GLU 2017 International Workshop on Grounding Language Understanding
25 August 2017, Stockholm, Sweden

Vision-based Active Speaker Detection in Multiparty Interactions

Kalin Stefanov, Jonas Beskow, Giampiero Salvi

KTH Royal Institute of Technology
TMH Speech, Music and Hearing
Stockholm, Sweden

kalins@kth.se, beskow@kth.se, giampi@kth.se

Abstract

This paper presents a supervised learning method for automatic visual detection of the active speaker in multiparty interactions. The presented detectors are built using a multimodal multiparty interaction dataset previously recorded with the purpose to explore patterns in the focus of visual attention of humans. Three different conditions are included: 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 also presents an evaluation of the active speaker detection method in a speaker dependent experiment showing that the method achieves good accuracy rates in a fairly unconstrained scenario using only image data as input. The main goal of the presented method is to provide real-time detection of the active speaker within a broader framework implemented on a robot and used to generate natural focus of visual attention behavior during multiparty human-robot interactions.

Index Terms: machine learning, active speaker detection, multiparty human-robot interaction

1. Introduction

Natural and effective human-robot interaction requires robots to produce humanlike nonverbal signals. One such nonverbal signal used by humans in daily face-to-face interactions is the eye-gaze, or more broadly, the focus of visual attention. The focus of visual attention can provide several cues, for example, it can regulate who is allowed to speak when and coordinate the changes in the roles on the conversational floor (speaker, addressee, bystander), known as *footing*. Since clear conversational roles in face-to-face communication are vital for smooth and effective interaction, a robot which is aware of the established roles in real-time could avoid misunderstandings or talking over other participants. On the other hand, the focus of visual attention can provide cues for important events or objects in the space shared during the interaction.

The context of the work presented in this paper is a framework aimed at recognizing (human) and generating (robot) real-time focus of visual attention to facilitate more natural and effective communication. A key component of this framework is a system which can keep track of the active speaker in real-time while imposing as little constraints as possible on the interaction. Therefore, in this paper we describe a method of detecting who is speaking based solely on visual input. For the purpose of efficient real world human-robot interaction, we have two main requirements for the method. 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. 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 person by the time it is recognized by the speech recognizer. The second requirement is that the method should make as little assumptions as possible for the environment in which the system will operate. Such assumptions can be noise-free environment, known number of participants, or known spatial configuration.

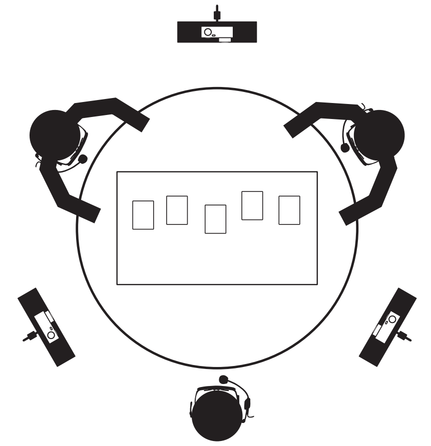

Figure 1: *Spatial configuration of the setup and the location of different sensors used in the dataset.*

The problem of identifying the active speaker is an important and recurring one in many areas, and different applications place different requirements on the solutions.

Audio-only speaker identification, know as *speaker diarization*, is the process of finding segments in the input audio associated with different speakers. Speaker diarization has received a fair amount of attention from researchers in the past. A comprehensive review of the recent research in the field is done in Miro et al. [1].

Audio-visual speaker identification approaches, on the other hand, attempt to combine information from both audio and video. Nock at al. [2] explore the application of audio-visual synchrony for active speaker localization in broadcast videos, Friedland et al. [3] present an audio-visual approach for unsupervised speaker localization in meetings, and Zhang et al. [4] propose a boosting-based multimodal speaker detection algorithm for distributed meetings. An information theoretical approach exploiting mutual correlations to associate an audio source with regions of a video stream was demonstrated by Fisher et al. [5], while Slaney and Covell [6] showed that audio-visual correlation may be used to automatically find the correct temporal synchronization between audio and a talking face. More recently researchers have employed deep architectures (Convolutional Neural Networks and Recurrent Neu-

