Music Teaching Robot Platform for Children with Autism

*Note: Sub-titles are not captured in Xplore and should not be used

1st Huanghao Feng
dept. name of organization (of Aff.)
name of organization (of Aff.)
Denver, USA
huanghao.feng@du.deu

2nd Mohammad H.Mahoor
dept. name of organization (of Aff.)
name of organization (of Aff.)
Denver, USA
m.mahoor@du.edu

***Abstract*—Music, performed by instrument, is created to express emotions. However, learning how to play can be a different story, even to those who have talent in music. Facial expressions could be a reliable representation of emotions from individuals for most of the cases. However, it could be challenge for specific populations such as Autism Spectrum Disorders (ASDs), a grouping of disorders characterized by profound difficulties with social interaction, have demonstrated impairments in facial emotion recognition. Indeed, Kanner (1943) originally described autism as a 'disorder of affective contact', and the current DSM-IV-TR diagnostic criteria for ASDs include items related to deficits in identifying and processing emotions: "marked impairments in the use of multiple nonverbal behaviors, such as ...facial expression..." and "lack of social or emotional reciprocity" (APA 2000). There are three different perspectives dealing with affect in multimedia, namely, expressed emotions, felt emotions and expected emotions, Electrodemal activity (EDA) signals which is considered as a periphery**

***Index Terms*—Social Robot, Autism, Music Teaching, Automatic System**

I. Introduction

Individuals with autism spectrum disorder experience verbal and nonverbal communication impairments, including motor control, emotional facial expressions, and eye gaze attention. Oftentimes, individuals with high-functioning autism have deficits in different areas, such as (1) language delay, (2) difficulty in having empathy with their peer and understanding others emotions (i.e. facial expressions recognition.), and more remarkably (3) joint attention (i.e. eye contact and eye gaze attention). Autism is a disorder that appears in infancy [**?**]. Although there is no single accepted intervention, treatment, or known cure for ASDs, these individual will have more successful treatment if ASD is diagnosed in early stages. People who suffer from autism might also have several other unusual social developmental behaviors that may appear in infancy or childhood. For instance children with autism show less attention to social stimuli (e.g. facial expressions, joint attention), and respond less when calling their names. Compared with typically developing children, older children or adults with autism can read facial expressions less effectively and recognize emotions behind specific facial expressions or the tone of voice with difficulties [**?**]. In contrast to TD individuals, children with autism (i.e. high-functioning, Asperger syndrome) may be overwhelmed with social signals such as facial behaviors and expression and complexity of them and they suffer from interacting with other individuals, therefore they would prefer to be alone. That is why it would be difficult for individuals with autism to maintain social interaction with others [**?**].

Identify applicable funding agency here. If none, delete this.

II. Related Works

A. Autism

Individuals with autism spectrum disorder experience verbal and nonverbal communication impairments, including motor control, emotional facial expressions, and eye gaze attention. Oftentimes, individuals with high-functioning autism have deficits in different areas, such as (1) language delay, (2) difficulty in having empathy with their peer and understanding others emotions (i.e. facial expressions recognition.), and more remarkably (3) joint attention (i.e. eye contact and eye gaze attention). They might also have several other unusual social developmental behaviors that may appear in infancy or childhood. For instance children with autism show less attention to social stimuli (e.g. facial expressions, joint attention), that is why it would be difficult for individuals with autism to maintain social interaction with others [**?**].

Turn-taking is a type of organization in conversation and discourse where participants speak one at a time in alternating turns. In practice, it involves processes for constructing contributions, responding to previous comments, and transitioning to a different speaker, using a variety of linguistic and non-linguistic cues. [**?**] While the structure is generally universal, [**?**] that is, overlapping talk is generally avoided and silence between turns is minimized, turn-taking conventions vary by culture and community. [**?**] Conventions

vary in many ways, such as how turns are distributed, how transitions are signaled, or how long is the average gap between turns. In many contexts, conversation turns are a valuable means to participate in social life and have been subject to competition. [**?**] It is often thought that turn-taking strategies differ by gender; consequently, turn-taking has been a topic of intense examination in gender studies. While early studies supported gendered stereotypes, such as men interrupting more than women and women talking more than men, [**?**] recent research has found mixed evidence of gender-specific conversational strategies, and few overarching patterns have emerged. [**?**]

Motor control is the systematic regulation of movement in organisms that possess a nervous system. Motor control includes movement functions which can be attributed to reflex, [**?**]. Motor control as a field of study is primarily a sub-discipline of psychology or neurology. While the modern study of motor control is an increasingly interdisciplinary field, research questions have historically been defined as either physiological or psychological, depending on whether the focus is on physical and biological properties, or organizational and structural rules. [**?**] Areas of study related to motor control are motor coordination, motor learning, signal processing, and perceptual control theory.

