Towards Audio-Visual On-line Diarization Of Participants In Group Meetings

Hayley Hung, Gerald Friedland

Towards Audio-Visual On-line Diarization Of Participants In Group Meetings

Hayley Hung[1] and Gerald Friedland[2]

[1] IDIAP Research Institute, Martigny, Switzerland
[2] International Computer Science Institute (ICSI), Berkeley, USA

Abstract. We propose a fully automated, unsupervised, and non-intrusive method of identifying the current speaker audio-visually in a group conversation. This is achieved without specialized hardware, user interaction, or prior assignment of microphones to participants. Speakers are identified acoustically using a novel on-line speaker diarization approach. The output is then used to find the corresponding person in a four-camera video stream by approximating individual activity with computationally efficient features. We present results showing the robustness of the association on over 4.5 hours of non-scripted audio-visual meeting data.

1 Introduction

Conventional speaker diarization aims to segment an audio signal into speaker-homogeneous regions to address the question of 'who spoke when?' [13]. In recent years these approaches have become sufficiently robust and are performed without prior knowledge using only a single distant microphone (SDM) as input. However, being a completely unsupervised process, the output of such systems consists only of labels like 'speaker_1'. Traditionally, diarization is solved offline where only fully finished recordings can be processed. This article presents a system that tries to solve the task audio-visually in real-time and on-line (i.e. incrementally and on-the-fly). Rather than labeling the speaker regions with numbers, they are associated with videos of the corresponding participant. A direct application of this is a remote meeting scenario: knowing who is currently active both in the video and audio stream and where speakers are located in the meeting room is desirable for retrieval, compression, and video editing.

This article is organized as follows: Section 2 discusses related work; Section 3 describes the meeting data; Section 4 presents our audio-visual diarization system; we show our results in Section 5 and conclude in Section 6.

2 Related Work

Common approaches to audio-visual speaker identification involve identifying lip motion from frontal faces [3], [7], [6], [9], [10], [11], [14], [15]. Therefore, the underlying assumption is that motion from a person comes predominantly from the motion of the lower half of their face. This is further enforced by artificial audio-visual data of short duration, where only one person speaks. In these scenarios, natural conversation is not possible so problems with overlapping speech are not

considered. In addition, gestural or other non-verbal behaviors associated with natural body motion during conversations are artificially suppressed.

Nock et al. [9] presents an empirical study to review definitions of audio-visual synchrony and examine their empirical behavior. The results provided justifications for the application of audio-visual synchrony techniques to the problem of active speaker localization in broadcast video. Vajaria et al. [16] presents a system that combines audio and video on a feature-level. Although their method improved the clustering performance over audio-only clustering, their approach was only tested using a laboratory two-speaker scenario. Zhang et al. [20] presented a multi-modal speaker localization method using a specialized satellite microphone and omni-directional camera. Though the results seem comparable to the state-of-the-art, the solution requires specialized hardware, which is not desirable in practice. Noulas et al. [10] integrated audio-visual features for online audio-visual speaker diarization using a dynamic Bayesian network (DBN) but tests were limited to two-person camera views. Tamura et. al. [15] demonstrate that the different shapes the mouth can take when speaking facilitates word recognition under tightly constrained test conditions (e.g., frontal position of the subject with respect to the camera while reading digits).

In a real scenario the subject behavior is not controlled and, consequently, the correct detection of the mouth is not always feasible using their method. Hung et al. [5] proposed an offline method for performing audio-visual association of video streams and speaker clusters. This work showed that audio-visual streams could be associated without the need for fine-grained spatially dependent pixel-based descriptors. While the total length of the data set that was used was considerably larger than those discussed above, the method used meeting lengths of 5 minutes to perform the association and finally only 21 data points were tested.

