Interspeech 2018
2-6 September 2018, Hyderabad

Improving Response Time of Active Speaker Detection using Visual Prosody Information Prior to Articulation

Fasih Haider[1,2] , *Saturnino Luz*[1], *Carl Vogel*[2] *and Nick Campbell*[2]

[1]IPHSI, University of Edinburgh, UK [2]ADAPT Centre, Trinity College Dublin, Ireland
{Fasih.Haider, S.Luz}@ed.ac.uk, {vogel, nick}@scss.tcd.ie

Abstract

Natural multi-party interaction commonly involves turning one's gaze towards the speaker who has the floor. Implementing virtual agents or robots who are able to engage in natural conversations with humans therefore requires enabling machines to exhibit this form of communicative behaviour. This task is called active speaker detection. In this paper, we propose a method for active speaker detection using visual prosody (lip and head movements) information before and after speech articulation to decrease the machine response time; and also demonstrate the discriminating power of visual prosody before and after speech articulation for active speaker detection. The results show that the visual prosody information one second before articulation is helpful in detecting the active speaker. Lip movements provide better results than head movements, and fusion of both improves accuracy. We have also used visual prosody information of the first second of the speech utterance and found that it provides more accurate results than one second before articulation. We conclude that the fusion of lip movements from both regions (the first one second of speech and the one second before articulation) improves the accuracy for active speaker detection.

Index Terms: social signal processing, human-computer interaction, situated interaction, active speaker detection, visual prosody

1. Introduction

Dialogue has two main components, one is verbal (the actual spoken content), and other is non-verbal (e.g. prosody, gaze). In order for a machine to be able to manage these two components when engaged in a dialogue with several humans it needs to be able to detect which speaker holds the floor. If the machine is to be seen as a believable participant in a communication [1], it should seem to turn its visual attention towards the current active speaker, and achieve realistic production and attunement to gaze, lip and head movements. An active speaker detection system can be used in a robot to aid the generation of the multimodal output (moving its head or gaze towards the speaker) particularly in situated interactions [2, 3, 4, 5, 6]. In Human-Human interaction, it is observed that the listener turns their gaze towards the speakers around 30-–80% of the time [7]. Hence, from the social robotics perspective, it is useful to detect the active speaker as soon as possible to enable the robot to turn gaze/head towards the speaker to show that it is attending to the speaker. In particular, it is useful if one may anticipate who the next active speaker will be, in order to accelerate this process.

This study continues the authors' past work [8, 9] which demonstrated the use of lip and head movements during speech articulation for active speaker detection but did not assess the discriminative power of visual prosody data captured just before and/or after articulation. In this study, we propose methods for detection of active speakers through use of visual prosody information one second before/after speech articulation and also evaluate the visual prosody information of the first second of the speech utterance. The system architecture is depicted in Figure 1. The system processes visual information before articulation from the memory buffer as soon as Voice Activity Detection (VAD) detects 10 ms of voice activity. The proposed methods are a step towards decreasing the response time of a robot in generating multimodal attention towards the user in situated interactions and experimental findings help in understanding the discrimination power of visual prosody for those regions (one second before/after the articulation). To the authors' best knowledge, it is the first automatic active speaker detection system with input from one camera which uses the visual information particularly head movements before articulation. Moreover, it is the first study which demonstrates the discrimination power of visual prosody (one second before and after articulation) for active speaker detection.

Multiple studies explore the use of visual information and its fusion with acoustic information to increase the performance of voice/speaker detection. Takeuchi et al. extract the low-level visual descriptors (optical flow vectors) from the mouth region for speech activity detection [10]. Viola et al. use the appearance and motion cues of humans [11] for speaker detection, with a dataset collected in a distributed meeting setting using smart rooms [12]. Some studies the audio and visual features to detect the speaker in videos and human-machine interaction [13, 14, 15, 5]. However, these studies do not focus on improving the response time of speaker detection.

