MULTI-MODAL SPEAKER DIARIZATION OF REAL-WORLD MEETINGS USING COMPRESSED-DOMAIN VIDEO FEATURES

Gerald Friedland‡∗, Hayley Hung§ and Chuohao Yeo♦

‡Int. Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA
§IDIAP Research Institute
Rue Marconi 19
CH-1920 Martigny
♦UC Berkeley
Dept. of EECS
Berkeley, CA 94720, USA

ABSTRACT
Speaker diarization is originally defined as the task of determining “who spoke when” given an audio track and no other prior knowledge of any kind. The following article shows a multi-modal approach where we improve a state-of-the-art speaker diarization system by combining standard acoustic features (MFCCs) with compressed domain video features. The approach is evaluated on over 4.5 hours of the publicly available AMI meetings dataset which contains challenges such as people standing up and walking out of the room. We show a consistent improvement of about 34% relative in speaker error rate (21% DER) compared to a state-of-the-art audio-only baseline.

Index Terms— Speaker extraction, multi-modal, compressed domain features

1. INTRODUCTION
Speaker diarization is traditionally defined as the task of estimating “who spoke when” given a single-source audio track and no prior knowledge of any kind. The problem has been investigated extensively in the speech community and NIST has conducted evaluations in order to measure the accuracy of various algorithms on different datasets. While progress has been made in these evaluations over the years, the performance of audio-only algorithms may have reached a plateau. At the same time, the application scenarios and the availability of increased amounts of video data suggests that a way of advancing the field is to include video data as well. This is not only a more natural approach, but is also supported by prior studies in social psychology [1]. In particular, it has been observed that in human conversations, speaking turns are also arbitrated by non-verbal cues such as body movements [1]. This has, in part, inspired some prior work into augmenting speaker diarization with visual cues, although typically under constrained laboratory conditions (see Section 2).

In this paper, we present an audio-visual approach capable of handling real-world meeting data where people walk around, leave the room, stand up, etc. We show that in our approach, video data can provide a substantial improvement to a state-of-the-art audio-only system without adding a significant amount of computational complexity. This is achieved using video features directly derived from MPEG-4 compressed data which is generated directly from recordings using modern video cameras.

The article is organized as follows: we first present prior related work in Section 2 and an overview of the underlying audio-only system in Section 3. Section 4 describes the compressed-domain video features while Section 5 presents our approach for multi-modal feature fusion. We then describe our experimental setup and results in Section 6 before concluding in Section 7.

2. RELATED WORK
A related class of works investigates the problem of audio-visual synchrony [2, 3], where given an audio or speech signal, the aim is to localize its source in video (or vice versa). However, in many cases, it is assumed that the audio or speech signal has already been segmented and experiments are carried out on audio-visual clips of less than 30s where only one of two people in the scene speaks at any instant. In addition, such work tends to rely on the audio-visual synchrony of lip and/or jaw motion and speech, leading to constrained test scenarios where the subjects try not to move their heads and must face the camera frontally for accurate localization of the mouth region. Our proposed system makes no such assumptions about the initial temporal segmentation of the audio data since it is trying to perform speaker segmentation. Also, our video data includes highly varying natural poses where people’s heads rarely face the camera frontally, and on many occasions the entire mouth region can be occluded or obscured by extreme head-pose angles (see Figure 4).

Another related class of works is that of online multimodal speaker detection [4], where given multiple audio and video streams (say from multiple microphones and cameras), one would like to know if someone is speaking and where that person is. However, such a system does not perform speaker
identification and thus is unable to answer the question of “who spoke when”. Furthermore, it requires a structured microphone array and an omni-directional camera system. In contrast, our proposed system makes use of a single far-field microphone and any collection of available uncalibrated cameras.

The most related class of works would be those that investigate the problem of interest, i.e. multi-modal speaker diarization [5], who used meeting data with non-frontal faces and full body motion was captured. Vajaria et al. uses an agglomerative clustering approach that is very similar to ours except that they concatenate the audio and video features together as an early-fusion approach; incidentally, they found that the combination of audio and video does not improve their diarization performance when compared to using audio or video alone. In contrast, by modeling audio and video features separately, we are able to improve upon audio-only speaker diarization when video features are also used. In addition, Vajaria et al. tested on data with two people while we test on 4-person meetings.

A common weakness of many of the above works is that the data used, in particular the recorded video, is highly constrained in terms of conversation flow and actions that subjects can take. In addition, with the exception of the CUAVE corpus used by Nock et al. [2], the data used are non-standard and thus not amenable to comparisons. In this paper, we evaluate our approach on over 4.5 hours of publicly available data [6] capturing five different exclusive groups of four individuals in a meeting scenario where participants could behave naturally.

3. AUDIO-ONLY SPEAKER DIARIZATION

A typical speaker diarization system conceptually performs three tasks: first, discriminate between speech and non-speech regions; second, detect speaker changes to segment the audio data; and third, group the segmented regions together into speaker-homogeneous clusters. For the experiments in this paper, we use the ICSI Speaker Diarization Engine [7], which takes an agglomerative clustering approach to perform both segmentation of the audio track into speaker-homogeneous time segments and grouping of these segments into speaker-homogeneous clusters.

