Dialocalization: Acoustic Speaker Diarization and Visual Localization as Joint Optimization Problem

GERALD FRIEDLAND, International Computer Science Institute
CHUOHAO YEO, Institute for Infocomm Research
HAYLEY HUNG, IDIAP Research Institute

The following article presents a novel audio-visual approach for unsupervised speaker localization in both time and space and systematically analyzes its unique properties. Using recordings from a single, low-resolution room overview camera and a single far-field microphone, a state-of-the-art audio-only speaker diarization system (speaker localization in time) 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 speech regions 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. We call this process “dialocalization.” 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 localization at little incremental engineering and computation costs. The combined algorithm has different properties, such as increased robustness, that cannot be observed in algorithms based on single modalities. The article describes the algorithm, presents benchmarking results, explains its properties, and systematically discusses the contributions of each modality.

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 Applications—Computer conferencing, teleconferencing, and videoconferencing

General Terms: Experimentation

Additional Key Words and Phrases: speech, visual localization, speaker diarization, multimodal integration

ACM Reference Format:
DOI = 10.1145/1865106.1865111 http://doi.acm.org/10.1145/1865106.1865111

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

H. Hung and G. Friedland were supported by the Swiss IM-2 and the EU-funded AMIDA project. C. Yeo was supported by an A*STAR fellowship.

Authors’ addresses: G. Friedland, International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704-1198; email: fractor@icsi.berkeley.edu; C. Yeo, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis (South Tower), Singapore 138632; H. Hung, IDIAP Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland.

Permission to make digital or hard copies of part or all 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 show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2010 ACM 1551-6857/2010/11-ART27 $10.00
DOI 10.1145/1865106.1865111
http://doi.acm.org/10.1145/1865106.1865111

ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 6, No. 4, Article 27, Publication date: November 2010.
unified [McGurk and MacDonald 1976]. Experiments show that by considering input from multiple sensors, perceptual problems can be solved more robustly and even faster [Hershenson 1962]. In computer science, however, synergistic use of data encoded for different human sensors has not always lived up to its promise.

We present a system where different modalities are combined to jointly tackle two problems at once, each traditionally solved using a single modality. The problem of acoustic speaker diarization is to determine “who spoke when” given a single or multisource audio track but no other information of any kind (number of speakers, text or language spoken, or amount of speech in the audio). Visual localization is the problem of finding regions of interest (in our case speakers) defined by certain properties (here activity) in an image or video. In the joint optimization problem, which we call “dialocalization,” the algorithm detects speech regions in audio-visual feature space, clusters them according to speaker, and then assigns image regions corresponding to the speakers to the speech segments, that is, finding two-dimensional locations of speakers in the camera image. Different research communities have tackled diarization and localization separately as the solution(s) to both problems have broad applicability in many domains, such as automatic meeting and conference transcription, summarization, and retrieval. From a cognitive standpoint, however, it seems natural to tackle the problem as a joint optimization problem since humans concurrently track events both in time and space.

In this article, we present our algorithm and evaluate its performance quantitatively. In addition to what was published in our prior work [Friedland et al. 2009c], here we also discuss its properties and systematically analyze the contributions of each modality.

Using the example of a practical meeting scenario, this article demonstrates that the combined tracking of speakers in time and space makes sense and also improves the performance of a current state-of-the-art audio-only speaker diarization system (“who spoke when”). The bimodal localization of the speakers in the video (“where is the speaker?”) has higher robustness against partial occlusions than the same approach used only in the visual domain. We therefore view this system as a successful example of synergistic multimodal integration. A major point in this work is the focus on a practical real-world scenario where an off-the-shelf method or a single set of specialized features will not easily solve the problem. 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 × 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 estimating behavioral types (e.g., dominance estimation [Hung et al. 2008; 2010]).