The approaches discussed above were tested on very limited data sets (which are not always publicly available) and were often recorded in highly constrained scenarios where individuals were unable to move or talk naturally. In general, the speakers face the camera frontally and do not talk over or interrupt each other. This article presents tests on over 4.5 hours of publicly available data [2], capturing 5 different exclusive groups of 4 individuals in a meeting scenario where participants could behave naturally. We propose an audio-visual on-line diarization system where the impact of associating the speech and video streams of meetings of shorter length on the performance is investigated. In contrast to previous methods which combine audio and video sources in the early stages of the speaker diarization process, we present a late fusion approach where noisy video streams are associated with estimated speaker channels. Importantly, the presented approach considers a modification to the visual activity features in [5] to account for cases when participants are beyond the field of view of the cameras, affording a larger more challenging data set.

3 Meeting Data

We used a subset (4.5 hours) of the publicly available AMI [2] meeting corpus where non-scripted meeting data was recorded. In our experiments, five teams of four participants were asked to design a remote control device over a series

of sessions which encouraged natural interactions. A microphone array and four cameras were set in the center of the room. Each camera captures the visual activity of a single seated participant, who is assigned a seat at the start of each meeting session. Participants are requested not to change seats during the session. No other people enter or leave the meeting during the session so there are always only 4 interacting participants. Each person also wore a headset and a lapel microphone, which we used to observe the performance difference of our system under less noisy audio conditions. Side-view and rear cameras were also capturing video data but were not used for feature extraction. A plan view of the meeting room is shown in Figure 1. Ground truth speaker segmentations were automatically generated by thresholding the speaker energy from the headset microphones (1: speaking, 0: silence). We found these produced better results than the provided ground truth, where temporal alignment errors from human judgments were possible.

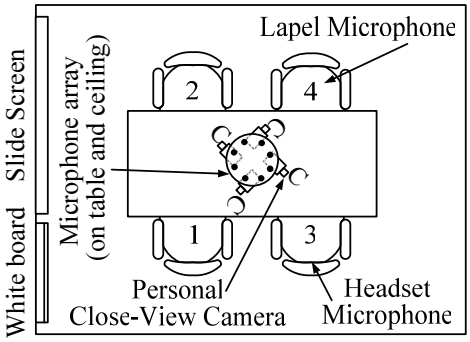


Fig. 1. Plan of the experimental meeting room.

4 Audio-Visual Diarization

Our goal is to (*i*) segment live-recorded audio into speaker-homogeneous regions to answer the question 'who is speaking now?' and (*ii*) show the corresponding video of the speaker. For the system to work live and on-line, the question must be answered on intervals of captured audio and video that are as small as possible, and performed in at least real-time. The following section presents our approach, which is also summarized in Figure 2. The on-line speaker diarization system has been described in detail in [17]. Audio is processed as 19th-order Mel Frequency Cepstral Coefficients (MFCC). A frame-length of 30 ms is used, with a step size of 10 ms. Speaker models are represented using Gaussian Mixture Models (GMMs). The system has two steps: (*i*) training, (*ii*) recognition.

4.1 On-line Audio Speaker Diarization

Unsupervised Training: To bootstrap the creation of models, we use the speaker diarization system proposed by Wooters et al. [18] in the first meeting to automatically estimate the number of speakers and their associated speaker models. The ICSI Speaker Diarization System has competed in the NIST evaluations of the past several years and established itself well among state-of-the-art

