SPEECH/NON-SPEECH DETECTION IN MEETINGS FROM AUTOMATICALLY EXTRACTED LOW RESOLUTION VISUAL FEATURES

Hayley Hung *

Idiap Research Institute, Martigny, Switzerland

Silèye O Ba *

Lab-STICC, Telecom Bretagne, Brest, France

ABSTRACT

In this paper we address the problem of estimating who is speaking from automatically extracted low resolution visual cues in group meetings. Traditionally, the task of speech/non-speech detection or speaker diarization tries to find "who speaks and when" from audio features only. In this paper, we investigate more systematically how speaking status can be estimated from low resolution video We exploit the synchrony of a group's head and hand motion to learn correspondences between speaking status and visual activity. We also carry out experiments to evaluate how context through the observation of group behaviour and task-oriented activities can help to improve estimates of speaking status. We test on 105 minutes of natural meeting data with unconstrained conversations and compare with state of the art audio-only methods.

Index Terms— visual focus of attention, speaker detection.

1. INTRODUCTION

Traditionally, voice activity detection or speaker diarization has been used predominantly in the speech processing community as a preprocessing step for tasks such as speech recognition and other more semantically high-level tasks such as dialogue act recognition. Recently, it has been shown that non-verbal behaviour using just the extracted speaker turn patterns can yield useful and robust features for tasks such as estimating dominance [1], or roles [2] in conversations. Such research shows that semantically high-level information need not be extracted fron verbal cues. From a privacy perspective, it would be highly desirable to estimate who is speaking without the need to record audio content from private conversations, which is perceived to be quite invasive even if only prosodic features are extracted. In addition, in a large cocktail party or during corporate team-work exercises there can be much background noise which makes distinguishing voices robustly difficult.

This paper addresses the problem of estimating speakers in four-participant meetings when only low resolution video data is available for testing, and audio-visual data for training. We present and compare novel approaches using both supervised and unsupervised models that investigate the extent that head and hand motion can be used to aid the estimation of who is speaking. In addition, we introduce supervised models that integrate different contextual cues such as the visual focus of attention (VFOA) of participants in the meeting and also activity on a slide screen to estimate speaking status. The main emphasis and novel contribution of our work is to study how head and hand activity features can be correlated with speaking status. Previous work which estimate how either gestures are related semantically to speech [3] have used high resolution features. To our knowledge, there has been no work that presented how low resolution visual features from the upper torso contribute to the the estimation of speaking status; we consider low resolution video to contain faces which are captured at around 20 pixels in height.

*This work was funded by the Swiss NCCR IM2.

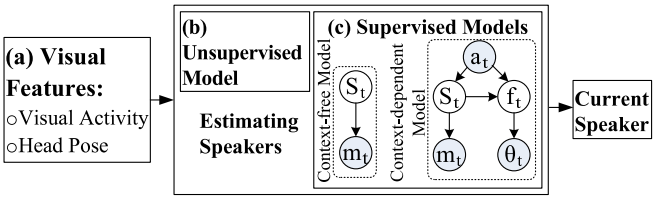


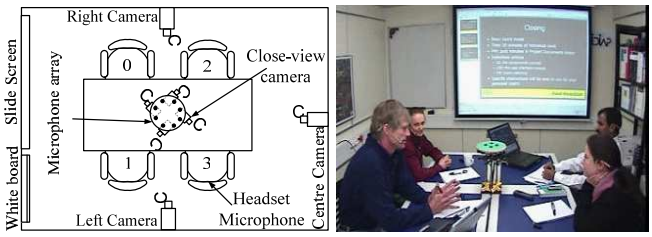
Fig. 1. Flow diagram of our approach.

In the audio domain, speech detection from independent head-set microphones (IHM) has been attempted by Wrigley et al. [6]. They used a Gaussian mixture model (GMM) to represent acoustic features which was then used to train an ergodic hidden Markov model (HMM) to estimate 4 classes related to speech and cross talk in meetings. When head-set microphones are not available, speaker diarization can be used. This tries to identify 'who spoke when' [7], which is typically approached by using an unsupervised agglomerative clustering method to identify regions of speech (after filtering out non-speech), and estimate the number of speakers. However, it is vulnerable to errors during periods of overlapping speech, even when multiple audio sources are used to estimate delays between captured audio signals. One solution is use visual cues to solve the problem audio-visually [8, 9, 10] but improvements are not always consistent so it is difficult to conclude how when they are useful.