We first present related work on audio or audio-visual speaker localization 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 multimodal integration is done. Section 7 then presents the visual localization approach. In Section 8, we discuss how the use of video features improves speaker diarization performance, compare it with other alternatives and present a quantitative evaluation of the speaker localization. Section 9 then presents an analysis of the contributions of each of the modalities. We conclude with the limits of the approach in Section 10 and lay out directions for future work in Section 11.
2. RELATED WORK

Audio speaker diarization is the task of finding “who spoke when” [Reynolds and Torres Carrasquillo 2005]. 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. It is usually approached as either one-step or a two-step audio clustering. In two-stage speaker diarization approaches, the speaker segmentation step initially detects speaker change points and is essentially a two-way decision problem. At each point, a decision on whether this is a speaker change point or not needs to be made. Subsequently, the speech segments, each of which contains only one speaker, are then clustered using either top-down or bottom-up approaches. A very popular method is the use of Bayesian Information Criterion (BIC) [Chen and Gopalakrishnan 1998] and Gaussian Mixture Models of frame-based cepstral features (see Section 4). Alternatively, many state-of-the-art speaker diarization systems, including the ICSI Speaker Diarization engine, use a one-stage approach—they unify the segmentation and clustering parts into a single step using agglomerative hierarchical clustering. 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 in terms of a unique ID.

Traditional ways of tackling visual localization include methods such as activity detection, object segmentation, tracking, and face detection. Many of these algorithms can be found in the Open CV software package. Areas where work has been done on localization of objects include, but are not limited to, robotics [Simon et al. 2001], surveillance tasks [Haritaoglu et al. 2000], or traffic applications [Beymer et al. 1997]. Numerous algorithms have also been developed in the MPEG-4 research community, for example [Chien et al. 2001]. Last but not least, the authors have worked on localization of instructors in a recorded lecture [Friedland et al. 2007]. Except for a few examples which we will discuss in the next paragraph, none of the methods work on multimodal cues. Also, most localization methods use specialized training to derive models of the objects to localize them. The method presented here learns the visual models in an unsupervised fashion using inference from the audio models which are also trained in an unsupervised way. Supervised methods have the disadvantage of the models being channel sensitive (room reverberation in the audio domain, lighting conditions in the visual domain)

Common approaches to audio-visual speaker identification or localization involve identifying lip motion from frontal faces [Nock et al. 2003; Noulas and Krose 2007; Chen and Rao 1996; Fisher et al. 2000; Fisher and Darrell 2004; Rao and Chen 1996; Siracusa and Fisher 2007], [Tamura et al. 2004]. In these approaches, 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 nonverbal behaviors associated with natural body motion during conversations are artificially suppressed, for example, for the CUAVE database [Patterson et al. 2002]. 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 to carry out robustly.

Nock et al. [2003] present 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.
[2006] presented a multi modal speaker localization method using a specialized satellite microphone and omnidirectional camera. Though the results seem comparable to the state-of-the-art, the solution requires specialized hardware, which is not always practical. Noulas and Krose [2007] integrated audio-visual features for online 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. The Ph.D. thesis by Hospedales [2008] investigates Bayesian modeling approaches for multimodal integration using the example of audio/visual localization. However, the experiments are carried out on a small dataset with two front-facing persons speaking in a laboratory setting. The tracking algorithm presented in Gatica-Perez et al. [2007] is based on multiple cameras and the incorporation of audio input. This study relies on the shape and structure of human heads.

It is important to note that in conversational scenarios, even if we cannot detect mouth motion, faces, or heads directly, other forms of body behavior, for example, head gestures, are also visible manifestations of speech [McNeill 2000]. 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 in Vajaria [2006], Hung and Friedland [2008], and Campbell and Suzuki [2006] that show the feasibility of the idea. The experiments in Vajaria et al. [2008], were the first to consider body movement as an aid for diarization and also speaker localization. Their results comparing the effects of audio and video modalities on their diarization algorithm provided some interesting scenarios in which diarization performance could be affected under differing modalities. They also test on meeting scenarios where the participants are able to move around freely in the room. We build on this preliminary work by providing a more systematic and in-depth study of how differing modalities, source locations and conditions, and also meeting activities and behaviors can affect the dialocalization performance.

These approaches, however, have never assumed audio-visual diarization and localization as a single, unsupervised joint optimization problem. This was achieved recently in [Friedland et al. 2009a] but this study was performed using multiple close-view cameras. The approach was extended in Friedland et al. [2009a] to only use a only single, low-resolution overview camera and, most importantly, to show visual speaker localization as a by-product. This article extends on the work by presenting a more thorough analysis of the properties of the algorithm and discussing the contribution of each modality in a realistic meeting scenario.

