Fusing Audio and Video Information for Online Speaker Diarization

Joerg Schmalenstroeer, Reinhold Haeb-Umbach

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

Abstract

In this paper we present a system for identifying and localizing speakers using distant microphone arrays and a steerable pan-tilt-zoom camera. The scenario at hand assumes audio streams to be processed in real-time to get the diarization information "who spokes when and where" with only short delays. Our new idea is to fuse the acoustical and visual observations directly within the Viterbi decoder to improve the diarization process. In contrast to standard Viterbi decoder implementations, we use a time variant transition matrix generated from speaker change hypotheses and location information. This allows a simultaneous segmentation and classification of the audio stream. Experiments show, that video information enables a substantial improvement of the diarization results.

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

1. Introduction

Speaker diarization is the task of annotating an audio signal with the information "Who speaks When?". This we extended with position information, which improves on the one hand the speaker diarization process itself and on the other hand allows speaker localization and tracking [1]. In ambient communication scenarios or telephone conferences an additional knowledge source becomes available: the video stream of a camera.

The starting point of our studies is a communication system consisting of multiple microphone arrays and a steerable pan-tilt-zoom camera, whose focus is controlled according to the video and audio information collected by the system. Therefore the video stream is scanned on a frame-by-frame basis for faces in different scale levels, whereas the computational demands are reduced by a skin color segmentation. Detected faces are focused automatically and zoomed in to improve the results of the face identification, which is processed by the Fisher-Faces method from [2].

We consider a scenario, where a single person or multiple persons use the system for an audio-visual communication with a distant person. The camera is automatically focused on the actual active speaker or, in case of no active speaker, on a detected face. The speaker diarization process estimates the position and the identity of the speakers and passes the information to the ambient system, which may react on this context information, implying that all gathered information have to be available in approximately real-time.

Thus the main difference between speaker diarization in acoustic scene analysis for ambient intelligence and speaker diarization in broadcast news or recorded meetings is the temporal limitation. On the one hand we have the disadvantage to neglected mulit-stage batch procedures and iterative approaches to meet the requirements, but on the other hand we can assume that the system has prior knowledge on the users.

In our scenario the acoustic scene analysis is based on multiple spatially distributed and wall-mounted microphone arrays. The audio signals in this setup suffer from noise and reverberation, but nethertheless enable the localization of speakers. If the assumption holds that users are spatially separated, the obtained position information can be used to greatly improve the diarization performance as shown in [3].

We model the diarization process as a Hidden Markov Model (HMM), where each state represents a speaker. The observation probability of a state is given by the combination of the Gaussian mixture models (GMMs) for speaker identification and the models for face identification. In contrast to the integrated approach in [4], we use time-variant state transition probabilities estimated from position information and Bayesian Information Criterion (BIC) based speaker change hypotheses. A Viterbi decoder with a latency limited partial traceback combines the acoustical knowledge with the visual information and thus delivers a fused version of the multi-modal input.

In the next section we give a system overview, introducing the available knowledge sources and their probabilistic modelling. Section 3 describes the module for controlling the steerable camera and section 4 is about our face identification system. Speaker diarization is introduced in section 5. Experimental results are presented in section 6 and we finish with some conclusions.

2. System overview

The system, as depicted in Figure 1, consists of two parts working in parallel, which are connected and synchronized via a shared memory (SHM) approach. The upper part processes the video stream delivered by a webcam for detecting and identifying faces, as well as controlling the camera focusing. The lower part handles the audio signals for speaker localization and diarization. Remark that the audio processing is done at a constant sampling rate of 16 *kHz* and the video processing runs at a variable frame rate, depending on the video stream itself and the camera. 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 relies on the actually stored information in the SHM.

All knowledge sources in the system are probabilistically modeled, namely position information from adaptive beamforming, speaker change information from the Bayesian Information Criterion (BIC), voice activity detection (VAD), Gaussian mixture models (GMM) for speaker identification, and the face detection and identification information.