B. Electrodermal activity (EDA) and Emotion Classification

Emotion is an intense mental experience often manifested by rapid heartbeat, breathing, sweating, and facial expressions. Emotion recognition from these physiological signals is a challenging problem with interesting applications such as developing wearable assistive devices and smart human-computer interfaces. This paper presents an automated method for emotion classification in children using electrodermal activity (EDA) signals. The time-frequency analysis of the acquired raw EDAs provides a feature space based on which different emotions can be recognized. To this end, the complex Morlet (C-Morlet) wavelet function is applied on the recorded EDA signals. The annotation process is performed considering the synchronicity between the children's facial expressions and the EDA time sequences. Various experiments are conducted on the annotated EDA signals to classify emotions using a support vector machine (SVM) classifier. The quantitative results show that the emotion classification performance remarkably improves compared to other methods when the proposed wavelet-based features are used.

EDA has been used as an effective and reproducible electrophysiological method for investigating sympathetic nervous system function [**?**], [**?**], [**?**], [**?**]. Note that the sympathetic nervous burst changes the skin conductance, which can be traced by analyzing the EDA signals [**?**], [**?**], [**?**]. The Q-sensor is a convenient wireless-based EDA device with no need for cables, boxes, or skin preparation. This device can track three types of data including EDA, temperature, and acceleration at the same time [**?**]. It is worth mentioning that as of today, there has been no published work on emotion classification using the EDA signals collected by this dataset collected at the Georgia Institute of Technology [**?**].

EDA signals are nonstationary and noisy; hence, wavelet-based analysis of EDA signals has been considered in the literature [**?**], [**?**] either as a pre-processing step or a feature extraction approach for emotion classification. [**?**] used a set of wavelet coefficients representing EDA features together with heart rate signal to increase the percentage of correct classifications of emotional states and provide clearer relationships among the physiological response and arousal and valence. [**?**] used a feature space based on the discrete wavelet transform (DWT) of the EDA signal to distinguish subjects suffering social anxiety disorder (SAD) and a control group. Using MLP and DWT features, they achieved a classification accuracy of 85%.

Physiological responses have been identified as reliable indicators of human emotional and cognitive states. This section is dedicated to review some existing methods used for human emotion recognition based on various physiological responses, such as facial expression and other types of bio-signals.

A wearable glass device was designed by [**?**] to measure both electrodermal activity (EDA) and photoplethysmogram data for emotion recognition purposes. A built-in camera was also used in this device for capturing partial facial expression from the eye and nose area. This approach obtains remarkable performance in facial expression recognition in the subject-dependent cases. However, for subject-independent cases, it results in different accuracies across different types of emotions, which is an undesirable feature.

Several emotion classification methods have been presented in the literature using different bio-signals [**?**], [**?**], [**?**], [**?**]. Due to the variety of the signals used in these methods, different approaches have been designed to comply with their specific characteristics. Analysis of variance (ANOVA) and linear regression [**?**] are the commonly used methods to extract features from bio-signals and to recognize different emotional states. These methods are based on the assumption of a linear relationship between the recorded signals and emotional states. A fuzzy-based classification method [**?**] has been used in to transform EDA and facial electromyography (EMG) to valence and arousal states. These states were then used to classify different emotions.

Support Vector Machine (SVM) is a well-known supervised learning algorithm that has extensively been used for pattern classification and regression [**?**]. The SVM classifier tends to separate dataset by drawing an optimal hyperplane

between classes such that the margin between them becomes maximum. The samples of each class that are located within the margin are called support vectors and play the main role in calculating the parameters of the hyperplanes between the corresponding classes. Machine learning algorithms such as SVM, linear discriminant analysis (LDA), and classification and regression tree (CART) have been employed for emotion classification purposes. For instance, in several works including [**?**], [**?**], the authors combined various types of bio-signals such as ECG, skin temperature (SKT), HR, and Photoplethysmogram (PPG) for emotion classification purposes. [**?**] proposed unsupervised clustering methods for emotion recognition. Their method benefited from several features obtained from different body responses such as SC, HR, and EMG. They showed that only a few statistical features such as the mean and standard deviation of the data can be relevant identifiers for defining different clusters.

