Response to Name: A Dataset and A Multimodal Machine Learning Framework towards Autism Study

Wenbo Liu*‡, Tianyan Zhou*‡, Chenghao Zhang*‡, Xiaobing Zou† and Ming Li*‡
*SYSU-CMU Joint Institute of Engineering, School of Electronics and Information Technology,
Sun Yat-sen University
†The Third Affiliated hospital, Sun Yat-sen University
‡Department of Electrical and Computer Engineering, Carnegie Mellon University

***Abstract*—In this paper, we propose a "Response to Name Dataset" for autism spectrum disorder (ASD) study as well as a multimodal ASD auxiliary screening system based on machine learning. ASD children are characterized by their impaired interpersonal communication abilities and lack of response. In the proposed dataset, the reactions of children are recorded by cameras upon calling their names. The responsiveness of each child is then evaluated by a clinician with a score among 0, 1 and 2 following the Autism Diagnostic Observation Schedule (ADOS). We then develop a rule-based multimodal framework to quantitatively evaluate each child. Our system involves speech recognition based automatic name calling detection, face detection/alignment, head pose estimation, and considers the response speed, eye contact duration and head orientation to output the final prediction. Compared to existing work, our dataset characterizes a more precise and detailed scoring system with clinical trial standards, as well as a more spontaneous setting by incorporating less lab-controlled sessions with dynamic/cluttered environments, multi-pose mobile captured videos, and flexible number of accompanying adults. Experiments show that our machine predicted scores align closely with human professional diagnosis, showing promising potential in early screening of ASD, and shedding light on future clinical applications.**

1. Introduction

Autism spectrum disorder represents a group of lifelong neurodevelopmental disorders characterized by ongoing social problems, repetitive behaviors, as well as limited interests or activities of the subjects. ASD can not be fully cured, and early intervention so far presents the best way to achieve positive longitudinal outcomes. With the rising number of children suffering autism spectrum disorder (ASD), early detection is becoming increasingly important to maximize the gain of early intervention. Current most widely used assessment methods include the Autism Diagnostic Observation Schedule-Generic (ADOS-G) [1] and the revised version ADOS-2 [2]. However, the interactive, human-in-loop nature of these methods not only requires the accompany and administration of clinically trained professionals, but also demands well-controlled protocols. Their time and labour consuming procedures present hindrance to early diagnosis and intervention despite the high validity.

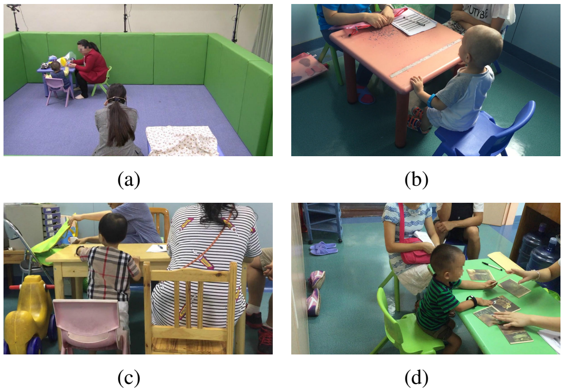
(a) (b) (c) (d)

Figure 1. Examples of different sessions from the dataset.

Recent development in machine learning has led to its broad applications in neural science, psychology study, general health care, emotion recognition [3], [4], [5] and behavior analysis [6], providing data-driven means to intelligently analyze ASD patterns and making it possible to conduct sophisticated ASD studies with large-scale multimodal data. Previous behavioral studies indicate that ASD individuals tend to show abnormal visual attention [7], [8], [9], [10], [11], [12]. As a result machine learning based approaches [13], [14] were proposed to identify ASD children based on eye movement patterns. Others include studies that use machine learning to identify ASD children based on motor abnormalities, or use machine learning to optimize the diagnosis process [15], [16], [17] with fewer test items.