The facial dynamics of a person can be used to detect the voice, and can also be helpful in predicting an active speaker from a set of subjects facing a robot. In previous studies, lip movements are considered an important signal which can increase the speech intelligibility [16, 17, 18]. While head movements have been less explored in this kind of task, perception studies have found that head movements are correlated with prosody and improve the speech intelligibility [19], as well as reveal prosodic structure [20]. Ishii et al. use manually annotated mouth opening transition patterns after the subject stops speaking to predict the next speaker in a meeting [21] and also

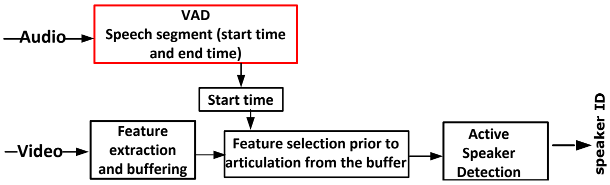
VAD
Speech segment (start time and end time)
Audio
Start time
Video
Feature extraction and buffering
Feature selection prior to articulation from the buffer
Active Speaker Detection
speaker ID

Figure 1: *The proposed system architecture for active speaker detection.*

10.21437/Interspeech.2018-2310

use respiration signal to predict the next speaker [22]. In contrast, our study focuses on automatic detection of active speakers without relying on knowledge of previous speakers. Murai proposed a speaker predicting system for teleconferencing [23] which uses an image prior to articulation for predicting the speaker. The system proposed by Murai needs a miniature camera and microphone for each speaker to detect who is going to speak along with a high resolution camera which turns and focus the active speaker (smart room setting) in a video conference. However, this study uses only one camera and microphone.

Although in previous studies, we have demonstrated that head and lip movements during speech articulation can predict the active speaker with promising results [8, 9], the response time issues and evaluation of visual prosody information just before/after articulation are left open. Cech et al. report on an active speaker detection system for a humanoid robot that uses audio and visual information of four microphones and two cameras [5]. Some studies propose models to detect speakers in videos [13, 14, 15], where the objective is not to generate a real-time multimodal output (gaze and head) of a robot. In those settings, a quick response is not needed as much as in multimodal interaction with robots.

2. Data Collection

An audio-visual dataset [8, 9] was collected in a task-free dialogue setting. Four participants (3 males and 1 female) converse with the "machine", but they are not allowed to speak with each other directly. They are free to gesture, display emotions, etc. so long as this does not change their location. The machine perspective is simulated by a fifth person (S0) in a separate location, using video conferencing software, and his face is displayed on the computer screen visible to the other four. The motivation for using full facial information is to simulate a humanoid robot/avatar that can handle and generate social signals and behaviours. The recording session consisted of two parts. In part one, the subjects (in-front of the camera) ask questions from the machine (S0) one by one through video conferencing, and the machine (S0) answers them. In the second part, the machine (S0) asks the questions from the other participants. Some sample questions are as follows: 1. Where is the nearest train station? 2. How can I reach to the football ground? 3. Where can I find a place for lunch?

A high-definition JVC video camera was used for recording the session. It recorded the video with a frame rate of 25 fps, and the duration of the video (dialogue) is 21 minutes. The distance between speakers and machine interface (S0) was approximately 2 meters. The ELAN annotation software was used for the annotation of speech segments of each subject [24]. The total conversation resulted in 81, 62, 58 and 18 speech segments for S1, S2, S3 and S4 respectively. However, since subject S4 (a male speaker) has very few speech segments, we ignored S4's data. Then we selected those speech segments which have at least 2-second pause before articulation that results in 50, 49, and 41 speech segments for S1, S2 and S3 receptively. Overall we selected 41 speech segments from each subject so as to use the balanced dataset for experimentation. The dataset is further divided into two subset to explore the generalizability of this study. One subset is used for Pearson correlation test and the other used for classification task.

3. Feature Extraction

The FaceAPI SDK [25] was used for tracking of facial landmarks and head coordinates for every speaker. FaceAPI is a commercially available software (a product of Seeing Machines) capable of tracking head pose and lip location as well as the location of jaw, eyebrows and eyes. Features used in this study are the lips' inner height, outer height and width (in meters) which are calculated by the position (x, y and z in meters) of face landmark ID numbers 101, 104, 202, 206, 200 and 204 as shown in Fig. 2 and the head rotation along x, y and z-axis (in radians). To calculate the feature set, we take the arithmetic average of the first derivative of lip (inner height, outer height and width) and head coordinates (x,y and z coordinates) for each frame. Then we calculate the mean and standard deviation values for 25 consecutive frames for 'going to speak', 'silence' and 'speech' regions as depicted in Figure 3. As a result, the final feature vector contains four features for each subject (2 for lip and 2 for head movements).

