USING AUDIO AND VISUAL CUES FOR SPEAKER DIARISATION INITIALISATION

Giulia Garau and Hervé Bourlard[*]

Idiap Research Institute - CP592, 1920 Martigny, Switzerland
Ecole Polytechnique Federale de Lausanne - 1015 Lausanne, Switzerland

ABSTRACT

In this paper we present a novel approach to audio visual speaker diarisation (the task of estimating "who spoke when" using audio and visual cues) in a challenging meeting domain. Our approach is based on the initialisation of the agglomerative speaker clustering using psychology inspired visual features, including Visual Focus of Attention (VFoA) and motion intensities. This method, providing initial speaker clusters of high purity, achieved consistent improvements over the widely adopted linear initialisation method. Moreover, the initialisation using both visual and Time Delay of Arrival (TDoA) cues was also investigated in conjunction with the multi–stream combination of acoustic and visual features (MFCC, TDoA, VFoA, motion intensity, and head pose likelihoods). This speaker diarisation framework allowed to successfully integrate three feature streams, further exploiting the complementarity between multimodal cues.

Index Terms— Audio Visual speaker diarisation, clustering initialisation

1. INTRODUCTION

The goal of speaker diarisation is estimating "who spoke when" [1]. This paper investigates the task of speaker diarisation, using both audio and visual cues, in the meeting domain. Unconstrained meetings are an interesting and challenging domain, both from an acoustic and visual point of view: 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). Speaker diarisation of meetings has been usually performed using audio features only [2, 1, 3], while fewer studies focused on the use of audio and visual cues [4, 5, 6].

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) [2]. The most commonly used technique for the latter task is the bottom-up hierarchical agglomerative clustering, where initial speaker clusters are iteratively merged according to their similarity until there are no more clusters to merge. The choice of the initial clusters was shown to be very important, since a bad initialisation propagates through the iterative resegmentation and clustering process [3, 7]. Linear initialisation, where the speech data are uniformly partitioned into K clusters, can lead to initial clusters containing data from multiple speakers and, therefore, wrong speaker models leading to mistakes during the agglomerative cluster merging. K-means initialisation, performed on MFCCs, was compared to linear initialisation by Ajmera et al. [8] without finding any improvement, while consistent improvements using a modified version of this algorithm, the segmental k-means, where reported in [7]. An interesting approach referred to as friends and enemies initialisation, aiming at improving the purity of the initial clusters, was proposed by Anguera et al. [3]. Approaches exploiting the spatial information, carried by the Time Delays of Arrival (TDoA) obtained from a microphone array, were investigated by Koh et al. [9] and Luque et al. [10]. In both works the TDoA distributions are analysed to find initial speaker clusters, which are subsequently resegmented and clustered using MFCC features.

In this paper, we address the speaker diarisation initialisation problem using both the spatial information captured by the time delays and the speaking status information carried by visual cues. To the best of our knowledge this is the first work which investigates the initialisation of an audio visual speaker diarisation system. Moreover it is the first study where visual features are used as an initialisation cue. In previous speaker diarisation studies, visual cues were mostly integrated during the clustering phase: for instance Friedland et al. [4] combined motion features derived from the compressed video and skin detection with a state of the art MFCC based system. Similarly, in [11] we investigated the use of Visual Focus of Attention information for speaker diarisation using a multi–stream approach. In the present paper we will adopt an experimental setup similar to the one adopted in [11], based on unconstrained 4 participant meetings. In [4, 11] visual features were integrated with audio features using a multi–stream feature combination approach. In the present work, we focus on using visual information to find initial clusters. Moreover, we investigate the use of TDoAs to initialise the audio visual speaker diarisation system presented in [11].

We investigated two types of visual features as initialisation cues: Visual Focus of Attention (VFoA) derived features and motion intensity features. VFoA features are motivated by language and social psychology studies on the role of gaze in a conversation [12]: listeners are likely to look at the person talking and they request turn shifts using gaze; speakers are likely to look at the addressed person and to shift their attention towards the next speaker before a speaker turn occurs. Motion intensity features take into account speaker's movement for speech production and gestures [13]. VFoA and motion intensity features, providing a measure of who is most likely to speak, can be exploited to find an initial cluster set; this initial clustering can be refined using the speaker diarisation system.

