# Visual Speaker Localization Aided by Acoustic Models

Gerald Friedland
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA
fractor@icsi.berkeley.edu

Chuohao Yeo
University of California Berkeley
Dept. of EECS
Berkeley, CA 94720, USA
zuohao@eecs.berkeley.edu

Hayley Hung
IDIAP Research Institute
P.O. Box 592
CH-1920 Martigny
Switzerland
hhung@idiap.ch

## ABSTRACT

The following paper presents a novel audio-visual approach for unsupervised speaker locationing. Using recordings from a single, low-resolution room overview camera and a single far-field microphone, a state-of-the art audio-only speaker localization system (traditionally called speaker diarization) is extended so that both acoustic and visual models are estimated as part of a joint unsupervised optimization problem. The speaker diarization system first automatically determines the number of speakers and estimates "who spoke when", then, in a second step, the visual models are used to infer the location of the speakers in the video. The experiments were performed on real-world meetings using 4.5 hours of the publicly available AMI meeting corpus. The proposed system is able to exploit audio-visual integration to not only improve the accuracy of a state-of-the-art (audio-only) speaker diarization, but also adds visual speaker locationing at little incremental engineering and computation costs.

## Categories and Subject Descriptors

H5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Signal analysis, synthesis, and processing*; H.4.3 [**Information Systems Applications**]: Communications ApplicationsComputer conferencing, teleconferencing, and videoconferencing

## General Terms

Multimodal Integration

## Keywords

speech, visual localization, speaker diarization, multimodal integration

## 1. INTRODUCTION

Research in cognitive psychology suggests that the human brain is able to integrate different sensory modalities, such as sight, sound, and touch, into a perceptual experience that is coherent and unified [13]. Experiments show that by considering input from multiple sensors, perceptual problems can be solved more robustly and even faster [7]. In computer science, however, synergistic use of data encoded for different human sensors has not yet lived up to its promise.

The following article presents a system where different modalities are combined to jointly tackle two problems, each traditionally solved using a single modality. The combination of the two modalities leads to a higher robustness than a current state-of-the-art audio speaker diarization, which aims to answer the question "who spoke when". Furthermore, the visual models and the output of the speaker diarization allows for a bi-modal localization of the speakers in the video ("where is the speaker?"). We view this system as a successful example of multimodal integration in computer science: a unimodal state-of-the art system gains improvements in accuracy and extends its capabilities by adopting an additional mobility without requiring either a fundamental redesign of an existing algorithm or significantly increasing its computational complexity.

For portability, low cost and ease of deployment, we have designed our system to require as inputs only audio from a single microphone and video from a low-resolution web camera. We used an annotated dataset that contains 4.5 hours of real-world meetings for evaluations; our proposed system only uses a single far-field audio channel and a single camera view with a resolution of $352 \times 288$ pixels. We think that portability would especially be important for content analysis of meetings or other events that are captured using a web cam. The algorithm presented here has many uses as a front-end processing step for other high-level analysis tasks, such as behavioral analysis (eg. dominance estimation [12, 10]).

In this paper, we present our proposed algorithm, discuss its properties, and evaluate its performance quantitatively. We first present related work on audio/visual speaker locationing and diarization in Section 2. In Section 3, we describe the meeting dataset that we used for evaluation to provide a context for our system. Section 4 presents the underlying speaker diarization approach, Section 5 discusses the video features that we use and Section 6 describes how multi-modal integration is done. Section 7 then presents the visual locationing approach. In Section 8 we then discuss how the use of video features improves speaker diarization performance, compare it with other alternatives and present a quantitative evaluation of the speaker locationing. We conclude and lay out directions for future work in Section 9.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
*MM'09,* October 19–24, 2009, Beijing, China
Copyright 2009 ACM 978-1-60558-608-3/09/10 ...$5.00.

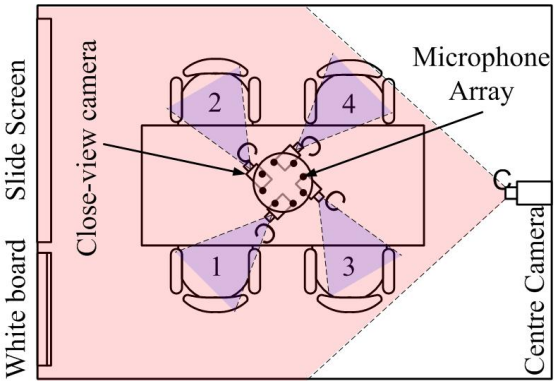


**Figure 1: Plan view of the meeting room set up. For the experiments presented in this article, data from a single audio-channel of the microphone array and the rear-video camera were combined.**

## 2. RELATED WORK

Audio speaker diarization is the task of finding "who spoke when", as defined and proposed by Reynolds et al. [21]. It can involve a single audio source or multiple audio sources. Single source solutions rely heavily on accumulating good models for each speaker and are location independent. When multiple sources are available, beamforming-based solutions can be used to enhance and localize the sound source, leading to the possibility of jointly identifying speakers and estimating their locations. However, this still only provides audio information about who is speaking.

