INTERSPEECH 2009 BRIGHTON

Fusing Audio and Video Information for Online Speaker Diarization

Joerg Schmalenstroeer, Martin Kelling, Volker Leutnant, Reinhold Haeb-Umbach

Department of Communications Engineering
University of Paderborn, Germany
{schmalen,kelling,leutnant,haeb}@nt.uni-paderborn.de

10.21437/Interspeech.2009-338

Abstract

In this paper we present a system for identifying and localizing speakers using distant microphone arrays and a steerable pan-tilt-zoom camera. Audio and video streams are processed in real-time to obtain the diarization information "who speaks when and where" with low latency to be used in advanced video conferencing systems or user-adaptive interfaces. A key feature of the proposed system is to first glean information about the speaker's location and identity from the audio and visual data streams separately and then to fuse these data in a probabilistic framework employing the Viterbi algorithm. Here, visual evidence of a person is utilized through a priori state probabilities, while location and speaker change information are employed via time-variant transition probablities. Experiments show that video information yields a substantial improvement compared to pure audio-based diarization.

Index Terms: speaker diarization, face identification, acoustic scene analysis

1. Introduction

Speaker diarization is traditionally concerned with answering the question "Who speaks When?" Recently it has been proposed to extend the problem to also retrieve speaker position information [1][2]. While on the one hand this improves the speaker diarization accuracy itself, there are on the other hand a number of applications which could benefit from this additional position information. For example, a user-adaptive interface could select the most appropriate I/O-device according to the user's location and adapt its behavior based on the identity of the user. Other interesting applications are ambient communication [3] or advanced video conferencing, where a conversation to a remote partner or partners is carried out in such a way that both the camera and the beam formed by the microphone arrays automatically focus on the current speaker.

The goal of this paper is to devise algorithms for joint audio-visual speaker segmentation, identification and localization to be used in one of the aforementioned applications. The speaker diarization process estimates the position and the identity of the speakers and passes the information to the system, which may react upon this context information, implying that the diarization has to be conducted with the lowest possible latency. This renders many of the methods proposed for diarization of broadcast news unsuitable, such as multi-stage batch processing and iterative refinements of segmentation and clustering results, as there usually an offline scenario is considered [4][5]. On the other hand we can assume that prior knowledge about the users is available for our applications. To be specific we assume that acoustic models of the speakers' voices and visual models of their faces are available.

We are given a setup consisting of multiple spatially distributed and wall-mounted microphone arrays and a steerable pan-tilt-zoom camera. On the video signal processing side the video stream is first scanned on a frame-by-frame basis for faces in different scale levels. Here a skin color segmentation is applied to reduce the computational demands. Detected faces are focused automatically and zoomed in to improve the results of the face identification, which is based on the method of Fisher-Faces [6]. On the audio side, the multiple microphone arrays serve to locate the active speaker. If the assumption holds that users are spatially separated, the obtained position information can be used to greatly improve the diarization performance as was shown in [7].

The diarization process is based on a Hidden Markov Model (HMM), where each state represents a speaker. The result of the visual face identification is used as a priori speaker probabilities to be combined with the likelihood obtained from the Gaussian Mixture Models for speaker identification. This together serves as the "observation probability" associated to a HMM state. Further, time-variant state transition probabilities are employed which are estimated from position information and Bayesian Information Criterion (BIC) based speaker change hypotheses. A Viterbi decoder with a latency limiting partial traceback is then applied to find the most likely state sequence, i.e. the final diarization result, thus fusing audio and visual information in a probabilistic framework.

In the next section we give a system overview, introducing the available knowledge sources and their probabilistic modelling. Sections 3 and 4 describe the module for controlling the steerable camera and the face identification system. HMM-based audio-visual speaker diarization is introduced in section 5. Experimental results are presented in section 6 and we finish with some conclusions.