One of the key symptoms of ASD patients is their impaired interpersonal communication ability. Studies showed that individuals with ASD respond to their Names differently from typical developing (TD, which means non-ASD) ones, and the decreased tendency to name response has been found across studies [18]. In addition, such criteria is widely adopted in early screening and diagnostic assessments to identify early signs of autism. In particular, a clinician calls the name of a child, and scores based on whether and how quick a clear response can be observed. This motivates us to

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

reproduce similar evaluation process with vision and learning techniques in order to reduce the required human interaction and expertise. Our major contribution in this work is twofold: First, we propose a "Response to Name Dataset" where adults and children interact in a relatively spontaneous way. As illustrated in Fig. 1, the collected videos contains both more lab-controlled sessions with cleaner background and HD cameras, as well as less controlled ones with cluttered backgrounds, more dynamic camera poses and mobile video capturing devices. By incorporating different sessions we hope that the children can react in a more natural way, and that methods can be tested with more challenging real world scenarios. This also makes future methods developed on this dataset possible to address retrospective home video analysis without requiring that the family comes to the lab venue. Second, we propose a machine learning based multimodal responsiveness assessment framework. Our hope is to provide an non-subjective framework with minimized human interaction and expertise. While currently designed framework largely follows clinical tradition by looking into response time and duration, the proposed system is able to support future extended frameworks with abundant additional features, such as head pose, motion and pose-wise duration.

2. Related Work

Recently, machine learning is playing an increasingly important role in ASD and psychology studies. The work of [15], [16], [17] used machine learning to simplify the test items in conventional clinical assessment methods in an optimized way. Their goal is to reduce unnecessary items while maintaining similar testing validity compared to the original schemes. Liu et al. [13], [14] proposed a bag-of-words (BoW) framework to represent face scanning patterns and used SVM to identify high-functioning ASD children. It was shown that the BoW model is an effective representation that helps to discover underlying ASD patterns. Crippa et al. [19] developed a machine learning method to identify low-functioning ASD children aged 2-4 based on upper-limb movement. Their method shows a possible motor signature of ASD that may be potentially useful in identifying patients.

Possibly one of the most related literature is the work proposed by Bidwell and Essa et al. [20]. The authors established a dataset containing 50 recorded "response to name" sessions, and explored markerless child head tracking with a camera recording from the top. The proposed method marks an important effort towards machine learning based "response to name" assessment. Our work in this paper seeks to further boost the application potential of the assessment system by differing from [20] in several aspects: First, our proposed dataset is relatively less lab-controlled with more dynamic indoor scenes, video poses, video qualities as well as flexible number of accompanying adults. While this presents additional challenge to our task, we believe it is necessary for developing assessment methods with stronger robustness and less input requirements. Second, we perform automatic name calling detection based on speech processing, while similar task is done manually in [20]. Third, our work adopts a more fine-grained responsiveness scoring scheme following clinical standards. Instead of using binary scoring, we follow ADOS with scores among 0, 1 and 2, where "0" indicates clear response, "1" partial response, and "2" no response.

3. The Proposed Dataset

3.1. Participating subjects

We recruited two groups of children: 22 ASD ones aged 2-3, and 21 TD, age-matched ones. Both groups are diagnosed by professional clinicians, and children in the ASD group are further evaluated comprehensively following ADOS. Before each experiment, we have obtained informed consent from children's parents. An overview is shown in Table 1.

TABLE 1. Overview of the participating subjects

Group	Average Age	Male	Female
ASD (n=22)	2.33	16	6
TD (n=21)	2.5	19	2

3.2. Collection procedure

The dataset collection procedure is ADOS-inspired, and follows similar protocol as the "response to name" probes in the ADOS assessment. In the experiment, each child is given a simple test where the child sits at a table playing with the accompanying adults or with toys. A doctor sits behind the child and calls his or her name in a natural tone. If the child turns around showing eye contact with the doctor, the test is completed. Otherwise, the doctor will repeat the call for additional 2 times. The whole process is recorded by a camera placed approximately behind the doctor (above the shoulder). The camera is expected to capture the eye contact from the child once he or she responds to the name calling.

