IDIAP RESEARCH REPORT

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

Hayley Hung Silèye O. Ba

Idiap-RR-20-2009

JULY 2009

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Speech/Non-Speech Detection in Meetings from Automatically Extracted Low Resolution Visual Features

Anon
Anon
anon@corporation.com

Anon
Anon
anon@corporation.com

ABSTRACT

In this paper we address the problem of estimating who is speaking from automatically extracted low resolution visual cues from group meetings. Traditionally, the task of speech/non-speech detection or speaker diarization tries to find who speaks and when from audio features only. Recent work has addressed the problem audio-visually but often with less emphasis on the visual component. Due to the high probability of losing the audio stream during video conferences, this work proposes methods for estimating speech using just low resolution visual cues. We carry out experiments to compare how context through the observation of group behaviour and task-oriented activities can help improve estimates of speaking status. We test on 105 minutes of natural meeting data with unconstrained conversations.

1. INTRODUCTION

As visual sensors become cheaper, recording meetings and transmitting them live has become a reality for many. Instrumented meeting rooms can have both audio and video sensors but there is always the possibility of failure to record one of the modalities or periods of drop-out. This paper addresses the problem of estimating speakers in four-participant meeting conversations when only low resolution video data is available. We compare both supervised and unsupervised models and investigate to what extent different contextual cues can be used to aid the estimation of who is speaking.

There has been research studying how either head or hand gestures are related semantically to speech [12, 9], but to our knowledge, there has been no study of 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. Estimate this basic unit of turn-taking when only video is available is useful for analysing semantically high-level non-verbal group behaviour e.g. who is dominant [6].

Speaker locationing using visual focus of attention (VFOA) has been addressed for two to three-person scenarios by Siracusa et al. [16] with good results but they used high resolution audio-visual sensors. Rienks et al. [14] 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.

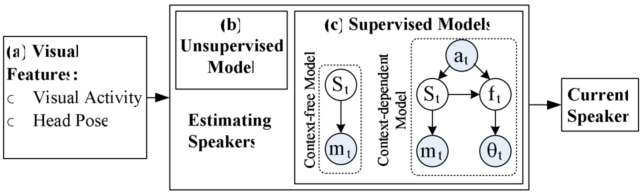


Figure 1: Flow diagram of our approach.

In the audio domain, speaker diarization tries to identify 'who spoke when' [13]. The task is an unsupervised agglomerative clustering method that identifies regions of speech (after filtering out non-speech), and also estimates the number of speakers. The task should be invariant to speaker location when using either a single or multiple microphones. It is vulnerable to errors during periods of overlapping speech, even in cases where the multiple audio sources can be used to estimate delays between captured audio sources. One solution is to try and solve the problem, audio-visually [11, 18, 3] but the visual stream can be given much less weighting than the audio features for empirical reasons but we do not know why this is or how the visual features contribute.

Much previous work that exploit temporal correspondences between speech and vision have tended to test on scenarios where the primary assumption is that the motion from the mouth is the principal visual manifestation of speech [10, 17, 11]. However, there is much evidence from both social psychology [8] and also computational methods [5, 18] to suggest that in group conversations, speaking can manifest itself in much broader body motions, which psychologists suggest aid the cognitive communicative processes.

A summary of our approach is shown in Figure 1 and is descriptions are arranged:- Section 2 describes the meeting data that we use; Section 3 (Figure 1 (a)) describes both the motion features and estimates of head pose and contextual features that were extracted from the video streams; Section 4 (Figure 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; Section 6 provides some concluding remarks.

2. DATA

We used meeting data from the Augmented Multi-Party Meeting (AMI) corpus captured in an instrumented meeting room (see Figure 2). These meetings were created from teams of four who were asked to design a remote control.

WOODSTOCK '97 El Paso, Texas USA

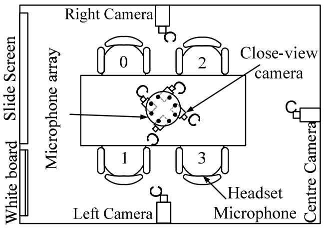


Figure 2: Plan view of the meeting room.

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 Figure 3(a)). There are also 2 side-view cameras which capture 2 people at a time, as shown in Figure 3 and a centre camera that captures everyone and the slide screen. In terms of audio sensors, each participant wears a headset microphone and also there is a microphone array set on the table around which the participants sit.

