Audio-based Emotion Estimation for Interactive Robotic Therapy for Children with Autism Spectrum Disorder

Jonathan C. Kim[1], Paul Azzi[2], Myounghoon Jeon[3], Ayanna M. Howard[4], and Chung Hyuk Park[5]
[1,5] Department of Biomedical Engineering, The George Washington University, Washington, DC 20052 U.S.A.
(Tel : +1-202-994-5147; E-mail: jonkim@gwu.edu and chpark@gwu.edu)
[2] Department of Electrical and Computer Engineering, The George Washington University, Washington, DC 20052 U.S.A.
(E-mail: pazzi@gwmail.gwu.edu)
[3] Department of Cognitive and Learning Science, Michigan Technological University, Houghton, MI 49931 U.S.A.
(Tel : +1-906-487-3273; E-mail: mjeon@mtu.edu)
[4] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30032 U.S.A.
(E-mail: ayanna.howard@ece.gatech.edu)

Abstract - Recently, efforts in the development of speech recognition systems and robots have come to fruition with an overflow of applications in our daily lives. However, we are still far from achieving natural interaction between humans and robots, given that robots do not take into account the emotional state of speakers. The purpose of this research is to develop an automatic emotion classifier integrated with a robot, such that the robot can understand the emotional state of a human user by analyzing the speech signals from the user. This becomes particularly relevant in the realm of using assistive robotics to tailor therapeutic techniques towards assisting children with Autism Spectrum Disorder (ASD). With the number of children being diagnosed with ASD on the rise, finding new, affordable, and accessible means of therapy and assistance has become more of a concern. Improving audio-based emotion prediction for children with ASD will allow for the robotic system to properly assess the engagement level of the child and modify its responses to maximize the quality of interaction between the robot and the child and sustain an interactive learning environment.

Keywords – Assistive Robotics, Autism Spectrum Disorder, Speech Analysis, Affective Computing

1. Introduction

Autism spectrum disorder (ASD) is a neurological disorder that can, to varying extent, bring social, communication, and behavioral challenges [1]. The number of cases has increased in children born between 1992 to 2002 with 1 in 150 children being diagnosed in 1992 to 1 in 68 children with ASD in 2002. On average, as of 2014, Autism services cost U.S. citizens between $236-262 billion annually [2]. These various services include school district costs towards servicing special needs children, including children with ASD [3]. Studies have shown that early diagnosis and intervention can save these national costs by as much two-thirds.

For this purpose, our on-going research efforts [4,5] have developed interactive robotics to engage in emotional and social interactions with children with ASD. Our interactive robotic framework consists of two types of robotic systems: a humanoid robot (Robotis Mini) with the capability of gesture representations and an iOS-based mobile robot (Romo) capable of conveying emotion through facial expressions and voice. The humanoid robot displays dynamically varied body movements and gestures, while the mobile robot displays facial cues corresponding to specific emotions, as shown in Figure 1. Using these two robots together allows for easy singling-out and articulation of emotions to autistic children. This reduces the complexity of human emotional expressions, in which multiple emotional cues can be coexisting, while our robotic framework can simplify the channel for emotional interaction. A human's body movements, when coupled with contradicting facial cues, can often complicate a child with ASD's ability to distinguish the intended emotion and lead to sensory overloads.

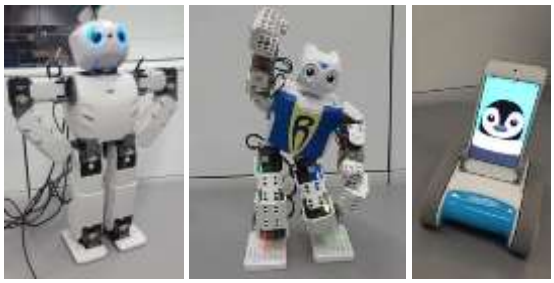

Fig. 1. Robotic systems used in our interactive robotic therapy sessions: Robotis OP2, Robotis Mini, and Romo

Our robots interact with children using pre-programmed scenarios, gestures, or games, as well as emotions the child is expressing while interacting with the robot. For example, if a child is crying during the session with the either robot,

the robot should appear to be aware and change the way it is interacting in order to comfort the child. This is where automatic emotion classification through audio and speech analysis becomes important to the robotic system. Moreover, this robotic system aims to integrate music into the learning environment in hopes of observing if and how music could further help children in relating body movements and gestures to specific emotions.