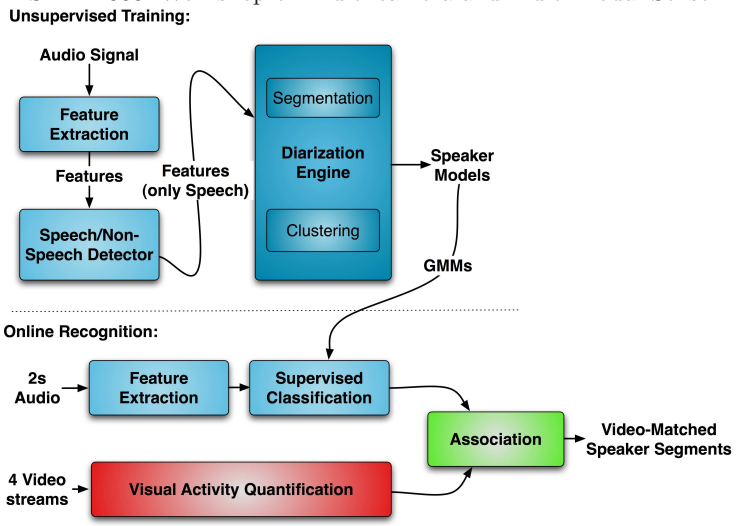


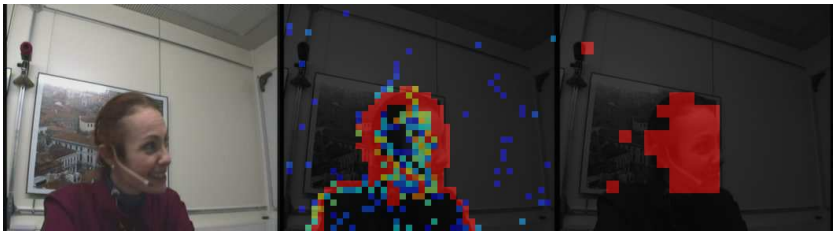
Fig. 2. Summary of the on-line audio-visual diarization algorithm.

systems[3]. This system works well but is an offline approach and usually needs at least 10-15 minutes of audio to work with optimal robustness. The voice is recorded and converted to 19-dimensional MFCC features. A speech/non-speech detector is run. The speech segments are then used to train a Gaussian Mixture Model (GMM). The number of Gaussians were determined empirically. In order to be able to cope with potentially difficult room conditions, e.g. air-conditioning noise, we also trained an additional 60-second room-specific non-speech model.

The output of the offline diarization is used to create models for the on-line diarization. We use the first 60 seconds of accumulated speech of each speaker to train a model for each of them. This enables an unsupervised learning approach that requires no manual intervention. Once models have been created, they are added to the pool of speaker models and can be reused for all subsequent meetings. In addition to speaker models, we also train a non-speech model which captures background noise and channel effects.

Recognition: In recognition mode, the system records and processes chunks of audio as follows. First, Cepstral Mean Subtraction (CMS) is implemented to reduce stationary channel effects [12]. While some speaker-dependent information is lost, according to our experiments performed, the major part of the discriminant information remains in the temporally varying signal. In the classification step, the likelihood for each frame is computed against each set of Gaussian Mixtures obtained in the training step. From our previous experiments on larger meeting corpora, [17], we decided to use two-second chunks of audio. Thus, a total of 200 frames are examined for each classification decision. This introduces a latency of about 2.2 seconds after a person has started talking. The decision on whether a segment belongs to a certain speaker or the non-speech model is reached using majority vote on the likelihoods of a frame belonging to a GMM. If the audio segment is classified as speech, we compare the winning speaker model

[3] (http://www.nist.gov/speech/tests/rt/rt2007/)



(a) (b) (c)

Fig. 3. Compressed domain video feature extraction. (a) Original image, (b) Residual coding bit-rate, (c) skin-colored regions. The use of compressed domain features allows very efficient processing.

against the second best model by computing the likelihood ratio. We use this as an indicator of the confidence level. In our experiments, we assume that there are speaker models for all possible speakers so we used the highest confidence level to indicate the most likely speaker. For a more realistic case, it is possible to apply a threshold to the confidence level to detect an unknown speaker but this currently requires manual intervention.

Offline audio speaker diarization can lead to more clusters than speakers since the method is data-driven. Due to the robustness of our on-line speaker diarization algorithm, while more clusters than participants can be generated in the offline training phase, in the on-line stage, noisy or extraneous clusters have much lower likelihoods, so they are never selected as likely speaker models. We found in our experiments that the number of recognized clusters and actual participants was always equal.

