Look Who's Talking

Visual Identification of the Active Speaker in Multi-party Human-robot Interaction

Kalin Stefanov
KTH Royal Institute of Technology
Lindstedtsvägen 24
Stockholm, Sweden
kalins@kth.se

Akihiro Sugimoto
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda
Tokyo, Japan
sugimoto@nii.ac.jp

Jonas Beskow
KTH Royal Institute of Technology
Lindstedtsvägen 24
Stockholm, Sweden
beskow@kth.se

ABSTRACT

This paper presents analysis of a previously recorded multi-modal interaction dataset. The primary purpose of that dataset is to explore patterns in the focus of visual attention of humans under three different conditions - two humans involved in task-based interaction with a robot; the same two humans involved in task-based interaction where the robot is replaced by a third human, and a free three-party human interaction. The paper presents a data-driven methodology for automatic visual identification of the active speaker based on facial action units (AUs). The paper also presents an evaluation of the proposed methodology on 12 different interactions with an approximate length of 4 hours. The methodology will be implemented on a robot and used to generate natural focus of visual attention behavior during multi-party human-robot interactions.

CCS Concepts

•Human-centered computing → Human computer interaction (HCI); •Computing methodologies → Activity recognition and understanding; Machine learning;

Keywords

Multi-modal interaction, Human-robot interaction, Active speaker identification

1. INTRODUCTION

Successful multi-party human-robot interaction requires keeping track of the active speaker. In this paper we describe a methodology of detecting who is speaking based solely on visual features. For the purpose of efficient real world human-robot interaction, we have two main requirements. The first one is that we should be able to make decisions in *real-time* (possibly with a short lag), which in practice means that the system should not require any future information beyond a limited fixed lookahead window. In a spoken human-robot interaction system, in practice, it is sufficient if the system can classify each detected utterance as coming from a particular speaker by the time it is recognized by the speech recognition module. The second requirement is that the classification should be *independent* of the speaker.

The basic problem of identifying the source speaker in a video is a recurring one in the area of multi-modal interfaces, and different applications place different requirements on the solutions. An information theoretical approach exploiting mutual correlations to associate an audio source with regions of a video stream was demonstrated by [5], while [10] showed that audiovisual correlation may be used to automatically find the correct temporal synchronization between audio and a talking face. A general pattern recognition framework was used by [3]. The above approaches are all evaluated on small amounts of data, and usability in real world scenarios has not been demonstrated. Real world applications where visual speaker identification has been mostly explored, include speaker diarization ([2],[8] and [6]) and video conference management [14]. [7] applied visual activity (the amount of movement) and focus of visual attention as features to determine who is the current speaker on real meeting room corpus data, however the results were lower than audio-only diarization. Most methods primarily use visual features from the lower half face, often directly calculated from pixel values, but body movement has also been shown to increase recognition rates, [12] and [1]. We believe that the challenge of identifying the active speaker in more dynamic and cluttered environments remains. We would like to derive an approach that can handle such type of interactions. For example, we do not want to impose limitations such as specific hardware arrangement or participants' location in the environment.

Since one main requirement of our application is that it should be speaker independent, it is convenient to use derived features in the form of speaker independent facial parameters (AUs) as input features, in contrast to the methods referenced above, which typically use pixel-based feature representations of some form. In addition, AU features are efficiently calculated which is important for the real-time applicability of the system.

The rest of the paper is structured as follows; Section 2 briefly introduces the dataset used for evaluation of the proposed methodology, and Section 3 provides a description of

ASSP4MI'16, November 16 2016, Tokyo, Japan

DOI: http://dx.doi.org/10.1145/3005467.3005470

the methodology. We then describe two main experiments conducted and the results obtained in Section 4. Finally, we conclude and discuss directions for future work in Section 5.

