Validation of Vocal Prosody Modifications to Communicate Emotion in Robot Speech

Joe Crumpton
Distributed Analytics and Security Institute
Mississippi State University
Starkville, MS, USA
joe.crumpton@msstate.edu

Cindy L. Bethel
Social, Therapeutic and Robotic Systems Lab
Department of Computer Science and Engineering
Mississippi State University
Starkville, MS, USA
cbethel@cse.msstate.edu

***Abstract*—This research investigated the use of MARY, an open source speech synthesizer, to convey the emotional intent of a robot through the robot's vocal prosody. The robot's pitch, pitch range, speech rate, and volume were varied to convey anger, fear, happiness, and sadness. The results indicate participants recognized the intended emotions at a rate statistically higher than chance whether they provided their own free choice word or used a forced choice survey to describe the emotion presented by the robot. As expected, the participants correctly transcribed more of the words spoken by the robot when the vocal prosody modifications performed were small (sadness) than when the vocal prosody modifications were large (anger and fear).**

Keywords*—*Affective, Emotive & Conative Aspects of Collaboration

I. INTRODUCTION

Most users of computers still use keyboards and mice for input and reading screens for output. The fact that robots are typically mobile makes these forms of input/output more problematic. Just as speech is typically used by humans in close proximity to communicate, robots and their human users will likely use speech to communicate. Prasad *et al.* point out that even communication between robots would ideally be via voice while the robots are in the presence of people [1], [2]. One area of synthesized speech that could be improved to sound more natural is the use of varying vocal prosody. Vocal prosody refers to non-linguistic attributes such as average pitch, pitch range, volume, and speech rate [3]. One use of vocal prosody by humans is to communicate the emotions felt by the speaker [4]–[6].

This experiment investigates the use of MARY [7], an open source speech synthesizer, to produce a robot voice that can convey emotions through varying vocal prosody attributes. Section II presents a brief review of the communication of emotions through vocal prosody and a review of previous research on the use of vocal prosody to communicate emotion by robots. Section III describes the equipment used in the experiment along with the study protocol. Section IV presents the results from the experiment and Section V contains a discussion of those results. Section VI presents conclusions and plans for future research.

TABLE I
EMOTIONS COMMUNICATED VIA VOCAL PROSODY [9]–[11]

Emotion	Average Pitch	Pitch Range	Timing	Loudness
Happiness	High	Large	Moderate	High
Surprise	High	Large	Slow	Moderate
Sadness	Low	Small	Slow	Low
Anger	High	Large	Fast	High
Disgust	Low	Small	Moderate	Low
Fear	High	Small	Fast	High

II. RELATED WORK

Human speech is typically characterized by fluctuations in vocal prosody and its attributes that convey the speaker's emotions to listeners. There have been attempts at incorporating these types of attributes in robotic speech to convey emotional content to people with whom the robot interacts.

A. Communicating Emotions via Vocal Prosody

Listeners can often recognize the emotional state of the person speaking through the speaker's vocal prosody [4]–[6]. The *Big Three* of vocal prosody: pitch, timing, and loudness [8] are among the the characteristics of speech determined to correlate with the communication of emotions. The vocal prosody characteristics that are typically used to express Ekman's *Big Six* emotions: happiness, surprise, sadness, anger, disgust, and fear [9]–[12] are listed in Table I. Note that pitch is involved in two of the attributes: average pitch and pitch range. Average pitch is the average value of the fundamental frequency (F_0) of a speech segment. Pitch range is the difference between the highest and lowest pitches contained in a speech segment.

B. Emotional Speech Synthesizers

The first attempts to convey emotions through synthesized speech were limited because changing the vocal prosody attributes such as pitch range and pitch contour of generated speech was not supported by speech synthesizers [13]. Some current speech synthesizers, such as MARY (Modular Architecture for Research on speech sYnthesis) [7], were designed with the generation of expressive speech as a goal. MARY

allows specification of the vocal prosody parameters through markup of its input and through speech synthesis request parameters [14]. Several commercially available speech synthesizers such as Acapela Group's Acapela and Cereproc's CereVoice text-to-speech engines contain voices claimed to portray different emotions. However, the authors have not found any validation tests for such speech synthesizers showing that listeners actually perceive the emotion being portrayed by the generated speech.

C. Use of Varying Vocal Prosody by Robots

The use of vocal prosody to communicate emotion in a robot voice has been a subject of past research. Tielman *et al.* modified a robot's speech through arousal and valence parameters while the robot was quizzing and being quizzed by a child [15]. However, the researchers did not validate that the children were correctly interpreting the emotional intent of the robot's speech. While the acoustical correlates of different emotions are known (see Table I), researchers should check that the intended emotions were actually communicated to the study participants. Research on the use of emotion by agents and on-screen characters was similarly criticized by Beale and Creed [16].