This paper is organised as follows. In Section 2 we describe the speaker diarisation engine. In Section 3 we outline the data used. In Section 4 we define the audio and visual features. In Section 5 we outline how these features are used for the initialisation. Finally in Section 6 we report experimental results drawing some conclusions in Section 7.

[*] 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 for the Speech/Non-Speech detection output, Dr. Sileye Ba and Dr. Jean-Marc Odobez for the automatic VFoA estimation output.

2. SPEAKER DIARISATION ENGINE

The work presented in this paper is based on the ICSI speaker diarisation system [2]. This system uses the following bottom-up agglomerative clustering approach. Speaker clusters are modelled with an ergodic Hidden Markov Model (HMM). Each state (corresponding to a single speaker cluster) is modelled as 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 substate is repeated several times. In the audio only speaker diarisation system the GMMs are trained on MFCC features, while separate GMMs are trained on multiple feature streams during the audio visual diarisation. The first step of the ICSI speaker diarisation system is the Speech/Non-Speech detection [2]; then, processing only the speech frames, K initial clusters are created using either linear initialisation (uniformly partitioning the speech frames in $K = 16$ clusters of equal length) or adopting one of the initialisation methods outlined in section 5. After the initial speaker clusters are formed, the corresponding GMM is trained for each speaker model. Three processing steps are then iterated: Viterbi decoding using the current ergodic HMM, training of a new GMM for each speaker cluster using the newly estimated segmentation, and cluster 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 (θ) and the sum of the log likelihoods of the models of the two clusters (θ_a and θ_b) modelled independently. It is also assumed that the complexity of the model θ is equal to the sum of the complexities of the models θ_a and θ_b [8].

The integration of multiple feature streams (e.g. two streams) is performed by training separate GMMs for each stream. The two streams are combined both during Viterbi segmentation and clustering, computing the total log likelihood as a weighted sum of the likelihood of the two separate models. In our experiments the first stream is always represented by MFCCs and, being this the most informative modality for speaker diarisation, a weight of 0.9 was assigned to MFCCs while 0.1 was adopted for the additional feature streams.

3. DATA

Experiments were performed on a subset of the AMI meeting corpus [14][1]. This multimodal collection of four participant meetings was recorded in rooms instrumented with a set of synchronised recording 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 poses, 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 11 meetings [2], 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). We can distinguish between static meetings, where people seat during the entire meeting, and dynamic meetings, where people leave their seat to go to the whiteboard or the slide–screen.

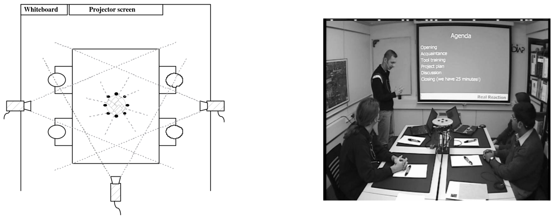

Fig. 1. Meeting room setup.

4. MULTIMODAL FEATURES

In this section we outline the features used both during the initialisation and the speaker clustering process.

Audio features: Beamforming was adopted to reduce the $d = 8$ microphone array signals (Section 3) to a single channel with enhanced sensitivity in the direction of the desired signal. To perform this task we used the *Beamformit* tool[3] [15], based on the delay and sum algorithm. The far-field signals are first enhanced using Wiener filtering. Then a reference channel is chosen so that the average cross-correlation with the other channels is maximised. With respect to this reference channel we computed $(d - 1)$ time delays of arrival using the GCC–Phat cross-correlation and used them for delay and sum beamforming. An acoustic feature vector f_A, comprising 19 MFCCs, was extracted from the beamformer output. We also used the beamformer TDoAs to initialise the multi–stream speaker diarisation system as outlined in section 5.

VFoA features: The assumption behind the adoption of these features for speaker diarisation is that while listening people are more likely to look at the person which is speaking. Therefore, we defined VFoA features as a measure of the number of persons who are looking at each meeting participant [11]. We experimented both with the manually annotated VFoA and with the VFoA automatically estimated using the system outlined in [16], adopting the most simple VFoA estimator based only on meeting participant head poses. In [11] we showed that this VFoA system, not being biased by any other cue, was the best performing system when used to extract features for speaker diarisation.