2. DATASET

The analysis presented in this paper is based on a newly recorded multi-modal, multi-party dataset [11]. The main purpose of that dataset is to explore patterns in the focus of visual attention of humans under three different conditions: two humans involved in task-based interaction with a robot; the same two humans involved in task-based interaction where the robot is replaced by a third human, and a free three-party human interaction. The dataset contains two parts: 6 sessions, each of which is with duration of approximately 30 minutes, and 9 sessions, each of which is with duration of approximately 40 minutes. Both parts of the dataset are rich in modalities and recorded data streams. They include the streams of three Kinect v2 devices (color, depth, infrared, body and face data), three high quality audio streams, three high resolution GoPro video streams, touch data for the task-based interactions and the system state of the robot. In addition, the second part of the dataset introduces the data streams from three Tobii Pro Glasses 2 eye trackers. The language of all interactions is English and all data streams are spatially and temporally aligned. All interactions in the dataset occur around a round table and the participants are seated. Figure 1 illustrates the spatial configuration of the setup.

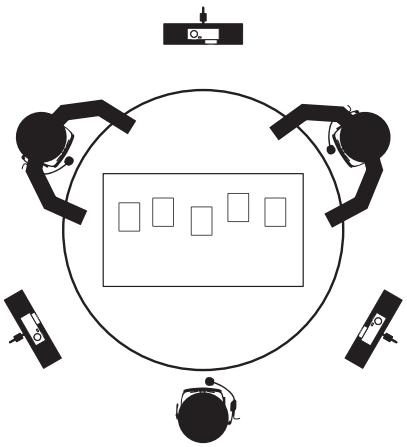

Figure 1: Spatial configuration of the setup and the location of different sensors used in the dataset.

3. METHODOLOGY

Our goal is to model two distinct behaviors during face-to-face interactions - *speaking* and *not speaking* and build an efficient and accurate classifier which can distinguish the current state of the participant given features extracted in real time. The information used for classification will vary both in space and time, therefore we use a spatio-temporal approach to modeling. One such approach is the Hidden Markov Model (HMM), [9] and [13], which is a double stochastic process governed by:

- an underlying Markov chain with a finite number of states;
- a set of random functions, each associated with one state.

In discrete time instants, the process is in one of the states and generates an observation symbol according to random function corresponding to that state. Each transition between the states has a pair of probabilities, defined as follows:

- transition probability, which provides the probability for undergoing transition;
- output probability, which defines the conditional probability of emitting an output symbol from a finite alphabet when given the state.

For the classification problem, the goal is to classify the unknown class of an observation sequence O into one of C classes. If we denote C different models by λ_c, $1 \leq c \leq C$, then the observation sequence is classified into class c^*, where $c^* = \text{argmax}_{c \in C} P(O|\lambda_c)$.

The generalized topology of an HMM is a fully connected structure, known as an *ergodic* model, where any state can be reached from any other state. We employ an 8 state fully connected HMM, as depicted in Figure 2, to model 2 different classes - speaking and not speaking.

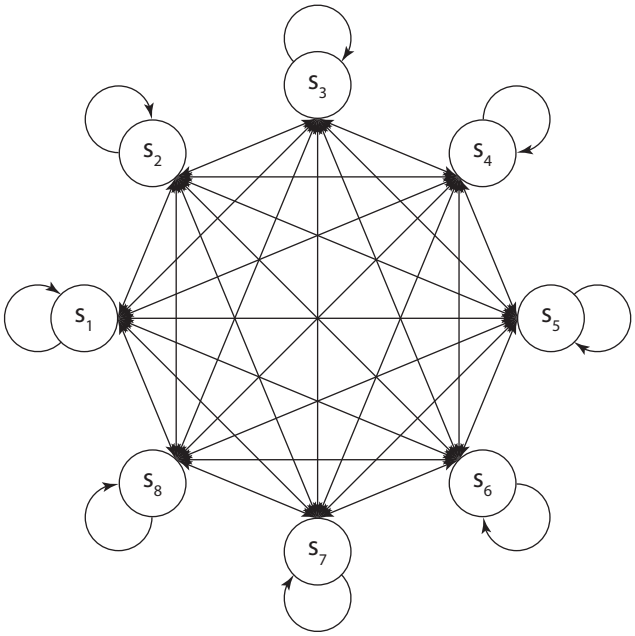

Figure 2: An 8 state fully connected HMM.

