FINDING PRESENTATIONS IN RECORDED MEETINGS USING AUDIO AND VIDEO FEATURES

Jonathan Foote, John Boreczsky, and Lynn Wilcox

FX Palo Alto Laboratory
3400 Hillview Avenue
Palo Alto, CA 94304
{foote, johnb, wilcox}@pal.xerox.com

ABSTRACT

This paper describes a method for finding segments in video-recorded meetings that correspond to presentations. These segments serve as indexes into the recorded meeting. The system automatically detects intervals of video that correspond to presentation slides. We assume that only one person speaks during an interval when slides are detected. Thus these intervals can be used as training data for a speaker spotting system. An HMM is automatically constructed and trained on the audio data from each slide interval. A Viterbi alignment then resegments the audio according to speaker. Since the same speaker may talk across multiple slide intervals, the acoustic data from these intervals is clustered to yield an estimate of the number of distinct speakers and their order. This allows the individual presentations in the video to be identified from the location of each presenter's speech. Results are presented for a corpus of six meeting videos.

1. INTRODUCTION

Many meetings contain slide presentations by one or more speakers, for example, the weekly staff meetings at FXPAL and the lab meetings at Xerox PARC. These meetings are often recorded for future review and reuse. For browsing and retrieval of such meetings, it is useful to locate these start and end time of presentations. If an agenda is provided for the meeting, presentations can be automatically labeled using the agenda information. This allows presentations to be easily found by presenter and topic. Thus meeting videos can be automatically indexed, browsed, and retrieved by content.

The system described here uses automatic image recognition in concert with audio-based speaker identification to precisely locate presentations within video recordings of meetings. We assume that the video recording of a presentation contains intervals where slides are being displayed, in addition to camera shots of the speaker and audience. We also assume that a single speaker is talking during an interval when slides are displayed. Slide intervals are automatically detected, and the audio in these regions is used to train a speaker spotting system. The audio is then resegmented using the speaker spotting system, yielding a sequence of single speaker intervals. Since a presentation by a given speaker can span multiple intervals, these speaker intervals are clustered by audio similarity to find the number and order of the speakers giving presentations in the video. After clustering, all data from a single speaker can be used as training data for speaker-identification and -segmentation techniques as in [1,2].

1.1 The task of locating presentations.

At FX Palo Alto Laboratory, weekly staff meetings are held in a conference room outfitted with multiple video cameras and microphones. Meetings start with general announcements from management and staff, then proceed to presentations by individual lab members. Presentations are usually given by one person and include graphics such as overhead or computer slides, and there is usually more than one presentation in a meeting. A camera person switches between the cameras in the room, providing shots of the presenters, audience, as well as presentation materials for the video recording. The video is MPEG-encoded, and made available to staff via the company intranet.

Because the audio comes from multiple ceiling microphones rather than lapel or other close-talking mikes, speaker identification (SID) becomes particularly difficult. Practically all SID techniques use some sort of audio spectral measure, such as mel-frequency cepstral coefficients, to characterize a particular speaker [3]. Far-field microphones in all real-world environments pick up speech both directly and reflected from environmental features such as walls, floors, and tables. These multipath reflections introduce comb-filtering effects that substantially alter the frequency spectrum of the speech. This problem is worsened by mixing signals from multiple microphones (as is common practice in teleconferencing systems). Additional effects due to room resonances will also color each microphone's frequency response. Both resonance and comb-filter effects change drastically and unpredictably with a speaker's position in the room. This makes current speaker-identification methods, where a sample of training speech is used to train a speaker model, particularly ill-suited to a far-field microphone environment. The spectral changes due to the acoustic environment can approach the same order of magnitude as the spectral differences between speakers.

To avoid the inevitable mismatch between training and test data due to unpredictable room acoustics, this system essentially obtains training data from the test data by extracting segments that were likely uttered by a single speaker. This is done by assuming a single speaker's speech is correlated with the display of presentation visuals such as slides. (In our domain, this assumption is usually, but not completely, accurate as there are frequently questions, laughter, or other interjections during a given slide interval.) Other video analyses, such as single-face or news-anchor detection, could be used in a similar manner. If face recognition is possible, it could augment or replace the audio clustering used to associate video intervals with particular speakers.