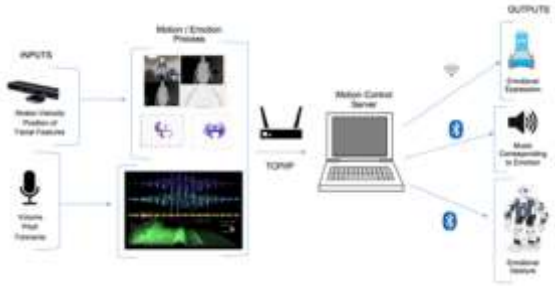

Fig. 2. Flow chart showing metrics robotic system uses to determine the appropriate system response

2. Automatic Emotion Classification

The ultimate goal of this research is to integrate an automatic emotion classifier with a robot for interactions with children in autistic spectrum. As an initial step to achieve the goal, we focus on constructing an automatic emotion classifier.

2.1 Database

In this paper, the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database was employed to extract emotional speech features to train an emotion classifier. The database was collected from 10 subjects (five males and five females), and two subjects form a pair for dyadic conversions. Each pair performed about 30 recording sessions which last about five minutes each. The five conversation pairs performed 71 scripted sessions and 80 spontaneous sessions in total. The total duration of recorded sessions is about 12 hours, and the audio sampling rate of the corpus is 16 kHz [6].

The dialogues were segmented at the turn level. In total the database contains 10039 turns with an average duration of 4.5 seconds, and the average number of words per turn is 11.4. Loosely speaking, the turn-level segmentation can be also viewed as the utterance level segmentation, where the speaker utters a thought or idea. The average duration of words in the database is about 400 ms; this gives the average speaking rate of the subjects 150 words-per-minute, which is also the average rate for English speakers in general [7].

The turn-level segments of the data were annotated with two different approaches, namely categorical and dimensional annotations. Three human evaluators annotated categorical emotions as neutral state, happiness, sadness, anger, surprise, fear, disgust, frustration, and excitement. Dimensions of valence, activation, and dominance were scaled from 1 to 5 by three human evaluators. The authors of the database employed the self-assessment manikin (SAM) to evaluate the corpus in emotional dimensions. The emotional dimensions were evaluated from 1 (negative) to 5 (positive) for valence (pleasure); 1 (low) to 5 (high) for activation (arousal); and 1 (weak) to 5 (strong) for dominance.

As suggested in [8], the five levels of the emotional dimensions were grouped into three due to the sparsity of data in the extremes of the scale range. The first level contains ratings in the range [1, 2], the second level contains ratings in the range (2, 4), and the third level contains ratings in the range [4, 5].

2.2 Speech Feature Extraction and Projection

One of the most popular speech feature extraction toolkits is openSMILE [9]. The openSMILE has been used by many speech researchers, especially for emotion classification in speech. The openSMILE toolkit extracts up to 6,373 acoustic features from speech signals. The openSMILE feature extractor provides energy, spectral, and voicing-related low-level descriptors, along with their statistical and regression measures [9].

It was reported that a multi-temporal analysis approach would improve the emotion classification accuracy [10]. However, one of our primary goals in this paper is to implement the automatic emotion classification in real time, and the computational coast of the multi-temporal approach would be burdensome for the real-time implementation. In the work of [10], the phrase-level emotion classification shows the highest performance rate; however, the phrase-level analysis would delay outputting the classification results, and it would not be near real-time processing. It is important to analyze the emotion in near real-time, such that a robot can react/respond in a spontaneous manner. As shown in [10], the performance rate of the 800 ms analysis approach is slightly below the phase-level approach, it was chosen in this work. Since the average the average speaking rate of English speakers is 150 words-per-minute in general, the 800-ms approach corresponds to analyzing two words per window [10].