The proposed dataset contains sessions with dynamic and challenging scenarios. Our expectation is to offer the children a more natural environment so that their reaction may better reveal their psychological status. As a result we do not impose too much control on the input video. Besides allowing dynamic camera poses, cluttered backgrounds and multiple accompanying adults, some scenarios contain ununiform illumination and occluded faces, and may be recorded by a smart phone rather than professional HD cameras.

3.3. Scoring

We ask a psychological clinician to evaluate each recorded video with a diagnostic score based on ADOS. In general, a clear and fast response and eye contact gives the score of 0, a response after multiple calls gives 1, while no response after all 3 calls gives 2. Each "response to name test" is

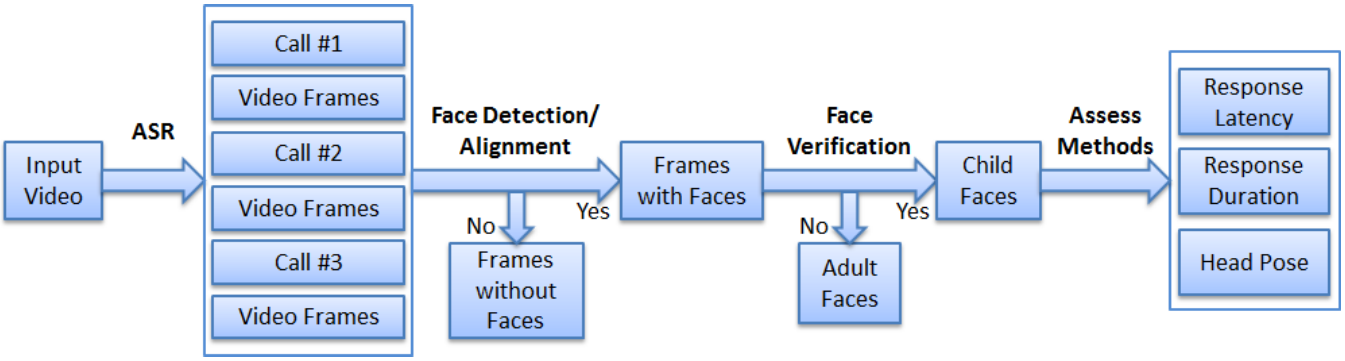

Figure 2. The proposed multimodal machine learning framework towards "response to name"

associated with one score. It should be noted that although having considerable correlation, the above assessment is an evaluation of the responsiveness to name calling, not an evaluation of autism risk.

4. A Multimodal Assessment Framework

An overview of the proposed multimodal "response to name" assessment framework is shown in Fig. 2. In this work, both speech processing and vision based methods are incorporated to minimize the human annotations in assessing the responsiveness. In particular, we simultaneously consider response speed, response duration as well as head pose information to jointly determine a predicted score. We hope such method can help to discover atypical response patterns effectively.

4.1. Automatic name calling detection

A core task in "response to name" experiments is to locate the time stamps of name callings such that response latency can be measured. In [20], name calling is annotated manually and incorporates unnecessary human interaction. We therefore propose an automatic name calling detection system based on automatic speech recognition (ASR). In particular, we designed an ASR system based on Kaldi [21], a toolbox widely used for speech recognition. The acoustic model is trained from 1000 hours of Mandarin telephone conversations with the chain model setup in [21]. In our experiment, the name of each child is first registered by our ASR system. The registered name is then matched with the ASR recognized speech signals throughout the video to locate the name callings.

4.2. Face detection and alignment

Another important task is to localize the child's head since it is the major body part which conveys actions of response and eye contact. Response and eye contact are often characterized by facing towards the interacting person. Such assumption becomes particularly valid when children suddenly called from behind while being highly focused. As a result, face detection presents a good method in signaling children's response when being called. In this work, we use the DLib [22] implementation of the face alignment methods proposed by Kazemi et al. [23] to simultaneously detect and align the faces. Besides detecting faces, the algorithm returns 68 landmarkers which later will be used to compute the head pose. An example of the detected face landmarkers is shown in Fig. 3.