The requirement to be able to identify the active speaker in real time imposes limitation on the features that we can use to build the models. We use features that can be efficiently extracted during the interaction, specifically, the results presented in this paper are based on facial animation units (AUs), which are automatically calculated by the Kinect v2 SDK. The exact features from the Facial Action Coding System (FACS) [4] are,

- JawOpen, JawSlideRight;
- LipPucker, LipStretcherRight, LipStretcherLeft, LipCornerPullerLeft, LipCornerPullerRight, LipCornerDepressorLeft, LipCornerDepressorRight, LowerlipDepressorLeft, LowerlipDepressorRight;
- LeftcheekPuff, RightcheekPuff;
- LefteyeClosed, RighteyeClosed, RighteyebrowLowerer, LefteyebrowLowerer.

For training and testing the models we use 51 dimensional feature vector with 17 raw AU feature values, 17 first order AU differences (the difference between the feature value at time step t and at time step $t-1$) and 17 second order AU differences (the differences between the first order differences). The observations are modeled as multivariate Gaussian distribution with the full covariance matrix. We have tested the diagonal covariance matrix which results in significantly lower performance (assuming feature independence in this case is not justifiable). The number of hidden states is motivated by experiments on the data and was chosen to be 8 with accuracy and efficiency in mind. We remark that the number of AUs is 17 because this is the total number detected by Kinect. Furthermore, we remark that we do not assume only one active speaker (or speaker at all), the models are meant to be used on all visible faces when speech is detected, therefore we also allow overlaps.

4. EXPERIMENTS

We have conducted two main experiments to test our modeling approach - speaker dependent and speaker independent experiments. The acoustic signal is passed through an automatic voice activity detector (VAD) which produces intervals of speech and no speech. The silence threshold for the VAD is fixed to 200ms. We then use a window with a fixed length to extract the observation sequences from the synchronized Kinect data for both classes. The presented results are based on three different window lengths - 200ms, 500ms and 1s. The data for the AUs in the first three sessions of the dataset is missing, therefore the results presented next are based on 12 interactions (not all 15).

4.1 Speaker Dependent

The speaker dependent experiment is designed as follows. We use a 10-fold cross validation procedure to build 10x2 different subject specific models, 2 models (positive and negative) per subject with the same topology as depicted in Figure 2. We train the models using 90% of the subject's data, for speaking and not speaking classes, and use the rest of the available data for that subject to test the models. Figure 3 illustrates the results of this experiment for 12 interactions, 3 subjects per interaction.

The confusion matrices for window length of 1s, included in Table 1, show that in the speaker dependent case the negative (not speaking) class is modeled slightly better than the positive (speaking) class by the 8 state HMM. Specifically, 77.3% of the negative instances are correctly classified and 75.6% of the positive instances are correctly classified.

The class distribution is skewed, with the negative (not speaking) class having more instances. Specifically, the negative class for window length of 1s is 67.6%. The average accuracy of the proposed methodology in this case is 76.8% which is a significant improvement compared to choosing the dominant class every time.

4.2 Speaker Independent

We designed the speaker independent experiment in the following way. First we built 36x2 subject specific models for each of the two classes. Then given all data for subject n we made classification with the rest 35x2 models (excluding the two models for the current subject). The final class assignment for the observation sequence is calculated by a majority vote between all 35x2 models (the majority voter takes the votes after $\text{argmax}\, P(O|positive, O|negative)$ for 35x2 models). Figure 4 illustrates the results of this experiment for 12 interactions, 3 subjects per interaction, for window length of 1s.

The confusion matrices for window length of 1s, included in Table 2, show that in the speaker independent case the positive (speaking) class is modeled better than the negative (not speaking) class by the 8 state HMM. Specifically, 43.6% of the negative instances are correctly classified and 74.8% of the positive instances are correctly classified.