Common approaches to audio-visual speaker identification involve identifying lip motion from frontal faces [16], [17], [3], [5], [4], [20], [22], [23]. Therefore, the underlying assumption is that motion from a person comes predominantly from the motion of the lower half of their face. In addition, gestural or other non-verbal behaviors associated with natural body motion during conversations are artificially suppressed e.g. for the CUAVE database [19]. Most of the techniques have involved identifying one or two people in a single video camera only where short term synchrony of lip motion and speech are the basis for audio-visual localization. In a real scenario the subject behavior is not controlled and, consequently, the correct detection of the mouth is not always feasible. Here audio and visual modalities are combined.

Nock et al. [16] presents an empirical study to review definitions of audio-visual synchrony and examine their empirical behavior. The results provide justifications for the application of audio-visual synchrony techniques to the problem of active speaker localization in broadcast video. Zhang et al. [28] 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 practical. Noulas et al. [17] integrated audio-visual features for on-line audio-visual speaker diarization using a dynamic Bayesian network (DBN) but tests were limited to discussions with two to three people on just two short test scenarios. Tamura et al. [23] 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).


**Figure 2: Two frames from the meeting video corpus which was used for the experiments. The meeting participant were free to move in the room. The faces are hard to detect, as in a natural scenario participants are rarely looking frontally into the camera. The frame on the right shows a partial occlusion of the fourth participant.**

It is important to note that in conversational scenarios, even if we cannot detect mouth motion directly, other forms of body behavior e.g. head gestures are also visible manifestations of speech[15]. While there has been relatively little work on using a person's global body movements for inferring speaking status, some studies have been carried out by Vajaria et al. [24, 25], Hung et al. [9, 11], and Campbell and Suzuki [1]. These approaches, however, have never assumed audio/visual diarization as a single, unsupervised joint optimization problem. This was achieved recently by Friedland et al. [6] but this study was performed using multiple close-view personal cameras. This article extends the approach by only using a single, low-resolution overview camera and, most importantly, by presenting the idea of visual speaker locationing as a by-product. We also test on meeting scenarios where the participants are able to move around freely in the room.

## 3. AMI MEETING CORPUS

In this paper, we use a subset of 12 meetings (4.5 hours) from the Augmented Multi-Party Interaction (AMI) corpus [2]. This subset contains the most comprehensively annotated meetings in the corpus, and is preferable since it allows for the quantitative evaluation of meetings analysis algorithms and the comparison of different approaches to the same task on a common dataset. Thus, it is widely used, for example by [12].

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. The participants therefore interacted naturally, including talking over each other.

Data was collected in an instrumented meeting room (see Figure 1), which contains a table, slide screen, white board and four chairs. While participants were requested to return to the same seat for the duration of a meeting session, they could move freely throughout the meeting. Different audio sources of varying distance to the speaker, and different

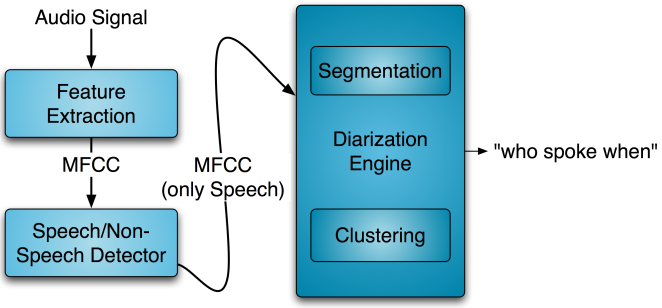


**Figure 3: Block diagram illustrating the traditional speaker diarization approach: As described in Section 4, an agglomerative clustering approach combines speaker segmentation and clustering in one step.**

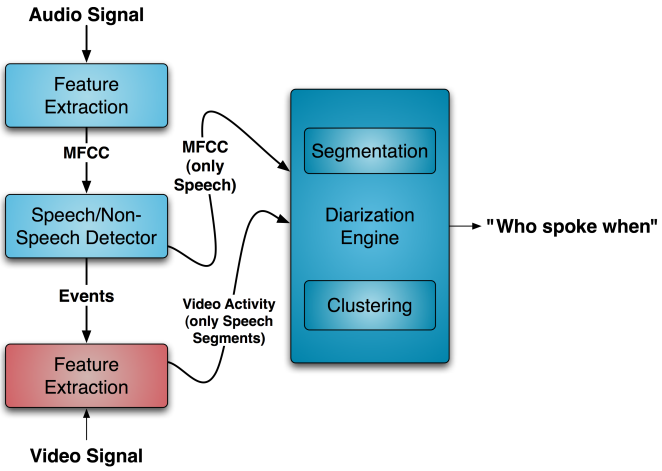


**Figure 4: Block diagram illustrating the extension of the traditional speaker diarization approach: The video activity vector is calculated on the speech segments and integrated into the speaker diarization process as described in Section 6.**

video sources of varying views and fields-of-view represent audio-visual data of varying quality which is useful for robustness testing.