Read and Belpaeme have investigated the use of vocal prosody by robots in non-linguistic utterances to convey emotions [17]–[20]. Their research has shown people do attribute emotions to the non-linguistic utterances of robots [17], [18]. But, the specific emotion conveyed is a result of the participant's observation of the robot's interactions, not the particular sound of the non-linguistic utterance made by the robot [18], [19]. Therefore, Read and Belpaeme recommend the use of non-linguistic utterances in addition to, not as a replacement for, natural language speech by a robot [20].

III. Experiment

This section details the equipment (robot and speech synthesizer), study procedure, and the experimental design for the experiment investigating the use of vocal prosody to convey emotions in robot speech.

A. Robot

The Survivor Buddy robot [21] was used during this experiment. The robot consists of a small monitor (with webcam and microphone) mounted to an arm. The arm uses Dynamixel actuators to provide four degrees of freedom. The robot is usually mounted to a mobile base and the robot was developed at Texas A&M University to investigate human-robot interaction related to disaster responses. For this experiment the robot sat on a table in front of the participant. Fig. 1 shows the arrangement of the participant and robot. This experiment was conducted using the *Wizard-of-Oz* technique [22]. The output of the robot's microphone and webcam was streamed to a separate room where a robot operator used a custom interface to control the robot's actions and its responses to the participant.

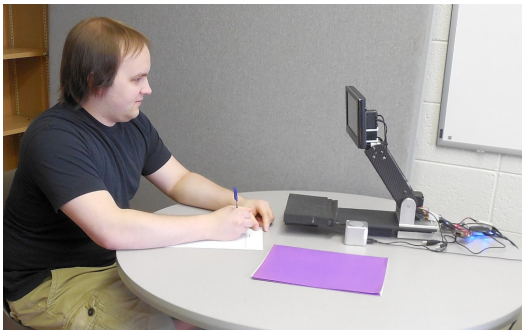

Figure 1. Participant interacting with the Survivor Buddy robot

Several steps were taken to avoid the communication of emotion by the robot's appearance and the content of the robot's speech. The "face" displayed on the Survivor Buddy's monitor (shown in Fig. 2) was based on Apple's Finder icon. The smile used in the original icon was removed from the image in an effort to avoid implying the robot was happy. Semantically Unpredictable Sentences (SUS) [23] were used as the text for the robot to speak during the study. Sets of SUS were originally proposed as the text to use in speech synthesizer intelligibility testing [23]. A SUS consists of a sentence made up of commonly used short words where the words are placed in a sentence structure according to their part of speech. The sentence structures used to create sets of SUS are shown in Fig. 3. The constructed sentences have no contextual meaning and therefore the sentences do not convey an emotion. During pilot studies [2] the ability of individual words to imply an emotion even when appearing in a meaningless sentence was noted. The words used in the current set of SUS were selected to avoid symbolism and the unintentional portrayal of emotion. Examples of semantically unpredictable sentences used in the study are:

- The front fact owned the chair.
- Grab the food or the sea.
- The case joined the chance that jumped.

A set of forty semantically unpredictable sentences was generated for this experiment. Each participant heard a subset of twenty sentences (four sentences in each of the five vocal prosody modifications described in the next section) during the study.

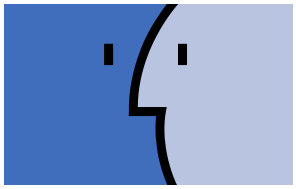

Figure 2. Image displayed as the face on the Survivor Buddy monitor

B. Speech Synthesizer

MARY (Modular Architecture for Research on speech sYnthesis) was the speech synthesizer used as Survivor Buddy's

Determiner + Noun + Verb (intransitive) + Preposition + Determiner + Adjective + Noun

Determiner + Adjective + Noun + Verb (transitive) + Determiner + Noun

Verb (transitive) + Determiner + Noun + Conjunction + Determiner + Noun

Determiner + Noun + Verb (transitive) + Determiner + Noun + Relative Pronoun + Verb (intransitive)

Figure 3. Sentence structures for Semantically Unpredictable Sentences [23]

```
<maryxml version="0.5" xml:lang="en-US">
   <p>
   <prosody
      rate="50%"
      contour="(0%,+0st)(30%,-0.5st)
         (50%,-2.0st)(70%,-3.0st)(100%,-4.5st)">
            The town came for the fast store.
   </prosody>
   </p>
</maryxml>
```

Figure 4. MaryXML for "Sad" vocal prosody

voice in this experiment [7]. MARY is an open source speech synthesizer designed to facilitate research in emotional speech synthesis. MARY was selected for use in this experiment because it is the only speech synthesizer that the authors have found that supports the modification of the pitch, pitch contour, pitch range, loudness, and speed of the synthesized speech. Many speech synthesizers such as Microsoft's .NET speech synthesizer and Nuance Communication's Dragon Mobile claim to support Speech Synthesis Markup Language and its prosody element. But the documentation for those two speech synthesizers states that prosody attributes such as pitch contour, pitch range, and duration are ignored when generating speech [24], [25]. Until the prosody modifications made in this experiment are actually supported by other speech synthesizers, the results of this experiment only apply to the MARY speech synthesizer.