10.21437/GLU.2017-10

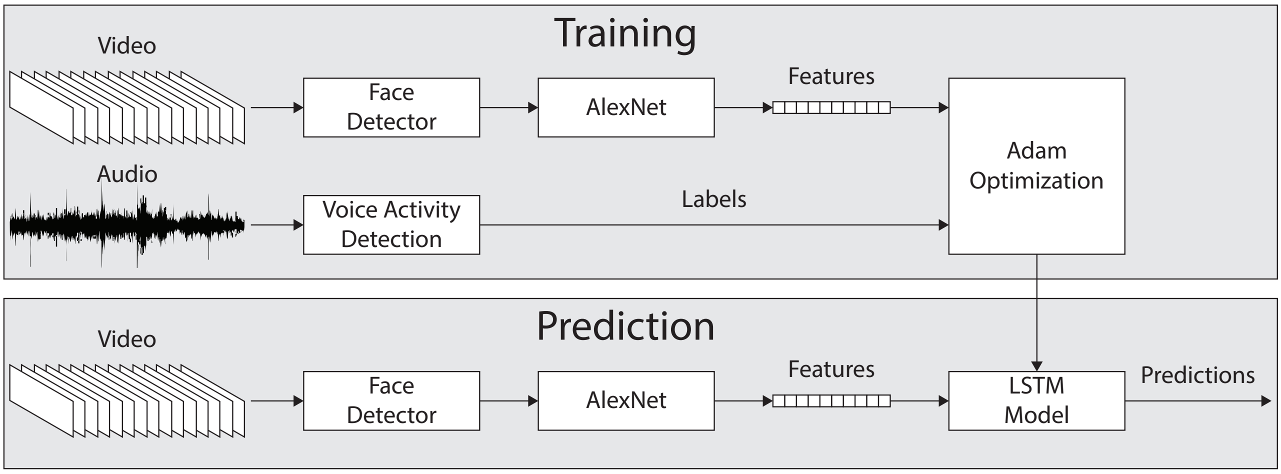

Figure 2: *System overview.*

ral Networks) to learn active speaker detectors from audio-visual input. Ren et al. [7] propose a multimodal Long Short-Term Memory (LSTM) model to learn shared weights between modalities (audio and video) and apply it to speaker naming in TV shows and Hu et al. [8] propose a Convolutional Neural Network (CNN) framework to learn the fusion function of face and audio cues.

Other approaches include a general pattern recognition framework used by Besson and Kunt [9]. Hung and Ba [10] 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. Action units (AU) were used as input features to Hidden Markov Models (HMM) in Stefanov et al.[11]. Vajaria et al. [12] showed that body movements can increase recognition rates.

The above approaches are either evaluated on small amounts of data, or usability in real-time settings has not been demonstrated. Furthermore, in many cases a certain spatial configuration is assumed and the relative location of the used sensors is known. Finally, the goal is usually an offline video/audio analysis task, such as semantic indexing and retrieval of TV broadcasts or meetings, or video/audio summarization. We believe that the challenge remains when it comes to identifying the active speaker in more dynamic and cluttered environments. For the purpose of generating robot's focus of visual attention based on the information of the active speaker we do not want to impose limitations such as specific hardware arrangement or participants' location in the environment. In this paper we present a method which has several desirable characteristics for such types of scenarios - 1) it works in real-time, 2) it does not assume specific spatial configuration, and 3) the possible number of (simultaneous) active speakers is free to vary during the interaction.

2. Method

The goal is to build a system that is able to detect in real-time the status (i.e. *speaking* and *not speaking*) of all visible faces in a multiparty interaction. Furthermore, the system should be able to achieve good accuracy rates given only the visual information, i.e. the RGB color data. Finally, the system should be able to generalize to unseen data.

We use a supervised learning approach to construct the active speaker detectors, an overview of the system is presented in Figure 2. In the training stage we use two inputs - video and audio. In the top part of the training pipeline, the video is fed into a face detection module [13] which attempts to locate all visible faces in the current frame. The output of the face detection module is the RGB color data for all found faces. This information is then fed into a CNN, in this case AlexNet [14], which calculates an n-dimensional feature vector for each face image. In the bottom part of the training pipeline, the audio is fed into a Voice Activity Detection (VAD) [15] module, which creates speech and non-speech intervals from the acoustic signal. The intervals detected by the VAD module are then used as labels. Finally, the features and the labels are combined into m-frame long segments which are used by a gradient-based optimization procedure [16] to adjust the weights of the hidden layers of an LSTM model [17]. The LSTM model includes several hidden layers - at the bottom is an input layer which sends frame segments of a certain length to the next LSTM layer. The LSTM layer is followed by a stack of time distributed dense layers with decreasing output sizes. Time distributed means that the same weights of the fully-connected dense layer are applied to each frame in the input sequence (sequence input - sequence output). The final layer in the model outputs a probability distribution over the possible outcomes (*speaking* and *not speaking*). Details on the network architecture are given in Section 3.2. Since the network performs binary classification, and the output is calculated with Softmax activation function, the detection of the active speaker happens when the corresponding probability exceeds 0.5. The evaluation is performed by computing the accuracy of the predictions on frame-by-frame basis.