As mentioned earlier, we wish for our system to be portable, low-cost, and easy to deploy. Therefore, it must be able to function using just a single-microphone input and a low-resolution web camera. The system is tested using a single far-field audio channel from the microphone array and a scaled-down image of the overview camera (352 × 288 pixels). The close-up cameras and the microphone array were only used for comparison purposes, as described in Section 6.

Figure 2 shows some sample snap-shots of the meeting recordings by the overview camera and points out some of the limitations: e.g. (i) the faces are mostly too small to be tracked by off-the-shelf face detectors; (ii) people walk around and also lean backwards and forwards, thus changing their appearance drastically; and (iii) participants are also sometimes occluded.

# 4. AUDIO SPEAKER DIARIZATION

The following section outlines the traditional audio-only speaker diarization approach. We use a state-of-the-art diarization engine [26] that performed very well in the last NIST RT evaluations.

## 4.1 Feature Extraction

Wiener filtering is first performed on the audio channel for noise reduction. The HTK toolkit[1] is used to convert the audio stream into 19-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) which are used as features for diarization. A frame period of 10 ms with an analysis window of 30 ms is used in the feature extraction.

## 4.2 Speech/Non-Speech Detection

The speech/non-speech segmentation [8] proceeds in three steps. At each step, feature vectors consisting of 12 MFCC components, their deltas and delta-deltas, and zero-crossings are used.

In the first step, an initial segmentation is created by running the Viterbi algorithm on a Hidden Markov Model (HMM) with Gaussian Mixture Model (GMM) emissions that have been trained on Dutch broadcast news data to segment speech and silence. In the second step, the non-speech regions are split into two clusters: regions with low energy and regions with high energy. A new and separate GMM is then trained on each of the two new clusters and on the speech region. The number of Gaussians used in the GMM is increased iteratively and re-segmentation is performed in each iteration. The model that is trained on audio with high energy levels is added to the non-speech model to capture non-speech-like sounds such as music, sound effects, slamming doors, paper rustling, etc. In the final step, the speech model is compared to all other models using the Bayesian Information Criterion (BIC). If the BIC score is positive, the models are considered to be trained on speech data. In this case, a new speech model is trained on the data of both speech and sound regions and the original model is discarded.

## 4.3 Speaker Segmentation and Clustering

In the segmentation and clustering stage of speaker diarization, an initial segmentation is generated by randomly partitioning the audio track into $k$ segments of the same length. $k$ is chosen to be much larger than the assumed number of speakers in the audio track. For meetings data, we use $k = 16$. The procedure for segmenting the audio data takes the following steps:

1. Train a set of GMMs for each initial cluster.
2. Re-segmentation: Run a Viterbi decoder using the current set of GMMs to segment the audio track.
3. Re-training: Retrain the models using the current segmentation as input.
4. Select the closest pair of clusters and merge them. 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

[1] http://htk.eng.cam.ac.uk/

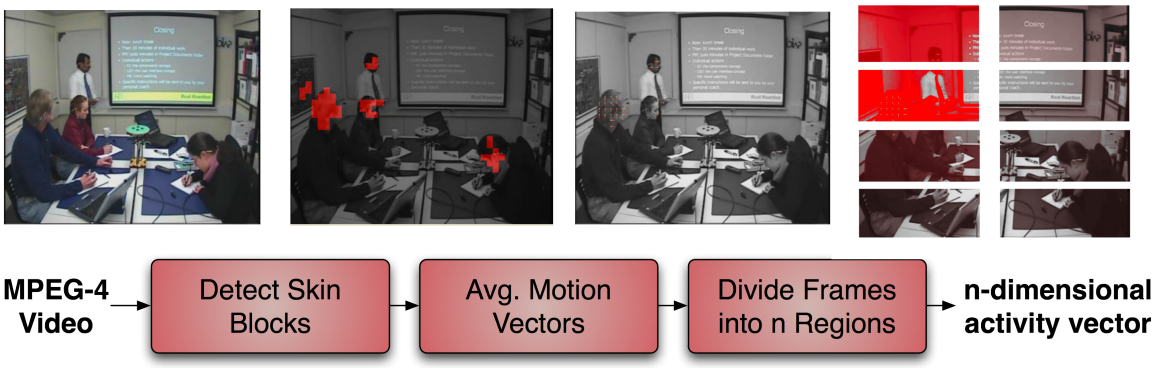


**Figure 5: Compressed-domain video features: Original, detected skin-color blocks, motion vectors, averaged motion vectors per region.**

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.