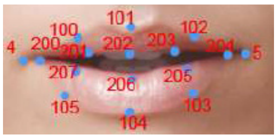

Figure 2: *The face tracking API lip tracking points [25]*

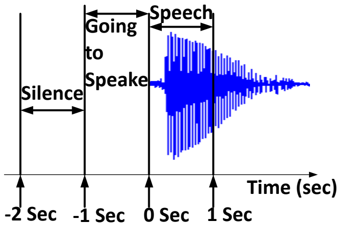

Figure 3: *Regions of interest for 'Going to Speak' (GTSR), Silence (SilR) and Speech (SR)*

4. Pearson correlation test

In line with evidence of low variation in electrical potential up to one second before the speech articulation in the human brain [26] we hypothesise that visual prosody also shows some characteristics (e.g. subject is going to start moving his or her lips for articulation) that are manifest during the one second before articulation. To validate this assumption, we performed Pearson correlation test using a subset of the corpora. For the test, we considered 20 speech segments for each subject and extracted features (mean value of the rate of change in lip and head movements) from the regions shown in Figure 3. We have 20 instances of Silence Region (SilR), Going To Speak Region (GTSR) and Speech Region (SR) for each subject. We defined SR as the first second of a speech utterance, and the GTSR is defined as a time window of one second before SR. The SilR is defined as a time window of one second before GTSR. All these regions are concatenated as depicted in Figure 3. The results of Pearson correlation test along with the null hypothesis are described below. This test helps us in finding the correlation between visual prosody of SR, GTSR and SilR.

Hϕ^1: There is no correlation between the visual prosody of SilR and SR data.

The Pearson correlation failed to rejected this null hypothesis at the $p_{\text{GTSR-SilR}} < 0.05$ significance level in all cases, as depicted in Table 1.

Hϕ^2: There is no correlation between the visual prosody of SR and GTSR data.

For this hypothesis, the Pearson correlation test rejected the null hypothesis ($p_{\text{GTSR-SR}} < 0.05$) in 4 out of 6 cases as depicted in Table 1. Only for S1 head and S2 lip data, the test was unable to reject the null hypothesis ($p_{\text{GTSR-SR}} > 0.05$).

Hϕ^3: There is no correlation between the visual prosody of GTSR and SilR data.

For Hϕ^3, the Pearson Correlation test rejected the null hypothesis ($p_{\text{GTSR-SilR}} < 0.05$) in 2 out of 6 cases as depicted in Table 1. S3 lip and S1 head data showed statistically significant correlation ($p_{\text{GTSR-SilR}} < 0.05$).

Table 1: *Pearson Correlation test results (statistical significance (p) and correlation coefficient (r)) for Silence Region (SilR), Speech Region (SR) and Going To Speak Region (GTSR).*

Feature	Subject	SilR-SR		GTSR-SR		GTSR-SilR	
		$r_{\text{SilR-SR}}$	$p_{\text{SilR-SR}}$	$r_{\text{GTSR-SR}}$	$p_{\text{GTSR-SR}}$	$r_{\text{GTSR-SilR}}$	$p_{\text{GTSR-SilR}}$
Lip	S1	0.164	0.489	**0.668**	**0.001**	0.297	0.204
	S2	0.065	0.786	-0.089	0.709	0.067	0.781
	S3	0.211	0.372	**0.626**	**0.003**	**0.627**	**0.003**
Head	S1	0.078	0.743	0.238	0.313	**0.687**	**0.001**
	S2	0.443	0.051	**0.656**	**0.002**	-0.151	0.525
	S3	-0.070	0.769	**0.791**	**0.000**	0.116	0.625

From these correlation tests, we conclude that: 1) the GTSR is highly correlated with the SR and this correlation is statistically significant ($p < 0.05$) in 4 out of 6 cases. It suggests that every speaker has some visually detectable means (i.e. head and/or lip movements) of communicating their intention to speak; 2) the GTSR is less correlated with the SilR than SR, and this correlation is statistically significant ($p < 0.05$) in 2 out of 6 cases; 3) the data shown no significant correlation between SR and SilR.

5. Active speaker detection experiments

Given that SR information seems better correlated with GTSR than SilR, we created feature sets that reflected this fact, and trained models for an automatic active speaker detection system using classification methods. We have performed three experiments using three different feature sets for classification as described below:

Experiment One: In this experiment, we have extracted the features one second before articulation (GTSR, see Figure 3).

Experiment Two: In this experiment, we have extracted the features one second after articulation. (SR, see Figure 3).

Experiment Three: In this experiment, we have fused the previous two experiments features.

