Investigating the use of Visual Focus of Attention for Audio-Visual Speaker Diarisation

Giulia Garau, Sileye Ba, Hervé Bourlard and Jean-Marc Odobez
Idiap Research Institute - CP592, 1920 Martigny, Switzerland
Ecole Polytechnique Federale de Lausanne - 1015 Lausanne, Switzerland
{giulia.garau, sileye.ba, herve.bourlard, jean-marc.odobez}@idiap.ch

ABSTRACT

Audio-visual speaker diarisation is the task of estimating "who spoke when" using audio and visual cues. In this paper we propose the combination of an audio diarisation system with psychology inspired visual features, reporting experiments on multiparty meetings, a challenging domain characterised by unconstrained interaction and participant movements. More precisely the role of gaze in coordinating speaker turns was exploited by the use of Visual Focus of Attention features. Experiments were performed both with the reference and 3 automatic VFoA estimation systems, based on head pose and visual activity cues, of increasing complexity. VFoA features yielded consistent speaker diarisation improvements in combination with audio features using a multi-stream approach.

Categories and Subject Descriptors: H.3.1 [Content Analysis and Indexing]: Indexing methods .

General Terms: Experimentation. **Keywords:** Audio-visual speaker diarization, Visual Focus of Attention.

1. INTRODUCTION

The goal of speaker diarisation is estimating "who spoke when" [10]. A robust speaker diarisation approach is beneficial for social signal processing applications such as: dominance detection, automatic role recognition, and addressing [5]. Most speaker diarisation systems work in two steps: the audio stream is classified into speech and non-speech segments (speech-non speech detection), then, the speech segments produced by the same speaker are grouped (clustering)[12]. In this paper we will address the 2^{nd} task in the challenging meeting domain, employing both audio and video cues during clustering. Meetings are an interesting and challenging domain both from an acoustic and visual point of view due to the presence of noise and the natural interaction between participants. Meeting participants have variable length speaker turns, their voices sometimes overlap, and they can move freely in the room (for example to go to the whiteboard), or they can turn their head while speaking, making lip movement detection challenging.

While audio only speaker diarisation was widely investigated [12, 10], audio-visual speaker diarisation is a novel domain especially when applied to the unconstrained meeting task. Noulas and Krose [6] investigated an on-line multimodal speaker diarisation system based on dymamic Bayesian networks and audio-visual mutual information in a constrained setting (videos of two seating persons speaking in turns). An interesting two steps real-time multimodal system to analyse group meetings was proposed by Otsuka et al. [8]. Speaker diarisation is performed by clustering/classifying the microphone array time delays; then, the delay clusters are associated to individual faces, combining face tracking and sound source localisation (assuming that people are seating all the time). Friedland et al. [4] addressed speaker diarisation using video features derived from the compressed data and skin detection in combination with a state of the art Mel Frequency Cepstral Coefficient (MFCC) based system. Experiments on a subset of the AMI meeting corpus [3] resulted in an improved speaker diarisation system.

Our experimental setup is similar to the one adopted by Friedland et al. [4], based on unconstrained 4 participant meetings. Note that our data are more challenging than those used in [6, 8] (where participants were assumed always seated in front of the camera). The main contribution of this paper is to exploit the role of gaze in coordinating turn-taking, by adopting a novel feature set based on Visual Focus of Attention (VFoA) to improve the speaker diarisation. VFoA features are motivated by language and social psychology studies on the role of gaze in a conversation [7, 11]: listeners are more likely to look at the person who is talking and they request turn shifts using gaze; speakers are likely to look at the person they are addressing and to shift their attention towards the next speaker before a speaker turn occurs. VFoA features are directly integrated during speaker clustering, differently from Otsuka et al. [8] where speaker diarisation and VFOA estimation were performed as separate tasks. We experimented both with the VFoA obtained from manual annotation and with the automatically estimated VFoA. For comparison we also investigated motion intensity features, which take into account both speaker's movement for speech production and the speaker's use of gestures to maintain the conversation floor [9]. Our motion features aim at measuring global motion activities in each closeup differently from [4] (based on skin detection), capturing the upper body.

MM'09, October 19–24, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-608-3/09/10 ...$5.00.

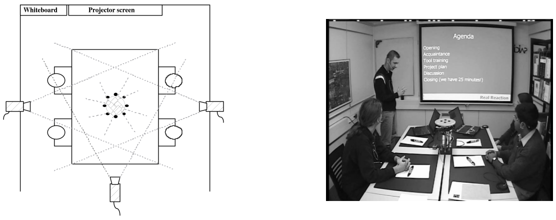


Figure 1: Meeting room setup.