3. Experiments

3.1. Data

The method presented in Section 2 is built and evaluated on a multimodal multiparty dataset described in [18]. The main purpose of the 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

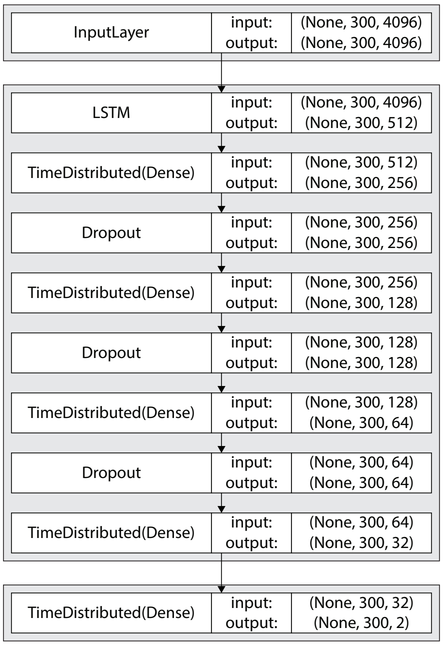

Figure 3: *Model overview. The first cell is the type of the layer. The second cell specifies the input and the output. The last cell is the size of the input and the output in the following format - number of batches, batch size (number of time steps/frames), frame size.*

which is with duration of approximately 30 minutes, and 9 sessions, with duration of approximately 40 minutes each. 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. Finally, there are 24 unique participants. Figure 1 illustrates the spatial configuration of the setup.

3.2. Experimental Setup

All results are based on random sampling of the available data for each subject, where $\approx 80\%$ is used for training, $\approx 15\%$ is used for testing, and $\approx 5\%$ is used validation. The video input is generated by the Kinect directed at the subject under consideration and the audio input is generated by his/her close-talking microphone. The networks are trained and evaluated with 300 frame (10 sec) long segments without overlaps. The total size of the data used for training is 1424700 frames ($\approx$ 13 hours), the total size of the data used for testing is 266100 frames ($\approx$ 2.5 hours), and the total size of the data used for validation is 89400 frames ($\approx$ 1 hour). We have trained and evaluated a separate network for each of the 24 subjects.

For training the networks we use Adam optimizer with default parameters ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$) and categorical crossentropy loss function. The rest of the model parameters and the model architecture are illustrated in Figure 3. All dense layers (densely-connected neural network layers) use a Rectified Linear Unit (ReLU) activation function except the final one (the network output layer) which uses a Softmax activation function. Each subject specific network is trained for 100 epochs and the reported results are for the state of the network in which it has the best performance on the validation set. The system is implemented in Keras [19] with TensorFlow [20] backend. During the prediction stage only the RGB color data is used as input. When evaluating the networks' performance, we use 0.5 as a threshold for assigning a class to each frame-level prediction.

Table 1: *Results per subject.*

Subject	Accuracy (%)	Subject	Accuracy (%)
1	**87.77**	13	77.81
2	79.43	14	79.87
3	73.14	15	79.72
4	72.15	16	78.81
5	81.59	17	75.56
6	**65.87**	18	76.08
7	78.85	19	78.03
8	75.87	20	75.83
9	84.75	21	78.93
10	85.10	22	85.79
11	71.81	23	77.85
12	83.38	24	73.00
		Mean	78.21
		Std	5.07

4. Results

This section presents the main (speaker dependent) numerical results of the evaluation and provides a discussion on the performance with illustrations of several cases. We have summarized the results per subject in Table 1. The table provides the accuracy rate per subject, as well as the mean accuracy over all subjects. The *easiest* and the *hardest* subjects are marked in boldface font. From the table we can see that the accuracy varies between subjects with the lowest results around 66% and the highest around 88%. This inter subject variability is further confirmed by the high standard deviation.