Much previous work that exploit temporal correspondences between speech and vision have tended to assume that the motion from the mouth is the principal visual manifestation of speech [11, 8]. However, there is much evidence from both social psychology [12] and computational methods [13, 9] to suggest that speaking in conversations can manifest itself in broader body motions, which psychologists suggest aid cognitive communicative processes [12].

Speaker locationing using visual focus of attention (VFOA) has been addressed for two to three-person scenarios by Siracusa et al. [4] with good results but they used high resolution audio-visual sensors. Rienks et al. [5] used magnetic sensor information to estimate the speaker based on just each person's VFOA in discussion-only scenarios. They found that human judgements performed significantly worse than computational modelling of the same features which suggests that using VFOA alone may not be sufficient.

A summary of our approach is shown in Fig. 1 and is descriptions are arranged:- Section 2 describes the meeting data that we use; Section 3 (Fig. 1 (a)) describes both the motion features and estimates of head pose and contextual features that were extracted from the video streams; Section 4 (Fig. 1(b) and (c)) describes the different methods we use to estimate the speakers from the visual cues; Section 5 shows and discusses the experiments that were carried out; and Section 6 provides some concluding remarks.



(a) (b)

Fig. 2. (a) Plan view of the meeting room. (b) Centre camera view.

2. DATA

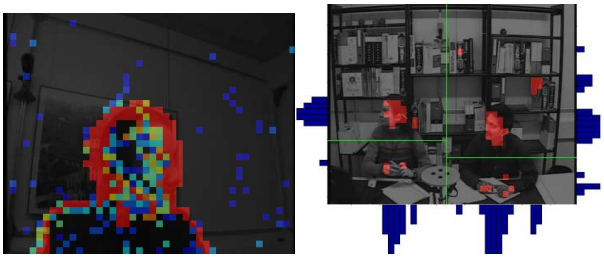
We used meeting data from the Augmented Multi-Party Meeting (AMI) corpus captured in an instrumented meeting room (see Fig. 2). These meetings were created from teams of four who were asked to design a remote control. The meetings were not scripted and team members had free use of a slide screen for presentations. The meeting room has 4 close-view cameras, capturing each individual participant (see Fig. 3(a)). There are also 2 side-view cameras which capture 2 people at a time, as shown in Fig. 3 and a centre camera that captures everyone and the slide screen. Each participant wears a headset microphone and there is also a microphone array set on the table around which the participants sit.

3. FEATURE EXTRACTION

3.1. Estimating Visual Activity

From the social psychology literature, we know that people move when they talk [3]. Therefore, measuring individual visual activity can give us a good indication of whether someone is speaking or not. We estimate body motion in the close-view video streams by extracting visual activity features directly from the compressed domain. The approach is based on features proposed by Yeo and Ramchandran [14]. The videos we use have been compressed with MPEG4 encoding using a group-of-picture (GOP) with a {I-P-P-...} structure; the first frame (I) is intra-coded, and the rest (P) are predicted. During encoding, motion vectors are obtained by matching blocks between consecutive frames using motion compensation. The difference between the source and predicted pixels (residue) needs to be encoded. The number of bits required to encode the quantised DCT coefficients is called the residual coding bitrate. An example of the residual coding bitrate extracted from a close-view camera is shown in Fig. 3(a). At each frame, the visual activity of each person is the average of the residual coding bitrate over all skin regions in each frame. Skin-colour regions are detected by modelling the distribution of the DCT of the chrominance coefficients in the UV colour space using a GMM [15].