5. If no pair with a positive difference is found, the algorithm stops, otherwise the algorithm repeats from step 2.

A more detailed description can be found in [26].

The result of the algorithm consist of a segmentation of the audio track with $n$ clusters and an audio GMM for each cluster, where $n$ is assumed to be the number of speakers.

### 4.4 Using multiple audio streams

The algorithm can be slightly modified to use multiple audio tracks as input (presumably from a far-field microphone array). Beamforming is first performed as a pre-processing step[2] to produce a single noise-reduced audio stream from the multiple audio channels by using a delay and sum algorithm. In addition, as part of its processing, beamforming also estimates time-delay-of-arrival (TDOA) between each microphone and a reference microphone in the array. The TDOA features contain information about the location of the audio source, and can be used as an additional feature in the clustering system. Separate GMM models are estimated from these TDOA features. In the Viterbi decoding and in the BIC comparison, a weighted combination of the MFCC and TDOA likelihoods is used. We will be using a similar mechanism for audio/visual integration (see Section 6).

## 5. VIDEO FEATURES

There has been evidence in literature (see Section 2) to suggest that body movement correlates with speech activity of a person. To provide video features for speaker diarization, we use frame-based visual activity features which can be efficiently extracted from compressed videos as indicated in [12]. In particular, we use block motion vector magnitude obtained from the compressed video bitstream as proposed by [27] (see Figure 5) to construct an estimate of personal activity levels as follows.

[2]In our work, we used BeamformIt, an open-source software to perform beamforming. See: http://www.xavieranguera.com/beamformit/

Each video frame is gridded into 4 × 2 non-overlapping subframes of equal size (see Figure 5). While we also experimented with other partitioning schemes, we found this to work the best. In each of the 8 subframes, the average motion vector magnitude over detected skin-color blocks is calculated and used as a measure of individual visual activity for that subframe. Note that the averaging over estimated skin blocks is done to reduce the effect of background clutter and mitigate pose and scale variations. These values from all subframes are averaged over 400 ms and stacked into an 8-dimensional vector. They are used as the video feature vector for all frames in the 400 ms region.

To detect skin blocks, we implement a block-level skin-color detector working mostly in the compressed domain (see Figure 5). A GMM is used to model the distribution of $(U, V)$ chrominance coefficients of skin-tone in the YUV colorspace [14], where each Gaussian component is assumed to have a diagonal covariance matrix. In the Intra-frames, we compute the likelihood of observed chrominance DCT DC coefficients according to the GMM and threshold it to detect skin-color blocks. Skin blocks in the Inter-frames are inferred by using motion vector information to propagate skin-color blocks through the duration of the group-of-picture (GOP).

Motion vectors and DCT coefficients are block-based and already computed during video compression. Compared to extracting higher resolution pixel-based features such as optical flow, compressed domain features are much faster to extract, with a run-time reduction of up to 95% [27].

## 6. MULTIMODAL INTEGRATION

As discussed earlier, audio features are extracted using a window of 10 ms while video features are extracted using a window of 400 ms. For the purpose of multi-modal integration, we duplicate the video features for each 10 ms audio frame within the corresponding 400 ms video analysis window.

The approach we chose for combining the compressed-domain video features and MFCC audio features is similar to the one proposed by Pardo et al [18] for acoustic feature integration. During every agglomerative clustering iteration (see Section 4), each speaker cluster is modeled by two GMMs, one for the audio MFCC features and one for the video activity features, where the number of mixture components varies for each feature stream. We determined experimen-

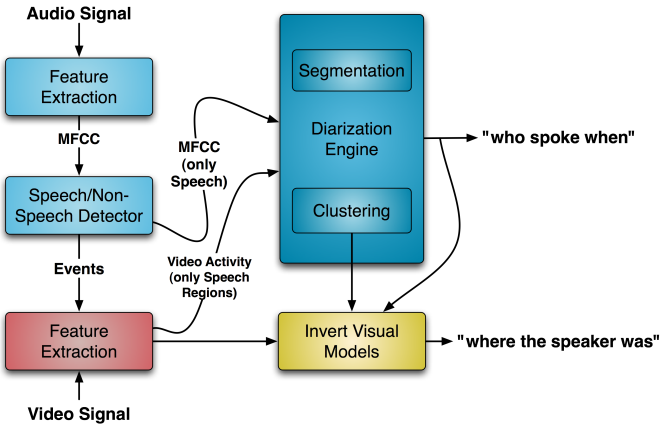


**Figure 6: Schematic of the system presented in this article. As discussed in Section 7, a second pass is added to the multimodal diarization engine that enables the inference of the locations of the speakers.**