In Table 2 we present the cumulative (for all subjects) confusion matrix of the networks output. The rows in the matrix represent the original labels (generated by the VAD module) and the columns represent the predictions of the networks. We use *pos* and *neg* to denote the *speaking* and *not speaking* class, respectively. The top number in each cell is the number of test frames. The mean accuracy is 80.75% and the balanced accuracy is 77.46%. The table also illustrates that the dataset used is unbalanced, where the number of *not speaking* frames is significantly larger than the number of *speaking* frames. This problem is partly addressed by assigning different weights on the frames during training of the networks (a correctly detected *speaking* frame has a higher weight than correctly detected *not speaking* frame).

Finally, Figure 4 illustrates the output of two networks for some examples in the data. The x-axis corresponds to time and the y-axis to the class label. We use 0 to denote the *neg* (*not speaking*) class and 1 to denote the *pos* (*speaking*) class. The output of the network corresponds to $P(\text{pos})$ and varies contin-

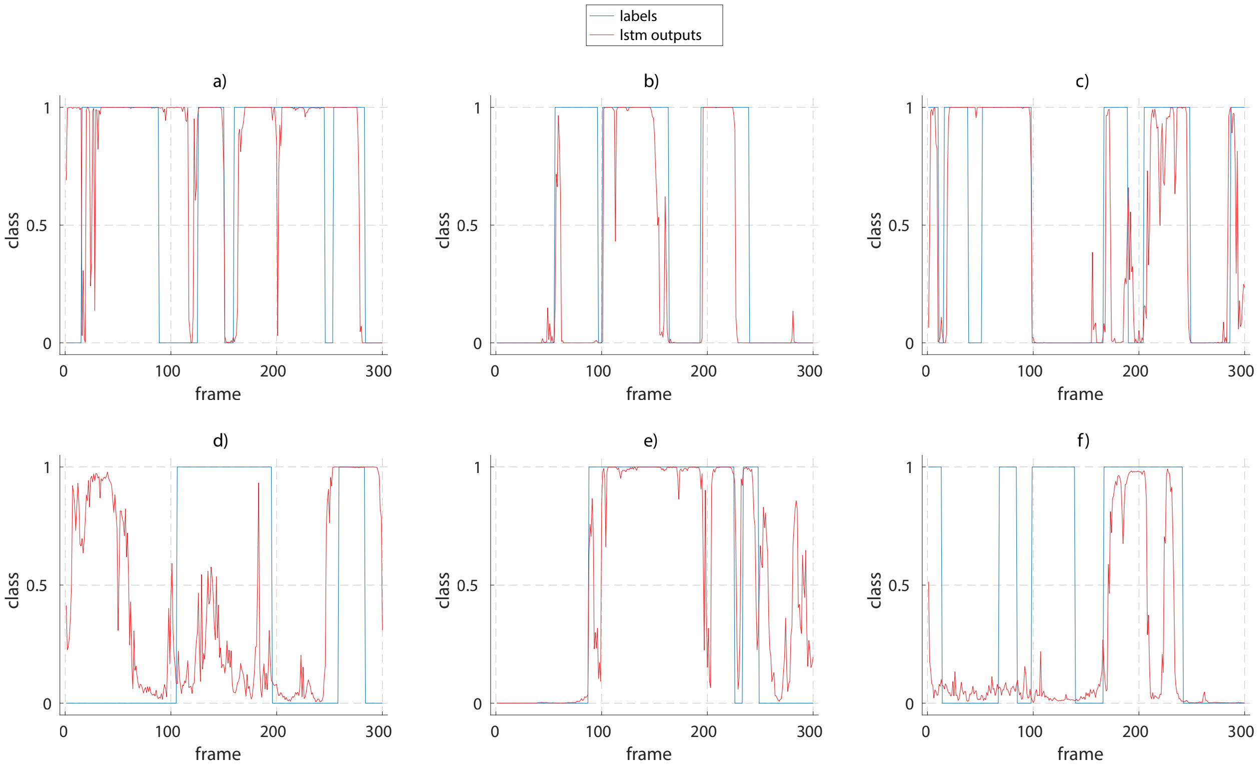

Figure 4: *Example outputs of two networks.*

Table 2: *Cumulative confusion matrix.*

		predictions pos	neg
labels	pos	59466 (67.81%)	28225 (32.19%)
	neg	22990 (12.89%)	155419 (87.11%)

uously between 0 and 1.

In the top row, subfigures a), b), and c), besides the accurate predictions of the network, we present cases which show the ability of the network to quickly switch between decisions even for brief intervals. Furthermore, in most cases, the network output comes very close to the target value (that is, the $P(\text{pos})$ is either very close to 0 or to 1). The subfigures also show the granularity of the target signal obtained by the VAD module. We can observe brief dips in the target signal which are not necessarily desirable. In real world application one might not want to switch the focus of visual attention for 200ms. This also illustrates the limitations of the used VAD module to generate the target labels.