Head pose likelihood features: These features are directly extracted from the head pose tracker employed by the VFoA estimation system [16]. Head pose features are obtained as: $f_{headpose}(k,t) = \sum_{i \neq k} \frac{P(O(i,t)|S_k)}{N-1}$ where $P(O(i,t)|S_k)$ is the probability of the observed head pose of meeting participant i at time t given that his focus S_k is participant k, and N is the number of meeting participants.

Motion Intensity features: These are extracted on each of the four closeup videos as the average of the pixel by pixel difference of subsequent gray images [17]. Analysing the whole image, we keep into account the fact that people tend to gesticulate more while they are speaking [13].

5. MULTIMODAL INITIALISATION

Similarly to [10] we initialise the agglomerative speaker diarisation system using clusters obtained from the TDoAs estimated during the beamforming. In addition we investigated an initialisation approach based on the visual features outlined in section 4: motion intensities and VFoA features. In all the three cases only the speech frames selected by the speech/non-speech detection were processed.

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

[2] With respect to [11] we had to remove meeting IS1003b because only one channel of the microphone array is available

[3] `www.icsi.berkeley.edu/~xanguera/beamformit/`.

When the linear initialisation is used the same amount of data is assigned to each cluster, and the same initial number of Gaussians per cluster is employed. When the initial clusters are chosen using cues such as TDoAs, VFoAs and motion intensities the distribution of the data between clusters is quite imbalanced. A criterion to select the initial complexity of the cluster models (number of Gaussians) is therefore necessary. While Luque et al. [10] adopted a criterion based on the amount of speech frames per Gaussian, in this work we adopt the Bayesian information criterion, maximising the expression:

$$j_{init} = \arg\max_{j} \log \mathcal{L}(X_i|M_j) - \lambda \frac{1}{2} j \log(N_i) \qquad (1)$$

where $\log \mathcal{L}(X_i|M_j)$ is the log likelihood of the data X_i of cluster i, given the model M_j of X_i with complexity of j Gaussians, N_i is the number of frames of X_i and λ is a penalty factor (10 in all our experiments).

TDoA initialisation: Spatial information carried by TDoAs is exploited by using k-means to cluster the time delays (considering the speech frames only) forming k initial clusters. The number of sound direction clusters (i.e. the TDoA clusters) is unknown a priori; therefore when performing k-means we have set $k = 16$, equal to the number of initial clusters adopted by the linear initialisation. It is known that the result of k-means clustering itself highly depends on its initialisation. We obtained the initial centroids by selecting the highest ordered peaks from the histograms of each of the $(d-1)$ TDoA sequences, obtaining k centroids of $(d-1)$ coordinates.

Motion Intensities and VFoA initialisation: Active speakers tend both to gesticulate/move the most, and to be looked the most by listeners. Initial clusters were therefore selected by finding for each time-frame respectively the closeup with the highest motion and the seat with the highest VFoA feature (i.e. the participant receiving the strongest focus of attention). Since in the motion based initialisation no motion can be estimated for people which are momentarily not in front of the closeup, we introduced a new cluster every time somebody is not in front of the camera. A better solution would have been to measure the motion intensity of participants while at the whiteboard or at the slidescreen. However in this case the video resolution is not sufficient to detect facial movements.

6. RESULTS

Speaker diarisation performances, in terms of the Diarisation Error Rate (DER), were evaluated using the tools provided by NIST [4]. DER is defined as the sum of the Speech/Non-Speech error and the speaker error percentage ($DER = SpNsp + Spkrerr$), i.e. the percentage of frames which were classified correctly as speech but assigned to the wrong speaker. The average Speech/Non-Speech detection error ($SpNsp$) is shared across all the experimental setups presented in this paper ($SpNsp = 13.9\%$); thus we can only aim at reducing the speaker errors.