The class distribution is skewed, with the negative (not speaking) class having more instances. Specifically, the negative class for window length of 1s is 67.6%. The average accuracy of the proposed methodology in this case is 53.7% which is significantly lower compared to choosing the dominant class every time.

5. CONCLUSIONS

We have presented a methodology that attempts to identify the active speaker during different types of multi-party interactions: collaborative task-based interactions with and without a robot and free three-party human interactions. This methodology involves efficient and automatic extraction of visual features, facial action units, and Hidden Markov Models spatio-temporal modeling. The evaluation performed shows that the not speaking class is more complex than the speaking counterpart when expressed in the selected feature space.

A desired property of the proposed approach is speaker independence. The conducted experiments show that the proposed methodology can exhibit an independence property for one of the classes modeled - the performance for the speaking class is the same for both experiments. However, the performance in the speaker independent case for the not speaking class considerably degraded from the one in the speaker dependent case.

Future work will involve the definition of different topology for the not speaking class (currently we use the same 8 state fully connected HMM for both classes), and investigation of other approaches to spatio-temporal data modeling. Furthermore, we plan to implement an additional design of the speaker independent experiment, where only two models are built from all data for $n-1$ subjects and tested on the data for the left out subject. We will also investigate additional features from the acoustic channel, which can be efficiently extracted during the interactions in order to increase the overall system accuracy. The introduction of new features will be accompanied with analysis of their individual contribution to the task of successful active speaker identification.

6. REFERENCES

[1] S. Alexanderson, J. Beskow, and D. House. Automatic Speech/Non-Speech Classification Using Gestures in Dialogue. In *Swedish Language Technology Conference*, 2014.

[2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker Diarization: A Review of Recent Research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2012.

[3] P. Besson and M. Kunt. Hypothesis Testing for Evaluating a Multimodal Pattern Recognition Framework Applied to Speaker Detection. *Journal of NeuroEngineering and Rehabilitation*, 2008.

[4] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. 1978.

[5] J. W. Fisher, T. Darrell, W. T. Freeman, and P. Viola. Learning Joint Statistical Models for Audio-Visual Fusion and Segregation. In *Advances in Neural Information Processing Systems*, 2000.

[6] G. Friedland, C. Yeo, and H. Hung. Visual Speaker Localization Aided by Acoustic Models. In *ACM International Conference on Multimedia*, 2009.

[7] H. Hung and S. O. Ba. Speech/Non-Speech Detection in Meetings from Automatically Extracted Low Resolution Visual Features. Technical report, 2009.

[8] H. J. Nock, G. Iyengar, and C. Neti. Speaker Localisation Using Audio-Visual Synchrony: An Empirical Study. In *Image and Video Retrieval*, 2003.

[9] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of IEEE*, 1988.

[10] M. Slaney and M. Covell. Facesync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks. In *Advances in Neural Information Processing Systems*, 2000.

[11] K. Stefanov and J. Beskow. A Multi-party Multi-modal Dataset for Focus of Visual Attention in Human-human and Human-robot Interaction. In *Language Resources and Evaluation Conference*, 2016.

[12] H. Vajaria, S. Sarkar, and R. Kasturi. Exploring Co-Occurence Between Speech and Body Movement for Audio-Guided Video Localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2008.

[13] J. Yamato, J. Ohya, and K. Ishii. Recognizing Human Action in Time-Sequential Images Using Hidden Markov Model. In *Computer Vision and Pattern Recognition*, 1992.