2. SPEAKER DIARISATION

The work presented in this paper is based on the ICSI speaker diarisation system [12]. This system uses the following bottom-up agglomerative clustering approach. Speaker clusters are modelled with an ergodic Hidden Markov Model (HMM), where each state (corresponding to each speaker cluster) is associated to a sequence of hidden substates sharing the same gaussian mixture model (GMM): in order to enforce a minimum duration constraint of 2.5 seconds the same state is repeated several times. In the audio only speaker diarisation system the GMMs modeling each substate are trained on MFCC features. The first step of the ICSI speaker diarisation system is the Speech/Non-Speech detection [12]; then, the speech frames are uniformly partitioned forming K speaker clusters of equal length (in our experiments $K = 16$). After the initial speaker clusters are formed and the corresponding GMM is trained for each speaker model, three processing steps are iterated: Viterbi decoding using the current ergodic HMM, training of a new GMM for each speaker cluster using the newly estimated segmentation, and clusters merging . For each iteration the most similar cluster pair is found according to a score based on the Bayesian Information Criterion (BIC) measuring the difference between the log likelihood of the model trained jointly on the data belonging to the two clusters and the sum of the log likelihoods of the models of the two clusters modelled independently.

3. DATA AND EXPERIMENTAL SETUP

Experiments were performed on a subset of the AMI corpus [3][1]. This multimodal collection of meetings was recorded in rooms instrumented with a set of synchronised devices as shown in Figure 1.

We used the 8-element circular table-top microphone array for audio feature extraction, the two side-cameras to extract head pose and the four individual closeup cameras to extract motion activity features. Although using individual microphones would simplify the speaker diarisation task, a microphone array setup is more portable and less noticeable by meeting participants. We selected the 12 meetings, which include the manual VFoA annotation. These meetings offer a variety of challenges both from the audio and the video point of view (overlapping speech, moving speakers and poor head resolution).

4. AUTOMATIC VFOA ESTIMATION

The visual focus of attention of each participant is described in our data in terms of 8 possible targets: 4 labels corresponding to each meeting participant, 3 targets corresponding to objects in the room (table, whiteboard and projection screen) and the unfocused label (when the meeting participant is not looking at any of the above defined targets). The automatic VFoA system used in this paper aims at finding the visual target of interest for each meeting participant while seating. We experimented with 3 different VFoA estimation systems, extracted using the Ba et al. system [2], relying on several graphical model structures and integrating different cues playing the role of context. The first system (referred to as VFoA(1)) does not use any context: only the participant head pose is used to estimate his focus. The second system (automatic VFoA(2)) exploits a slide activity cue to detect when looking at slides is more likely and remove ambiguities (when the same head pose can be used to look at different targets). The third system (automatic VFoA(3)) exploits in addition visual activities at each seat and the whiteboard to detect who are more visually active, and hence more likely to speak, and thus the visual focus of others. While including more context always improves the overall VFoA recognition rate, the recognition rate of different targets (people, slide screen, table, etc.) varies due to the use of different contexts, and thus target priors. Therefore the three systems result in different behaviours when VFoA is used to improve speaker diarisation.

5. AUDIO VISUAL FEATURES

Audio features: We performed our experiments in the Multiple Distant Microphone (MDM) task. Beamforming was used to reduce the MDM signals to a single channel with enhanced sensitivity in the direction of the desired signal. To perform this task we used the *Beamformit* tool[2] [1], based on the delay and sum algorithm. On the beamformer output we computed 19 MFCCs as acoustic features f_A.

VFoA features: The assumption behind the adoption of VFoA features for speaker diarisation is that while listening people are more likely to look at the person which is speaking. Therefore, we define VFoA features as a measure of the number of persons who are looking at each meeting participant. We also performed smoothing to denoise the VFoA. Our VFoA features are computed for each frame t and each person i as follows:

$$f_V(i,t) = f_{vfoa}(i,t) = \frac{\sum_{\tau=t-\frac{w_{vfoa}}{2}}^{\tau=t+\frac{w_{vfoa}}{2}} \left(\frac{\sum_{k\neq i} VFoA(k,i,\tau)}{N-1}\right)}{w_{vfoa}} \qquad (1)$$

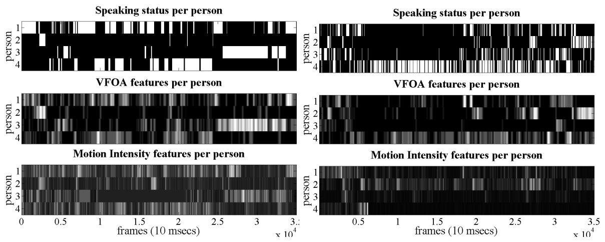
where $N = 4$ is the number of meeting participants i. $VFoA(k,i,t)$ is 1 if the VFoA target of participant k is i at time t, otherwise $VFoA(k,i,t) = 0$, and w_{vfoa} is the smoothing window width.