When only 2 side-view cameras are available, 2 people are captured at a time so we implement a compressed domain, modified version of [20] to automatically divide the frame into two halves, thus separating the two people. For each frame, we construct a horizontal profile by accumulating the number of detected skin-colour blocks in each column, (see bottom of Fig. 3(b)). K-means clustering is used to find the locations of the two peaks. The cluster centres are initialised to the locations of the peaks found in the previous frame. The boundary, shown as a green vertical line, is the mid-point between the two peaks. We could have simply divided the side-view cameras equally in two but the automatic division provides a more robust estimate if one person leans towards the other to grab something or the position of the seats are changed. Once the left and right region of each camera-view is separated, we treat the two portions of the frame as two separate video streams. We average the residual coding bitrate over the skin-colour blocks in the relevant half of the frame to get the $Halves$ feature.



(a) (b)

Fig. 3. (a): Example of the residual coding bitrate in the close-view cameras. High: red; Low:blue. (b): Horizontal and vertical profiles of the skin blocks (red) and the estimated boundaries (green lines) between the two people and their respective head and hand regions in the side-view camera.

Taking the average residual coding bitrate over the side view cameras led to noisier estimates of visual activity than extracting them from the close-view cameras since head and hand motion is captured. Head and hand motion tends to be asynchronous for a speaker so averaging over all skin colour regions biases the activity values. To prevent this, we first take the average residual coding bitrate for the head ($Head$) and hand ($Hand$) regions before taking the maximum; $MaxHH = max$(Head,Hand). Head and hand motion is extracted by dividing each half of the side view using a vertical profile of the skin-colour blocks of the frame as shown by the vertically oriented histograms in Fig. 3(b). The dividing points between the head and hand regions are estimated in a similar fashion as before and are shown by green horizontal lines in Fig. 3(b); an upper bound is used to remove spurious detections of skin colour.

3.2. Estimating Head Pose

The head pose of each person is used to estimate people's VFOA in the pixel domain. To estimate people's head location and pose we rely on a Bayesian formulation of the tracking problem solved through particle filtering techniques as proposed by Ba et al. [16]. We applied the head tracking method in the side-view cameras. Given an initial head location, the tracker iteratively estimates people's head location and pose. At each time t, the tracker outputs the head locations in the image plane and the poses θ_t for each person.

3.3. Slide Change Detection

To detect the slide changes we used the method which works in the compressed domain proposed by Yeo and Ramchandran in [14]. A slide change variable is used to model contextual information for visual attention recognition. Slide changes, captured by the centre camera (see Fig. 2(b)), correspond to temporally localised peaks of the residual coding bitrate in the region corresponding to the projection screen. Thresholding the amount of residual coding bitrate in the projection screen area gives the slide change instants. Given that the slide changed at time t, we build a slide activity variable a_t that stores the time that has elapsed since the last slide change.

4. ESTIMATING SPEECH ACTIVITY FROM VIDEO

4.1. Unsupervised Model

A simple method of estimating speech activity is based on findings indicating that those who speak tend to move more [13, 9]. We implement a simple algorithm that estimates whether someone is the principal speaker based on who moves the most over a sliding time window. To ensure that a person's motion is consistently high and not just very high for a short period of time, we count the number of times someone's motion is the highest during the window (see Algorithm 1). Each speaker's visual activity is normalised by their maximum before applying the algorithm to ensure that the system is not biased to people who tend to move more in general.

foreach *t in Window* **do**
  $i = argmax_{k\in\{1..4\}}(m_t^k)$;
  $Votes^i = Votes^i + 1$;
end
$j = argmax_{k\in\{1..4\}}(Votes^k)$;
$S_t^j = 1$;

Algorithm 1: Estimating speaking, S, from visual activity, m.

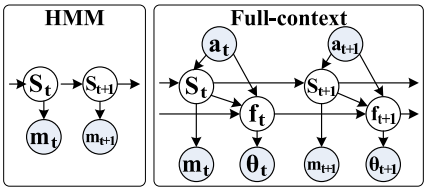


Fig. 4. Supervised models; *HMM* and *Full*.

4.2. Supervised Models

The unsupervised method can only estimate one person speaking at a time so will not account for periods of overlapping speech. Another method is to just assume that each person is likely to move when they speak; we infer that they speak when their visual activity is above a certain threshold. Before thresholding, each individual visual activity stream is normalised by dividing by the maximum for that meeting session. We will refer to this method later as $Thres$.