It is also important to note that the data that we use includes overlapping speech. These periods are automatically ignored when the speaker models are generated to ensure they remain as clean as possible. Work has been carried out to address overlapping speech in offline diarization systems but involve a second pass over the diarized audio signal, which would not be feasible for an on-line and real-time system [1].

4.2 Visual Activity from Compressed Video

The association of the video streams with the diarization output is based on frame-based visual activity features. We re-used some of the video processing which is already applied for video compression to estimate the visual activity of each person very efficiently. The method we use is detailed in [19]. To construct an estimate of personal activity levels we extracted the residual coding bit-rate, which was found to be the most discriminative (see Figures 3 (b)) for speech-visual activity association. For each camera view we estimate a participant's activity level by implementing a block-based skin-color detector working mostly in the compressed domain, to estimate head and hand regions as illustrated in Figure 3 (c). To do this, we use a GMM to model the distribution of chrominance coefficients [8] in the YUV colour-space. To determine the skin-colour blocks in the Intra-frames, the likelihood of the mean of the chrominance coefficients from the GMM are thresholded. These blocks in the Inter-frames are inferred using

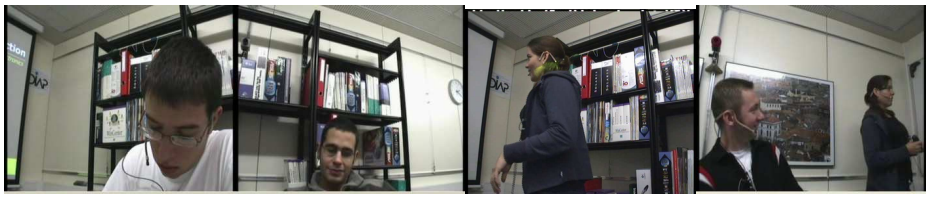


Fig. 4. Possible pose variations and ambiguities captured from the video streams.

motion vector information to propagate them through the duration of the group-of-picture (GOP). For each frame the average residual coding bit-rate over all the estimated skin blocks is calculated and used as an estimate of individual visual activity. Compared to extracting many higher resolution pixel-based features such as optical flow, compressed-domain features take 95% less time to run.

Though the video cameras capture mostly, a close-view of each person, natural body poses can lead to highly varying activity as illustrated in Figure 4. Also, on occasion, someone who presents at or walks to the whiteboard or slide screen could also be captured by their non-corresponding camera. In addition, people move even when they are not speaking, so associating these visual activity features with the estimated speaker clusters is challenging. From these examples of natural body poses, conventional methods of audio-visual association that try to locate an individual's lip motion would be an inefficient use of resources and is likely to lead to quite a noisy signal. We show in this work, that despite the coarse and noisy representation of a person's visual activity, their body motion is still well correlated with when they speak.

The video capture results in four streams; one for each camera. Each of the four streams were represented using the average residual coding bit-rate with filtering of the skin-colored regions. Using this feature, we are only able to test on the meetings where all the participants were always seated and therefore in view of their corresponding camera. To measure the motion activity when the person was not seated (e.g. presenting or talking at the slide screen or white board), the visual activity feature was modified. When the person was not detected in the camera view, the maximum visual activity value for that person was used. To detect whether a person was standing or not, we used a threshold of the total number of skin-color blocks in one frame. We shall refer to this as the free case, compared to the original seated only (raw) features.

4.3 Associating Audio and Video Channels

We use a similar method to Hung et al. [5] to associate and evaluate the speaker and video streams. The pair-wise correlation between the speaker clusters and visual activity features is computed. Then a greedy association between pairwise feature streams is performed, where the pair with the highest correlation is matched first. Then both corresponding streams are eliminated from the matrix and the procedure repeats until all visual activity channels is associated with a speaker cluster. Figure 5 shows the algorithm in more detail.