In general, a larger number of features does not always result in better classification. It is important to reduce the dimensionality of the feature set, not only to speed up the classification process, but also to optimize classification performance. Feature projection algorithms are often employed for this reason. Feature projection algorithms use statistical methods to reduce the dimension of the features by applying linear transformation. One popular feature projection algorithm is principal components analysis (PCA). PCA finds the optimal orthogonal linear transformation matrix that preserves the subspace with the largest variance without paying any particular attention to the underlying class structure [11].

To obtain the optimal number of the principal components, we increase the number of the components by 10 each iteration. For each iteration, SVMs were employed to calculate unweighted accuracy (UWA) over 10 subjects using a leave-one-out cross-validation (LOOCV) technique. The averaged unweighted accuracy (UWA) was measured as defined in Eq. (1).

$$\text{UWA} = \frac{1}{M}\sum_{m=1}^{M}\frac{\#\text{ of hits in class } m}{\#\text{ of instances in class } m}, \qquad (1)$$

where M is the number classes. The level of chance in classifying M classes is $1/M$, and in our case of classifying the three levels of emotional states, the level of chance is 1/3.

The choice of the kernel function of SVMs is important in both the classification performance and the computational cost. Since the size of the IEMOCAP dataset is quite large both in the feature dimension and the number of instances, a linear kernel method was chosen as suggested in [12].

The results of sweeping the number of principal components from 10 to 200 are shown in Fig. 3. For classifying the levels of arousal and valence, Fig. 3 shows trends of increases in UWA as the number of components increases. In the case of dominance, a certain or pattern is not observed. It is known that classifying the levels of dominance is relatively difficult than the other two dimensions, and the speech acoustic features may not be the best for modeling the levels of dominance. Similar results in classifying the level of dominance are reported by others [8, 10]. Despite the importance of dominance dimension, due to its unpromising results, no further analysis in dominance dimension is carried out in this paper.

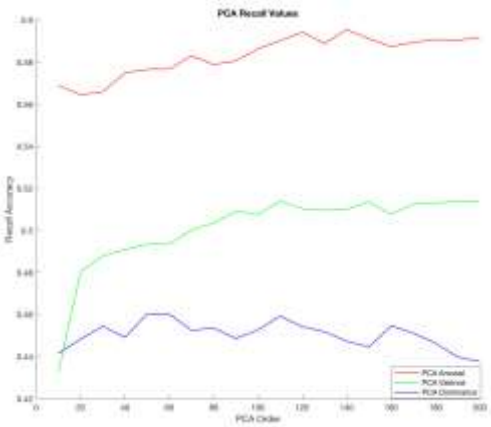

Fig. 3. UWA of PCA when the number of components are swept from 10 to 200 components

As shown in Fig. 3, it is suggested to use a large number of principal components; however, the trade-off is between the computational cost and subtle increases in accuracy rate. After around 150 principal components, the increases in the accuracy is very subtle. Throughout the rest of this paper, 150 principal components are used.

3. Real-time Implementation

3.1 Speaker Normalization

Since the recording environment and channel conditions of the IEMOCAP data are different from the real-world data collecting conditions in this research, a novel normalization method is proposed in this seciton. In previous work, a general speaker normalization method has been employed to resolve the expressivity variations across the speakers [8, 10, 13]. However a problem with the speaker normalizaiton is its assumption on the data distribution of the emotional states of each speaker. The assumption is that the data of each speaker has a similar distribution over the emotional states. For example, if the dataset of a particular person has significantly more "highly" aroused data than other speakers, such a speaker normalization method will be biased, and the classifier of the person will degrade. To overcome the issue, we propose a speaker normalization method, wherein only a few samples from the neutral emotional state of each speaker are used for normalization. This approach can be considered as a "configuration stage." The hypotheis is that if a machine learning algorithm knows what a person sounds like when the person is in a "neutral" state, and the data (features) are normalized in such a manner, the machine learning algorithm's prediction would improve. This method does not assume data distribution properties, but requires a configuration stage for a new speaker. The method is performed by the following steps:

1) Extract speech features from speech data
2) Perform PCA for feature dimension reduction.
3) Randomly select a subset of samples from the neutral state of each speaker.
4) For each speaker, calcualte the means and the variances of the selected data in the reduced feature dimensions.
5) For each speaker, subtract the means then divide the variances from all the data of the speaker.
6) Train the classifier.

When running the classifier with a new speaker, the steps 1-5 are performed in the same manner. To do so, a few samples of neutral state from the speaker must be collected. Now, the question is how many samples are sufficient. Using the IEMOCAP dataset, we increased the size of the data for normalization from 1 min to 10 mins. The results are shown in Fig. 4.

As shown in Fig. 4, trends of increases in the performance rate as the size of randomly selected samples increases in classification of activation and valence. Although a certain trend or pattern is not clearly observed in classifying the levels of dominance, the proposed method is promising. It is known that classifying the levels of dominance is relatively difficult than the other two dimensions, and the speech acoustic features may not be the best for modeling the levels of dominance.

Using the general speaker normalization with an 800 ms analysis window, the reported UWAs for classifying the three levels of activation and valence are 59.7 % and 51.2%, respectively [10]. As expected, the proposed method outperforms the general speaker normalization method. By normalizing data for each speaker using 3 mins of neutral data, the UWA is 62.9% and 52.7% for activation and valence, respectively.

To test whether or not this improvement is statistically significant, a paired t-test was performed. For classifying the three levels of activation, the proposed method

improved the UWA by 3.2 percentage points with a *p*-value less than 0.01. For classifying the levels of valence, the proposed method improved the UWA by 1.5 percentages points with a *p*-value less than 0.05. Since the *p*-values are less than 0.05 for both the cases, the improvement is statistically significant.

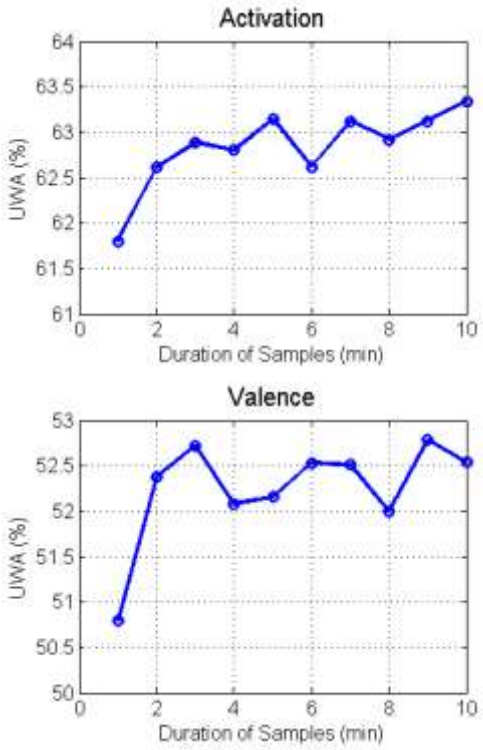

Fig. 4. Unweighted accuracies (UWA) when the duration of samples for normalization increases from 1 min to 10 mins.

Table 1. shows the UWA for each speaker when the duration of the randomly selected samples for normalization is 3 mins. The emotion classification accuracies are noticeably higher with female over male subjects. Much psychology and sociology literature reports that women are more emotionally expressive than men [14]. The findings in the emotion classification difference between the genders do not attempt to confirm their studies on expressivity; rather the current findings are supported by them.