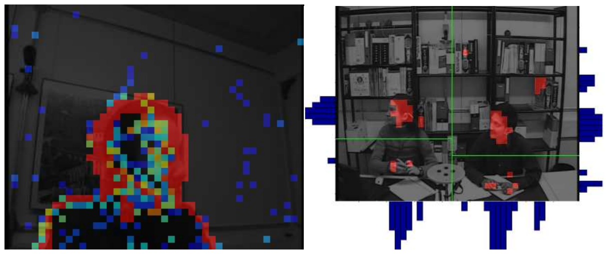
3. FEATURE EXTRACTION

3.1 Estimating Motion

We estimate body motion in the close-view video streams by extracting visual activity features directly from the compressed domain. The videos we use have been compressed using MPEG4 encoding. We use the method proposed by Yeo and Ramchandran [19]. We use the residual coding bitrate to represent visual activity levels. For our dataset, we used an MPEG-4 encoder with 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. There is rarely a perfect match and the difference is the residual. The number of bits required to encode the quantised DCT coefficients of the residual signal is called the residual coding bitrate. An example of the residual coding bitrate extracted from a close-view camera is shown in Figure 3(a). At each frame, the visual activity of each person is the average of the residual coding bitrate over each frame and regions of skin. Skin-colour regions are detected by modelling the distribution of the DCT of the chrominance coefficients in the UV colour space using a Gaussian Mixture Model [7].

When only 2 side-view cameras are available, 2 people are captured at a time so we need to automatically divide the frame into two halves the person on the left and right hand side. For each frame, we construct a horizontal profile by accumulating of the number of detected skin-colour blocks in each column, (see bottom of Figure 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 automatic dividing 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.

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 hand motion is captured and head motion is less pronounced. Head



(a) (b)

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

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 Figure 3(b). The estimated dividing points between the head and hand regions are shown by green horizontal lines in Figure 3(b) where an upper bound is used to remove spurious detections of skin colour in the background.

3.2 Estimating Head Pose

The head pose of each person is used to estimate people's VFOA. 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. [1]. We applied the tracking method to track heads in the side-view cameras. Given an initial head location, the tracking method iteratively estimate people's head location and pose. At each time t, the tracker outputs the head locations in the image plane and the head poses θ_t of 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 [19]. Slide changes, captured by the centre camera, correspond to temporally localised peaks of the residual coding bitrate in the image area 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. The variable is used to model contextual information for visual attention recognition.

4. ESTIMATING SPEAKING FROM VIDEO

4.1 Unsupervised Model

A simple method of estimating speaking activity is based on findings indicating that those that speak tend to move more [5, 18]. 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. Each speaker's visual activity is normalised by their maxi-

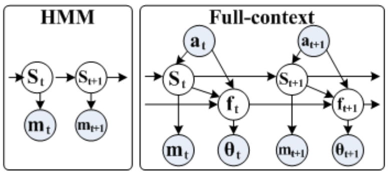


Figure 4: Supervised models;*HMM* and *Full*.

mum before applying the algorithm. We will refer to this method as $HighSus$. The algorithm is shown below:

foreach t *in* $Window$ **do**
  $i = Argmax_{k \in \{1..4\}}(VisualActivity[t, k]);$
  $Votes[i] = Votes[i] + 1;$
end
$j = Argmax_{k \in \{1..4\}}(Votes[k]);$
$HighSus[j] = 1;$

Algorithm 1: Estimating speakers using visual activity.

4.2 Supervised Models

The $HighSus$ scheme can only estimate one person speaking at a time so it will not during periods of overlapping speech. Another method is to just assume that each person is likely to move when they speak and therefore 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 which use the slide screen. Also, the VFOA has been shown to be an important cue for estimating who is speaking [14]. 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 underlying hypothesis is that people's attention is more likely to converge on the speaker rather than the others. People's visual attention plays the role of contextual information for estimating the speaking status.