To assess the performances of the three proposed initialisation methods, we measured the Average Cluster Purity (ACP) before the actual speaker diarisation took place. The ACP, defined for the first time by Ajmera et al. [18], measures how well a cluster is limited to only one speaker. The upper part of Figure 2 reports the ACP for each meeting for various initialisation methods, while the lower part reports DER results after reclustering using the MFCC based diarisation system. The ACP graph shows that all the proposed initialisation methods outperformed the linear initialisation across the entire testing set. This is also evident on the final diarisation output, where the proposed approaches outperformed the linear initialisation (i.e. resulted in lower DERs) on most of the recordings. We can also notice that the initialisation mostly affects dynamic meetings, while static meetings are less influenced (e.g. static meetings IS1001c, IS1008a, IS1008d at the bottom of Figure 2). In particular highly dynamic meetings (where participants go frequently to the whiteboard), such as IS1006b and IS1006d, benefit the most from a visual feature based initialisation. Finally the reference VFoA initialisation resulted in the best DER and ACP.

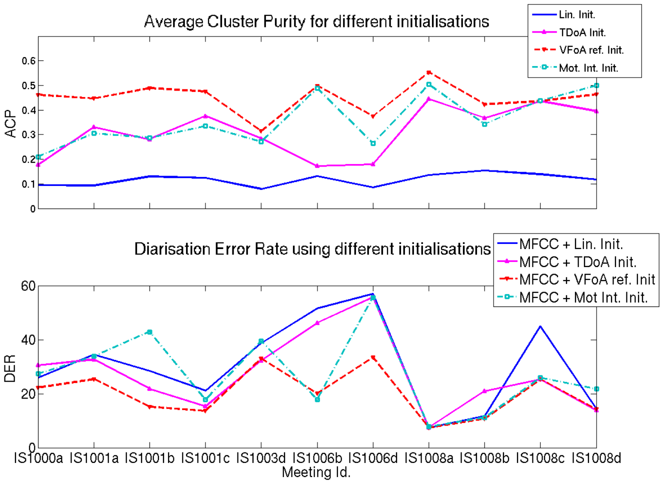

Fig. 2. Initialisation Average Cluster Purity (how well a cluster is limited to only one speaker) using different cues (linear, TDoA, VFoA and motion intensity based) and diarisation error rate of the MFCCs only system using the above mentioned initialisations.

Results in terms of DER forcing the speaker diarisation system to provide the true number of speakers (which can be evinced from the video recordings) are reported in Table 1. Detailed results are reported in square brackets for static (i.e. the participants are seated all the time) and dynamic meetings (i.e. there is a lot of activity at the whiteboard).

Most of the combinations using the linear initialisation ($1st$ row, 2–6 column) resulted in an improvement compared to the baseline MFCC only system ($1st$ row, $1st$ column). An exception is represented by the TDoA combination where an increase in DER is observed for dynamic meetings, on which the spatial information is more difficult to be exploited. On the MFCC only diarisation system (first column of table 1) all the multimodal initialisation approaches resulted in an overall improvement over the linear initialisation.

Further improvements can be observed when multimodal combination is also performed on top of the three proposed initialisation systems. In particular the adoption of TDoA features both for the initialisation and the multi–stream diarisation resulted in the best DER for static meetings (11.2%) employing automatically extracted cues.

The multi–stream combination of MFCCs and headpose features on top of a motion intensity based initialisation provides the best automatic diarisation performances for dynamic meetings (30.3%).

The best results (including manually annotated cues) were achieved by using the reference VFoA initialisation (third row), obtaining excellent performances both for static (10.9%) and dynamic meetings (20.9%) combining MFCCs with TDoAs.

[4] www.nist.gov/speech/tests/rt/2006-spring/index.html

	Combined feature streams					
	MFCC	MFCC	MFCC	MFCC	MFCC	MFCC
	—	TDoA	VFoA ref.	VFoA auto.	Motion Intensity	Headpose
	All [Stat. Dyn.]	All [Stat. Dyn.]	All [Stat. Dyn.]	All [Stat. Dyn.]	All [Stat. Dyn.]	All [Stat. Dyn.]
Linear Initialisation	31.0 *[14.7 36.5]*	31.0 *[12.9 37.0]*	28.0 *[21.8 30.1]*	28.3 *[13.0 33.4]*	28.6 *[13.0 33.8]*	26.6 *[13.1 31.0]*
TDoA Initialisation	27.7 *[12.6 32.7]*	31.1 *[**11.2** 37.7]*	**25.9** *[11.8 30.6]*	26.2 *[12.2 30.9]*	25.9 *[12.0 30.5]*	28.3 *[11.5 33.9]*
VFoA Reference Initialisation	19.7 *[12.1 22.2]*	**18.4** ***[10.9 20.9]***	18.6 *[11.5 21.0]*	19.8 *[11.9 22.5]*	19.0 *[11.7 21.5]*	18.9 *[11.5 21.4]*
Motion Intensity Initialisation	27.2 *[16.6 30.7]*	26.6 *[11.8 31.5]*	25.5 *[14.5 29.2]*	28.9 *[18.4 32.5]*	30.5 *[19.2 34.3]*	26.9 *[16.6 **30.3**]*