3. AMI MEETING CORPUS

The experimental setup for the approach presented here is as follows. We used a subset of 12 meetings (4.5 hours) from the Augmented MultiParty Interaction (AMI) corpus [Carletta et al. 2005]. This subset contains the most comprehensively annotated meetings in the corpus, and is preferable since it allows for the quantitative evaluation of meeting analysis algorithms and the comparison of different approaches to the same task on a common dataset. Thus, it is commonly used by different researchers [Boakye et al. 2008; Ba and Odobez 2009; Garau et al. 2009; Reidsma 2008]. This work investigates an unsupervised approach; therefore, there was no need to split the data into test and training set, which would have made the quantitative results less comparable to the numbers reported by other researchers.

The AMI corpus consists of audio-visual data captured from 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
Fig. 1. Plan view of the meeting room set up. The system presented in this article only needs data from a single audio-channel of the microphone array and the rear-video camera. The close-view cameras and multiple microphones were only used for the comparison experiments presented in Section 9.

Fig. 2. Two frames from the meeting video corpus which was used for the experiments. The meeting participants 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.

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, a 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 from the speaker, and different video sources of varying views and fields-of-view were recorded and represent audio-visual data of varying quality.

As mentioned earlier, we wish 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 snapshots of the meeting recordings by the overview camera and points out some of the limitations: for example, (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 [Wooters and Huijbregts 2007] that performed very well in the 2007 and 2009 NIST Rich Transcription evaluations.

4.1 Feature Extraction

Wiener filtering [Adami et al. 2002] is first performed on the audio channel for noise reduction. The HTK toolkit\(^3\) is used to convert the audio stream into 19-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) [Mermelstein 1976] 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

We use the same speech/nonspeech segmentation as in Wooters and Huijbregts [2007], which is explained in Huijbregts [2008]. It is an HMM/GMM approach originally trained on broadcast news data that generalizes well to meetings.

4.3 Speaker Segmentation and Clustering

In the segmentation and clustering stage of speaker diarization, an initial segmentation is generated by uniformly 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. Resegmentation: Run a Viterbi decoder using the current set of GMMs to segment the audio track.
3. Retraining: Retrain the models using the current segmentation as input.

\(^3\)http://htk.eng.cam.ac.uk/
(4) Select the closest pair of clusters and merge them. At each iteration, the algorithm checks all possible pairs of clusters to see if there is an improvement in BIC scores when the clusters are merged and the two models replaced by a new GMM trained on the merged cluster pair. The clusters from the pair with the largest improvement in BIC scores, if any, are merged and the new GMM is used. The algorithm then repeats from the resegmentation step until there are no remaining pairs that when merged will lead to an improved BIC score.

A more detailed description can be found in Wooters and Huijbregts [2007].

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\(^4\) 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 the literature (see Section 2) to suggest that body movement correlates with speech activity of a person. At the same time, one can assume that in a meeting, sounds sources are mostly speakers. Therefore, to provide video features for speaker diarization, we use subframe-based visual activity features which can be efficiently extracted from compressed videos as indicated in Hung and Friedland [2008]. In particular, we use block motion vector [Richardson 2003] magnitude obtained from the compressed video bitstream as proposed by Yeo and Ramchandran [2008] (see Figure 4) to construct an estimate of personal activity levels as follows.

Each video frame is gridded into \( 4 \times 2 \) nonoverlapping subframes of equal size (see Figure 4). While we also experimented with other partitioning schemes, we found this to work the best. In each of the 8

\(^4\)In our work, we used BeamformIt, an open-source software to perform beamforming. See: http://www.xavieranguera.com/beamformit/
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 4). A GMM trained on a separate dataset is used to model the distribution of \((U, V)\) chrominance coefficients of skin tone in the YUV colorspace [McKenna et al. 1998], where each Gaussian component is assumed to have a diagonal covariance matrix. In the Intraframes, we compute the likelihood of observed chrominance DCT DC coefficients [Richardson 2003] 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% [Yeo and Ramchandran 2008].