4.3. Face verification

Since we allow accompanying adults, sometimes they may also respond to name calling with their faces detected. As a result, multiple faces from both adult and child can simultaneously appear in the same frame, and face verification is needed to distinguish the desired children's faces. Again we register each child's face in the system and verify every detected face based on the method proposed in [24]. We also formulate the verification problem as a structured sequence prediction problem with temporal information incorporated. This helps to stabilize and improve the verification performance under certain situations.

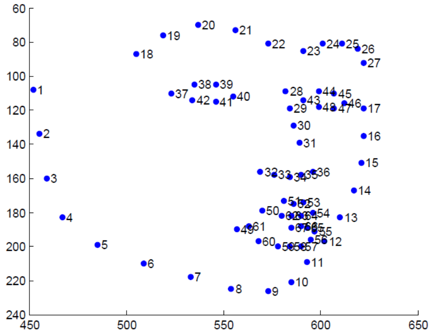

Figure 3. The facemarker demonstration

4.4. "Response to name" assessment

With the detected child faces, we are able to propose a rule-based "response to name" assessment framework.

Similar to conventional clinical diagnosis, our framework is based on the following two basic assumptions:
Assumption 1: A clear response should happen with relatively small latency upon calling.
Assumption 2: A clear response should last for a certain length of duration.

The response latency in the first assumption can be naturally modeled as:

$$latency = T_f - T_{c1}, \tag{1}$$

where T_f and T_{c1} indicate the time stamp of the first detected face and the beginning of the first call, respectively. The response duration in the second assumption can also be modeled as:

$$duration = N_f/\text{frame rate}, \tag{2}$$

where N_f represents the total number of frames containing detected child faces.

We observe that head pose also presents an important source of response information. Not only is the pose correlated with the clearness of response, but also it can provide head motion information that will benefit future assessment methods. Unfortunately, the DLib face alignment algorithm does not provide such information. We therefore propose a light and effective real-time pose estimation based on the detected landmarkers. We first propose a robust feature to effectively encode the pose information. Suppose there are n landmarkers whose coordinate in the frame is represented as (x_i, y_i), the head pose feature can be computed as:

$$\begin{array}{c} f = [x_1' - x_2', ..., x_1' - x_n', x_2' - x_3', ..., x_{n-1}' - x_n' \\ y_1' - y_2', ..., y_1' - y_n', y_2' - y_3', ..., y_{n-1}' - y_n'] \end{array}, \tag{3}$$

where x_i and y_i are the normalized relative coordinates of the ith marker with respect to the left and top most land markers:

$$\begin{array}{c} x_i' = \dfrac{x_i - \min\{x_i\}}{\max\{x_i\} - \min\{x_i\}} \\ y_i' = \dfrac{y_i - \min\{y_i\}}{\max\{y_i\} - \min\{y_i\}} \end{array}. \tag{4}$$

Since the above feature looks into the differences of all non-repeated pairwise landmarker combinations, the feature dimension can be high. This significantly adds computational costs to our head pose algorithm. We note that a large portion of the dimensions in fact contains redundant and non-informative information. As a result we apply PCA to the extracted features and maintain the top 20 dimensions with the highest energy. Given a training set with labeled head pose angles and any testing set, we perform face detection/alignment on both sets and extract the above head pose features. We then regress the head pose on the test set by referring to the top K nearest training samples with majority pose voting.

Upon obtaining the head pose information we incorporate it in the duration estimation by weighting each frame with the following biased Gaussian kernels:

$$duration = \sum_{i \in \mathbf{F}} k(\theta_{x,i}, \theta_{y,i})/\text{frame rate}, \tag{5}$$

$$k(\theta_x, \theta_y) = \exp(\frac{\theta_x^2}{2\sigma_x^2}) \exp(\frac{\theta_y^2}{2\sigma_y^2}), \tag{6}$$

where $\theta_{x,i}$ and $\theta_{y,i}$ are the horizontal and vertical head pose angles of the ith detected child face. In our experiment, we set σ_x and σ_y respectively to 45 and 60.