2. System Overview

The overall system consists of an audio and a video processing subsystem, which are connected and synchronized via a shared memory (SHM), see Fig. 1. The video stream delivered by a webcam is processed in order to detect and identify faces, which in turn is used to control the camera's orientation and focus. The lower part handles the audio signals for speaker localization and diarization. Note that the audio processing is done at a constant sampling rate of 16 *kHz*, while the video processing runs at a variable frame rate, which depends on the speed of the network. Information about identified faces is stored in the shared memory and is overwritten each time a new picture has been processed. In the meantime the audio processing uses the currently stored information in the SHM.

All knowledge sources relevant for the camera control and diarization purposes are modeled probabilistically, namely po-

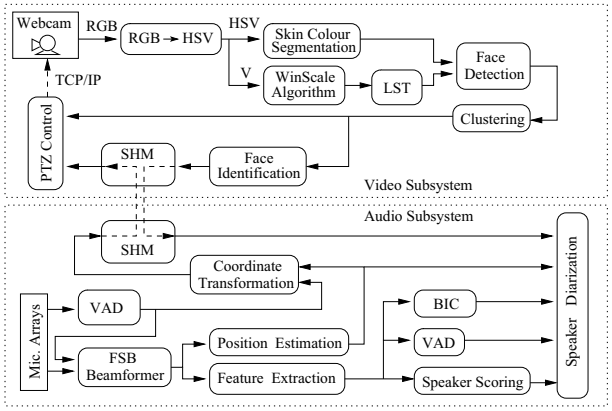

Figure 1: *System overview*

sition information from adaptive beamforming, speaker change information from the Bayesian Information Criterion (BIC), voice activity information (VAD) , Gaussian mixture models (GMM) for speaker identification, and the face detection and identification information. In the following we will describe the individual components.

We apply a Filter-and-Sum Beamformer (FSB) [8] on each microphone array for signal enhancement. The FSB can be viewed as a "self-steering" Delay-and-Sum Beamformer, as the adaptation of the filter coefficient is based on an eigenvalue decomposition, which delivers the Direction-of-Arrival of the dominant sound source as a byproduct. Our experimental setup as shown in Figure 2 uses one T-shaped 4-element microphone array (Array_1), mounted between the display and the webcam, which estimates an azimuth angle α_1 and a tilt angle β towards the speaker. Two ancillary 2-element arrays (Array_2 and Array_3) deliver additional angles α_2 and α_3. The position of the speaker in the x/y plane is retrieved from the estimated angles α_i of the three arrays by calculating the centroid of the intersection points (s_{12}, s_{13}, s_{23}). According to the frame rate of the beamformer a position estimate is obtained every $10\ ms$. The empirical variance computed from the 50 estimates within a window of $0.5\ s$ is used as a feature $x^{pos}(k)$ in the subsequent diarization process, as will be described further below. We model $p(x^{pos}|c)$ by a Gaussian for simplicity, whose parameters are estimated from training data. Here c is a binary variable, which indicates the presence ($c = 1$) or absence ($c = 0$) of a speaker change in the observed temporal window.

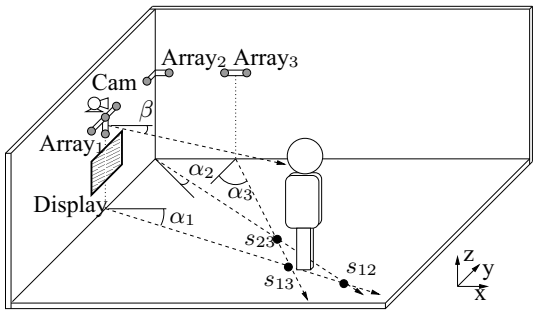

Figure 2: *Experimental setup and room layout*

The feature vector x^{sid} to be used for speaker identification and speaker chance detection is computed by the ETSI advanced feature extraction (AFE) front-end [9] applied to the enhanced beamformer output signal. We use a 42 dimensional vector consisting of 13 MFCCs ($c_0 \dots c_{12}$) and a voicedness feature and their first and second order derivatives.