We apply a Filter-and-Sum Beamformer (FSB) [5] on each microphone array for signal enhancement, which as a byprod-

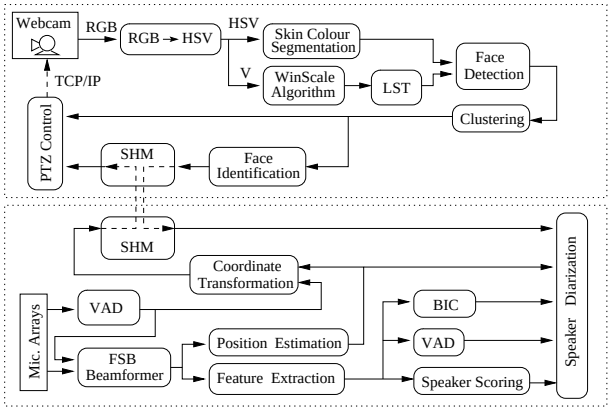

Figure 1: *System overview and knowledge sources*

uct can be used for estimating the Direction-of-Arrival information for the dominant sound source. Our experimental setup as shown in Figure 2 uses one T-shaped microphone array (Array_1), mounted between the display and the webcam, for estimating the tilt angle β towards the speaker. The postion of the speaker 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}) of the on the floor projected directions. The variance of the position estimates $x^{pos}(k)$ in a time window of $0.5\ s$ is used as a feature in the diarization process. The parameters of the Gaussian $p(x^{pos}|c)$ are estimated from training data, using the binary variable c, which indicates the presence ($c = 1$) or absence ($c = 0$) of a speaker change in the observed time frame.

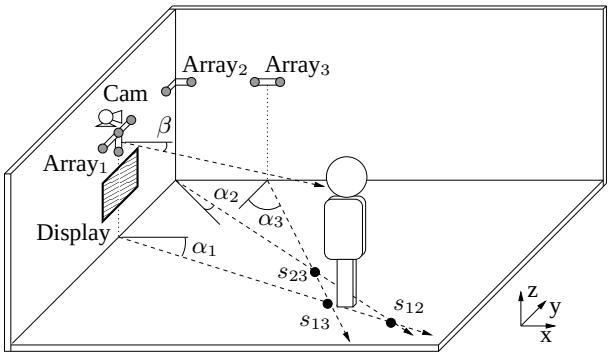

Figure 2: *Experimental setup and room*

The feature vector x^{sid} for speaker identification and speaker chance detection is computed by an ETSI advanced feature extraction (AFE) front-end [6] applied to the enhanced beamformer output signal. We use a 42 dimmensional vector consisting of 13 MFCCs ($c_0 \dots c_{12}$) and a voicedness feature [7] and their first and second order derivatives.

Upon the feature vectors in a sliding window of length $0.6\ s$ a BIC value to hypothesize speaker changes is computed [8]. Its variance x^{bic} in a time frame of $0.5\ s$ is used in the probabilistic model $p(x^{bic}|c)$ for speaker change detection.

Also the feature vectors are utilized in the speaker scoring, where for each user i a GMM $p(x^{sid}|\Omega = i) := p(x^{sid}|\Omega_i)$ is evaluated. The GMMs are trained on user-specific audio data by Bayesian adaptation from a universal background model (UBM) for speakers.

The system needs two different types of voice activity detections. One is required for controlling the adaptation of the beamformers and the other is used in the speaker diarization process. Adaptation of the FSB filters is only performed in the case of active speakers and suffers from low energy signals or false alarms. Thus we employed an energy based voice activity detection for this task. Contrary to beamforming speaker diarization requires a VAD that bridges short speech pauses and acts like a VAD for speech recognition. The voice activity detection from the ETSI AFE could either be an appropriate choice for the task, or a modified version of the energy based VAD. In both cases the VAD information is represented by $P(V|x^{vad})$ being in the range between 0 (absence of speech) and 1 (presence of speech), whereas the feature x^{vad} depends on the type of VAD.

3. Camera control

The webcam is controlled by the module *PTZ Control* considering 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 in Cartesian coordinates and the tilt angle β are transformed into pan, tilt and zoom information and passed via the shared memory to the camera control module. Inconsistencies between audio and video localization are treated as follows: If only audio or only video 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 more the active speaker by favouring the audio information.

4. Face detection and identification

In this scenario users communicate via voice and video with each other. Thus the assumption is justified that they act cooperative to a certain extend. The deployed face detector is limited to upright faces looking towards the camera, which is fulfilled in most cases, as the camera is mounted above the display showing the far-end communication participant.

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 [9], such that faces could be detected in different sizes. Each subframe is processed with a local structure transformation (LST) as proposed in [10] and scanned for faces with a detection cascade as suggested by Viola and Jones in [11].