The method we use is based on the work of Ba et al [15] who were trying to esimate VFOA from contextual cues; we use their head pose observation model and hidden state dynamical model. Let us denote by $S_t = (S_t^1, ..., S_t^4)$ the speaking states of the four meeting participants. $S_t^k = 1$ 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 three other persons, the table, the white board, the projection screen and unfocused when the person is not visually focused on any of the previously mentioned targets. a_t is an observation variable built from the detected slide change that stores the time that has passed since the last slide change. $\theta_t = (\theta_t^1, ..., \theta_t^4)$ are the head pose observation for each person. And finally $m_t = (m_t^1, ..., m_t^4)$ are each person's visual activity 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 Figure 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 (pdf) $p(S_{1:T}, f_{1:T}|m_{1:T}, \theta_{1:T})$ 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 Figure 4(b)). 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 evolution and dependence of the the visual attention state given people's speaking statuses 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

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%.

First, we compare the performance of the *HighSus* method (Section 4.1) using different parts of the body as shown in the upper part of Table 1. We discuss the results in terms of F-measure only due to space limitations; the precision and recall are provided for interest. Higher resolution features extracted from the 4 individual close-view cameras (labelled as *CloseHead*) are included for interest. 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

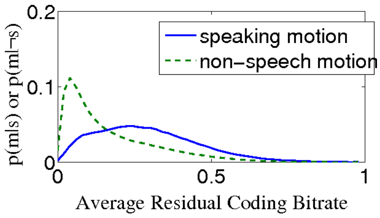


Figure 5: Probability distributions of $MaxHH$ given $S_t = 1$ and $S_t = 0$.

between the side view and close-view cameras. The *Hand* feature performed worst but the *Head* feature performed comparably well to the *MaxHH* feature.

		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
	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: Summary of all results using precision (P), recall (R) and F-measure (F).

For the basic thresholded models, we selected a threshold where 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 basic thresholded method performs worse than using *HighSus*. When *HMM* is used, the performance increases considerably but the *Head* feature still performs less well than *MaxHH*. When the *Full* model is used, the performance increases again and this time the *Head* feature performs slightly better. Closer inspection of the results revealed that the *Head* feature did not perform consistently better than *MaxHH*. Meetings where *MaxHH* performed better contained people who used their hands more for speaking than other activities such as writing.

We compare our video-only methods with two different audio-only methods. The first is a speech/non-speech detector which estimates whether the wearer of the microphone is speaking which assumes that each potential speaker wears their own microphone [2]. This method is referred to as *Audio*4. 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 [4] and is referred to as *Audio*1. 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.

Surprisingly, the results show that the diarization results improve on video-only approaches with *Audio*4 performing the best. On closer inspection, there are a few meetings segments where the the *Full* model out-performed the *Audio*1 method by almost 20% in absolute terms. If longer meeting segments are used, the clustering performance 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 speaking 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 or results show that it could be a reasonable substitute if periods of audio data are missing. 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 whether it is possible to identify the number of participants in the meeting from video-only methods.

7. REFERENCES

[1] 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*, pages 9–16, 2005.

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

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

[4] 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.

[5] 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.

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

[7] S J McKenna, S Gong, and Y Raja. Modelling facial colour and identity with Gaussian mixtures. *Pattern Recognition*, 31(12):1883–1892, 1998.

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

[9] L.P. Morency, I. Kok, and J. Gratch. Predicting Listener Backchannels: A Probabilistic Multimodal Approach. *Lecture Notes in Computer Science*, 5208:176–190, 2008.

[10] H J Nock, G Iyengar, and C Neti. Speaker localisation using audio-visual synchrony: An empirical study. In *CIVR*, pages 488–499, 2003.

[11] A Noulas and Ben J. A. Krose. On-line multi-modal speaker diarization. In *ICMI*, pages 350–357, New York, USA, 2007. ACM.

[12] 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.*, 9(3):171–193, 2002.

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

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

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

[16] M Siracusa, K Wilson, J Fisher, and Trevor Darrell. A multi-modal approach for determining speaker location and focus. In *ICMI*, pages 77–80, 2003.

[17] M.R. Siracusa and J.W. Fisher. Dynamic dependency tests for audio-visual speaker association. In *ICASSP*, 2007.

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

[19] C Yeo and K Ramchandran. Compressed domain video processing of meetings for activity estimation in dominance classification and slide transition detection. Technical Report UCB/EECS-2008-79, EECS Department, University of California, Berkeley, Jun 2008.