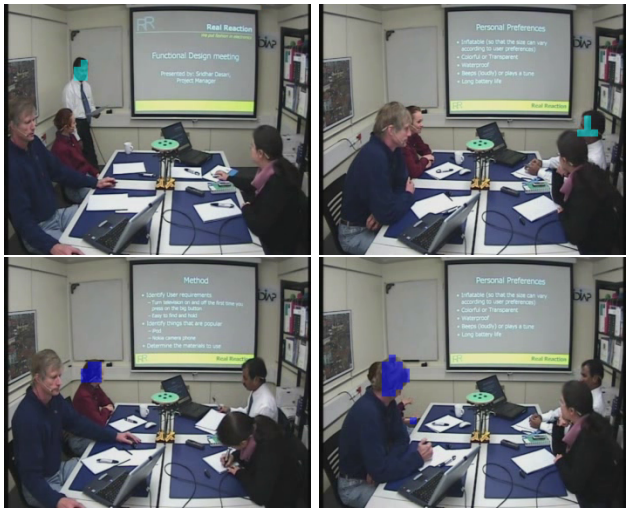

**Figure 7: The result of the visual locationing step: Speakers are identified using different colors and their movements are highlighted when they talk. Speakers may be located even when they are partially occluded (see row below).**

tally that 5 Gaussian components for the audio data and 2 Gaussian components for the video data give the best results. We assume that the two sets of features are conditionally independent given a speaker. In the segmentation step (which uses Viterbi decoding) and in the merging step (which compares BIC scores), we use a weighted sum of the log-likelihood scores of the two models. In other words, the combined log-likelihood score of the audio-visual observation for a particular frame is defined as:

$$\log p(x_{MFCC}, x_{VID}|\theta_i) \doteq$$
$$(1-\alpha)\log p(x_{MFCC}|\theta_{i,MFCC}) + \alpha \log p(x_{VID}|\theta_{i,VID}) \quad (1)$$

where $x_{MFCC}$ is the 19-dimensional MFCC vector, $x_{VID}$ is the 8-dimensional visual activity feature vector, $\theta_{i,MFCC}$ denotes the parameters of a GMM trained on MFCC features of cluster $i$, and $\theta_{i,VID}$ denotes the parameters of a GMM trained on video features of cluster $i$. $\alpha$ is a parameter that is used to weigh the contributions of each feature stream. In the extreme case where $\alpha = 0$, video features would not play a role. Experimentally we found $\alpha = 0.1$ to provide the best results. Note that due to differently scaled log-likelihood values for audio and video, the weighting does not necessarily imply a priority of one modality over the other. Section 8 presents detailed results.

It is possible to treat audio and video data as part of the same optimization problem to help improve the diarization task. The system has been submitted as part of the NIST Rich Transcription evaluation (multimodal condition) and is currently being evaluated on the NIST benchmark data. However, the combined training of audio and video models allows for more than just improved accuracy. In addition to these quantitative improvement, the next section will present qualitative improvements that cannot be achieved by adding more audio channels or features.

## 7. VISUAL LOCATIONING

Before describing the visual locationing procedure, let us recall how multi-modal speaker diarization is done. The audio and video features are used to create visual and acoustic models $\theta_{i,VID}$ and $\theta_{i,MFCC}$ for each speaker cluster. The classification is then performed by calculating the combined log-likelihoods as given by Equation (1). In other words, for each frame,

$$speaker = \quad argmax_i \; p(x_{MFCC}, x_{VID}|\theta_{i,MFCC}, \theta_{i,VID}) \quad (2)$$

In the audio stream, the log-likelihoods are computed based on the cepstral features; in the video stream the log-likelihoods are computed based on the average activity values in one of the 8 regions in the video. As we see from Section 8, using both $x_{MFCC}$ and $x_{VID}$ gives better speaker diarization performance than just using $x_{MFCC}$ alone.

Now that audio and video models are given and one can calculate an estimate of the current speaker, it is also possible to infer the location of the current speaker in the video. This is done by performing a second processing pass over the video (Figure 6 illustrates the idea). In this second pass over the video, the likelihood for each subframe of belonging to the current speaker is computed using the learned visual GMMs $\theta_{i,VID}$. The detected skin-color blocks that are in the subframe with highest likelihood of belonging to the active speaker are tagged for visualization or further processing. Figure 7 shows some sample frames where different speakers are marked using different colors. We use a region growing approach to compensate for faces and hands crossing subframe borders.