Motion Intensity features: They are extracted on each of the four closeup videos as the average of the pixel by pixel difference of subsequent gray images [13], smoothed using an averaging window. The obtained four dimensional feature vectors $f_V(t) = f_{mot}(t)$ are then projected through principal component analysis. The use of the whole closeup images has two advantages: it does not rely on face tracking, being thus more robust and computationally efficient, and it allows to capture gestures too. Using the whole image, we also keep into account the fact that people tend to gesticulate more while they are speaking [9].

Correlation of VFoA features and the speaking status: The VFoA features $f_{vfoa}(i,t)$ and the motion fea-

[1] Available from http://corpus.amiproject.org

[2] www.icsi.berkeley.edu/˜xanguera/beamformit/.



(a) IS1008a (b) IS1006d

Figure 2: Comparison of the Speech/Non-Speech status (top), reference VFoA features $f_{vfoa}(i,t)$ (middle) and Motion Intensity features $f_{mot}(i,t)$ (bottom).

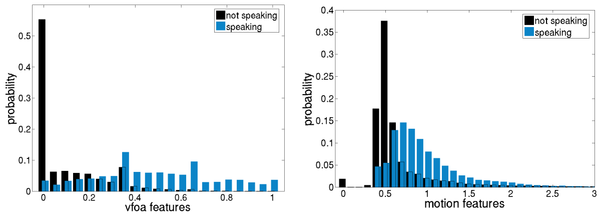


Figure 3: Distributions of $f_{vfoa}(i,t)$ and $f_{mot}(i,t)$.

tures $f_{mot}(i,t)$ are compared to the speaking status for each participant i in Figures 2(a) and 2(b) for an exerpt of meeting IS1008a and IS1006d respectively. Meeting IS1008a is a static meeting and there is a strong correlation between the speaking status and both the motion intensity and the VFoA features. During meeting IS1006d, participant 4 goes at the whiteboard (after around 600 frames) and the correlation with the speaking status is only true for the VFoA features, but not for motion features (being these computed on the participant closeup). VFoA features, being computed as the received focus of attention (they are "passive"), are independent of the position of the speaker and complementary to motions. The correlation of the speaking status and both VFoA and motion intensity features is also evident from their probability distributions shown in Figure 3.

6. MULTI-STREAM COMBINATION

During the audio visual diarisation process, the integration of multiple feature streams is performed by training separate models for each audio and video stream [12]. Audio and video streams are assumed to be independent and synchronous so that the log probabilities given a cluster c_k can be expressed as:

$$\log\left[p(f_A, f_V|c_k)\right] = \gamma_A * \log\left[p(f_A|c_k)\right] + (1-\gamma_A) * \log\left[p(f_V|c_k)\right].$$

In our experiments the combination was performed both during Viterbi segmentation (where the log-likelihood is given by the weighted sum of the log-likelihoods of each feature stream) and during clustering (computing the BIC distance between clusters as a weighted combination of the BIC scores of each stream). Moreover being audio the most important modality for speaker diarisation we assigned, on both steps, a weight of $\gamma_A = 0.9$ to MFCCs and a weight of $\gamma_V = 1 - \gamma_A = 0.1$ to video features.

7. RESULTS

Performances, in terms of the Diarisation Error Rate (DER), were evaluated using the tools provided by NIST [3]. DER is defined as the sum of the Speech/Non-Speech error and the speaker error percentage ($DER = SpNsp + Spkrerr$), that is the percentage of frames which were classified correctly as speech but assigned to the wrong speaker. The average Speech/Non-Speech detection error ($SpNsp$), reported in the first column of Table 1, is shared across all the experimental setups presented in this paper; thus we can only aim at reducing the speaker errors. Table 1 (columns 2–7) reports experimental results in terms of DERs forcing the system to provide the true number of speakers (4 in this dataset), which can be evinced from the video recordings. In fact this information can be used as a speaker clustering stopping criterion. We also report in brackets DERs using the BIC stopping criterion (see Section 2). Detailed results are reported for static meetings, where people seat during the entire meeting (IS1001c, IS1003b, IS1008a, IS1008d), and dynamic meetings, where people leave their seat to go to the whiteboard or to the slide-screen. The second column reports the performances of the baseline audio only speaker diarisation system obtaining a DER of 31.46%.