4.5. Score prediction

We perform grid search on both $latency$ and $duration$, and optimize a decision tree to predict the responsiveness score.

5. Experiment Result

5.1. Angle Estimation on the head pose database

We first conduct experiments on the head pose database. The dataset contains face images of 15 persons with variations of vertical angle and horizontal angles from -90 to +90 degrees. We split the dataset into two parts, faces of 10 persons as training set and 5 persons as testing set. the absolute error of the 480 detected test faces is shown in Table 2.

TABLE 2. The accuracy of head pose estimation framework

	Vertical angle	**Horizontal angle**
Range	-90 - +90	-90 - +90
Mean Abs Error	8.09	6.18

5.2. Result on the proposed database

Besides considering latency and duration jointly, we also consider applying the rule-based decision tree separately on each feature. The table3 shows the prediction accuracy of different methods. One could see that jointly considering both features returns the best performance. In addition, Table 4 shows the confusion matrix of the proposed method.

TABLE 3. Accuracy of different rules

Rule	**Latency**	**Weighted Duration**	**Combined**
Accuracy	90.7%	90.7%	**93%**

We also visualize the samples with respect to latency and duration on a 2-D plane. One could see that the features can clearly separate samples with different ground truth scores.

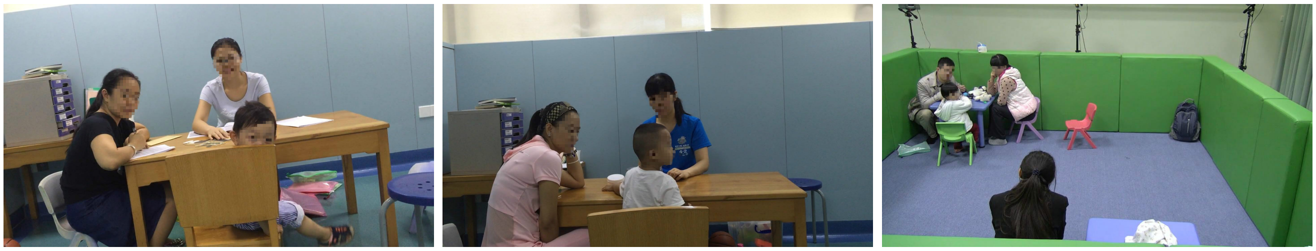

Figure 4. Examples of failure cases.

TABLE 4. Confusion matrix of our proposed method

		Clinician Score		
		0	1	2
Predicted Score	0	**24**	0	0
	1	0	**2**	0
	2	2	1	**14**

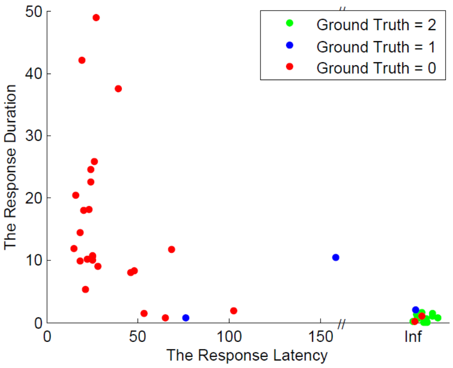

Figure 5. The relationship between ground truth and estimated duration and latency

5.3. Failure case study

We note that 3 children were incorrectly identified. The major reason lies in the failure face detection due to heavy occlusion or far away distance. Example failure cases are shown in Fig. 4. These children indeed have positive responses, while face detector failed in such cases and led to false final classification.

6. Conclusion

In this paper, we proposed a "Response to Name Dataset" for autism study. We also proposed a novel multimodal machine learning based assessment framework to automatically predict the responsiveness scores. Experimental results show that scores predicted by the proposed framework align closely with professional human diagnosis.

Acknowledgment

This research was funded in part by the National Natural Science Foundation of China (61401524, 61773413), Natural Science Foundation of Guangdong Province (2014A030313123), Natural Science Foundation of Guangzhou City (201707010363), the Fundamental Research Funds for the Central Universities (15lgjc12), Science and Technology Development Foundation of Guangdong Province, National Key Research and Development Program (2016YFC0103905) and IBM Faculty Award.