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. [2007] 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 experimentally 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:

$$\log p(x_{MFCC}, x_{VID}|\theta_i) = (1 - \alpha) \log p(x_{MFCC}|\theta_i, MFCC) + \alpha \log p(x_{VID}|\theta_i, VID).$$

(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 \(i\).

The parameter \(\alpha\) is used to weigh the contributions of each feature stream. In the extreme case where \(\alpha = 0\), video features would not play a role. Using inverse-entropy weighting [Misra et al. 2003] and supervised tuning with the Nelder-Mead method [Lagarias et al. 1999] as a control, we found \(\alpha = 0.1\) to provide the best results. Figure 7 in Appendix A shows a plot of different values for \(\alpha\) correlated with the Diarization Error Rate (see Section 8). We kept the value for \(\alpha\) constant throughout all the experiments presented in this article and also for the generalization experiment presented in Table XI shown in Appendix A. We conclude that changes to the value of the \(\alpha\)-parameter inside a certain range have only little influence on the results.

ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 6, No. 4, Article 27, Publication date: November 2010.
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 SPEAKER LOCALIZATION

Before describing the visual localization method, 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, $\text{speaker} = \text{arg} \max_i p(x_{MFCC}, x_{VID}|\theta_i,_{MFCC}, \theta_i,_{VID})$. (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 each of the eight regions in the video. As we shall 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 5 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
Fig. 6. The result of the visual localization step: Speakers are identified using different colors and their movements are highlighted when they talk. As explained in Section 9.4, speakers may be located even when they are partially occluded (see lower row).

belonging to the active speaker are tagged for visualization or further processing. Figure 6 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 and the visual models for the current speaker \( \theta_{\text{speaker,VID}} \), we first find the subframe with the highest likelihood of being occupied by the current speaker using:

\[
\text{location}(\text{speaker}) = \arg\max_j p(x_{\text{VID}}(j) | \theta_{\text{speaker,VID}}(j)),
\]

where \( x_{\text{VID}}(j) \) refers to the visual activity of the \( j \)th subframe, and \( \theta_{\text{speaker,VID}}(j) \) refers to the visual model of the \( j \)th subframe (with some abuse of notation). Practically, we simulate \( n \) feature vectors \( x_{\text{VID}} = (c_i) \), with \( i = 0..n \), \( n \) equals the number of subframes and \( c_i = 0 \), except for one particular \( c_i \) which is the subframe in question and contains the original compressed domain feature value for the subframe.

All detected skin-color blocks in subframe \( \text{location}(\text{speaker}) \) 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 localization of the speakers in a video. The runtime of the approach is about 0.1 × realtime.
Table I. Comparison of the Diarization Error Rate (DER) for audio-only diarization (baseline) and the proposed multi-modal system. Table VII in Appendix A shows the per-meeting results.

<table>
<thead>
<tr>
<th>Meeting ID</th>
<th>Audio-only</th>
<th>Multimodal</th>
<th>Relative Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>32.09 %</td>
<td>27.52 %</td>
<td>14.14 %</td>
</tr>
</tbody>
</table>

This “visual localization using acoustic models” example shows that the proper integration of acoustic and visual data can lead to new synergistic effects: not only the accuracy of the diarization improved but a new capability is 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 (DER), which is defined by NIST\(^5\). The 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\(^6\). 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 RT ’07 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 I presents the results of the multimodal clustering in comparison to an audio-only clustering for each meeting in the experiment. The average DER contains 12.20% Speech/Non-Speech Error for both cases. Per-meeting results, as well as results on an additional dataset, are shown in Appendix A.

8.2 Evaluation of Visual Localization

The accuracy of the system presented in this article depends on the following four factors: the speech/nonspeech error, the initial estimation of the speaker, the accuracy of the subframe 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 localization 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 localization step by evaluating the correct subregion assignment from the estimated speaker. 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


\(^6\)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/
G. Friedland et al.

Table II. Comparison of the Localization Error Rate (LER) for the proposed multimodal localization system and a face detection baseline system based on OpenCV and frontal-face videos (refer to Section 8.2). Table VIII in Appendix A shows the per-meeting results.