VFoA features: In the 3^{rd} column performances combining MFCCs and reference VFoA features are reported. This system provides a relative improvement of around 13% compared to the baseline audio only system. Results on the use of the automatic VFoA are reported in columns 4, 5, and 6. Interestingly, from the 3 VFoA systems, the one using only head pose is by far the best, providing an overall 10% relative improvement w.r.t. the MFCC only system. Indeed, although this system performs worse for VFoA estimation, it treats all the targets equally since it is not biased by any other information. This might be the reason why VFoA(1) performs better than the other two automatic systems, which are biased by priors on the slide change or on the participant visual activities. For the automatic VFoA features larger improvements are observed on static meetings (26% reduction for VFoA(1)), for which the accuracy of the estimated VFoA is also higher. In fact automatic VFoA estimation in dynamic meetings is problematic for several reasons: when a participant is near the slide-screen to make a presentation, the two targets might be confused. Since the presenter is the main speaker (standing presentation occurs 33% of the time on average), this confusion can significantly affect the results. In addition, being VFoA estimated only for seating participants, we have one less measurement to compute our VFoA features during these presentations. For dynamic meetings the best performances are achieved by the reference VFoA features (17% relative improvement). VFoA features, being computed as a measure of how many persons are looking at each meeting participant are independent on the speakers position in the room and are well correlated to the participant speaking status, even when he/she is not in front of the closeup camera.

Motion Intensity features: The last column of Table 1 reports the DER for the combination of motion intensity features with MFCCs. This provided an overall 7% relative DER reduction. Similar DER reductions are achieved both on static and dynamic meetings using 4 clusters while using the BIC offers larger improvements on static meetings.

[3] `http://www.nist.gov/speech/tests/rt/2006-spring/`

Meeting	Type	SpNSp	MFCC only	MFCC+ ref. VFoA	MFCC+ autom. VFoA (1)	MFCC+ autom. VFoA (2)	MFCC+ autom. VFoA (3)	MFCC+ motion intensity
IS1000a	D.	13.5	**25.8** [37.1]	27.6 [26.3]	30.9 [31.0]	32.1 [32.1]	30.4 [30.4]	28.9 [37.2]
IS1001a	D.	17.9	34.4 [34.4]	32.8 [32.4]	34.3 [34.3]	34.7 [34.9]	35.2 [35.2]	34.3 [**31.3**]
IS1001b	D.	8.5	**28.4** [28.4]	30.1 [29.4]	**28.5** [28.9]	37.5 [37.5]	37.2 [37.2]	32.7 [33.0]
IS1003d	D.	20.2	38.7 [38.9]	**38.2** [37.5]	53.4 [54.4]	39.6 [38.4]	47.7 [55.3]	39.9 [39.9]
IS1006b	D.	11.8	51.6 [52.4]	25.0 [25.3]	**18.0** [22.4]	52.2 [55.8]	42.0 [25.7]	41.6 [24.9]
IS1006d	D.	24.7	**56.9** [52.9]	59.9 [54.5]	57.3 [65.9]	61.4 [58.5]	70.9 [76.6]	62.2 [62.2]
IS1008b	D.	9.1	21.1 [23.6]	**10.6** [22.0]	11.1 [12.3]	12.5 [13.4]	13.1 [13.1]	**10.8** [10.8]
IS1008c	D.	23.4	26.6 [26.6]	25.9 [26.3]	45.1 [39.4]	39.3 [39.6]	25.2 [26.0]	**25.0** [26.9]
IS1001c	S.	8.2	21.0 [20.8]	19.1 [19.3]	15.7 [15.9]	16.3 [21.7]	35.0 [35.0]	**15.3** [15.3]
IS1003b	S.	14.0	32.7 [32.7]	**15.7** [15.7]	17.8 [18.7]	34.2 [34.2]	**17.0** [17.0]	32.2 [17.8]
IS1008a	S.	6.8	**7.3** [8.5]	35.8 [27.5]	7.6 [21.8]	8.5 [8.5]	7.7 [7.7]	8.0 [8.0]
IS1008d	S.	12.7	14.4 [15.8]	**14.3** [14.3]	14.5 [14.7]	14.6 [14.6]	14.6 [14.6]	14.5 [15.5]
Dynamic	D.	15.4	36.5 [36.6]	**30.1** [31.0]	33.4 [34.7]	38.6 [38.9]	37.2 [36.3]	33.8 [32.6]
Static	S.	10.6	19.1 [19.8]	20.3 [18.6]	**14.2** [17.4]	18.6 [20.0]	19.0 [19.0]	17.7 [**14.4**]
ALL		14.0	30.6 [31.5]	**27.1** [27.3]	**27.5** [29.4]	32.5 [33.1]	31.7 [31.0]	28.9 [**26.5**]