In the bottom row, subfigures d), e), and f), we show cases where the method is less accurate. In some of the examples the network emits high probability of *speaking* even though the label tells otherwise. In other cases, the activity is somewhere in between 0 and 1 not reaching the desired decision. We also present another problem associated with the VAD module. At the end of the segment in d), the network's output *fitted* the target signal with small advance and delay. This can be attributed to the way the VAD module generates many speech/non-speech intervals. Usually, the intervals start before and end after the actual speech, causing the network to learn to activate the output for a longer period of time than the actual speech act. This phenomenon is, however, only visible in plot d) among our examples.

5. Conclusions

In this paper we have proposed and evaluated a method for automatic detection of the active speaker in multiparty interactions based solely on visual input. Although similar to other methods proposed in the literature, in our approach we try to reduce the assumptions about the environment to a minimum. We allow the different speakers to speak simultaneously as well as to be all silent. We do not assume a specific number of speakers, and we estimate the probability of speaking independently for each speakers, thus allowing the method to be used as is, even if the number of speakers is changed during the interaction.

Furthermore, we evaluate our system on fairly large dataset including some challenging examples. For example, around half of the time the participants interact with a touch surface and they look down while talking, making the feature extraction after face detection more difficult.

The method performs fairly well on a speaker dependent fashion, reaching 78% accuracy on a frame-by-frame evaluation metric. The presented results will serve as a baseline for comparison with future extensions of the described method.

Future work will include extending to method to obtain speaker independent results and comparison with audio-visual approaches proposed in the literature.

6. Acknowledgements

Some of the GeForce GTX TITAN cards used for this research were donated by the NVIDIA Corporation. Some of the computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Parallel Data Center (PDC) at KTH. This research was partly supported by the CHIST-ERA project IGLU.

7. References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[2] H. J. Nock, G. Iyengar, and C. Neti, *Speaker Localisation Using Audio-Visual Synchrony: An Empirical Study*. Springer Berlin Heidelberg, 2003, pp. 488–499.

[3] G. Friedland, C. Yeo, and H. Hung, "Visual Speaker Localization Aided by Acoustic Models," in *Proceedings of the 17th ACM International Conference on Multimedia*. ACM, 2009, pp. 195–202.

[4] 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*, vol. 10, no. 8, pp. 1541–1552, 2008.

[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 13*. MIT Press, 2001, pp. 772–778.

[6] 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 13*. MIT Press, 2001, pp. 814–820.

[7] J. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan, "Look, Listen and Learn - a Multimodal LSTM for Speaker Identification," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, pp. 3581–3587.

[8] Y. Hu, J. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang, "Deep Multimodal Speaker Naming," *Computing Research Repository*, vol. abs/1507.04831, 2015.

[9] P. Besson and M. Kunt, "Hypothesis Testing for Evaluating a Multimodal Pattern Recognition Framework Applied to Speaker Detection," *Journal of NeuroEngineering and Rehabilitation*, vol. 5, no. 1, p. 11, 2008.

[10] H. Hung and S. O. Ba, "Speech/Non-Speech Detection in Meetings from Automatically Extracted Low Resolution Visual Features," Tech. Rep., 2009.

[11] K. Stefanov, A. Sugimoto, and J. Beskow, "Look Who's Talking: Visual Identification of the Active Speaker in Multi-party Human-robot Interaction," in *Proceedings of the 2Nd Workshop on Advancements in Social Signal Processing for Multimodal Interaction*, ser. ASSP4MI'16. ACM, 2016, pp. 22–27.

[12] H. Vajaria, S. Sarkar, and R. Kasturi, "Exploring Co-Occurrence Between Speech and Body Movement for Audio-Guided Video Localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1608–1617, 2008.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–511–I–518.

[14] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105.

[15] G. Skantze and S. Al Moubayed, "IrisTK: A Statechart-based Toolkit for Multi-party Face-to-face Interaction," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ser. ICMI'12. ACM, 2012, pp. 69–76.

[16] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Computing Research Repository*, vol. abs/1412.6980, 2014.

[17] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] K. Stefanov and J. Beskow, "A multi-party multi-modal dataset for focus of visual attention in human-human and human-robot interaction," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*, 2016.

[19] F. Chollet *et al.*, "Keras," 2015. [Online]. Available: https://github.com/fchollet/keras

[20] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: http://tensorflow.org/