<table>
<thead>
<tr>
<th>Meeting ID</th>
<th>LER (OpenCV)</th>
<th>LER (our approach)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>58.96%</td>
<td>26.4%</td>
</tr>
</tbody>
</table>

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 subframe per speaker is always modeled correctly by the system. Given no spatial ambiguity, that is, all subframes are only occupied by exactly one speaker, the system would 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 Localization Error Rate (LER) as the fraction of 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.

To get an idea how a canonical unimodal localization algorithm would perform, we compared the obtained LER with an approach based on the OpenCV face detector [Baker and Matthews 2004]. This baseline approach (which is described further in Huynh [2008]) uses a face detector to find the faces in the frontal videos of the meeting corpus (also referred to as “side-view cameras”) which, in contrast to the single-camera view used by us, contain almost no occlusion. It then uses the same motion activity features (refer Section 5) to determine if the face is talking by taking the maximum of all measured activities. This is done on a frame-by-frame basis with a smoothing window, the same length (400 ms) as the window used in our approach. Table II shows the results.

The variability of both DER and LER in the results (compare Tables I and II) reflects the variability of real-world meetings and the complexity of the task. Errors in the skin detector, overlapping speech and motion regions, speech segments with no perceivable motion, errors in the speech/nonspeech detection, or occlusions are only examples of the reasons that all contribute to the localization error. Also, applying a different $\alpha$ parameter to individual meetings might change the behavior. However, the philosophy of the speaker diarization tasks forbids for meeting-specific training, we therefore did not investigate this hypothesis further.

9. CONTRIBUTIONS OF EACH MODALITY

In order to further the understanding for the advantages and disadvantages of a multimodal approach, the following section discusses experiments to evaluate the contributions of each modality to the different subtasks. We mainly focus on the diarization error since the localization error might be dependent on it so it may be more difficult to draw conclusions based on the Localization Error Rate alone.

9.1 Multimodal versus Unimodal

In a first test we wanted to know how well the diarization algorithm performed if only the video features were used and how well the localization would work if we only fed audio features into it (assuming constant but random image region assignment for each speaker). Not surprisingly, the localization error rate for an audio-only approach is 78% which can be interpreted as the random baseline. This
Table III. Comparison of the Diarization Error Rate (DER) for audio-only diarization (baseline) and different multistream systems. Table IX in Appendix A shows the per-meeting results.

<table>
<thead>
<tr>
<th>Meeting</th>
<th>1 mic</th>
<th>1 mic/1 cam</th>
<th>8 mic</th>
<th>1 mic/4 cams</th>
<th>8 mic/4 cams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average DER</td>
<td>32.09%</td>
<td>27.52%</td>
<td>27.56%</td>
<td>24.00%</td>
<td>19.52%</td>
</tr>
<tr>
<td>Runtime</td>
<td>1.0</td>
<td>1.4</td>
<td>2.2</td>
<td>1.3</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Table IV. Comparison of the Localization Error Rate (LER) for different multistream systems. Table X in Appendix A shows the per-meeting results.

<table>
<thead>
<tr>
<th>Meeting</th>
<th>1 mic/1 cam</th>
<th>1 mic/4 cams</th>
<th>8 mic/4 cams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average LER</td>
<td>26.45%</td>
<td>26.03%</td>
<td>25.45%</td>
</tr>
</tbody>
</table>

result, however, confirms the validity of the localization approach. The DER for the video only case is 72%. This is a surprisingly bad result since the video features seem to definitely help in combination with audio features. Our interpretation of this result is that the main contribution for the diarization comes from the audio part and that the video features can sometimes provide complementary information. Also, the speaker diarization approach as presented here has not been tuned to accept only video features as input.