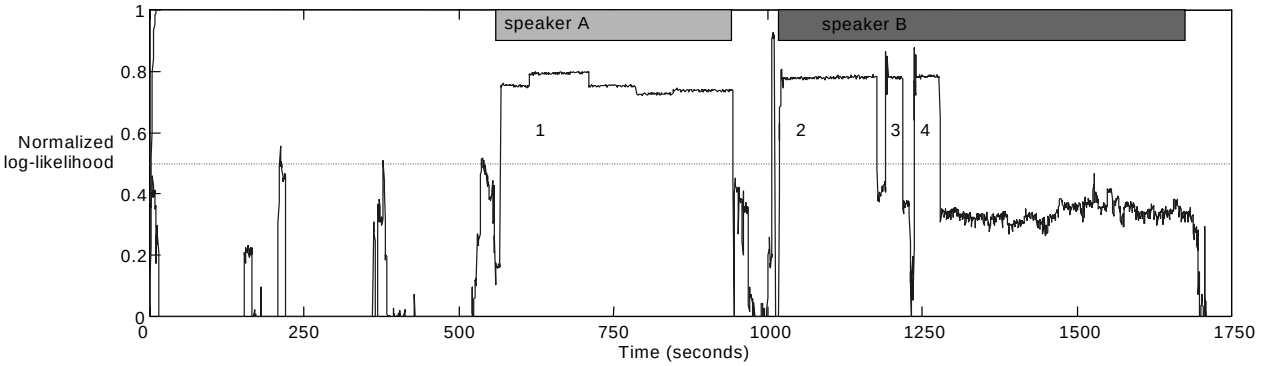

Figure 1. Slide likelihood, detected slide regions, and speaker presentation ground truth for an example meeting video

2. RELATED WORK

This work follows directly from that of Wilcox *et al*. [1,2]. This previous work was concerned with segmenting audio recordings only, thus there was no video channel to exploit. This system used uniform-duration windows as initial data for speaker clustering. If windows were too long, then chances of capturing multiple speakers were high, however too short a window resulted in insufficient data for good clustering. In the absence of additional cues, windows will often overlap a change in speaker, making them less useful for clustering. Other segmentation work has also based primarily on audio, for example, the meeting segmentation system using speech recognition from lapel microphones of Yu [4].

3. AUTOMATIC SEGMENTATION

The first step in the segmentation process is to locate slides in the video. This is done using the techniques of [5], which yield accurate estimates of when presentation graphics are displayed in the video. The original MPEG-1 video is decimated both in time, to two frames per second, and in space, to a 64×64 grayscale representation. Each reduced frame is then transformed, using the DCT transform. The transform is applied to the entire frame image, rather than to smaller sub-blocks as is typical for image compression. The transformed data is then reduced by projection onto its 100 principal components. This results in a compact feature vector (the 100 reduced coefficients) for each frame. A diagonal-covariance Gaussian model is trained on slide images from several unrelated meeting videos. This is used to generate a likelihood for each video frame, which measures the log-likelihood that the given frame is a slide. When thresholded at 1 standard deviation, this yields a robust estimate of when slides are shown in the video. As shown in Table 1, the slides were associated with presentations with 94% accuracy. Slide intervals of longer than 20 seconds are used as candidate speech intervals for the system. Figure 1 shows a plot of the slide log-likelihood for a staff meeting. There are 4 intervals that meet the criteria of being above the threshold (dotted line) for longer than 20 seconds: these are labeled 1 through 4. There were two presentations during this particular meeting, respectively given by two speakers labeled A and B. The extent of each presentation is indicated at the top of Figure 1; note that speaker B's presentation lasted more than twice as long as slides were displayed.