19th-order MFCC features, with a frame size of 10 ms, are extracted from the audio track. As a pre-processing step, a speech/non-speech classifier is used to filter out regions that do not contain speech. The non-speech regions are excluded from the agglomerative clustering.

The algorithm is initialized using $k$ clusters; $k$ is chosen to be much larger than the assumed number of speakers in the audio track. An initial segmentation is generated by randomly partitioning the audio track into $k$ segments of the same length. A Gaussian Mixture Model (GMM) is then trained on each of the $k$ initial segments. As classifications based on 10 ms frames are unreliable, a majority decision rule is applied across consecutive frames. We assume a minimum duration of 2.5 seconds for each speech segment. The algorithm then performs the following loop (see Fig. 1):

**Re-Segmentation:** Decide cluster membership of each minimum duration segment by computing the likelihood of belonging to each GMM.

**Re-Training:** Given the new segmentation of the audio track, train a new GMM for each segment.

**Cluster Merging:** Given the new GMMs, try to find the two models that most likely represent the same speaker. This is done by going over all possible pairs of clusters, and computing the difference between the sum of the Bayesian Information Criterion (BIC) scores of each of the models and the BIC score of a new GMM trained on the merged cluster pair. The clusters from the pair with the largest positive difference are merged, the new GMM is used and the algorithm repeats from the re-segmentation step. If no pair with a positive difference is found, then the algorithm stops.

A more detailed description can be found in [7].

4. COMPRESSED DOMAIN VIDEO FEATURES

To provide video features for speaker diarization, we use frame-based visual activity features that have been shown to correlate well with speaking activity patterns [8]. In particular, we use the motion vector magnitude (see Fig. 2) to construct an estimate of personal activity levels [9] as follows.

For each frame, the average motion vector magnitude over estimated skin blocks is calculated and used as a measure of individual visual activity for a camera view. Note that the averaging over estimated skin blocks is done to reduce the effect of background clutter and mitigate pose and scale variations. To detect skin blocks, we implement a block-level skin-color detector working mostly in the compressed domain (see Fig. 2). A GMM is used to model the distribution of $(U, V)$ chrominance coefficients of skin-tone [10] in the YUV

---

**Fig. 1.** The agglomerative clustering approach of the ICSI Diarization Engine is explained in Section 3. Retraining and re-segmentation ends when no more models can be merged as of the BIC score.

**Fig. 2.** Compressed-domain video features: From left, original, detected skin-color blocks, motion vectors.
the contributions of each feature stream. In the extreme case where $\alpha = 0$, video features do not play a role.

6. EXPERIMENTS

6.1. Data

To evaluate the performance of our proposed approach, we used a subset of 12 meetings from the Augmented Multi-Party Interaction (AMI) corpus [6]. This subset contains the most comprehensively annotated meetings in the corpus, therefore it was also preferred by many other authors and related work, for example [8]. The AMI corpus consists of audio-visual data captured of four participants in a natural meeting scenario. The participants volunteered their time freely and were assigned roles such as “project manager” or “marketing director” for the task of designing a new remote control device. The teams met over several sessions of varying lengths (15-35 minutes). The meetings were not scripted and different activities were carried out such as presenting at a slide screen, explaining concepts on a whiteboard or discussing while sitting around a table.

Data was collected in an instrumented meeting room (see Fig. 3), which contains a table, slide screen, white board and four chairs. While participants were requested to return to the same seat during the same recording, they could move freely during the meeting. Different audio and video sources recorded from mounted microphones and cameras on the table, and lapel microphones, or headsets were used to represent increasingly noisy versions of the audio-signal for robustness testing. For the experiments presented here, we used audio data from a single far-field microphone, and video data recorded from four cameras mounted on the center of the table. Fig. 4 shows some of the many different possible poses and poses in example snap-shots of the participants during the meetings.

6.2. Evaluation method

The output of a speaker diarization system consists of metadata describing speech segments with starting time, ending time, and speaker cluster name. It is evaluated against manually annotated ground truth. A dynamic programming procedure is used to find the optimal one-to-one mapping between the hypothesis and the ground truth segments so that the total overlap between the reference speaker and the correspond-
This paper presents an approach for performing multi-modal speaker diarization even further.

We believe that the speaker diarization performance of our multi-modal system could be further improved in a number of ways. First, a dynamic texture classifier could be used to identify motion events that are not of typical human behavior and thus not indicative of speaking turns. Second, more sophisticated action classification can be used to detect visual cues used for turn taking, such as nodding of a speaker’s head. Finally, we could modify the visual activity features to take into account when people are not seated, which should lead to a better correlation with the speech. Moreover, a wealth of statistical feature and model combination techniques could be tried to improve speaker clustering even further.

8. REFERENCES