In other words, given the current speaker *speaker*, the visual models $\theta_{speaker,VID}$ for the current speaker, we first find the subframe with the highest likelihood of being occupied by the current speaker using:

$$location(speaker) = \quad argmax_j \; p(x_{VID}(j)|\theta_{speaker,VID}(j)) \quad (3)$$

where $x_{VID}(j)$ refers to the visual activity of the $j$th subframe, and $\theta_{speaker,VID}(j)$ refers to the visual model of the $j$th subframe (with some abuse of notation). All detected

| Meeting ID | Audio-only | Multi-Modal | Relative Δ |
|---|---|---|---|
| IS1000a | 42.40 % | 31.82 % | 24.95 % |
| IS1001a | 39.40 % | 35.40 % | 11.26 % |
| IS1001b | 35.50 % | 35.75 % | −0.70, % |
| IS1001c | 30.40 % | 26.91 % | 11.48 % |
| IS1003b | 31.40 % | 16.87 % | 46.27 % |
| IS1003d | 56.50 % | 52.93 % | 6.31 % |
| IS1006b | 24.10 % | 16.29 % | 32.40 % |
| IS1006d | 60.40 % | 58.68 % | 2.84 % |
| IS1008a | 8.20 % | 4.57 % | 44.26 % |
| IS1008b | 10.10 % | 7.44 % | 26.33 % |
| IS1008c | 14.40 % | 12.74 % | 11.52 % |
| IS1008d | 32.30 % | 30.84 % | 4.52 % |
| Average | 32.09 % | 27.52 % | 14.14 % |

**Table 1: Per-Meeting comparison of the Diarization Error Rate (DER) for audio-only diarization (baseline) and the proposed multi-modal system. The DER contains a total of 12.20 % Speech/Non-Speech Error for both cases.**

skin-color blocks in subframe *location*(*speaker*) are then tagged as belonging to the current speaker. Since we use a diagonal-only covariance matrix in the video models, and given that the models were obtained without external training data, this step enables a completely unsupervised diarization and locationing of the speakers in a video. The runtime of the approach is about 0.1 ×realtime.

This "visual locationing using acoustic models" example shows that the proper integration of acoustic and visual data can lead to new synergistic effects: not only was the accuracy of the diarization improved but a new capability was added to the system at very little engineering cost.

# 8. QUANTITATIVE EVALUATION

## 8.1 Diarization Performance Improvements

The output of a speaker diarization system consists of meta-data describing speech segments in terms of starting time, ending time, and speaker cluster name. This output is usually evaluated against manually-annotated ground truth segments. 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 corresponding mapped hypothesized speaker cluster is maximized. The difference is expressed as Diarization Error Rate, which is defined by NIST[3]. The Diarization Error Rate (DER) can be decomposed into two components: Speech/non-speech error (speaker in reference, but not in hypothesis or speaker in hypothesis, but not in reference), and speaker errors (mapped reference is not the same as hypothesized speaker).

The Speaker Diarization System used for these experiments has competed in the NIST evaluations of the past several years and established itself well among state-of-the-art systems[4]. In order to evaluate the multimodal approach we scored it using the NIST scoring tools and compared it against other common testing conditions.

The baseline single-distant microphone system, as presented in the NIST RT07 evaluation, results in a DER of 32.09 %. The multimodal system as presented here, results in an accuracy improvement of 14 % relative in DER. Table 1 present the results of the multimodal clustering in comparison to an audio-only clustering for each meeting in the experiment.

| | 1 ch | 1 ch/1 cam | 8 ch | 1 ch/4 cams |
|---|---|---|---|---|
| DER | 32.09 % | 27.52 % | 27.55 % | 24.00 % |
| Relative | baseline | 14 % | 14 % | 25 % |
| Speed | 1.0 | 1.4 | 2.2 | 1.3 |

**Table 2: Comparison of the Diarization Error Rate (DER) for audio-only diarization (baseline) and different multistream systems. The DER contains a total of 12.20 % Speech/Non-Speech Error for all cases.**

In order to test the influence of the audio and video channels we ran two contrast conditions with the same engine, which are summarized in Table 2.

First, we tested how the engine would perform if, instead of audiovisual integration, we use all 8 microphone channels from the microphone array as described in Section 4.4. The accuracy of this approach is about the same as the audio/visual approach. However, the runtime is worse and requiring a microphone array instead of a camera is a much higher usability hurdle.

Adding further cameras, however, results in an about 25 % relative improvement compared to the baseline. For this experiment, we used the four closeup cameras in the meeting room and calculated the features as described in Section 5. However, instead of partitioning the video frames into 8 regions, we used the motion vectors of the entire camera frame (thus using a 4-dimensional feature vector instead of an 8-dimensional feature vector). The approach is described in [6].

