is to localize the speaker in the corresponding

video frames. For this, we use PCA on the video feature vector to identify regions of correlated motion. The first eigenvector (with the largest eigenvalue) finds the dominant motion modes. This can be used to identify portions of the images with motion. The subsequent eigenvectors can be used to segment these portions into individual persons. For instance, for a scene with two persons, the first eigenvector will have non-zero components corresponding to the moving parts of the image, in addition the sign of the eigenvector component can be used to further divide the moving portions into two parts. However, we cannot identify which of these parts to select based on just the first eigenvector. The second eigenvector which captures the next dominant mode of motion correlation and is orthogonal to the first eigenvector can be used to identify the portion from the dominant speaker. Mathematically, if $\mathbf{v}_1$ is the largest eigenvector and $\mathbf{v}_2$ is the second largest eigenvector, then the selected region is given by

$$\arg \max_{+, -} \{ |\mathbf{v}_2^T | \mathbf{v}_1^+ |, |\mathbf{v}_2^T | \mathbf{v}_1^- | \} \quad (6)$$

where $\mathbf{v}_1^+$ and $\mathbf{v}_1^-$ are the positive and negative parts of the eigenvector, denoting the two possible motion blobs, and T is the transpose operator.

We also implemented the mutual information (MI) approach [6] for localization as a baseline. This technique finds image pixels whose intensities strongly correlate with the audio energy.

6. Results

The techniques were evaluated on four video clips, each approximately 5 minutes long. In each clip, two participants take turns at speaking. On an average, there are 30 change-points in each clip and each participant spoke for about half the time. The four clips were of varying difficulties. In the first clip, the participants stay more or less still when they are not speaking. The second clip is more natural where the listener has some spurious movements and often nods his head in agreement with the speaker. In the third clip, the conversation is similar to the second but a cell phone rings intermittently adding noise to the audio signal. In the fourth clip, in addition to audio noise, the participants intentionally move their hands and swivel on the chair while listening adding noise to the video signal. In all these clips, the participants do not explicitly look at the camera and as a result on an average a three-fourth view of one participant's face is visible while only the profile view of the other participants face is visible. While conversing, the participant's hands occasionally occlude their face. The clustering results and the localization performance of the two schemes are shown in the following sections.

6.1. Clustering Performance

BIC finds change-points in the feature vector streams. These change-points correspond to segment boundaries. The

Figure 2. Clustering performance obtained using audio, video and fused feature vectors.

individual segments are then clustered. Three clustering results are obtained by clustering the feature vector streams using their own change-points. A fourth result is obtained by clustering the audio vector stream using change-points detected from the fused feature vector stream (clustering audio using fused vector change-points - AF). The classification accuracy is computed as the percentage of total frames that were assigned to the correct cluster.

The audio performance reflects the difficulty of the four clips. In the first two clips, where no background noise is present, the average classification performance is 90%. In clips where the cell phone rings in the background, the audio performance drops to about 65%. A similar trend is visible for video. Performance drops from 90% in the first clip to about 82% in clips 2 and 3, where the conversation is natural. In clip 4, the performance drops to 66% on account of the intentional spurious movements. The fused signal does not provide any significant improvement. The system performance using the fused signal is similar to that obtained using the audio or video signal. However, clustering the audio signal with fused breakpoints (AF) shows robustness to audio noise and spurious video movements. The performance in the first two clips is similar to that of original audio clustering. However, in the third clip, where audio performance degrades (64%), the AF does not (84%). Interestingly, AF performs well on the fourth clip where both, audio and video signals fail. On further inspection we found that the audio and video signals fail at different times. The fused signal finds breakpoints that the audio signal missed. Reclustering audio with the fused breakpoints (AF) yields a much better performance.

Interestingly, the fused method did not perform better than the audio or video methods. It oversegments the signal which aids in clustering audio, however the nature of the segments are not amenable to clustering. This is because segments where the audio is from the same person, might have different video characteristics (the listener moves more than the speaker or different parts of the

Figure 3. Video localization within each Cluster. (a-b) sample images from each cluster. (c-d) averaged difference images for each cluster. (d-e) speaker localization using MI. (f-g) localization using eigenvectors.

speaker's body move) which prevent them from being clustered together.

6.2. Speaker Localization

In this section, we compare three approaches for speaker localization. Simply averaging the difference images of that cluster should ideally provide the location of the speaker. However because the listeners also move a little, localization by this method will not be perfect. We compare the MI based approach with our eigenvector approach for localization. Figures 3a and 3b show a representative frame from each cluster. In (a) the person on the right is speaking and in (b) the person on the left is speaking. Figures 3c and 3d show the average of the difference images overlaid on the original frames. It can be seen that in (c) the localization is more accurate than in (d). Figures 3e and 3f show that the MI approach fails to filter out the non-dominant speaker. Figures 3g and 3h show the principal eigenvector of the

cluster overlaid on the original images. The speakers are successfully localized in both clusters.

7. Conclusion

Given a video clip, locating different speakers and isolating their speech is a prerequisite to various other tasks. In this work we cluster the audio signal using information from both audio and video sources. Our method improved the clustering performance over audio only clustering. Once the audio clusters are found, we localize the speaker in the corresponding video frames. To localize the speaker, we used the eigenvectors of the video frames of that cluster. We find that the eigenvector approach was better suited for our dataset compared to the MI approach. Our technique does not make any assumptions about frontal faces, nor does it requires any training or *a priori* information. Thus it is readily adaptable to other scenarios. Also, since the feature vectors used are simple and the BIC technique is tweaked to improve speed, the approach is computationally efficient.

References

- [1] S. Chen and P. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *DARPA speech recognition workshop*, 1998.
- [2] S.-S. Cheng and H.-M. Wang. A sequential metric-based audio segmentation method via the bayesian information criterion. In *Proceedings of Eurospeech*, pages 945–948, 2003.
- [3] T. Darrell, J. Fisher, P. A. Viola, and W. T. Freeman. Audio-visual segmentation and “the cocktail party effect”. In *ICMI*, pages 32–40, 2000.
- [4] J. Fiscus, N. Radde, J. Garofolo, A. Le, J. Ajot, and C. Laprun. Rich transcription 2005 spring meeting recognition evaluation. In www.nist.gov/speech/publications/papersrc/rt05sresults.pdf.
- [5] J. Fisher and T. Darrell. Probabilistic models and informative subspaces for audiovisual correspondence. In *ECCV(3)*, pages 592–603, 2002.
- [6] J. Hershey and J. R. Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *NIPS*, 1999.
- [7] E. Kidron, Y. Schechner, and M. Elad. Pixels that sound. In *CVPR*, pages I: 88–95, 2005.
- [8] H. J. Nock, G. Iyengar, and C. Neti. Speaker localisation using audio-visual synchrony: An empirical study. In *CIVR*, 2003.
- [9] F. Quek, D. McNeill, R. Ansari, X.-F. Ma, R. Bryll, S. Duncan, and K. E. McCullough. Gesture cues for conversational interaction in monocular video. In *RATFG-RTS*, 1999.
- [10] S. Sarkar and P. Soundararajan. Supervised learning of large perceptual organization: Graph spectral partitioning and learning automata. 22(5):504–525, May 2000.
- [11] A. Tritschler and R. Gopinath. Improved speaker segmentation and segments clustering using the bayesian information criterion. In *Proceedings of Eurospeech*, 1999.