To the best of our knowledge there are a few works [**?**], [**?**] that have studied and compared different automated classification techniques for emotion recognition of children with autism using EDA signals. This motivated us to conduct this study using an existing dataset, which concentrates on emotion classification of ASD children based on the relationship between their facial expressions and the collected EDA signals.

III. Methodology

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections III-A–**??** below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Xylo-Bot: An Interactive Music Teaching System

A novel interactive human-robot music teaching system design is presented in this chapter. In order to make the robot play the xylophone properly, several things need to be done. First is to find a proper xylophone with correct timber; second, we have to arrange the xylophone in the proper position in front of the robot to make it visible and be reachable to play; finally, design the intelligent music system for NAO.

B. NAO: A Humanoid Robot

We used a humanoid robot called NAO developed by Aldebaran Robotics in France. NAO is 58 cm (23 inches) tall, with 25 degrees of freedom. This robot can conduct most human behaviors. It also features an onboard multimedia system including four microphones for voice recognition and sound localization, two speakers for text-to-speech synthesis, and two HD cameras with maximum image resolution 1280 x 960 for online observation. As shown in Figure 1, these utilities are located in the middle of the forehead and the mouth area. NAO's computer vision module includes facial and shape recognition units. By using the vision feature of the robot, the robot can see the instrument from its lower camera and be able to implement an eye-arm self-calibration system which allows the robot to have real-time micro-adjustment of its arm-joints in case of off positioning during music playing.

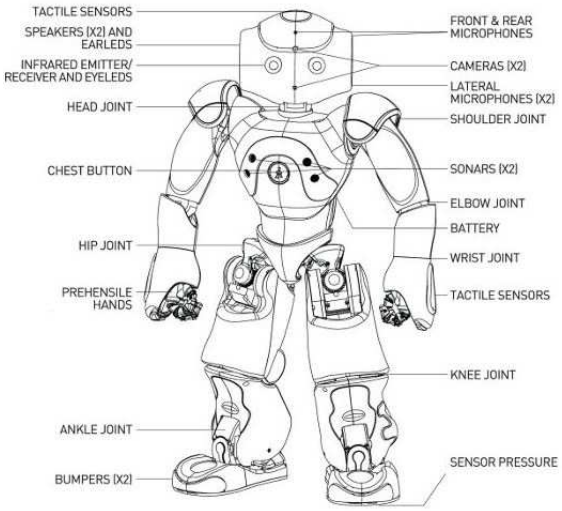

Fig. 1. A Humanoid Robot NAO: 25 Degrees of Freedom, 2 HD Cameras and 4 Microphones

The robot arms have a length of approximately 31 cm. Each arm has five degrees of freedom and is equipped with sensors to measure the position of each joint. To determine the pose of the instrument and the mallets' heads, the robot analyzes images from the lower monocular camera located in its head, which has a diagonal field of view of 73 degrees. These dimensions allow us to choose a proper instrument presented in the next section.

The four microphone locations embedded on the toy or NAO's head can be seen in Figure 1. According to the official Aldebaran documentation, these microphones have sensitivity of 20mV/Pa +/-3dB at 1kHz, and an input frequency range of 150Hz - 12kHz. Data will be recorded as a 16 bit, 48000Hz, 4 channels wav file which meets the requirements for designing the online feedback audio score system described below.

C. Accessories

The purpose of this study is to have a toy-size humanoid robot play music. Some necessary accessories needed to be purchased and made before the robot was able to play music. All accessories will be discussed in the following sections.