Evidence for speaker changes is obtained by computing BIC values from feature vectors in a sliding window of length $0.6\ s$ [10]. The variance of subsequent BIC values in a time window of length $0.5\ s$, is used as feature $x^{bic}(k)$. The parameters of the Gaussian $p(x^{bic}|c)$ are again obtained from training data.

The ETSI AFE feature vector is also utilized in the speaker scoring, where for each user i a GMM $p(x^{sid}|\Omega_i)$ is evaluated, where Ω_i indicates the i-th speaker. The GMMs are trained on user-specific audio data by Bayesian adaptation from a universal background model (UBM) for speakers.

The system needs a voice activity detection at two places, however with different requirements. A first VAD is needed for controlling the adaptation of the beamformers, as adaptation must only be performed in the case of an active speaker. Here, a low false alarm rate is required, since classifying a silence frame erroneously as active speech is detrimental, while a moderate missed hit rate is acceptable. This VAD is based on an energy criterion. A second VAD is used in the subsequent speaker diarization process as part of the Viterbi alignment. This VAD should be tuned towards a low missed hit rate as the speaker diarization VAD has similar requirements as an ASR VAD. We employ the VAD of the ETSI AFE here. The VAD information is represented by $P(V|x^{vad})$ being in the range between 0 (absence of speech) and 1 (presence of speech).

3. Camera Control

The webcam is controlled by the module *PTZ Control* utilizing location information from the audio part as well as information from the video stream. Each frame of the video stream is scanned for faces and the found face positions are passed to the control module. The audio localization information, which is available in Cartesian x/y coordinates and the tilt angle β are transformed into pan, tilt and zoom information and passed to the camera control module via the shared memory. Inconsistencies between audio and video localization are resolved as follows: If only acoustic or only visual evidence for a person is available, the camera is adjusted according to the available information source. If both modalities deliver inconsistent information, which means that a face is found in the actual picture, but the active speaker is localized outside the camera view, the camera holds the view angle for a few seconds and then focuses towards the active speaker by favouring the audio information.

4. Face Detection and Identification

In our envisaged applications we assume that the user communicates with the system or a remote partner via voice and video. Thus the assumption is justified that he will face the camera most of the times, as the camera is mounted above the display which shows the far-end communication partner. The deployed face detector is therefore currently limited to detect upright faces looking towards the camera.

Each frame of the video stream is transformed from RGB to HSV color space. On the V component (grey scale picture) the scan for faces is performed, which is limited to the areas of the frame, where skin color is detected. Skin color segmentation uses a histogram look-up approach with smoothing techniques for determining coherent skin color regions. The grey scale picture is scaled to subframes with different resolutions using the WinScale algorithm [11], such that faces can be detected in different sizes. Each subframe is processed with a local structure

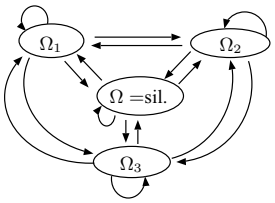

Figure 3: *Hidden Markov Model for speaker diarization*

transformation (LST) as proposed in [12] and scanned for faces with a detection cascade as suggested by Viola and Jones [13].

The approach described above tends to find multiple detections of a face in shifted positions or different scaling levels. Hence a clustering module based on a Leader-Follower method is deployed to merge multiple detections of single faces to unique size and position information of faces.

The face identification employs a principal component analysis (PCA) followed by a linear discriminant analysis (LDA) as proposed in [6]. At first an area around the middle point of the detected face is cut out of the grey scale picture and scaled such that its size fits 60×60 pixels resulting in a 3600 dimensional vector. Then a PCA is used to reduce the dimension of the feature vector from 3600 to 200 and subsequently a LDA reduces this to a feature vector $x^{vid}(k)$ with dimension "number of trained users minus one".