As reported in previous works, the confusion matrices of the proposed method in Tables 2 and 3 show that the classification task is relatively easier in the opposite extremes than in the midrange emotions. The results are again based on the proposed method, where the duration of randomly selected samples is 3 mins.

Each row of the confusion matrices represents the instances in an actual class normalized by the total number of the instances, and each column represents the normalized instance in a predicted class. The opposite extremes are infrequently confused with each other.

Table 1. Unweighted accuracies (UWA) for classifying the levels of activation and valence, when 3 mins of neutral state data are used for normalization.

Speaker (gender)	**Activation UWA**	**Valence UWA**
1 (F)	67.2	48.3
2 (M)	63.1	54.5
3 (F)	68.0	48.2
4 (M)	62.1	46.9
5 (F)	63.1	58.5
6 (M)	57.0	49.2
7 (F)	65.4	59.1
8 (M)	62.3	51.2
9 (F)	63.0	53.2
10 (M)	62.2	56.5
Overall	**62.9**	**52.7**

Table 2. Confusion matrix for classifying the three levels of valence.

	Neg'	Neu'	Pos'
Neg	48.5	34.8	16.7
Neu	14.6	63.5	21.9
Pos	18.0	36.0	46.0

Table 3 Confusion matrix for classifying the three levels of activation.

	Low'	Med'	High'
Low	73.9	20.8	5.3
Med	29.1	30.9	40.0
High	4.9	9.5	85.6

3.2 Real-time emotion classification

Based on the PAD emotional state model, all emotions can be represented using the dimensions of pleasure (valence), arousal (activation), and dominance. Therefore, three SVMs were created. This way each SVM would determine the level of expression for its own specific dimension it was trained on. By separating and analyzing the emotions by their PAD dimensions instead of predicting emotions as a single unit, accuracies for each SVM could be assessed in order to increase prediction accuracies individually. The data would then be mapped using the three PAD dimensions in order to determine the emotion being expressed. An example of the mapping is shown in Fig. 5.

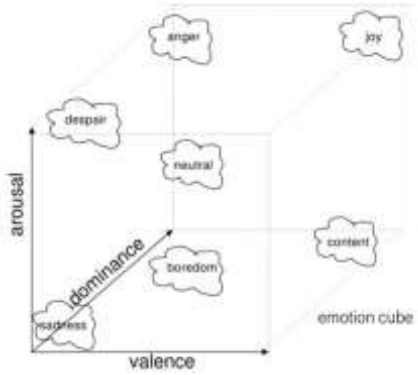

Fig. 5. Mapping of emotions from PDA dimensions

Using the MATLAB DSP Toolkit, the emotional classification program is able to read in live audio signals. The speech signal is continuously read and stored in an 800ms buffer. The data in this buffer is then sent into the openSmile Toolkit in order to extract features from it. Since the SVMs have been previously created, the classification program only needs to send in its extracted feature data into each of the three SVMs in order to get expression levels for valence, arousal, and dominance. These levels are measured from one to three; one being low; two being neutral; three being high. A three-point averaging filter is then implemented for each dimension so that the prediction values don't get influenced too heavily if a single 800ms frame acts as an outlier to the other two frames in the filter. Implementing the filter also allows for a smoother transition of predicted emotions. The original speech signal, predicted arousal and valence levels, and emotion mapping are displayed in a MATLAB GUI and updated every 800 ms. The GUI is shown in Figure 6. Dominance is not yet shown on in the GUI due to its low prediction accuracy.