[14] C. Zhang, P. Yin, Y. Rui, R. Cutler, P. Viola, X. Sun, N. Pinto, and Z. Zhang. Boosting-Based Multimodal Speaker Detection for Distributed Meetings. *IEEE Transactions on Multimedia*, 2008.

	neg	pos
neg	81.7 (464)	18.3 (104)
pos	17.3 (51)	82.7 (244)

Subject 1

	neg	pos
neg	75.4 (371)	24.6 (121)
pos	6.4 (21)	93.6 (309)

Subject 2

	neg	pos
neg	81.3 (650)	18.7 (150)
pos	39.7 (33)	60.3 (50)

Subject 3

	neg	pos
neg	76.4 (480)	23.6 (148)
pos	44.6 (82)	55.4 (102)

Subject 4

	neg	pos
neg	85.1 (536)	14.9 (94)
pos	61.7 (92)	38.3 (57)

Subject 5

	neg	pos
neg	82.5 (608)	17.5 (129)
pos	35.6 (37)	64.4 (67)

Subject 6

	neg	pos
neg	83.1 (446)	16.9 (91)
pos	14.5 (35)	85.5 (206)

Subject 7

	neg	pos
neg	81.9 (484)	18.1 (107)
pos	52.2 (96)	47.8 (88)

Subject 8

	neg	pos
neg	75.6 (403)	24.4 (130)
pos	35.0 (68)	65.0 (126)

Subject 9

	neg	pos
neg	72.4 (497)	27.6 (190)
pos	18.6 (114)	81.4 (500)

Subject 10

	neg	pos
neg	85.4 (962)	14.6 (165)
pos	31.1 (74)	68.9 (164)

Subject 11

	neg	pos
neg	74.5 (762)	25.5 (261)
pos	10.4 (34)	89.6 (292)

Subject 12

	neg	pos
neg	61.2 (472)	38.8 (299)
pos	23.3 (106)	76.7 (349)

Subject 13

	neg	pos
neg	76.5 (675)	23.5 (207)
pos	31.3 (110)	68.7 (241)

Subject 14

	neg	pos
neg	73.5 (767)	26.5 (277)
pos	36.0 (90)	64.0 (160)

Subject 15

	neg	pos
neg	88.3 (416)	11.7 (55)
pos	17.9 (77)	82.1 (354)

Subject 16

	neg	pos
neg	76.7 (447)	23.3 (136)
pos	20.1 (61)	79.9 (242)

Subject 17

	neg	pos
neg	75.0 (411)	25.0 (137)
pos	24.2 (70)	75.8 (219)

Subject 18

	neg	pos
neg	75.5 (497)	24.5 (161)
pos	22.4 (95)	77.6 (329)

Subject 19

	neg	pos
neg	79.4 (557)	20.6 (145)
pos	17.7 (59)	82.3 (274)

Subject 20

	neg	pos
neg	70.3 (479)	29.7 (202)
pos	25.0 (79)	75.0 (237)

Subject 21

	neg	pos
neg	86.4 (317)	13.6 (50)
pos	19.1 (30)	80.9 (127)

Subject 22

	neg	pos
neg	82.8 (288)	17.2 (60)
pos	5.2 (9)	94.8 (165)

Subject 23

	neg	pos
neg	63.4 (178)	36.6 (103)
pos	12.3 (27)	87.7 (193)

Subject 24

	neg	pos
neg	78.9 (557)	21.1 (149)
pos	18.3 (98)	81.7 (436)

Subject 25

	neg	pos
neg	80.8 (709)	19.2 (169)
pos	19.8 (59)	80.2 (239)

Subject 26

	neg	pos
neg	73.1 (529)	26.9 (195)
pos	34.8 (136)	65.2 (255)

Subject 27

	neg	pos
neg	88.6 (340)	11.4 (44)
pos	24.0 (69)	76.0 (218)

Subject 28

	neg	pos
neg	81.2 (341)	18.8 (79)
pos	11.2 (39)	88.8 (309)

Subject 29

	neg	pos
neg	78.6 (487)	21.4 (133)
pos	31.8 (41)	68.2 (88)

Subject 30

	neg	pos
neg	79.5 (601)	20.5 (155)
pos	28.9 (125)	71.1 (308)

Subject 31

	neg	pos
neg	76.6 (672)	23.4 (205)
pos	28.5 (82)	71.5 (206)