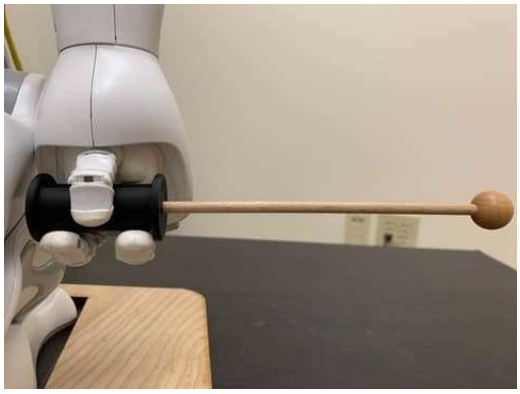

Fig. 2. Mallet Griper

Fig. 3. Instrument Stand Front View.

D. Xylophone: A Toy for Music Beginner

In this system, due to NAO's open arms' length, we choose a Sonor Toy Sound SM soprano-xylophone with 11 sound bars of 2 cm in width. The instrument has a size of 31 cm x 9.5 cm x 4 cm, including the resonating body. The smallest sound bar is playable in an area of 2.8 cm x 2 cm, the largest in an area of 4.8 cm x 2 cm. The instrument is diatonically tuned in C-Major/a-minor. For the beaters/mallets, we used the pair that came with the xylophone with a modified 3D printed grip (details in next subsection) to allow the robot's hands to hold them properly. The mallets are approximately 21 cm in length and include a head of 0.8 cm radius.

The 11 bars of the xylophone represent 11 different notes (11 frequencies) which covers approximately a one and a half octave scale starting from C6 to F7.

E. Mallet Gripper Design

According to the size of Nao's hands, we designed and 3D printed a pair of gripers to have the robot be able to hold the mallets properly. All dimensions can be found in Figure 2.

F. Instrument Stand Design

A wooden base was designed and laser cut to hold the instrument in the proper place for the robot to be able to play music. All dimensions can be found in Figure 3.

IV. Module-Based Acoustic Music Interactive System Design

In this section, a novel module-based robot-music teaching system will be presented. Three modules have been built in this intelligent system including module 1: eye-hand self-calibration micro-adjustment; module 2: joint trajectory generator; and module 3: real time performance scoring feedback. See Figure 4.

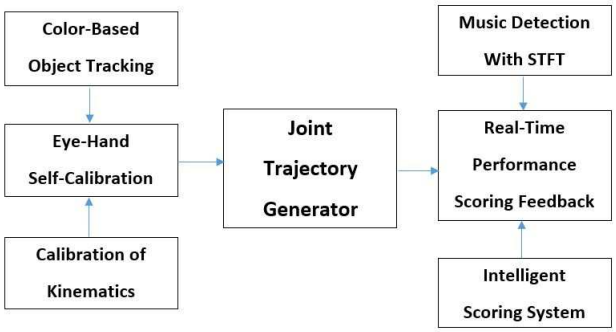

Fig. 4. Block Diagram of Module-Based Acustic Music Interactive System

A. Module 1: Eye-hand Self-Calibration Micro-Adjustment

Knowledge about the parameters of the robot's kinematic model is essential for tasks requiring high precision, such as playing the xylophone. While the kinematic structure is known from the construction plan, errors can occur, e.g., due to the imperfect manufacturing. After multiple rounds of testing, it was found the targeted angle chain of arms never actually equals the returned chain. We therefore used a calibration method to accurately eliminate these errors.

1) Color-Based Object Tracking: To play the xylophone, the robot has to be able to adjust its motions according to the estimated relative position of the instrument and the heads of the beaters it is holding. To estimate these poses, adopted in this thesis, we used a color-based technique.
The main idea is, based on the RGB color of the center blue bar, given a hypothesis about the instrument's pose, one can project the contour of the object's model into the camera image and compare them to actually observed contour. In this way, it is possible to estimate the likelihood of the pose hypothesis. By using this method, it allows the robot to track the instrument with very low cost in real-time. See Figure 5

B. Module 2: Joint Trajectory Generator

Our system parses a list of hex-decimal numbers (from 1 to b) to obtain the sequence of notes to play. It converts the notes into a joint trajectory using the beating configurations obtained from inverse kinematics as control points. The timestamps for the control points will be defined by the user in order to meet the experiment requirement. The trajectory is then computed using Bezier interpolation in joint space by the manufacturer-provided API and then sent to the robot controller for execution. In this way, the robot plays in-time with the song.

C. Module 3: Real-Time Performance Scoring Feedback

The purpose of this system is to provide a real-life interaction experience using music therapy to teach kids social skills and music knowledge. In this scoring system, two core features were designed to complete the task: 1) music detection; 2) intelligent scoring-feedback system.