– **Quantifying the distance between audio-visual streams (a)**: the pairwise correlation between each video, v_i, and audio stream, a_j, is calculated:-

$$\rho_{v_i,a_j} = \frac{\sum_{t=0}^{T} v(t) \cdot a(t)}{\sum_{t=0}^{T} v(t) \sum_{t=0}^{T} a(t)}, \quad \forall\{i,j\} \tag{1}$$

where T is the total length of the meeting and in our experiments t indexes the feature value at each frame. For our experiments, the frame rate used was 5 frames per second.
– **Selecting the closest audio-visual streams (b)**: the pair of audio and video streams with the highest correlation are selected.
– **Selection of the next closest audio-visual streams (c)** : the next best correlated pair of audio and video streams is selected.
– **Full assignment of audio and video streams**: step (c) is repeated until all audio-visual streams are associated.

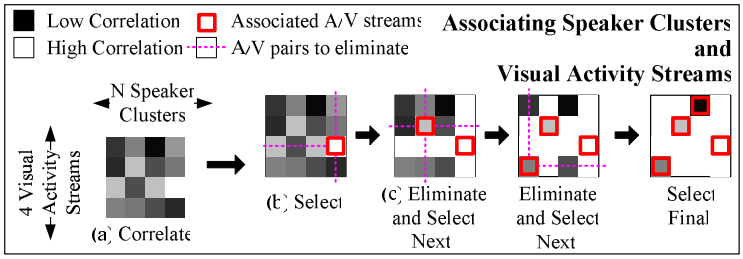


Fig. 5. Greedy Algorithm for ordered and discriminative pairwise associations between audio and video streams.(a) All pairwise combinations of the audio and video streams are correlated. (b) The pair with the highest correlation is associated first and then eliminated from the correlation matrix.

Other algorithms could have been used to associate the audio-visual streams. The reason for using this method was because the audio and visual streams can be very noisy, particularly in situations where a person doesn't speak much (but is still likely to move) or when estimations of whether someone is in view of the camera or not is vulnerable to inaccuracies (e.g. the rightmost snapshot in Figure 4). In such circumstances, it is easier to offset the problems of noisy estimates by using those that are more reliable. Therefore, the assumption is that by associating the better correlated features first, we are more likely to associate the less well-correlated streams accurately by a process of elminiation.

5 Experimental Results

5.1 On-line Diarization

The output of a speaker diarization system consists of meta-data describing speech segments in terms of start and end times, and speaker cluster labels. This output is usually evaluated against manually annotated ground truth segments. Note that the manually labeled ground truth was used here since the algorithm for calculating the DER also performs forced alignment to remove human errors in temporal alignment in the estimated speaker segments. The

manual alignments are created by using an automatically generates speaker segmentations, generated by using a threshold on the speaker energy from personal headset microphones as a starting point. The error is expressed as Diarization Error Rate which is defined by NIST (http://nist.gov/speech/tests/rt/rt2004/fall) and quantifies errors in terms of misses, false alarms, and speaker errors.

To validate the on-line approach presented here, we performed different experiments on the audio recordings of the AMI meetings (see Section 3). One meeting is selected at random for offline diarization so that speaker models can be generated. The on-line system is then used to classify this meeting and all other meetings with the same speaker. Another element that we investigated was how varying the signal-to-noise ratio (SNR) of the input audio source could affect performance. First, we used a signal obtained by mixing the four individual headset microphones (MH) and lapel microphones (ML) using a basic summation. Finally, a true far-field case (SDM) was used where a single microphone from the array on the table was used. Table 1 shows the results for the on-line audio diarization system and the corresponding SNR for each of the source types. We also show results using an off-line system for comparison. The off-line system uses the automated fast match approach (KLFM) of Huang et al. [4]. We observe a decrease in performance as the SNR decreases.