Subject 32

	neg	pos
neg	67.0 (300)	33.0 (148)
pos	25.1 (177)	74.9 (529)

Subject 33

	neg	pos
neg	87.2 (1267)	12.8 (186)
pos	21.9 (70)	78.1 (249)

Subject 34

	neg	pos
neg	77.0 (483)	23.0 (144)
pos	34.1 (349)	65.9 (674)

Subject 35

	neg	pos
neg	59.4 (729)	40.6 (498)
pos	22.4 (113)	77.6 (391)

Subject 36

Table 1: Confusion matrices for the speaker dependent experiment. The observation sequence window length is 1s, *neg* refers to the *not speaking* class and *pos* refers to the *speaking* class. The number in brackets represents instance count.

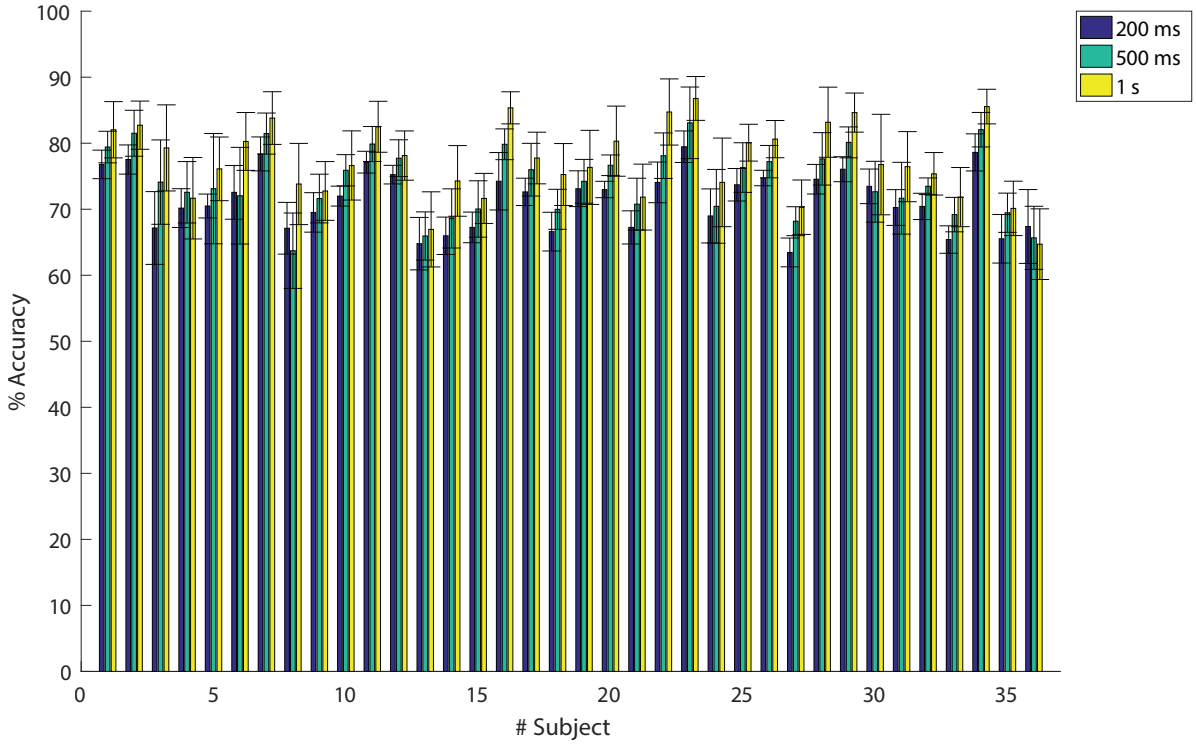

Figure 3: 10-fold cross validated results for the speaker dependent experiment. The accuracy is calculated according to $accuracy = \frac{tp+tn}{tp+tn+fp+fn}$, for 10 different train/test iterations and averaged to produce the bar graph for each subject for each fixed window length. The error bar illustrates the standard deviation.