Table 1. DER results for the whole dataset (All), and in square brackets for static (Stat.) and dynamic (Dyn.) meetings. The initialisation of the audio visual diarisation with four methods is reported by rows (row: 1-linear, 2-TDoA, 3-VFoA and 4-motion intensity based initialization). By column we report the DER for the MFCC only system (column 1) and combining MFCCs (using the multi–stream approach) with 5 sets of features (column 2–6: TDoA, VFoA reference and automatic features, motion intensity features and head pose likelihood features).

7. CONCLUSIONS

In this paper we investigated the initialisation of an audio visual agglomerative speaker diarisation system using psychology inspired visual cues. The explored visual cues include: Visual Focus of Attention based features (exploiting the fact that listeners tend to look at speakers the most) and motion intensity features (capturing speakers use of gestures and movement during speech production). These features were exploited using a majority voting criterion (the meeting participant who was looked at the most and the one with highest motion intensity respectively) to find initial clusters with an improved purity. Moreover initialisation based on the spatial information carried by the Time Delays of Arrival, estimated during microphone array beamforming, was also investigated. These three initialisation methods were studied in conjunction with an audio visual speaker diarisation system, combining through a multi–stream approach MFCCs and features based on TDoA, VFoA, motion intensities and head pose likelihoods.

Numerical experiments were performed on a challenging collection of meeting recordings. It was found that the information carried by TDoA, VFoA and motion intensity features can be successfully exploited to initialise the diarisation system. Further improvements can be obtained by employing a multi–stream combination of MFCC, TDoA, VFoA, motion intensity and head pose features during the agglomerative clustering. Therefore the multimodal initialisation not only provides consistent improvements by itself, but also positively interacts with the multi–stream combination, allowing the integration of more than two modalities.

This paper represents the first work where: 1) visual cues are exploited to initialise a speaker diarisation system; 2) multimodal initialisation is performed in conjunction with a multi–stream diarisation approach. We proposed an effective approach to initialise an audio visual speaker diarisation system, achieving significant improvements over a conventional linear initialisation.

8. REFERENCES

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

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

[3] X. Anguera, C. Wooters, and J. Hernando, "Friends and Enemies: A Novel Initialization for Speaker Diarization," in *Proc. ICSLP*, 2006.

[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] A. K. Noulas, G. Englebienne, and B. J. A. Krose, "Multimodam speaker diarisation," *CVIU*, 2009.

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

[7] O. Ben-Harush, I. Lapidot, and H. Guterman, "Weighted Segmental K-Means Initialization for SOM-Based Speaker Clustering," in *Proc. ICSLP*, 2008.

[8] J. Ajmera and C. Wooters, "A Robust Speaker Clustering Algorithm," in *Proc. ASRU*, 2003.

[9] E. C. W. Koh et al., "Speaker Diarization using Direction of Arrival Estimate and Acoustic Feature Information: the I2R-NTU Submission for the NIST RT 2007," in *Proc. Rich Transcription Spring Meeting Recognition Evaluation*, 2007.

[10] J. Luque, C. Segura, and J. Hernando, "Clustering Initialization Based on Spatial Information for Speaker Diarization of Meetings," in *Proc. ICSLP*, 2008.

[11] G. Garau, S. Ba, H. Bourlard, and J.-M. Odobez, "Investigating the Use of Visual Focus of Attention for Audio-Visual Speaker Diarisation," in *Proc. ACM Multimedia*, 2009.

[12] 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.

[13] E. Padilha and J. Carletta, "Nonverbal Behaviours Improving a Simulation of Small Group Discussion," in *Proc. of the 1st Nordic Symposium on Multimodal Communications*, 2003.

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

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

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

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

[18] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, "Unknown-Multiple Speaker Clustering Using HMM," in *Proc. ICSLP*, 2002.