3.1 Model Construction and Alignment

Once slide regions have been identified, a hidden Markov model (HMM) can be automatically constructed and trained. Figure 2 shows the HMM structure which models the time extent of the video. Each 'region' model represents the audio from the associated slide interval. It is assumed that the speaker will be speaking for a longer extent than the slides are displayed, as the video will switch between images of the speaker, audience, and the presented slides. The "filler" model represents audio assumed to be other than a presenter's speech. In the present system, the filler model is trained on silence, laughter, applause, and audience noise segmented from a set of training videos, as well as audio from first two minutes of the source video (which is assumed to not contain speech from the presentation speakers). The filler models, though multiply-instantiated, are identical. The region-specific models represent speech from the presentation speakers. Each region-specific model is trained on the audio from the slide interval associated with it. Concatenating a region model and an optional filler model results in a "interval unit," one for each detected slide interval. These are concatenated to result in the final model, which enforces the proper speaker order. Segmentation is performed using the Viterbi algorithm to find the maximum-likelihood alignment of the source audio with the full model [3]. This allows the time extent of the speakers to be determined, as it may differ substantially from the intervals in which slides are shown. In particular, it is common for the video to alternate between shots of the speaker, audience, and the presentation slides while the speaker is talking. In the current system, both filler and region models have a single state, and have single-mixture full-covariance Gaussian output distributions. Because models are single-state and single-mixture, they can be rapidly trained in one pass. Multiple-state or -mixture models may improve performance at the cost of more expensive training. Self-transitions are allowed with no penalty, resulting in an ergodic model that has no explicit time duration This allows a model to represent any given length of time with no probability penalty.

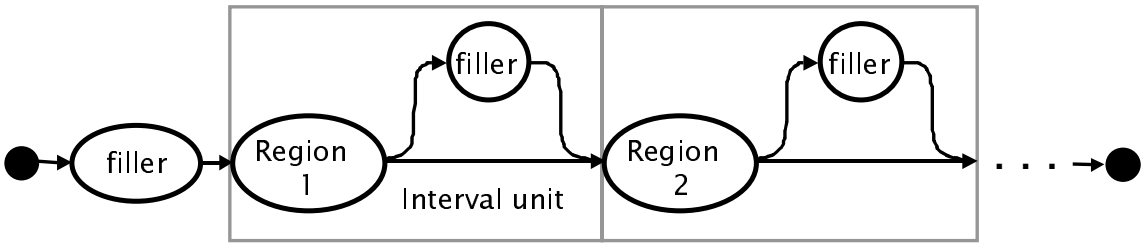

Figure 2. Hidden Markov model for segmentation forced alignment

3.2 Speaker Clustering

The next step is to cluster the candidate intervals to determine how many speakers have given slide presentations. In many cases, there are multiple adjacent intervals that correspond to the same speaker, for example the ones labeled 2, 3 and 4 in Figure 1. Clustering can be done using many techniques, for example the likelihood-ratio distance of Gish [6]. The clustering method used here is based on the non-parametric distance measure of [7]. MFCC-parameterized audio segments are used to train a supervised vector quantizer, using a Maximum Mutual Information criterion to find class boundaries. Once trained, segments are vector quantized, and a histogram is constructed of the bin distributions. This histogram serves as a signature of the audio file; if treated as a vector, the cosine between two histograms serves as a good measure of audio similarity. Figure 3 shows a distance matrix computed using this measure. This shows the audio similarity between 12 slide regions from a single meeting video. Each element *i, j* has been colored to show the difference between segment *i* and *j*, such that closer, hence more similar, distances are darker. From Figure 3, it is clear that there are several acoustically similar groups, each of which correspond to speech from a particular speaker. The exception is from segment 7, which corresponds to the titles from a video shown during the middle speaker's presentation. Such a distance matrix can be clustered to find similar intervals that correspond to a single speaker. Though any sort of hierarchical clustering can be used, the simple approach taken here was to enforce the time-adjacency of cluster members, by considering all adjacent segments to be part of the same cluster as long as none of their respective distances exceeded a threshold. For the segments of Figure 3, this resulted in 5 clusters as follows:

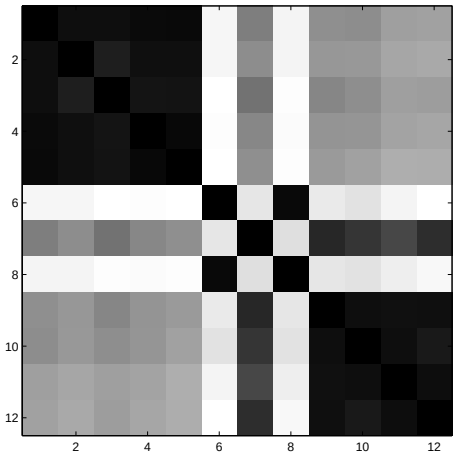

Figure 3. Intersegment acoustic distance matrix

```
(1,2,3,4,5) (6) (7) (8) (9,10,11,12)
```

The ground truth was that there were three presentations, so this clustering method has incorrectly segmented the second presentation into three, based on the audio distance. Because our ultimate application is finding indexes for video browsing, this is not a disastrous error: it might be desirable to find when the video was shown as well as when the presentation started. More sophisticated clustering methods could be used to ignore audio outliers, such as segment 7 of Figure 3, or other anomalous audio such as questions or applause. For example if questions are asked about a particular slide, the resulting interval might contain speech from many different speakers.

4. EXPERIMENTS

Six videotaped meetings containing slide presentations were used as a test corpus. Training data for audio filler models and slide images came from another set of videos. The six videos total length was 280 minutes, 21 seconds for an average length of about 45 minutes. Each video contained from one to five presentations, for a total of 16, though three presentation contained video as well as slides and most had audience questions or comments. Because presentations were typically longer than the duration of slide intervals, the presence of slides was a good indicator of a presentation, but not vice versa, as shown in Figure 4. Table 1 shows that slides were shown in the video for only about 25% of a typical presentation, thus finding presentations from slides alone would result in missing more than 75% of the presentation. The second row of Table 1 shows how speaker segmentation improves this: only about 5% of presentations were mis-identified as being other than presentations.

Features used	Missed	False Positive
Slides	0.745	0.058
Slides + Speaker segmentation	0.042	0.013

Table 1. Presentation classification errors by frame

From the 16 presentations, there were a total of 32 endpoints to detect (as well as additional endpoints from the video and anomalous audio). An endpoint was considered correct if it occurred within 15 seconds of the actual speaker's speech starting or ending. Table 2 shows the accuracy of endpoint location. Before clustering, there were 114 endpoints from the 57 slide intervals. Given the ground truth of 32 relevant endpoints to detect, and 26

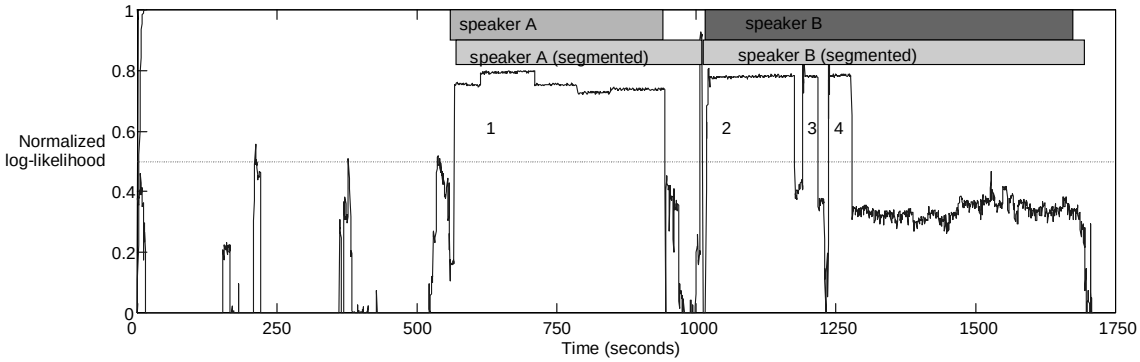

Figure 4. Slide likelihood, detected presentation speakers, and ground truth.