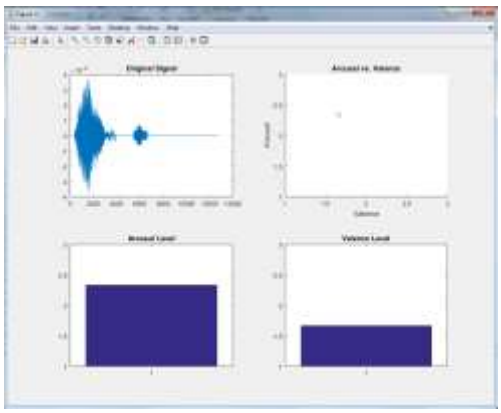
Fig. 6. Graphical output of audio analysis

4. Conclusion

To achieve natural interaction between humans and robots, it is crucial for a robot to obtain the capability of understanding emotional states from human responses. In this paper, we discussed a method of implementing an automatic emotion classifier using speech features, and a method of resolving variations of speakers by a novel method of speaker normalization. The ultimate goal of this research is to integrate the automatic emotion classifier with a robotic system for interactions with children with ASD in emotional communications. As an initial step to achieve the goal, we focused on the identification process of optimal features in vocal data and the automatic emotion classifier algorithms for real-time processing. The proposed real-time emotion classifier shows promising results, and the improvement in the unweighted accuracies compared to a previous method is statistically significant. We are currently collecting more data from children to expand the training sets and increase the accuracy.

In our future work, we will report emotion classification results using children's data (both from neurotypical group as well as ASD group) along with evaluations on child-robot interactions. Furthermore, similar approaches will be employed to classify music samples into emotional states and to investigate how a certain emotional music can increase the efficacy of the child-robot interactions.

Acknowledgement

This research is supported by the National Institutes of Health (NIH) under grant #R01-HD082914 through the National Robotics Initiative.

References

[1] *Autism Spectrum Disorder (ASD)*. Centers for Disease Control and Prevention, 17 Aug. 2015. Web. 16 Nov. 2015.

[2] Järbrink, Krister. "The economic consequences of autistic spectrum disorder among children in a Swedish municipality." Autism 11.5 (2007): 453-463.

[3] Reffert, Lori A., "Autism education and early intervention : what experts recommend and how parents and public schools provide" (2008). Theses and Dissertations. Paper 1224.

[4] R. Bevill, C. H. Park, H. J. Kim, J. W. Lee, A. Rennie, M. Jeon, and A. M. Howard, "Interactive robotic framework for multi-sensory therapy for children with autism spectrum disorder," in 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI, pp. 421–422), 2016.

[5] M. Jeon, R. Zhang, W. Lehman, S. Fakhrhosseini, J. Barnes, and C. H. Park, "Development and Evaluation of Emotional Robots for Children with Autism Spectrum Disorders." In International Conference on Human-Computer Interaction, pp. 372-376. Springer International Publishing, 2015.

[6] Busso, Carlos, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. "IEMOCAP: Interactive emotional dyadic motion capture database." *Language resources and evaluation* 42, no. 4 (2008): 335.

[7] J. R. Williams, "Guidelines for the use of multimedia in instruction," Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Chicago, IL, USA, vol. 42, no. 20, pp. 1447–1451, 1998.

[8] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan,"Context-sensitive learning for enhanced audiovisual emotion classification," Affective Computing, IEEE Transactions on, vol. 3, no. 2, pp. 184–198, 2012.

[9] F. Eyben, M.W¨ollmer, and B. Schuller,"OpenSMILE: The Munich versatile and fast open-source audio feature extractor," Proceedings of the International Conference on Multimedia, Singapore, pp. 1459–1462, 2010.

[10] Kim, Jonathan C., and Mark A. Clements. "Multimodal affect classification at various temporal lengths." *IEEE Transactions on Affective Computing* 6, no. 4 (2015): 371-384.

[11] A. M. Martinez and A. C. Kak, "PCA versus LDA," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 228–233, 2001.

[12] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," The Journal of Machine Learning Research, vol. 9, pp. 1871–1874, 2008.

[13] Kim, Jonathan C., and Mark A. Clements. "Formant-based feature extraction for emotion classification from speech." Telecommunications and Signal Processing (TSP), 2015 38th International Conference on. IEEE, 2015.

[14] A. M. Kring and A. H. Gordon, "Sex di erences in emotion: expression, experience, and physiology.," Journal of Personality and Social Psychology, vol. 74, no. 3, p. 686, 1998.