5.1. Classification Methods

The classification is performed using four different methods, namely Linear Discriminant Analysis (LDA), Naïve Bayes (NB), Nearest Neighbour (KNN with K=1) and Decision Trees (DT). These classifiers are employed in MATLAB[1] using the statistics and machine learning toolbox in the 10-fold cross-validation setting. LDA works by assuming that the feature sets of the classes to be discerned are drawn from different Gaussian distributions and adopting a pseudo-linear discriminant analysis (i.e. using the pseudo-inverse of the covariance matrix [27]). The NB model we employed also assumes a kernel distribution for the feature set. KNN and DT are non-parametric, and non-linear classification methods.

6. Results and Discussion

The classification is performed on subset of the dataset, Where each subject has 21 instances. In this case, blind guess and majority guess are the same (33.33%), and we set it as a baseline for the classification task. The results of experiment one are shown in Table 2. It is observed that the lip movements (42.86%) provide better results than the head (38.10%) and fusion of lip and head movements (47.62%) improves the performance. The LDA classifier provides the best results. However, the fusion of lip and head movements does not improve accuracy for NB and KNN, and overall lip movements provide better results than head movements. This probably reflect the fact that these models are more prone to being misled by irrelevant features than LDA, as is well known of KNN, for instance.

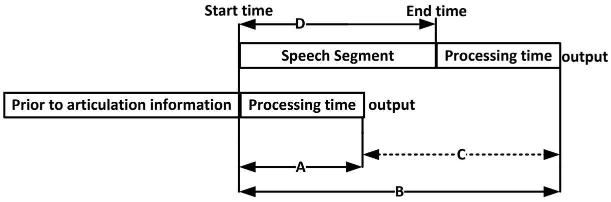

Figure 4: *Highlighted the proposed system response time (A), baseline response time (B), improvement in response time (C) and duration (couple of seconds) of Speech segment (D). The output is the predicted label and processing time is the time taken by a machine's processor for classification purpose.*

Table 2: *Accuracy (%) for experiment one (10-fold cross-validation): facial features one second before articulation.*

Feature	Baseline	KNN	DT	NB	LDA
Head	33.33	31.75	28.57	25.40	38.10
Lip	33.33	36.51	38.1	30.16	42.86
Fusion	33.33	28.57	41.27	30.16	**47.62**

The results of experiment two are shown in Table 3. It is observed that the lip movements (52.38%) provide better results than the head (46.03%) and fusion of lip and head movements (49.21%) cause a slight increase in accuracy. The LDA classifier provides the best results. The fusion of lip and head movements do not improve accuracy.

Table 3: *Accuracy (%) for experiment two (10-fold cross-validation): features one second after articulation.*

Feature	Baseline	KNN	DT	NB	LDA
Head	33.33	39.68	46.03	42.86	44.44
Lip	33.33	42.86	44.44	42.86	**52.38**
Fusion	33.33	36.51	46.03	42.86	49.21

The results of experiment three are depicted in Table 4. It is observed that the lip movements provide better results than head movements using LDA, DT and KNN classifiers. The fusion of lip and head movements improves accuracy for LDA.

From the above three experiments results, it is observed that the visual prosody one second before articulation provide good results for active speaker detection. This can be due to the fact

[1] http://uk.mathworks.com/products/matlab/ – Last verified June 2017

Table 4: *Accuracy (%) experiment Three (10-fold cross-validation): fused features.*

Feature	Baseline	KNN	DT	NB	LDA
Head	33.33	30.16	41.27	37.10	39.68
Lip	33.33	33.33	50.79	36.51	52.38
Fusion	33.33	30.16	42.86	33.87	**55.56**

that the subjects start moving their lips before articulation of speech.

We use a Venn diagram to visualise the range of classification overlaps of the best performing classifier (LDA) for each experiment. In Figure 5, the red circle (Target) represents the annotated labels, the yellow circle (Exp.2) represents the predicted labels by the lip movements in experiment two, the blue circle (Exp.1) represents the predicted labels by the fusion of head and lip movement in experiment one, and finally the green circle (Exp.3) represents the predicted labels by the fusion of head and lip movements in experiment three. It is observed that the yellow and green circles have the highest overlap (38 out of 63), and that both these circles have an overlap of 35 samples with the red circle (Target). It is also observed that there are 12 instances (4 of S1, 6 of S2 and 2 of S3) which have no overlap with any circle. There are 16 instances (2 of S1, 7 of S2 and 7 of S3) which have been detected by all the three experiments as depicted in Figure 5. We also compare the predictive accuracies of our three best results using the mid-p-value McNemar test (testcholdout[2]) with a null hypothesis that predicted labels of Exp.1, Exp.2 and Exp.3 have equal accuracy for predicting the target. The statistical test was unable to reject the null hypothesis ($p_{\text{Exp.1-Exp.2}} = 0.58$, $p_{\text{Exp.1-Exp.3}} = 0.29$ and $p_{\text{Exp.2-Exp.3}} = 0.66$) and shows that although GTSR provides less accurate results than SR and fusion of both regions but the difference is not statistically significant. Hence demonstrating that the GTSR and SR regions have similar characteristics for active speaker detection. In previous studies, we could distinguish the speech/non-speech frames and were also able to classify the active speaker with good accuracy using visual prosody information during speech articulation. However, the proposed methods were not developed for a quick response time, and the main objective was to evaluate the lip and head movements as discriminative features for active speaker detection in human-machine multi-party setting [8, 9].