endpoints were correctly located, this resulted in a recall of 0.81 with a precision of 0.23, thus most endpoints were found but less than one in four detected endpoints was likely to be correct. Clustering the 57 aligned segments yielded 23 clusters, which dramatically improved the precision by reducing the number of incorrect endpoints. Note that at least 2 of the detected endpoints were due to videos internal to a presentation, so the precision is unduly pessimistic. The non-ideal audio environment also caused clustering problems. Microphones are mounted in acoustic ceiling tiles near HVAC vents. Several presentations were mis-clustered due to the presence or absence of ventilation noise. This affected the acoustic signal enough that the same talker was clustered differently depending on the state of the ventilation system: several cluster boundaries occur exactly as the ventilation switches on or off.

Endpoint detection	Recall	Precision
Before clustering	0.81	0.23
After clustering	0.81	0.57

Table 2. Endpoint detection accuracy

5. FURTHER APPLICATIONS

The techniques presented here could be improved upon in a number of ways. More sophisticated acoustic models, such as multiple Gaussian mixtures, could improve speaker segmentation. Further improvements might be obtained by enforcing a duration model on each speaker, as in [2]. There is additional room for improvement in the clustering, as mentioned previously. We are currently investigating clustering segments based on video as well as audio features, under the assumption that a presenter's slides should have a similar composition and color scheme, as well as images of the presenters themselves. This would also allow us to identify anomalous regions of both audio and video due to videos being shown during presentations.

The same techniques used to segment and a single meeting can be applied across multiple meetings containing a similar set of speakers. This allows a catalog of presenters to be created. If this contains enough examples of the same speaker's speech across potentially different acoustic environments (room positions), a more robust position-independent speaker model could be trained. In addition, if speakers are identified in meeting agendas, speaker models could be automatically associated with names for subsequent identification and retrieval. Besides meeting videos, these methods are applicable to any domain where individual speakers can be associated with identifiable video characteristics. One example might be news broadcasts, where shots of news anchors can often be reliably identified by image composition and background [9].

6. ACKNOWLEDGMENTS

.Thanks to John Doherty for producing the meeting videos in our corpus, and to Andreas Girgensohn for the slide likelihood data.

7. REFERENCES

[1] Wilcox, L., Chen, F., Balasubramanian, V., "Segmentation of speech using speaker identification," in *Proc. ICASSP '94* Volume S1, pp. 161-164, 1994

[2] Kimber, D., and L. Wilcox, L., "Acoustic segmentation for audio browsers," in *Proc. Interface Conference*. Sydney, Australia, 1996.Rabiner, L., and Juang, B.-H., *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ, 1993

[3] Rabiner, L., and Juang, B.-H., *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ, 1993

[4] Yu, H., Clark, C., Malkin, R., Waibel, A., "Experiments in automatic meeting transcription using JRTk," in *Proc. ICASSP 98*, Volume 2, pp. 921-924.

[5] Girgensohn, A., and Foote, J., "Video frame classification using transform coefficients," submitted to *ICASSP '99*

[6] Gish, H., Siu, M-H., and Rohlicek, R., "Segregation of speakers for speech recognition and speaker identification", in *Proc. ICASSP 91*, pp. 873--86.

[7] Foote, J., "Content-based retrieval of music and audio," in *Proc. SPIE, Multimedia Storage and Archiving Systems II*, vol. 3229, C.-C. J. Kuo et al., Ed., 1997, pp. 138--147.

[8] Foote, J., J. Boreczky, J., Girgensohn, A., and L. Wilcox, "An intelligent media browser using automatic multimodal analysis," in *Proc. ACM Multimedia*, Bristol, UK, Sept. 1998.

[9] Ariki, Y., and Sakurai, M., and Sugiyama, Y., "Article extraction and classification of TV news using image and speech processing", in *Proc. International Symposium on Cooperative Database Systems for Advanced Applications (CODAS-96)*, 1996