Input Source	**SNR (dB)**	**DER : Online (%)**	**DER : Offline (%)**
Mixed Headset (MH)	31	18.49	33.78
Mixed Lapel (ML)	22	28.88	36.47
Far-field (SDM)	21	28.93	36.14

Table 1. Comparison of the performance of the on-line diarization system (audio only) on 4.5 hours of the AMI corpus for different input sources. The offline performance from [4]is also shown.

5.2 Audio-visual Association

For evaluation purposes we also perform the same greedy mapping procedure described in Section 4.3 with speaker clusters and individual ground truth speaker segmentations. Using the associations of the clusters to labeled speaker segments, a scoring criteria is enabled where the mapping is true only when the corresponding headset segmentation is associated with the correct visual activity channel through the corresponding speaker cluster, as shown in Figure 6. We also have the true mappings of the ground truth speaker segmentations to video streams to evaluate our association method.

Since the association is performed on a meeting basis, it is important to evaluate the performance similarly. Three evaluation criteria are used, to observe the difficulty in associating more channels correctly in each meeting. Hard (EvH), medium (EvM), and soft (EvS), criteria are used which assigns respectively a score of 1 for each meeting only when all, at least two, or at least one of the pairs of associated audio and visual streams is correct for each meeting. We refrain from evaluating on a participant basis since the meeting-based ordered mapping procedure, by definition, discriminates pairs that are easier to distinguish, as a means of improving the association from noisier channels which may have less observable activity.

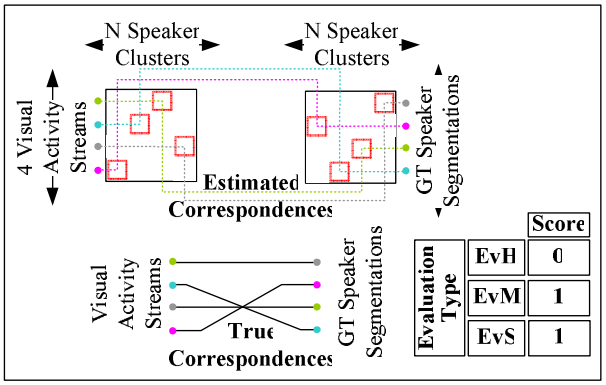


Fig. 6. Example meeting showing how the audio-visual associations are evaluated. The associations of video streams to speaker clusters is estimated. The speaker clusters and ground truth speaker segmentations are also associated. Using these associations as a look-up table, both the estimated and true video to ground truth speaker segmentations can be evaluated. The scores are shown for the three different evaluation criteria.

Firstly, the audio-visual association of speaker clusters to our raw, unlabeled visual activity features was calculated using only the meetings where everyone was fully seated (see Table 2. This consisted of 21 5-minute meetings. To increase the size of the data set, the window length was decreased incrementally. In addition, decreasing the window length allowed us to observe the extent of the decrease in performance if the audio-visual association if we were to go towards a sliding window approach to performing on-line and real-time audio-visual diarization. In addition, we tested on the diarized clusters that were generated using the three different audio source types: using mixed headset (MH), mixed lapel (ML) or the single microphone far-field case (SDM). The best performing association (100%) was reached when 5-minute meetings with the ideal MH source case were used. This showed considerable improvement over the results presented in Hung et al. [5] where the best EvH score was 43%. Overall, there was a consistent drop in performance as the window length decreased, and in general similar decreases were observed when the SNR dropped.

Length(# Meetings)	**Source**	**EvH**	**EvM**	**EvS**
5" (21)	**MH**	1	1	1
	ML	0.76	0.95	1
	SDM	0.76	0.95	0.95
2"30' (42)	**MH**	0.69	1	1
	ML	0.52	0.81	0.93
	SDM	0.4	0.86	0.98
1"15'(84)	**MH**	0.44	0.83	0.96
	ML	0.24	0.69	0.87
	SDM	0.24	0.68	0.93

Table 2. Results for the 1 hour 40 mins meetings were everyone was always seated.