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 [2]. At first an area around the middle

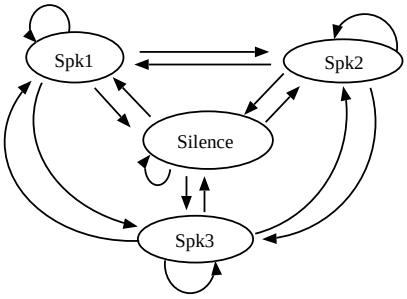

Figure 3: *Hidden Markov Model for speaker diarization*

point of the detected face is cut out the grey scale picture and scaled such that its size fits 60×60 pixel 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 timestep. Thus we get the a posteriori probability of the last ν observations to be:

$$p(\Omega_i|x_\nu^{vid}(k)) = p(\Omega_i|x^{vid}(k),\ldots,x^{vid}(k-\nu)) \quad (1)$$

$$= \frac{p(x^{vid}(k)|\Omega_i)\cdot p(\Omega_i|x_{\nu-1}^{vid}(k-1))}{\sum_j p(x^{vid}(k)|\Omega_j)\cdot p(\Omega_j|x_{\nu-1}^{vid}(k-1))} \quad (2)$$

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

5. 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 that the frame contains voice. For the observation probability we get

$$b_j(x(k)) = \begin{cases} p(\Omega_j|x^{sid}(k),x_\nu^{vid}(k))\cdot P(\mathrm{V}|x^{vad}) & \Omega_j : \text{spk} \\ p(\Omega_j|x^{sid}(k))\cdot(1-P(\mathrm{V}|x^{vad})) & \Omega_j : \text{sil} \end{cases} \quad (3)$$

Assuming that $x^{sid}(k)$ and $x_\nu^{vid}(k)$ are statistically independent, we find

$$p(\Omega_j|x^{sid}(k),x_\nu^{vid}(k)) = \frac{p(\Omega_j,x^{sid}(k)|x_\nu^{vid}(k))}{p(x^{sid}(k)|x_\nu^{vid}(k))} \quad (4)$$

$$= \frac{p(x^{sid}(k)|\Omega_j,x_\nu^{vid}(k))\cdot p(\Omega_j|x_\nu^{vid}(k))}{p(x^{sid}(k)|x_\nu^{vid}(k))} \quad (5)$$

$$= \frac{p(x^{sid}(k)|\Omega_j)\cdot p(\Omega_j|x_\nu^{vid}(k))}{p(x^{sid}(k))} \quad (6)$$

$$= \frac{p(x^{sid}(k)|\Omega_j)}{\sum_i p(x^{sid}(k)|\Omega_i)p(\Omega_i)}\cdot p(\Omega_j|x_\nu^{vid}(k)) \quad (7)$$

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

The new idea introduced in [1] is to form time variant transition probabilities from the available speaker change information. In this setup we derive speaker change information from BIC and 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

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

$$= \frac{p(x^{pos}(k)|c(k))P(c(k))}{p(x^{pos}(k))}\,\frac{p(x^{bic}(k)|c(k))P(c(k))}{p(x^{bic}(k))}\,\frac{1}{P(c(k))}. \quad (8)$$

Under the assumption that $P(c(k))$ is constant for all $c(k)$, the transition score can be simplified to

$$a_{ij}(k) = \frac{p(x^{pos}(k)|c(k))}{\sum_{c'} p(x^{pos}(k)|c')}\cdot\frac{p(x^{bic}(k)|c(k))}{\sum_{c'} p(x^{bic}(k)|c')}. \quad (9)$$

Transitions to or within the silence state requires 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 deployed on the trellis diagram of the unfolded state transition diagram of Figure 3 to find the single best state sequence, given the acoustical and visual observations:

$$\hat{s}_1^K = \underset{s_1^K}{\operatorname{argmax}} \sum_{k=1}^{K} \left(\log b_j(x(k)) + \kappa \log a_{ij}(k)\right). \quad (10)$$

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, 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 telephone conference, where two people are on one side of the system alternately talking. In the latter case the camera's focus has to switch between the users to focus 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. At time instance $7s$ a speaker change happens causing the camera panning towards the second speaker. The speaker location is found after just a short delay, but the camera panning and focusing takes a while until the new speaker is found and identified.