References

[1] C. Lord, S. Risi, L. Lambrecht, E. H. Cook Jr, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedułegeneric: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[2] K. Gotham, S. Risi, A. Pickles, and C. Lord, "The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity," *Journal of autism and developmental disorders*, vol. 37, no. 4, pp. 613–627, 2007.

[3] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 94–101.

[4] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 423–426.

[5] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 435–442.

[6] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim *et al.*, "Decoding children's social behavior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3414–3421.

[7] J. W. Tanaka and A. Sung, "The eye avoidance hypothesis of autism face processing," *Journal of autism and developmental disorders*, pp. 1–15, 2013.

[8] S. Weigelt, K. Koldewyn, and N. Kanwisher, "Face identity recognition in autism spectrum disorders: a review of behavioral studies," *Neuroscience & Biobehavioral Reviews*, vol. 36, no. 3, pp. 1060–1084, 2012.

[9] L. Yi, C. Feng, P. C. Quinn, H. Ding, J. Li, Y. Liu, and K. Lee, "Do individuals with and without autism spectrum disorder scan faces differently? a new multi-method look at an existing controversy," *Autism Research*, vol. 7, no. 1, pp. 72–83, 2014.

[10] W. Jones and A. Klin, "Attention to eyes is present but in decline in 2-6-month-old infants later diagnosed with autism," *Nature*, vol. 504, no. 7480, pp. 427–431, 2013.

[11] S. Wang, J. Xu, M. Jiang, Q. Zhao, R. Hurlemann, and R. Adolphs, “Autism spectrum disorder, but not amygdala lesions, impairs social attention in visual search,” *Neuropsychologia*, vol. 63, pp. 259–274, 2014.

[12] S. Wang, M. Jiang, X. M. Duchesne, E. A. Laugeson, D. P. Kennedy, R. Adolphs, and Q. Zhao, “Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking,” *Neuron*, vol. 88, no. 3, pp. 604–616, 2015.

[13] W. Liu, L. Yi, Z. Yu, X. Zou, B. Raj, and M. Li, “Efficient autism spectrum disorder prediction with eye movement: A machine learning framework,” in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 649–655.

[14] W. Liu, M. Li, and L. Yi, “Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework,” *Autism Research*, 2016.

[15] D. Bone, M. S. Goodwin, M. P. Black, C.-C. Lee, K. Audhkhasi, and S. Narayanan, “Applying machine learning to facilitate autism diagnostics: Pitfalls and promises,” *Journal of autism and developmental disorders*, pp. 1–16, 2014.

[16] J. Kosmicki, V. Sochat, M. Duda, and D. Wall, “Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning,” *Translational psychiatry*, vol. 5, no. 2, p. e514, 2015.

[17] M. Duda, J. Kosmicki, and D. Wall, “Testing the accuracy of an observation-based classifier for rapid detection of autism risk,” *Translational psychiatry*, vol. 4, no. 8, p. e424, 2014.

[18] A. S. Nadig, S. Ozonoff, G. S. Young, A. Rozga, M. Sigman, and S. J. Rogers, “A prospective study of response to name in infants at risk for autism,” *Archives of pediatrics & adolescent medicine*, vol. 161, no. 4, pp. 378–383, 2007.

[19] A. Crippa, C. Salvatore, P. Perego, S. Forti, M. Nobile, M. Molteni, and I. Castiglioni, “Use of machine learning to identify children with autism and their motor abnormalities,” *Journal of autism and developmental disorders*, vol. 45, no. 7, pp. 2146–2156, 2015.

[20] J. Bidwell, I. A. Essa, A. Rozga, and G. D. Abowd, “Measuring child visual attention using markerless head tracking from color and depth sensing cameras,” in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 447–454.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The kaldi speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[22] D. E. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[23] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[24] W. Liu, Y. Wen, Z. Yu, and M. Yang, “Large-margin softmax loss for convolutional neural networks,” *arXiv preprint arXiv:1612.02295*, 2016.