We now introduce a more complex model which takes into account other aspects of the meeting dynamics such as presentation activities using the slide screen and the VFOA, which has been demonstrated to be an important cue for estimating who is speaking [5]. The goal of the full-context model is to introduce in a principled fashion, information about people's visual attention to estimate their speaking status. The hypothesis is that a group's visual attention is more likely to converge on the speaker than on others so we use this contextual information for estimating the speaking status.

We use the the head pose observation model and hidden state dynamical model from our previous work [17]. We denote $S_t = (S_t^1, ..., S_t^4)$ to be the speaking states of the four meeting participants at time t. $S_t^k = 1$ when person k is speaking and 0 otherwise. $f_t = (f_t^1, ..., f_t^4)$ denotes the visual attention states of the four people. For each person, the set of possible visual attention targets has been discretised and restricted to a set of seven targets: the other three people, the table, the white board, the slide-screen and unfocused when the person is not visually looking aany of these targets. a_t is an observation variable built from the detected slide change that stores the elapsed time since the last slide change. $\theta_t = (\theta_t^1, ..., \theta_t^4)$ are the head pose observations for each person. Finally $m_t = (m_t^1, ..., m_t^4)$ are each person's estimated visual activity (see Section 3.1) over a window of fixed size centred on frame t.

Our goal is to jointly estimate the speaking states S_t and visual attention state f_t given the observations (see Fig. 4(a)). This problem can be posed in a probabilistic framework as finding the sequence of hidden states $S_{1:T}$ and $f_{1:T}$ that maximises the posterior probability distribution $p(S_{1:T}, f_{1:T}|m_{1:T}, \theta_{1:T})$ which according to the independence assumption implied by the graphical model displayed in Fig 4(a) can be factorised as:

$$p(S_0, f_0)\prod_{t=1}^{T} p(m_t|S_t)p(\theta_t|f_t)p(S_t, f_t|S_{t-1}, f_{t-1}, a_t) \quad (1)$$

The probability density function in Eq 1 is defined by five terms. The first is the initial state prior $p(S_0, f_0)$ that we modelled by a uniform distribution. The second is the motion observation model $p(m_t|S_t)$ (labelled as HMM in Fig 4(a)) modelling the relation between people's speaking observations and their speaking states. We factorise the motion observation model as $p(m_t|S_t) = \prod_{k=1}^{4} p(m_t^k|S_t^k)$ the probability $p(m_t^k|S_t^k)$ for each person when $(S_t^k = 1))$ and $(S_t^k = 0)$, is accumulated from training data (see Fig. 4(b)).

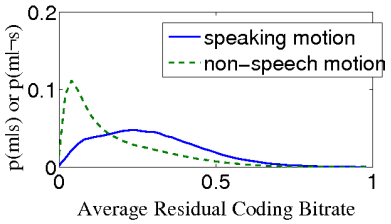


Fig. 5. Probability distributions of $MaxHH$ given $S_t = 1$ and 0.

The third term is the head pose observation model $p(\theta_t|f_t)$ relating people's head poses to their visual attention. For each visual target v, $p(\theta_t^k|f_t^k = v) = \mathcal{N}(\theta_t^k, \mu_{k,v}, \Sigma_{k,v})$ is modelled as a Gaussian distribution with mean and covariance $(\mu_{k,v}, \Sigma_{k,v})$. The parameters $(\mu_{k,v}, \Sigma_{k,v})$ can be either learned or predicted according to the geometry of the room. The last term is the dynamical model and represents the temporal evolution of the hidden states given the projection screen activities. It can be factorised as:

$$p(S_t, f_t|S_{t-1}, f_{t-1}, a_t) = p(f_t|f_{t-1}, S_t, a_t)p(S_t|S_{t-1}, a_t) \quad (2)$$

where $p(f_t|f_{t-1}, S_t, a_t)$ models the evolution and dependence of the the visual attention state given each person's speaking status and the slide screen activity. This term encodes the relationship between the visual attention and speaking behaviours. $p(S_t|S_{t-1}, a_t)$ models the evolution of the speaking status states given the time elapsed since the last slide change, representing the dependencies between the speaking statuses and the projection screen activities. The full context model will be referred to as $Full$.