In Table 1 some results of the experiments are listed to show the advantages and disadvantages of our approach. The

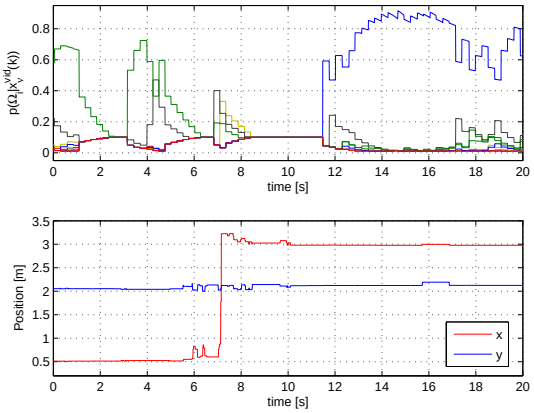

Figure 4: *Camera information and position estimation*

User	Faces obs.	Faces corr.	DER % Audio	DER % Fusion	time [min:sec]
A	94.21	89.36	94.86	99.38	3:30
B	90.31	74.57	92.58	98.93	3:26
C	79.46	83.99	98.17	99.87	3:15
D	89.95	100.00	81.61	99.71	3:04
E	89.14	19.50	99.07	90.64	2:55
F	77.71	92.15	99.30	99.52	3:12
G	58.61	91.02	68.33	82.94	2:16
D & A	68.02	85.47	73.32	89.59	3:09
A & B	74.56	89.74	71.97	93.97	5:19
C & A	72.02	82.71	69.77	89.80	3:21
F & A	49.21	89.84	66.85	89.27	3:38
Average	76.75	81.47	82.64	94.23	32:59

Table 1: *Experimental results*

face identification relies on the results of the face detection, which means that only detected faces in a upright position can be found and identified. Unfortunately users tend to move their face during conversations such that the face detector randomly delivers no detections or the camera has to follow the movement. During the focussing process no detections are available, which, especially in the multi user case, causes a low observation rate (Faces obs. in Table 1). The third column shows the rate of correct identifications of the found faces with an average value of 81.47%.

The performance of the diarization process is measured with the diarization error rate (DER), which is the percentage of correct labeled data. 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 performance is improved by 11.59% absolute, if the video data is incorporated.

The main disadvantage of our approach can be observed for the user "E". In case of false face identifications, here only 19.50% of the detected faces are correctly identified, the diarization error rate is increased by the false informations.

Another observation from the experiments is, that moving speakers or multiple switching speakers are worse diarized then 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. This has no effect on the identification rate of the faces, since the system only identifies detected faces. Altough in this case the system rarely detects and identifys a face, particularly the diarization benefits most from the additional information gathered from the frames.

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 streams. Thus a parallel segmentation and classification with low latency is feasible. Experiments showed substantial improvements for the speaker diarization task for single as well as for multiple users if face identification information is considered during the diarization. Further improvements may be realized by employing more advanced face identification and tracking techniques.

8. References

[1] J. Schmalenstrooer, R. Haeb-Umbach, "Joint Speaker Segmentation, Localization and Identification for Streaming Audio", Proc. Interspeech 2007, Antwerp, Belgium, Aug. 2007

[2] P. Belhumeur, J. Hespanha, D. Kriegman "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, Jul. 1997

[3] J. Schmalenstroeer, R. Haeb-Umbach, "Online Speaker Change Detection by Combining BIC with Microphone Array Beamforming", in *Proc. Interspeech 2006*, Pittsburgh, USA, Sept. 2006

[4] S. Meignier et al., "Step-by-Step and Integrated Approaches in Broadcast News Speaker Diarization", *Comput. Speech Lang.*, no. 20, pp. 303-330, Sept. 2005.

[5] E. Warsitz, R. Haeb-Umbach, S. Peschke, "Adaptive Beamforming Combined with Particle Filtering for Acoustic Source Localization", Proc. ICSLP 2004, Jeju, Korea, Oct. 2004

[6] ETSI ES 202 050 V1.1.3 , "ETSI Standard Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms", Nov. 2003

[7] A. Zolnay, R. Schlüter, H. Ney, "Extraction Methods of Voicing Feature for Robust Speech Recognition", in *Proc. EUROSPEECH*, Geneva, Sept. 2003

[8] M. Nishida, T. Kawahara, "Speaker Model Selection Based on the Bayesian Information Criterion Applied to Unsupervised Speaker Indexing", *IEEE Trans. on Speech and Audio Processing*, Vol. 13, no. 4, July 2005

[9] C. Kim, S. Seong, J. Lee, L. Kim, "WinScale: An Image-Scaling Algorithm Using an Area Pixel Model", IEEE Transactions on Curcuits and Systems for Video Technology, Vol. 13, No.6, 2003

[10] B. Froeba, C. Kueblbeck, "Face tracking by Means of Continous Detection", Proceedings IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2004

[11] P. Viola, M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", Proceedings IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, 2001