Most studies [28, 13] to date process the full speech segment acoustic and visual information ('D' region as depicted in Figure 4) and then assign each utterance a speaker label that added a latency and decrease the response time of a machine that may results in turning the gaze and head of a robot to the active speaker in its view after the subject finished speaking ('B' region as depicted in Figure 4). In a previous study, we evaluated the head and lip movements of the 'D' region for speaker detection and observed an average of 71.29% accuracy using lip movements on the same dataset used in this study [8]. While the accuracy for the 'D' region is better than GTSR and SR regions for speaker detection, the former will generate a multimodal output for a robot only after processing the full speech segment with a duration of some seconds (depending on the speech segment length, which is typically couple of second to around 20 seconds, plus processing time). The latter will generate the output after 0-1 second (plus processing time) of speech articulation. The current and previous study [8] both require an input from audio-VAD. The strength of the current study is its focus on quick response time ('A' region as depicted in Figure 4) which can increase the naturalness of a machine in a human-machine multi-party interaction. In another study [9], we proposed a visual active speaker detection system at frame level which do not need an input from audio-VAD to operate and detect the speaker at video segment level. This involved processing of consecutive 25 frames (1 second of video segment) with an overlap of 24 frames with neighbouring video segment, hence detecting who is speaking in each video frame, instead of speech segment level (detected by audio-VAD) using lip and head movements with GTSR treated as SilR [9]. While we observed high accuracy ($> 90\%$) in classifying video segments of active speaker, that study did not explicitly demonstrate the discrimination power of speech frames (video segments) one second before/after articulation, which this study covers. Based on the experimental findings (GTSR is more correlated with SR than SilR), we recommend that GTSR should be treated as SR instead of SilR for the development of visual active speaker detection systems for noisy environments.

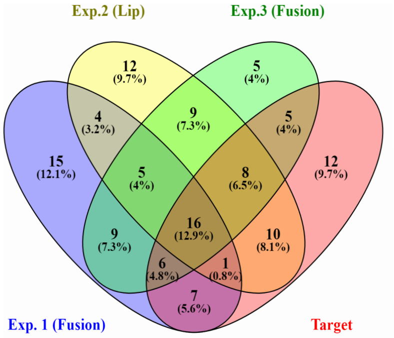

Figure 5: *Venn Diagram of the best results of three experiments and annotated labels (Target).*

7. Conclusion

The results show that the period of one second before speech articulation contains useful information about who holds the floor in a dialogue. The visual prosody features extracted from this region provide less accurate results than the speech region for the classification task but the difference is not statistically significant. The fusion of features from both regions improves performance in LDA and DT classifiers. Although these results are promising, one should interpret them with caution due to the small size of the dataset and the possibility of model overfitting, as suggested by the results of KNN, DT and NB on GTSR-only models. Possible future work is to evaluate the low-level visual descriptors (e.g. histogram of the gradient) extracted from 'going to speak' and 'speech' region for active speaker detection and its fusion with the audio features. Another possible future work is to detect the 'going to speak region' by using visual information only instead of relying on 10 ms of speech utterance.

8. Acknowledgment

This research is supported by EU H2020 project SAAM under grant number 769661 at the University of Edinburgh, UK, and "ADAPT 13/RC/2106" project (http://www.adaptcentre.ie/) in the Speech Communication Lab at Trinity College Dublin, the University of Dublin, Ireland. The data set is collected under the supervision of Dr. Samer Al Moubayed at the Department of Speech Music and Hearing, the Royal Institute of Technology, Stockholm, Sweden.