For each user a single Gaussian $p(x^{vid}|\Omega_i)$ is learned from training data and evaluated for the current observation. Consecutive observations in the same view angle of the camera are linked using the a posteriori class probabilities of the last time step as a priori class probabilities for the current time step. Thus we get the a posteriori probability of the the i-th user given the last ν observations $x_\nu^{vid}(k) = x^{vid}(k), \ldots, x^{vid}(k-\nu)$ to be:

$$P(\Omega_i|x_\nu^{vid}(k)) = \frac{p(x^{vid}(k)|\Omega_i) \cdot P(\Omega_i|x_{\nu-1}^{vid}(k-1))}{\sum_j p(x^{vid}(k)|\Omega_j) \cdot P(\Omega_j|x_{\nu-1}^{vid}(k-1))}. \quad (1)$$

To accommodate for errors and enforce stability, the posterior is lower-bounded by a minimum threshold.

5. Audio-Visual Speaker Diarization

Our speaker diarization is based on an ergodic Hidden Markov Model (HMM) and a Viterbi decoder. The HMM has one hidden state per user and an extra state for "silence". The observation probability of each state is given by the combination of the acoustic knowledge $p(x^{sid}(k)|\Omega_i)$ and the visual knowledge $P(\Omega_i|x_\nu^{vid}(k))$. In Figure 3 an example for three users is depicted. Since the acoustic user models are trained on speech data without silence parts and no voice activity detection with frame dropping is done upfront, the GMM likelihood must be multiplied with the probability $P(\mathrm{V}|x^{vad}(k))$ that the frame contains voice. For the acoustic observation probability we get

$$p'(x^{sid}(k)|\Omega_j) = \begin{cases} p(x^{sid}(k)|\Omega_j)P(\mathrm{V}|x^{vad}(k)) & \Omega_j : \text{spk} \\ p(x^{sid}(k)|\Omega_j)(1 - P(\mathrm{V}|x^{vad})(k)) & \Omega_j : \text{sil} \end{cases} \quad (2)$$

Furthermore in the case of silence $P(\Omega_j = \text{sil}|x^{sid}(k))$ is set to the average GMM-score value of the speaker models.

The information about the user obtained from face identification, eq. (1), is now used as a priori state probability in the decoder. To this end we define the following "observation probability"

$$b_j(x(k)) := p'(x^{sid}(k)|\Omega_j) \cdot P(\Omega_j|x_\nu^{vid}(k)) \quad (3)$$

In [1] we have proposed to form time variant transition probabilities from the available speaker change information. In this setup we derive speaker change information from BIC and from position information and we further assume that $x^{pos}(k)$ and $x^{bic}(k)$ are statistically independent. Employing the binary random variable $c(k)$, which is 1 if a speaker change occurs between the time instances $k-1$ and k and 0 else, it follows that

$$P(c(k)|x^{pos}(k), x^{bic}(k)) = \frac{p(x^{pos}(k), x^{bic}(k)|c(k))P(c(k))}{p(x^{pos}(k), x^{bic}(k))}$$
$$= \frac{p(x^{pos}(k)|c(k))P(c(k))}{p(x^{pos}(k))} \frac{p(x^{bic}(k)|c(k))P(c(k))}{p(x^{bic}(k))} \frac{1}{P(c(k))}. \quad (4)$$

Under the assumption of a uniform prior $P(c(k))$, the transition probability from state i to j can be simplified to

$$a_{ij}(k) = \frac{p(x^{pos}(k)|c(k))}{\sum_{c'} p(x^{pos}(k)|c')} \cdot \frac{p(x^{bic}(k)|c(k))}{\sum_{c'} p(x^{bic}(k)|c')}. \quad (5)$$

Transitions to or within the silence state require special treatment, as for the case of silence no position change information is available. Thus we define $a_{ij}(k) = P(c(k)|x^{bic}(k))$ for $j = sil$ and arbitrary values of i.