## 8.2 Evaluation of Visual Locationing

The accuracy of the system presented in this article depends on the following four factors:

- the speech/non-speech error,
- the initial estimation of the speaker,
- the accuracy of the sub-frame assignment,
- and the accuracy of the skin-patch detection

Unfortunately, there is no publicly available dataset that allows for the exact evaluation of visual speaker locationing and speaker diarization at once. In addition, we believe that the skin patch detection might not be required in future systems. Therefore we concentrated on evaluating the visual locationing step by evaluating the correct sub-region assignment from the estimated speaker. The approach is described as follows.

First we annotated the mapping between subframes and speakers for each meeting. A speaker is considered to be in a subframe if his or her face stays in it during the meeting for more than 5 seconds. This enables us to list the subframes occupied by each speaker during the meeting. Clearly, if the lists of subframes per speaker are disjoint the accuracy is the highest. The more spatial ambiguity, the lower the accuracy. We found that in all 12 meetings, the most-used

[3]http://nist.gov/speech/tests/rt/rt2004/fall

[4]NIST rules prohibit publication of any rankings. Please refer to the NIST website for further information: http://www.itl.nist.gov/iad/mig/tests/rt/

| Meeting ID | DER | LER |
|---|---|---|
| IS1000a | 31.82 % | 11.52 % |
| IS1001a | 35.40 % | 4.04 % |
| IS1001b | 35.75 % | 2.29 % |
| IS1001c | 26.91 % | 35.58 % |
| IS1003b | 16.87 % | 51.30 % |
| IS1003d | 52.93 % | 28.96 % |
| IS1006b | 16.29 % | 32.47 % |
| IS1006d | 58.68 % | 3.43 % |
| IS1008a | 4.57 % | 28.03 % |
| IS1008b | 7.44 % | 37.62 % |
| IS1008c | 12.74 % | 61.60 % |
| IS1008d | 30.84 % | 19.92 % |
| Average | 27.52 % | 29.40 % |

**Table 3: Per-Meeting comparison of the Diarization Error Rate (DER) and the Location Error Rate (LER) for the proposed multimodal locationing system (refer Section 8.2). The DER contains a total of 12.20 % Speech/Non-Speech Error.**

subframe per speaker is always modeled correctly by the system. Given no spatial ambiguity, i.e. all subframes are only occupied by exactly one speaker, the system would therefore always find the right mapping between speaker and location and vice versa in this data set.

In order to get a time-based measurement for how much the spatial ambiguity influences the final results, we define the Locationing Error Rate (LER) as the time a wrong subframe is selected relative to the total meeting time. The error is calculated by finding the location given the estimated speaker and then estimating the speaker given that location (as defined in Section 7). In order for the result to be correct, both speakers must match, otherwise the system is affected by an ambiguity. Table 3 shows the results together with the diarization error for each meeting. The variability of both DER and LER in the results reflects the variability of real-world meetings and the complexity of the task.

# 9. CONCLUSIONS AND FUTURE WORK

This article presents an algorithm that is an example of successful multimodal integration in computer science. By adding two steps to a state-of-the-art audio-only speaker diarization system, not only is the accuracy quantitatively improved, a new feature also adds qualitative improvement of the system. The algorithm as presented here uses very little assumptions and is able to cope with an arbitrary amount of cameras and subframes. The increased computational and engineering cost was kept low by adding computationally efficient features to an existing state-of-the art system.

As a result of training a combined audio and visual model, we found that the locationing algorithm has interesting properties that may not be observed by either audio-only or image-only locationing. They are discussed in the following.

Since speaker diarization is an unsupervised approach, audio-only diarization does not provide a means of identifying speakers beyond clustering numbers. Traditionally, the speaker regions are assigned to real names by performing speaker identification (using externally trained acoustic models) in a second step. Alternatively, speaker diarization might be performed as a supervised approach where the speakers in the audio recordings are known a-priori. While association with real names might be desirable in some cases, this is, of course, not possible without pre-trained models (either acoustic or visual). The audio-visual combination allows for a completely unsupervised approach that associates the cluster numbers to faces, and as such simulates what a human can do with a recording of a meeting of strangers that speak an unknown language.

Supervised or unsupervised visual locationing requires the use of models created from the image part of a video which makes them inherently dependent on the appearance of an object. Most locationing algorithms therefore show significant lack of robustness against unexpected visual changes in a video, such as change in lighting conditions, partial occlusions, total disappearance of the object, etc. Also, inaccurate modeling might results in the indistinguishability of two different objects. Combined audio/visual models are more robust against lighting changes, partial occlusion, or other uni-modal distortions. Figure 7 (bottom right) shows an interesting example: Even though the head is occluding the speaker in the upper left corner, the system still attributes the right location to (occluded) face and hands of the speaker.[5] Of course, if both the voice print and the appearance changes, there is nothing that can be done – even a human would most likely assume a new person.