1) A. Music Detection: Music, in the understanding of science and technology, can be considered as a combination of time and frequency. In order to make the robot detect a sequence of frequencies, we adopted the short-time Fourier transform (STFT) to this audio feedback system. This allows the robot to be able to understand the music played by users and provide the proper feedback as a music teaching instructor.

The short-time Fourier transform (STFT) , is a Fourier-related transform used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time. In practice, the procedure for computing STFTs is to divide a longer time signal into shorter segments of equal length and then compute the Fourier transform separately on each shorter segment. This reveals the Fourier spectrum on each shorter segment. One then usually plots the changing spectra as a function of time. In the discrete time case, the data to be transformed could be broken up into chunks or frames (which usually overlap each other, to reduce artifacts at the boundary). Each chunk is Fourier transformed, and the complex result is added to a matrix, which records magnitude and phase for each point in time and frequency. This can be expressed as:

Fig. 5. Color Detection From NAO's Bottom Camera Color Based Edge Detection.

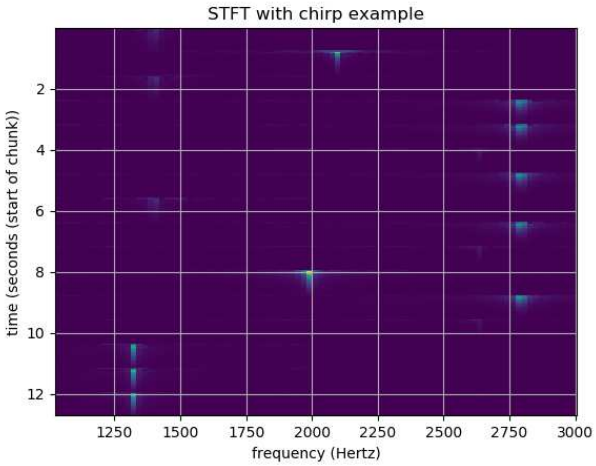

Fig. 6. Melody Detection with Short Time Fourier Transform

$$\{x[n]\}(m,\omega) \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n}$$

likewise, with signal x[n] and window w[n]. In this case, m is discrete and ω is continuous, but in most typical applications, the STFT is performed on a computer using the Fast Fourier Transform, so both variables are discrete and quantized.
The magnitude squared of the STFT yields the spectrogram representation of the Power Spectral Density of the function:

$$\text{spectrogram}\{x(t)\}(\tau,\omega) \equiv |X(\tau,\omega)|^2$$

After the robot detects the notes from user input, a list of hex-decimal number will be returned. This list will be used in two purposes: 1) to compare with the target list for scoring and sending feedback to user; 2) used as a new input to have robot playback in the game session as discussed in the next chapter.

2) B. Intelligent Scoring-Feedback System: In order to compare the detected notes and the target notes, we used an algorithm which is normally used in information theory linguistics called Levenshtein Distance. This algorithm is a string metric for measuring the difference between two sequences.

In our case, the Levenshtein distance between two string-like hex-decimal numbers a, b (of length $|a|$ and $|b|$ respectively) is given by $\text{lev}_{a,b}(|a|,|b|)$ where

$$\text{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) \\ \min \begin{cases} \text{lev}_{a,b}(i-1,j)+1 \\ \text{lev}_{a,b}(i,j-1)+1 \\ \text{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} \end{cases}$$

where $1_{(a_i \neq b_j)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise, and $\mathrm{lev}_{a,b}(i,j)$ is the distance between the first i characters of a and the first j characters of b.

Note that the first element in the minimum corresponds to deletion (from a to b), the second to insertion and the third to match or mismatch, depending on whether the respective symbols are the same. Table **??** demonstrates how to apply this principle in finding the Levenshtein distance of two words "Sunday" and "Saturday".

Based on the real life situation, we defined a likelihood margin for determining whether the result is good or bad:

$$likelihood = \frac{len(target) - lev_{target,source}}{len(target)}$$

where if the likelihood is greater than 66% 72%, the system will consider it as a good result. This result will be passed to the accuracy calculation system to have the robot decide whether it needs to add more dosage to the practice. More details will be discussed in the next chapters as it relates to the experiment design.

D. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 7", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Fig. 7. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.