We also ran the association algorithm over a larger data-set of 57 5-minute meetings where participants were not always in view of their camera. For this data-set, we used the modified version of the visual activity features to take into account when the participant was standing and probably at the white board or slide screen. The results are shown in Table 3 where a general drop in performance was observed compared to the fully seated meetings. We found that without modifying the visual activity features, performance was worse for the seated and standing meetings but when the free case was applied to just the seated meetings, the performance was worse compared to the raw feature case. This implies that the modification to account for when people are standing introduces some noise to the visual activity values when someone is standing since the estimate of whether someone is in view of the camera can be incorrect at times. Closer inspection showed cases where estimated skin-color blocks corresponded to regions of the background. Overall, for this larger data set, similar trends to those of the smaller data-set are also seen so shorter meetings had worse performance. However, almost all the SDM cases gave better performance compared to the corresponding ML case for this larger data-set.

Length(# Meetings)	**Source**	**EvH**	**EvM**	**EvS**
5" (57)	**MH**	0.65	0.87	0.91
	ML	0.38	0.71	0.88
	SDM	0.43	0.8	0.95
2"30' (114)	**MH**	0.54	0.8	0.89
	ML	0.35	0.65	0.83
	SDM	0.3	0.73	0.9
1"15'(228)	**MH**	0.34	0.65	0.86
	ML	0.18	0.56	0.77
	SDM	0.2	0.57	0.84

Table 3. Results for 4.5 hours of seated and standing meetings using the free visual activity features.

Figure 7 summarizes the effects of decreasing the window length, SNR, and tougher evaluation criteria on performance. A window of 37s is also included for interest. Note that while the performance using EvH decreases significantly, we found the audio-visual data of the more talkative participants tended to be more reliably associated with the correct video stream. This is probably due to the accumulation of better models for people who speak longer and would also lead to a more accurate estimate of their speaking patterns. For applications such as a remote meeting scenario, the audio-visual data of the more talkative participants are typically more relevant to the meeting context and would be more likely to be handled appropriately.

6 Conclusion

This article presented an experimental system for fully automatic, unsupervised, unintrusive, live speaker identification in meetings. Our on-line speaker diarization system performs better than off-line versions and speakers are identified well from corresponding video streams using a comparatively simple technique.

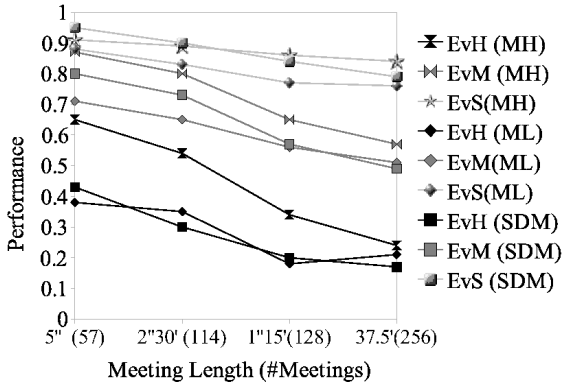


Fig. 7. Comparing performance across decreasing window length, SNR and with tougher evaluation. *EvH*: Hard eveluation strategy where all audio-visual streams in the meeting must be associated correctly; *EvM* Medium evaluation strategy where at least 2 of the audio-visual streams in the meeting must be associated correctly; *EvS* Soft evaluation strategy where at least 1 of the audio-visual streams in the meeting must be associated correctly.

The experiments were performed on a publicly available meeting database using a non-scripted, natural conversation scenario with four participants. Different noise conditions were tested by varying the input audio source. While these initial experiments performed the audio-visual association on long windows, we hope to shorten this latency by using semantically higher-level computationally efficient video features. We also believe the number of video streams could be reduced using a single omni-directional camera if participants were to remain spatially separated. Our experiments show the offline training phase used for the audio diarization could also be applied to the audio-visual association. Finally, it would also be interesting to see if the audio-video correlations of an individual's speech and visual activity patterns are consistent on different occasions and could therefore be used to identify meeting participants more robustly.