[2] https://uk.mathworks.com/help/stats/testcholdout.html (Aug 2017)

9. References

[1] A. Esposito, A. M. Esposito, and C. Vogel, "Needs and challenges in human computer interaction for processing social emotional information," *Pattern Recognition Letters*, vol. 66, pp. 41–51, 2015.

[2] J. Han, E. Gilmartin, and N. Campbell, "Herme, yet another interactive conversational robot," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 711–712.

[3] A. D. Christian and B. L. Avery, "Digital smart kiosk project," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co., 1998, pp. 155–162.

[4] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 119–155, 2003.

[5] J. Cech, R. Mittal, A. Deleforge, J. Sanchez-Riera, X. Alameda-Pineda, and R. Horaud, "Active-speaker detection and localization with microphones and cameras embedded into a robotic head," in *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*. IEEE, 2013, pp. 203–210.

[6] H. Sansen, M. I. Torres, G. Chollet, C. Glackin, D. Petrovska-Delacretaz, J. Boudy, A. Badii, and S. Schlögl, "The roberta ironside project: A dialog capable humanoid personal assistant in a wheelchair for dependent persons," in *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. IEEE, 2016, pp. 381–386.

[7] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta psychologica*, vol. 26, pp. 22–63, 1967.

[8] F. Haider and S. Al Moubayed, "Towards speaker detection using lips movements for humanmachine multiparty dialogue," *The XXVth Swedish Phonetics Conference (FONETIK)*, pp. 117–120, 2012.

[9] F. Haider, S. Luz, and N. Campbell, "Active speaker detection in human machine multiparty dialogue using visual prosody information," in *Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. Washington, D.C., USA: IEEE, 2016, pp. 1207–1211.

[10] S. Takeuchi, T. Hashiba, S. Tamura, and S. Hayamizu, "Voice activity detection based on fusion of audio and visual information," *Proceedings of the Auditory-Visual Speech Processing. AVSP (Norwich, UK)*, 2009.

[11] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[12] C. Zhang, P. Yin, Y. Rui, R. Cutler, P. Viola, X. Sun, N. Pinto, and Z. Zhang, "Boosting-based multimodal speaker detection for distributed meeting videos," *Multimedia, IEEE Transactions on*, vol. 10, no. 8, pp. 1541–1552, 2008.

[13] P. Chakravarty, S. Mirzaei, T. Tuytelaars *et al.*, "Who's speaking?: Audio-supervised classification of active speakers in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 87–90.

[14] V. Pavlović, A. Garg, J. M. Rehg, and T. S. Huang, "Multimodal speaker detection using error feedback dynamic bayesian networks," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 2. IEEE, 2000, pp. 34–41.

[15] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1589–1592.

[16] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[17] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[18] L. E. Bernstein, P. E. Tucker, and M. E. Demorest, "Speech perception without hearing," *Perception & Psychophysics*, vol. 62, no. 2, pp. 233–252, 2000.

[19] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," *Psychological Science*, vol. 15, no. 2, pp. 133–137, 2004. [Online]. Available: http://www.jstor.org/stable/40063940

[20] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*. IEEE, 2002, pp. 396–401.

[21] R. Ishii, S. Kumano, and K. Otsuka, "Analyzing mouth-opening transition pattern for predicting next speaker in multi-party meetings," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 209–216.

[22] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato, "Analysis of respiration for prediction of who will be next speaker and when? in multi-party meetings," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 18–25.

[23] K. Murai, "Speaker predicting apparatus, speaker predicting method, and program product for predicting speaker," Mar. 15 2011, uS Patent 7,907,165.

[24] H. Brugman, A. Russel, and X. Nijmegen, "Annotating multi-media/multi-modal resources with elan." in *Language Resources and Evaluation Conference (LREC)*, 2004, pp. 2065–2068.

[25] S. Machines, "Faceapi," *URL: http://www.seeingmachines.com*, 2009.

[26] D. W. McAdam and H. A. Whitaker, "Language production: Electroencephalographic localization in the normal human brain," *Science*, vol. 172, no. 3982, pp. 499–502, 1971.

[27] S. Raudys and R. P. W. Duin, "Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix," *Pattern Recognition Letters*, vol. 19, no. 5-6, pp. 385–392, Apr. 1998.

[28] P. Chakravarty, J. Zegers, T. Tuytelaars, and H. Van hamme, "Active speaker detection with audio-visual co-training," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ser. ICMI 2016. New York, NY, USA: ACM, 2016, pp. 312–316. [Online]. Available: http://doi.acm.org/10.1145/2993148.2993172