	neg	pos
neg	14.8 (84)	85.2 (484)
pos	0.3 (1)	99.7 (294)

Subject 1

	neg	pos
neg	88.6 (436)	11.4 (56)
pos	30.6 (101)	69.4 (229)

Subject 2

	neg	pos
neg	13.3 (106)	86.7 (694)
pos	4.8 (4)	95.2 (79)

Subject 3

	neg	pos
neg	29.9 (188)	70.1 (440)
pos	10.3 (19)	89.7 (165)

Subject 4

	neg	pos
neg	44.1 (278)	55.9 (352)
pos	17.4 (26)	82.6 (123)

Subject 5

	neg	pos
neg	20.8 (153)	79.2 (584)
pos	4.8 (5)	95.2 (99)

Subject 6

	neg	pos
neg	33.7 (181)	66.3 (356)
pos	7.5 (18)	92.5 (223)

Subject 7

	neg	pos
neg	23.2 (137)	76.8 (454)
pos	11.9 (22)	88.1 (162)

Subject 8

	neg	pos
neg	39.6 (211)	60.4 (322)
pos	11.8 (23)	88.2 (171)

Subject 9

	neg	pos
neg	23.7 (163)	76.3 (524)
pos	1.3 (8)	98.7 (606)

Subject 10

	neg	pos
neg	24.9 (280)	75.1 (847)
pos	21.8 (52)	78.2 (186)

Subject 11

	neg	pos
neg	41.8 (428)	58.2 (595)
pos	27.9 (91)	72.1 (235)

Subject 12

	neg	pos
neg	58.1 (448)	41.9 (323)
pos	43.1 (196)	56.9 (259)

Subject 13

	neg	pos
neg	39.1 (345)	60.9 (537)
pos	16.2 (57)	83.8 (294)

Subject 14

	neg	pos
neg	26.6 (278)	73.4 (766)
pos	14.0 (35)	86.0 (215)

Subject 15

	neg	pos
neg	42.3 (199)	57.7 (272)
pos	3.0 (13)	97.0 (418)

Subject 16

	neg	pos
neg	59.2 (345)	40.8 (238)
pos	29.0 (88)	71.0 (215)

Subject 17

	neg	pos
neg	83.6 (458)	16.4 (90)
pos	42.9 (124)	57.1 (165)

Subject 18

	neg	pos
neg	56.5 (372)	43.5 (286)
pos	31.4 (133)	68.6 (291)

Subject 19

	neg	pos
neg	75.5 (530)	24.5 (172)
pos	46.5 (155)	53.5 (178)

Subject 20

	neg	pos
neg	35.3 (240)	64.7 (441)
pos	15.2 (48)	84.8 (268)

Subject 21

	neg	pos
neg	38.2 (140)	61.8 (227)
pos	21.6 (34)	78.4 (123)

Subject 22

	neg	pos
neg	82.2 (286)	17.8 (62)
pos	56.3 (98)	43.7 (76)

Subject 23

	neg	pos
neg	72.3 (203)	27.7 (78)
pos	43.6 (96)	56.4 (124)

Subject 24

	neg	pos
neg	35.7 (252)	64.3 (454)
pos	36.7 (196)	63.3 (338)

Subject 25

	neg	pos
neg	57.8 (507)	42.2 (371)
pos	33.9 (101)	66.1 (197)

Subject 26

	neg	pos
neg	58.3 (422)	41.7 (302)
pos	46.0 (180)	54.0 (211)

Subject 27

	neg	pos
neg	62.8 (241)	37.2 (143)
pos	16.0 (46)	84.0 (241)

Subject 28

	neg	pos
neg	19.1 (80)	80.9 (340)
pos	4.3 (15)	95.7 (333)

Subject 29

	neg	pos
neg	16.8 (104)	83.2 (516)
pos	22.5 (29)	77.5 (100)

Subject 30

	neg	pos
neg	38.8 (293)	61.2 (463)
pos	22.6 (98)	77.4 (335)