The most important limit of the approach as presented here is the coarse granularity of the subframes. A larger amount of subframes might help to have fewer ambiguities (i.e. speakers in the same subframes). If we were to work with features at the block-level, we may need to perform more more explicit spatial clustering. However, given the correlation between speech and body motion, it may not be necessary to rely on appearance alone. Exploiting the synchrony of gestural motion with speech may already highlight the head, arm and hand regions of the speaking person. In addition, even if one of the regions of the body is occluded, the speaker may still be identified if parts of their moving body are visible.

The properties of the algorithm presented in this article suggest many ideas that could improve the accuracy and qualities of the system. Especially common challenges in speaker diariazation that seem to be very hard to tackle with audio-only approaches might be addresses multimodally using an extension of the presented method. Examples include the exact discrimination between speech and noise (non-speech), the detection of two or more speakers talking at the same time (overlap), and the detection and proper assignment of very short speech segments (smaller than about 0.5 sec), for example due to backchannels. Other interesting future work includes generalizing the system to work with other acoustic events, in addition to speech, that are correlated with visual features such as motion.

## Acknowledgments

Hayley Hung and Gerald Friedland are supported by the Swiss IM2 project and the EU-funded AMIDA project. Chuohao Yeo is sponsored by A*STAR.

# 10. REFERENCES

[1] N. Campbell and N. Suzuki. Working with Very Sparse Data to Detect Speaker and Listener

[5] This is enabled by the fact that the head is detected by the skin color detector.

Participation in a Meetings Corpus. In *Workshop Programme*, volume 10, May 2006.

[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] J. W. Fisher and T. Darrell. Speaker association with signal-level audiovisual fusion. *IEEE Transactions on Multimedia*, 6(3):406–413, 2004.

[5] J. W. Fisher, 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.

[6] G. Friedland, H. Hung, and C. Yeo. Multi-modal speaker diarization of real-world meetings using compressed-domain video features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, page (to appear), 2009.

[7] M. Hershenson. Reaction time as a measure of intersensory facilitation. *J Exp Psychol*, 63:289–93, 1962.

[8] M. Huijbregts. *Segmentation, Diarization, and Speech Transcription: Surprise Data Unraveled*. PrintPartners Ipskamp, Enschede, The Netherlands, 2008.

[9] H. Hung and G. Friedland. Towards audio-visual on-line diarization of participants in group meetings. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications in conjunction with ECCV*, Marseille, France, October 2008.

[10] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez. Estimating the dominant person in multi-party conversations using speaker diarization strategies. In *International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[11] 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.

[12] H. Hung, Y. Huang, C. Yeo, and D. Gatica-Perez. Correlating audio-visual cues in a dominance estimation framework. In *CVPR Workshop on Human Communicative Behavior Analysis*, 2008.

[13] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–48, 1976.

[14] S. J. McKenna, S. Gong, and Y. Raja. Modelling facial colour and identity with gaussian mixtures. *Pattern Recognition*, 31(12):1883–1892, 1998.

[15] D. McNeill. *Language and Gesture*. Cambridge University Press New York, 2000.

[16] 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.

[17] 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.

[18] J. Pardo, X. Anguera, and C. Wooters. Speaker Diarization For Multiple-Distant-Microphone Meetings Using Several Sources of Information. *IEEE Transactions on Computers*, 56(9):1189, 2007.

[19] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy. CUAVE: A new audio-visual database for multimodal human-computer interface research. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 2017–2020, 2002.

[20] R. Rao and T. Chen. Exploiting audio-visual correlation in coding of talking head sequences. *International Picture Coding Symposium*, March 1996.

[21] D. A. Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. In *Proc. of International Conference on Audio and Speech Signal Processing*, 2005.

[22] 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.

[23] S. Tamura, K. Iwano, and S. FURUI. Multi-Modal Speech Recognition Using Optical-Flow Analysis for Lip Images. *Real World Speech Processing*, 2004.

[24] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi. Audio segmentation and speaker localization in meeting videos. *International Conference on Pattern Recognition, 2006. ICPR 2006. 18th*, 2:1150–1153, 2006.

[25] H. Vajaria, S. Sarkar, and R. Kasturi. Exploring co-occurence between speech and body movement for audio-guided video localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 18:1608–1617, Nov 2008.

[26] C. Wooters and M. Huijbregts. The ICSI RT07s speaker diarization system. In *Proceedings of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop*, 2007.

[27] C. Yeo and K. Ramchandran. Compressed domain video processing of meetings for activity estimation in dominance classification and slide transition detection. Technical Report UCB/EECS-2008-79, EECS Department, University of California, Berkeley, Jun 2008.

[28] 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.