Acknowledgments

This research was funded by the EU project AMIDA the Swiss NCCR IM2, the German Academic Exchange Service (DAAD) and Singapore A*STAR. We also thank Chuohao Yeo (EECS, UCB) for providing the basic visual activity features and Oriol Vinyals (UC San Diego) for providing help with the on-line diarization experiments.

References

1. K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland. Overlapped speech detection for improved speaker diarization in multiparty meetings. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4353–4356, 2008.
2. J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraiij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, M. McCowan, W. Post, D. Reidsma, and P. Wellner. The ami meeting corpus: A pre-announcement. In *Joint Workshop on Machine Learning and Multimodal Interaction (MLMI)*, 2005.

3. T. Chen and R. Rao. Cross-modal Prediction in Audio-visual Communication. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 2056–2059, 1996.
4. Y. Huang, O. Vinyals, G. Friedland, C. Müller, N. Mirghafori, and C. Wooters. A fast-match approach for robust, faster than real-time speaker diarization. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2007.
5. H. Hung, Y. Huang, C. Yeo, and D. Gatica-Perez. Associating audio-visual activity cues in a dominance estimation framework. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Human Communicative Behavior*, Ankorage, Alaska, 2008.
6. J. W. F. III and T. Darrell. Speaker association with signal-level audiovisual fusion. *IEEE Transactions on Multimedia*, 6(3):406–413, 2004.
7. J. W. F. III, T. Darrell, W. T. Freeman, and P. A. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Conference on Neural Information Processing Systems (NIPS)*, pages 772–778, 2000.
8. S. J. McKenna, S. Gong, and Y. Raja. Modelling facial colour and identity with gaussian mixtures. *Pattern Recognition*, 31(12):1883–1892, 1998.
9. H. J. Nock, G. Iyengar, and C. Neti. Speaker localisation using audio-visual synchrony: An empirical study. In *ACM International Conference on Image and Video Retrieval (CIVR)*, pages 488–499, 2003.
10. A. Noulas and B. J. A. Krose. On-line multi-modal speaker diarization. In *Proc. International Conference on Multimodal Interfaces (ICMI)*, pages 350–357, New York, USA, 2007. ACM.
11. R. Rao and T. Chen. Exploiting audio-visual correlation in coding of talking head sequences. *International Picture Coding Symposium*, March 1996.
12. D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17(1-2):91–108, 1995.
13. D. A. Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.
14. M. Siracusa and J. Fisher. Dynamic dependency tests for audio-visual speaker association. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2007.
15. S. Tamura, K. Iwano, and S. FURUI. Multi-Modal Speech Recognition Using Optical-Flow Analysis for Lip Images. *Real World Speech Processing*, 2004.
16. H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi. Audio segmentation and speaker localization in meeting videos. In *International Conference on Pattern Recognition (ICPR)*, pages 1150–1153, 2006.
17. O. Vinyals and G. Friedland. Towards semantic analysis of conversations: A system for the live identification of speakers in meetings. In *Proceedings of IEEE International Conference on Semantic Computing*, August 2008.
18. C. Wooters and M. Huijbregts. The ICSI RT07s speaker diarization system. In *Proceedings of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop*, 2007.
19. C. Yeo and K. Ramchandran. Compressed domain video processing of meetings for activity estimation in dominance classification and slide transition detection. EECS Dept, UC Berkeley Technical Report, 2008.
20. C. Zhang, P. Yin, Y. Rui, R. Cutler, and P. Viola. Boosting-Based Multimodal Speaker Detection for Distributed Meetings. *IEEE International Workshop on Multimedia Signal Processing (MMSP) 2006*, 2006.