Table 1: From left to right: static/dynamic meetings, speech-non speech error, DER for MFCC only baseline and combination of MFCC using the multi-stream approach with reference VFoA features, VFoA features derived from the head pose only system (1), VFoA features derived from head pose and slide change system (2), VFoA features derived from the head poses, the motion activities and slide change system and motion intensity video features. In brackets the DER using the BIC criterion to stop the agglomerative clustering (otherwise DER is reported forcing the system to a number of clusters equal to the real number of speakers).

8. SUMMARY AND CONCLUSIONS

In this paper we investigated psychology inspired video features for audio-visual speaker diarisation of meetings. The visual focus of attention information was exploited, relying on the fact that while listening people tend to look at who is speaking most of the time. We also compared these features with the use of motion cues capturing the speakers use of gestures for conversation floor management.Experiments using manual and automatically estimated VFoA, and motion intensity features resulted in consistent improvements over the baseline audio only system. Interestingly it was found that to use automatic VFoA for speaker diarisation it is important to achieve good VFoA performances on human targets. Therefore in the future we will investigate new directions to improve the VFoA estimation system on these particular targets. Moreover VFoA and motion intensity features capture different modalities of nonverbal communication, thus we will investigate new ways of integrating them. One interesting direction might be to exploit the gaze-turn taking relationship to predict speaker transitions.

Acknowledgements: This work was supported by the EU 6th FWP IST Integrated Project AMIDA (Augmented Multiparty Interaction with Distant Access) and the NCCR Interactive Multimodal Information Management project (IM2). We thank Dr. Gerald Friedland at ICSI for the Speech/Non-Speech detection output and for his helpful advice.

9. REFERENCES

[1] X. Anguera, C. Wooters, and J. Hernando. Speaker diarization for multi-party meetings using acoustic fusion. In *Proc. ASRU*, 2005.

[2] S. Ba, H. Hung, and J.-M. Odobez. Visual Activity Context for Focus of Attention Estimation in Dynamic Meetings. In *Proc. of ICME*, 2009.

[3] J. Carletta et al. The AMI Meeting Corpus: A Pre-Announcement. *Proc. MLMI*, 2005.

[4] G. Friedland, H. Hung, and C. Yeo. Multi-Modal Speaker Diarization of Real-World Meetings using Compressed Domain Video Features. In *Proc. ICASSP*, 2009.

[5] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: a review. *Image and Vision Computing, Special Issue on Human Naturalistic Behavior*, In press, 2009.

[6] A. Noulas and B. Krose. On-Line Multi-Modal Speaker Diarisation. In *Proc. ICMI*, 2007.

[7] D. Novick, B. Hansen, and K. Ward. Coordinating Turn-Taking with Gaze. In *Proc. ICSLP*, 1996.

[8] K. Otsuka et al. A Realtime Multimodal System for Analysing Group Meetings by Combining Face Pose Tracking and Speaker Diarisation. *Proc. ICMI*, 2008.

[9] E. Padilha and J. Carletta. Nonverbal Behaviours Improving a Simulation of Small Group Discussion. In *Proc. of the 1st Nordic Symposium on Multimodal Communications*, pages 93–105, 2003.

[10] S. Tranter and D. Reynolds. An Overview of Automatic Speaker Diarization Systems. *IEEE Transactions on Audio, Speech and Language Processing*, 14, 2006.

[11] R. Vertegaal, R. Slagter, G. Van der Veer, and A. Nijholt. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proc. of ACM SIGCHI*, 2001.

[12] C. Wooters and M. Huijbregts. The ICSI RT07s Speaker Diarization System. *Proc. Rich Transcription Spring Meeting Recognition Evaluation*, 2007.

[13] M. Zobl, F. Wallhoff, and G. Rigoll. Action Recognition in Meeting Scenarios using Global Motion Features. In *Proc. PETS-ICVS*, 2003.