Subject 31

	neg	pos
neg	44.5 (390)	55.5 (487)
pos	22.2 (64)	77.8 (224)

Subject 32

	neg	pos
neg	53.8 (241)	46.2 (207)
pos	56.2 (397)	43.8 (309)

Subject 33

	neg	pos
neg	84.2 (1224)	15.8 (229)
pos	75.9 (242)	24.1 (77)

Subject 34

	neg	pos
neg	10.7 (67)	89.3 (560)
pos	7.6 (78)	92.4 (945)

Subject 35

	neg	pos
neg	42.0 (515)	58.0 (712)
pos	22.2 (112)	77.8 (392)

Subject 36

Table 2: Confusion matrices for the speaker independent experiment. The observation sequence window length is 1s, *neg* refers to the *not speaking* class and *pos* refers to the *speaking class*. The number in brackets represents instance count.

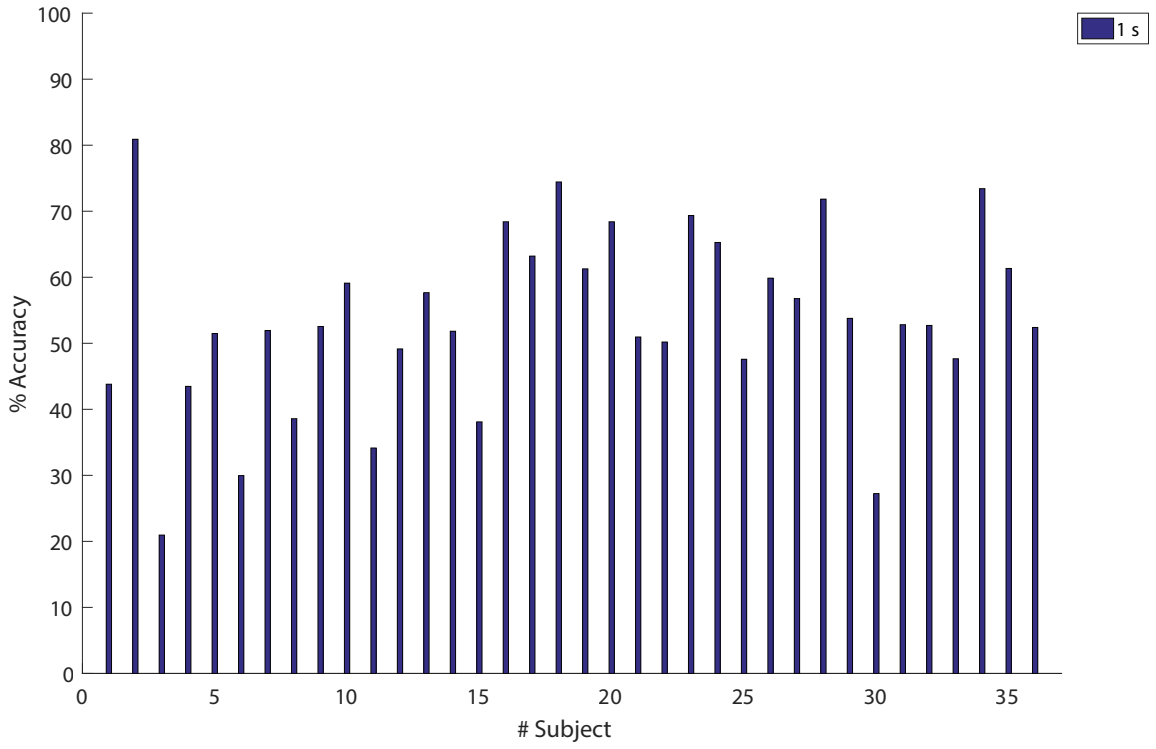

Figure 4: Majority vote results for the speaker independent experiment. The accuracy is calculated according to $accuracy = \frac{tp+tn}{tp+tn+fp+fn}$.