5. EXPERIMENTS

First, we compare the performance of the unsupervised method (Section 4.1) using different parts of the body, shown in the upper part of Table 1. We discuss the results in terms of F-measure; the precision and recall are provided for interest. The precision represents the false detections and the recall measures missed detections. The f-measures is the harmonic mean of the precision and recall.

We performed experiments on 105 minutes of meeting data consisting of 21 5-minute meetings with 4 groups of seated people. Boundary estimation was evaluated by using annotations of bounding boxes of speakers' heads. The error rate of finding the boundary between two people was 0.4%, where an error occured when the estimated boundary did not cleanly separate the bounding boxes of the two people. The error rate for dividing the head and hands was 0.5%.

Higher resolution features extracted from the 4 individual close-view cameras (labelled as *CloseHead*) are included for comparison. Going from *CloseHead* to *Halves* leads to a decrease of 3.5% in performance in absolute terms. As expected, separating the hand and motion with *MaxHH* performed better, leading to only a 0.5% drop in performance, despite a reduction in resolution between the side view and close-view cameras. The *Hand* feature performed worst but the *Head* feature performed well compared to using *MaxHH*.

For the basic $Thres$ models, we selected a threshold so that the precision and recall were approximately equal. We compare with both *Head* and *MaxHH* features. There is a significant improvement in performance when using *MaxHH* features but overall the $Thres$ method performed worse than the unsupervised method. When *HMM* is used, the performance increases considerably but *MaxHH* still performs better than *Head*. When the *Full* model is used, the performance increases again and but now the *Head* feature performs a bit better. Closer inspection of the results reveals that the

		P	R	F
Unsupervised	**Hands**	41.32	52.6	41.85
	Head	51.87	49.5	48.38
	MaxHH	50.72	50	48.49
	Halves	48.57	49.97	46.51
	CloseHead	58.22	45.9	49.02
Supervised	**Thres Head**	45.18	36.15	38.14
	Thres MaxHH	43.22	41.93	41.00
	HMM(MaxHH)	61.64	54.36	54.83
	Full(MaxHH)	62.24	54.54	55.19
	HMM(Head)	62.99	53.23	54.45
	Full(Head)	63.36	54.74	55.92
Audio-only	**Audio1**	71.26	60.87	63.38
	Audio4	55.43	80.21	81.62

Table 1. Performance as precision (P), recall (R) and F-measure (F).

Head feature did not perform consistently better than *MaxHH*. People who used their hands more for speaking than other activities such as writing occursed in meetings where *MaxHH* performed better.

We compare our video-only methods with two different audio-only methods. The first is the speech/non-speech detector proposed by Dines et al. [18] that assumed IHM conditions. This method is referred to as $Audio4$. The other is a more challenging scenario where only a single microphone from the array is used but the number of speakers is known beforehand. We use the "NoFM" method described in [19] and is referred to as $Audio1$. The diarization was performed on each 5-minute meeting segment. Using 5-minute segments is challenging for diarization systems since each speaker has little time in which representative speaker models can be accumulated. There is typically an improvement in performance when longer conversations are used.

The results show that the diarization results improve on video-only approaches with $Audio4$ performing the best. On closer inspection, there are a few meetings segments where the the $Full$ model out-performed the $Audio1$ method by almost 20% in absolute terms. If longer meeting segments are used, the clustering performance for $Audio1$ will improve so our experiments represent the worst-case scenario and show that in the presence of short data, visual features may be a good substitute for audio-only methods.