A Viterbi decoder is used to find the single best state sequence, given the acoustical and visual observations:

$$\hat{\Omega}_1^K = \underset{\Omega_1^K}{\operatorname{argmax}} \sum_{k=1}^{K} \left(\log b_j(x(k)) + \kappa \log a_{ij}(k)\right). \quad (6)$$

The Viterbi decoder is implemented with a partial traceback, starting from the state with the currently best score. It determines the unique part of the state history and delivers it as output. In the rare case of a missing unique trace and simultaneously exceeding the limit of maximum delay, which was set to $2\ s$ in our experiments, a traceback is forced and the trace with the highest score is chosen.

6. Experiments

Experiments were conducted in a room of size $3.5m \times 7.3m$ with a room reverberation time of $150\ ms$. The database for training the system contains the audio and video data of 10 users. First we considered a scenario where a single user is interacting with the system at a relatively fixed position as it is usually the case during a phone call. The second scenario is more like a video conference, where two people are on one side of the system talking in alternating order. In the latter case the camera's focus has to switch between the users to focus always on the active speaker.

In Figure 4 an example for a speaker change is depicted, showing the results of the acoustic based location information in Cartesian coordinates and the a posteriori probabilities $P(\Omega_j|x_\nu^{vid}(k))$ of the face detection (users $j = 1 \ldots 10$ indicated by colours). At time instance $7s$ a speaker change happens. The acoustic localization module immediately detects the change, as indicated by the plotted position information. The new speaker position is forwarded to the camera control. The camera starts panning and focusing and no face can be found during this period, as can be seen in the top figure by the fact that all posteriors $P(\Omega_j|x_\nu^{vid}(k))$ are equal in the time interval $8\ s - 11\ s$. After that obviously a face has been found, as one posterior is clearly larger than all others.

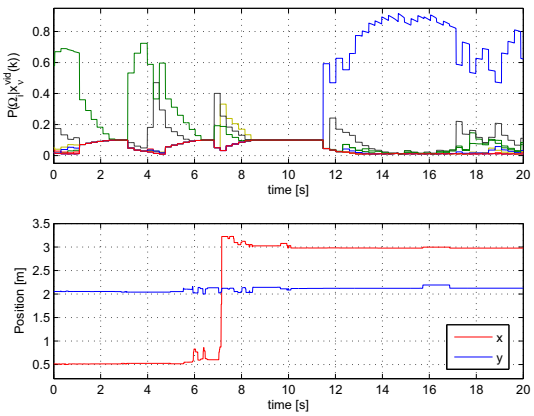

Figure 4: *Camera information and position estimation*

In Table 1 some results of the experiments are listed to show the benefits but also the limitations of our approach. The face identification relies on the results of the face detection, which means that only detected faces can be identified. Unfortunately users tend to move their face during conversations such that the face detector occasionally delivers no detections or the camera has to follow the movement. During the focussing process no detections are available, which, especially in the two-user case, causes a low observation rate (Faces obs. in Table 1). The third column shows the percentage of correct identifications given a face was found. Here, an average of 81.47% was obtained.

The performance of the diarization process is measured with the diarization error rate (DER), which is the percentage of incorrectly labeled frames [4]. A comparison between the standard approach using only acoustical information (DER, Audio) and the new approach using the fusion of acoustical and visual information (DER, Fusion) is given in the table. On average the diarization error rate is decreased by 11.77% absolute, if the video data is incorporated.

Using visual information can also be contraproductive, as can be observed for the user "E", for which the face identification accuracy was particularly low: only 19.50% of the detected faces are correctly identified. As a consequence the diarization error rate is increased compared to audio-only diarization.