9.2 Multimodal versus Multi-Sensor

In order to test the influence of the audio and video channels we ran different contrast conditions, which are summarized in Table III. Again, the DER contains a total of 12.20% Speech/Nonspeech Error for all cases. First, we tested how the engine would perform if, instead of audio-visual integration, we use all eight microphone channels from the microphone array using the combination of MFCC features and time-delay-of-arrival features 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 because correlating 8 channels of audio using beamforming is more complex than our compressed domain video features and, of course, the visual localization is much more challenging with an audio-only system. Adding further cameras, however, results in 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 almost 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 based on one frame from each camera instead of an 8-dimensional feature vector based on 8-subframes from one camera. The 4-camera approach is also described in [Friedland et al. 2009a]. As a last experiment, we tried running the system using all cameras and all microphones. For this experiment, we extended the log-likelihood combination defined in Equation (1) to three streams:

\[
\log p(x_{\text{MFCC}}, x_{\text{DELAY}}, x_{\text{VID}}|\theta_i) = \alpha \log p(x_{\text{MFCC}}|\theta_i, \text{MFCC}) + \beta \log p(x_{\text{DELAY}}|\theta_i, \text{DELAY}) + \gamma \log p(x_{\text{VID}}|\theta_i, \text{VID}),
\]

with \(\alpha + \beta + \gamma = 1\). In analogy to the values found in Section 6, we used \(\alpha = 0.8\) and \(\beta = \gamma = 0.1\). The result surpasses the results of all other combinations. Adding more sensors therefore seems to help in any domain independently which we interpret as an indication for the complementariness of the video and audio features. For practical purposes using less sensors is desirable and the most natural scenario is one camera sensor and one microphone (webcam).

The DER is much lower using four cameras compared to using only one camera and the localization task using four cameras should become much easier. Therefore, we measured the localization error rate for the relevant multi-sensor cases. The results are shown in Table IV but are all about the same,
Table V. Comparison of the Diarization Error Rate (DER) for one MFCC feature stream (baseline) and MFCC plus prosodic features (M+P) as described in [Friedland et al. 2009] versus adding one and four cameras to the combination.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>M+P</th>
<th>M+P+1 cam</th>
<th>M+P+4 cams</th>
</tr>
</thead>
<tbody>
<tr>
<td>DER</td>
<td>32.09%</td>
<td>20.33%</td>
<td>18.98%</td>
<td>16.64%</td>
</tr>
<tr>
<td>Relative</td>
<td>baseline</td>
<td>36%</td>
<td>41%</td>
<td>49%</td>
</tr>
<tr>
<td>Runtime</td>
<td>1.0</td>
<td>2.5</td>
<td>2.9</td>
<td>2.9</td>
</tr>
</tbody>
</table>

which we interpret as an indication that the localization error rate is not mainly determined by the DER but by other factors such as people’s behavior, for example, people may move without talking, as already hypothesized in Section 8.2.

9.3 Multimodal versus Multi-Feature Speaker Diarization

A fundamental counterargument when presenting multimodal algorithms is the critique that more features in the same modality might have helped as well so there is “no need” to investigate another modality. The approach in this article combines MFCC features with video features. It is well known, however, that feature combination approaches, such as as combining MFCCs with long-term audio features improve the accuracy of speaker diarization tremendously [Friedland et al. 2009b]. We therefore repeated the experiments presented in [Friedland et al. 2009b] using the same set of prosodic and long term features. Table V presents the results along with the computational speed factor. Again, the DER contains a total of 12.20 % Speech/Nonspeech Error for all cases. As can be seen, the combination of the ten prosodic and long term features and MFCCs results in a much better error rate than the combination of MFCC features and even 4 cameras. As expected and reported in Friedland et al. [2009b], long-time feature calculation is not as runtime-effective as our compressed domain video features and therefore the runtime cost per improvement is high. Also, using only combined audio-features means that localization is not possible.

Adding one and four cameras to the combination using equation 4 yields further improvement of the DER compared to the baseline: the total improvement is about 50 % relative. We, again, interpret this as an indication for the complementariness of the video domain and audio domain features.

9.4 Properties of the Dialocalization Algorithm

Apart from quantitatively evaluating the presented algorithm, we also manually investigated the quality of the output. This is discussed as follows. We found that the localization algorithm has properties that may not be observed by either audio-only or video-only localization. Supervised or unsupervised visual localization 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 localization 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 result in the indistinguishability of two different objects. Combined audio-visual models are more robust against lighting changes, partial occlusion, or other unimodal distortions. Figure 6 (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. However, this occurred because the head of the occluding speaker is detected by the skin color detector, an occlusion by a different object would not yield this result. Also, 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 different person.

ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 6, No. 4, Article 27, Publication date: November 2010.
Table VI. Robustness of the localization performance against lighting changes. A percentage of the video frames were randomly dimmed.

<table>
<thead>
<tr>
<th>Dimmed frames</th>
<th>0%</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LER</td>
<td>26.40%</td>
<td>31.03%</td>
<td>33.69%</td>
<td>42.38%</td>
<td>68.32%</td>
</tr>
</tbody>
</table>

In order to investigate this property even further, we artificially dimmed different percentages of random frames of the input video of the meeting IS1000a. The brightness of these frames was adjusted by a random number varying from 0% to complete darkness. Table VI shows the Localization Error Rate for different percentages of dimmed frames. As can be seen, the algorithm is fairly robust to changes in lighting up to about 50% of the frames. As a note, the Diarization Error Rate was insignificantly affected even with 75% of the frames being dimmed.

10. LIMITS OF THE APPROACH

The most important limit of the localization approach as presented here is the coarse granularity of the subframes. A larger number of subframes might help to decrease the Localization Error Rate by decreasing the number of ambiguities (i.e., speakers in the same subframes). Of course, this would also increase the complexity of the algorithm. If we were to work with features at the block level, we may need to perform 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.

Another limit of the approach presented here is the implicit dependence on the dataset due to several basic assumptions. Basically, it is assumed that the only moving objects in the video are (parts of) speakers defined by skin color. Also, speakers are mostly stationary. In a dance hall scenario, for example, the algorithm might not at all be able to learn the correlation between movement and sound source. Unfortunately, as of today, the availability of data sets that are annotated for both localization as well as diarization is quite limited. For that reason, further and more thorough investigation of the properties of the algorithm will require annotation of videos from different domains.

Except for better accuracy, we were unable to observe any distinct qualitative improvements to the diarization part of the algorithm. Overlapped speech is still not handled at all and nonlinguistic sounds, such as laughter or coughing, are still causes for errors. Another bottleneck of the approach is the speech/nonspeech detector. Since external noise, such as moving chairs or beeping computers can affect the speaker models, the DER is directly tied to the performance of the speech/nonspeech detection. Furthermore, since movement is assumed to be at the sound source, the error would propagate to the localization part of the algorithm.

11. 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, the visualization also adds qualitative improvement of the system. The algorithm as presented here uses very few assumptions and is able to cope with an arbitrary number of cameras and subframes. The increased computational and engineering cost is kept low by adding computationally efficient features to an existing state-of-the-art system.

Since speaker diarization is an unsupervised approach, audio-only diarization does not provide a means of identifying speakers beyond cluster identifiers. 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, which is widely known as speaker identification. While association with real names might be desirable in some cases, this is, of course, not possible without pretrained 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.

The properties of the algorithm presented in this article suggest many ideas that could improve the accuracy and qualities of the system. Increasing the number of streams by increasing both the number of sensors and the number of extracted features might improve the accuracy even further. Utilizing multiple CPU cores at the same time might help to cope with the increased computational complexity. Common challenges in speaker diarization that seem to be very difficult to tackle with audio-only approaches might be addressed in a multimodal fashion using an extension of the presented method. Examples include the exact discrimination between speech and noise (speech activity detection), the detection of two or more speakers talking at the same time (overlap detection), and the detection and proper assignment of very short speech segments (smaller than about 0.5 sec), which can be due to backchannels. Other interesting future work includes generalizing the system to work with other acoustic events, besides speech, that are correlated with visual features. Also using the system in other domains, such as broadcast news or university webcasting applications are possible future work paths. Initial experiments performed by the authors suggest that another extension possibility is the use of the proposed method in 3D localization algorithms, such as those used in Tele-Immersion scenarios.

ACKNOWLEDGMENTS
We thank Adam Janin and Mary Knox for very helpful input on this article and Bao-Lan Huynh for the baseline experiments with the OpenCV face detector.

REFERENCES


ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 6, No. 4, Article 27, Publication date: November 2010.


Received January 2010; revised May 2010; accepted July 2010