6. CONCLUSION

Our results show that it is possible to estimate speech activity from low resolution visual features using both supervised and unsupervised methods. We have also demonstrated that using the context of the meeting to estimate who is speaking improves the overall performance and stability of the estimates. We have shown that both the visual activity of the head and hands can contribute to estimates of speaking status. The video-only methods do not out-perform the audio-only methods but our results show that it could be a reasonable substitute if periods of audio data are missing or impractical to extract. In terms of on-line real-time systems, our unsupervised model is already able to work on-line but for the $Full$ model, further work is needed. It would be interesting to investigate the effect on estimation performance of behavioural constructs such as dominance when using the noisier estimates of speaking status.

Acknowledgments We thank Chuohao Yeo for his help on extracting the compressed domain features, Yan Huang for the extracted speaker diarization segementations, John Dines for providing extracted speech/nonspeech segmentations for our experiments, and Jean-Marc Odobez for his feedback on this article.

7. REFERENCES

[1] D Babu Jayagopi, H Hung, C Yeo, and D Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE TASLP*, 2008.

[2] S. Favre, H. Salamin, J. Dines, and A. Vinciarelli, "Role recognition in multiparty recordings using social affiliation networks and discrete distributions," in ICMI, 2008.

[3] F Quek, D McNeill, R Bryll, S Duncan, X-F Ma, C Kirbas, K E. McCullough, and Rashid Ansari, "Multimodal human discourse: gesture and speech," *ACM Trans. Comput.-Hum. Interact.*, vol. 9, no. 3, pp. 171–193, 2002.

[4] M Siracusa, K Wilson, J Fisher, and Trevor Darrell, "A multimodal approach for determining speaker location and focus," in *ICMI*, 2003.

[5] R. Rienks, R. Poppe, and D. Heylen, "Differences in head orientation between speakers and listeners in multi-party conversations," *International Journal HCS*, 2005.

[6] SN Wrigley, GJ Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Transactions on Speech and Audio Processing*, pp. 84–91, 2004.

[7] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *ICASSP*, 2005.

[8] A Noulas and Ben J. A. Krose, "On-line multi-modal speaker diarization," in *ICMI*, 2007, pp. 350–357, ACM.

[9] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi, "Audio segmentation and speaker localization in meeting videos," *ICPR*, vol. 2, pp. 1150–1153, 2006.

[10] G. Friedland, H.Hung, and C.Yeo, "Multi-modal speaker diarization of real-world meetings using compressed-domain video features," in *ICASSP*, April 2009.

[11] H J Nock, G Iyengar, and C Neti, "Speaker localisation using audio-visual synchrony: An empirical study," in *CIVR*, 2003

[12] D. McNeill, *Language and Gesture*, Cambridge University Press New York, 2000.

[13] H Hung, Y Huang, C Yeo, and D Gatica-Perez, "Associating audio-visual activity cues in a dominance estimation framework," in *CVPR Workshop on Human Communicative Behavior*, Ankorage, Alaska, 2008.

[14] C Yeo and K Ramchandran, "Compressed domain video processing of meetings for activity estimation in dominance classification and slide transition detection," Tech. Rep. UCB/EECS-2008-79, EECS Department, University of California, Berkeley, June 2008.

[15] S J McKenna, S Gong, and Y Raja, "Modelling facial colour and identity with Gaussian mixtures," *Pattern Recognition*, vol. 31, no. 12, pp. 1883–1892, 1998.

[16] S. O. Ba and J.-M. Odobez, "A Rao-Blackwellized mixed state particle filter for head pose tracking," in *ICMI Workshop on Multimodal Multiparty Meeting Processing*, 2005, pp. 9–16.

[17] S. O. Ba, Hayley Hung, and J.-M. Odobez, "Visual activity context for focus of attention estimation in dynamic meetings," in *ICME*, 2009.

[18] J Dines, J Vepa, and Thomas Hain, "The segmentation of multi-channel meeting recordings for automatic speech recognition," in *INTERSPEECH*, 2006.

[19] Y. Huang, O. Vinyals, G. Friedland, C. Müller, N. Mirghafori, and C. Wooters, "A fast-match approach for robust, faster than real-time speaker diarization," in *ASRU*, 2007.

[20] A. Jaimes, Posture and activity silhouettes for self-reporting, interruption management, and attentive interfaces, Intelligent user interfaces,2006, pp. 24–31.