Another observation from the experiments is, that the diarization of moving speakers or multiple switching speakers is more difficult than for single, fixed speakers. Users "A" to "F" are fixed speakers, user "G" is a moving speaker and the last four rows are experiments with switching speakers. If a user moves, the observation rate of his face decreases, as the camera needs time to focus on the new position. Altough in this case the system more seldom detects and identifies a face, the diarization accuracy still benefits quite strongly from the additional visual information gathered from the frames.

	Faces		DER %		time
User	obs.	corr.	Audio	Fusion	[min:sec]
A	94.21	89.36	5.37	0.62	3:30
B	90.31	74.57	7.90	1.07	3:26
C	79.46	83.99	1.83	0.13	3:15
D	89.95	100.00	21.97	1.06	3:04
E	89.14	19.50	0.93	10.90	2:55
F	77.71	92.15	2.06	0.93	3:12
G	58.61	91.02	33.34	18.63	2:16
D & A	68.02	85.47	28.32	11.56	3:09
A & B	74.56	89.74	29.22	7.75	5:19
C & A	72.02	82.71	31.41	11.32	3:21
F & A	49.21	89.84	32.69	10.76	3:38
Average	76.75	81.47	18.32	6.55	32:59

Table 1: *Experimental results*

7. Conclusions

In this paper we have presented our system for online speaker diarization based on distant microphone array audio data and video data obtained from a steerable pan-tilt-zoom camera. An HMM approach with probabilistic modelling of the knowledge sources in combination with a Viterbi decoder and a partial traceback implementation enables online processing of audio-visual streams. Thus a parallel segmentation and classification with low latency of about 2 s is feasible. Experiments showed substantial improvements for the speaker diarization task for single as well as for multiple users if face identification information is used as a priori information in the diarization process. Further improvements may be realized by employing more advanced face identification and tracking techniques.

8. References

[1] J. Schmalenstroeer and R. Haeb-Umbach, "Joint speaker segmentation, localization and identification for streaming audio," in *Proc. Interspeech'07*, Antwerp, Belgium, Aug. 2007.

[2] C. Busso *et al.*, "Smart room: Participant and speaker localization and identification," in *Proc. ICASSP'05*, Philadelphia, USA, Mar. 2005, pp. 1117–1120.

[3] S. Borkowski, T. Flury, A. Gerodolle, and G. Privat, "Ambient communication and context-aware presence management," *Communications in Computer and Information Science*, no. 11, pp. 391–396, 2008.

[4] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1557–1565, Sep. 2006.

[5] S. Meignier, D. Moraru, C. Fredouille, J. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech & Languagee*, vol. 20, no. 2-3, pp. 303–330, Jul. 2006.

[6] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[7] J. Schmalenstroeer and R. Haeb-Umbach, "Online speaker change detection by combining bic with microphone array beamforming," in *Proc. Interspeech'06*, Pittsburgh PA, USA, Sep. 2006.

[8] E. Warsitz and R. Haeb-Umbach, "Acoustic filter-and-sum beamforming by adaptive principal component analysis," in *Proc. ICASSP'05*, Philadelphia, USA, Mar. 2005.

[9] ETSI. (2002) ES 202 212 V1.1.1: Speech Processing, Transmission and Quality aspects (STQ); Distributed Speech Recognition; Advanced front-end feature extraction algorithm; Compression algorithms. [Online]. Available: http://www.etsi.org

[10] M. Nishida and T. Kawahara, "Speaker model selection based on the bayesian information criterion applied to unsupervised speaker indexing," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 583–592, Jul. 2005.

[11] C. Kim, S. Seong, J. Lee, and L. Kim, "Winscale: An image-scaling algorithm using an area pixel model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 6, pp. 549–553, Jun. 2003.

[12] B. Froeba and C. Kueblbeck, "Face tracking by means of continous detection," in *Proc. CVPRW'04*, Washington DC., USA, Mar. 2004, pp. 65–71.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. CVPR'01*, Kauai, Hawaii, Dec. 2001, pp. 511–518.