The Hidden Markov Model (HMM) based voice used in this study was generated from the CMU_ARCTIC speech synthesis database [26] for the female American English speaker `slt`. A HMM based voice (as opposed to a unit selection based voice) was used because MARY supports changing the vocal prosody parameters of generated speech when MARY is used with a HMM based voice. Table II shows the changes made to the standard voice to convey the different emotions. Only four (anger, fear, happiness, and sadness) of Ekman's *Big Six* emotions were used in the experiment. Disgust and surprise were omitted to reduce the number of emotion choices. Disgust and surprise were chosen to omit because they seemed the least likely emotions for a robot to need to convey in human-robot interactions. Neutral was used to label the vocal prosody meant to convey no emotion. Note the speech rate and volume of the "neutral" vocal prosody were set to values that allowed the speech rate and volume to be increased and decreased for the other vocal prosody manipulations to convey the four basic emotions used in the study without causing distortion in the generated speech. The values for the vocal prosody parameters were based on a review of the literature on the use of vocal prosody to communicate emotions in human speech [9]–[11] and modified through an iterative process based on the results from previous experiments using the MARY speech synthesizer to communicate emotions [2]. There is one difference between the values found in Table I and the specifications given in Table II. Typically anger is expressed by a higher than normal average pitch. Since fear and happiness are also conveyed by a higher than normal average pitch, the average pitch used in this experiment for anger is lower than normal. This change was motivated by the discussion of hot anger and cold anger by Pell *et al.* [27]. Hot anger (such as rage) is expressed by a high pitch and cold anger (such as a threat) is expressed by a lower pitch [27].

Fig. 4 gives the input used to instruct the MARY speech synthesizer to say the sentence *The town came for the fast store* using vocal prosody to convey "sadness". In addition to the markup shown in the figure, the changes to the average pitch (-30Hz), pitch range (70%), and volume (40%) were made through HyperText Transfer Protocol (HTTP) request parameters that accompanied the speech synthesis request to the MARY server.

C. Procedure

Each participant completed an informed consent form and a demographics survey. The demographics survey asked for several standard items such as gender, age, occupation, highest level of education, ethnicity, and race. The demographics survey also asked a set of questions concerning the participant's previous experience with technology and synthesized speech. The technology questions included asking the participant to rate their prior computer experience, prior robot experience, and video game experience. The questions about synthesized speech asked the participant if they used GPS units that gave spoken directions or if the participant used digital personal assistants such as Apple's Siri.

Short form versions of the Positive and Negative Affect Schedule (PANAS) [28], [29] and Big Five Inventory (BFI-10) [30], [31] were completed by the participant. The ten item PANAS survey results in a measure of the participant's positive and negative affect [29]. The ten item BFI-10 yields measure of the participant's personality in terms of five dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness [31]. Although there were not specific hypotheses concerning relationships between the participant's mood or personality and their ability to recognize emotions conveyed by vocal prosody, these measures were used to investigate if such correlations existed.

The researcher then explained the study instructions to the participant. During the study the robot would say a sentence. Based on the sound of the robot's voice, the participant would write one word that describes the emotion being conveyed by the robot. Then the participant transcribed the sentence said by the robot. The robot would repeat the sentence once automatically. The participant could ask the robot to repeat

TABLE II
CHANGES MADE TO STANDARD VOICE TO CONVEY EMOTIONS

Emotion	Pitch	Pitch Range	Pitch Contour	Speech Rate	Volume
Anger	-50Hz	120%	each word has a falling contour	95%	95%
Fear	+70Hz	20%	rising	100% with random pauses between words	70%
Happiness	+50Hz	200%	varies between -5% and +25%	varies between 70% and 90%	80%
Neutral	unchanged	unchanged	flat	85%	60%
Sadness	-30Hz	70%	falling	50%	40%

the sentence by saying "repeat". The robot would repeat the sentence as often as the participant requested. The participant could ask the robot to proceed to the next sentence by saying "next". This procedure was followed for the first ten sentences spoken by the robot.

The researcher would warn the participant that the sentences said by the robot would consist of real words, but the words would be in random order so the sentences would not make sense. The sentences were actually constructed using the semantically unpredictable sentence process outlined in section III-A. The participants were told that the words were in random order to avoid explaining what semantically unpredictable sentences are and the purposes for their use in this experiment. The researcher gave the participant a response sheet to use during the study and then the researcher would leave the room to avoid influencing the participant's responses. The robot introduced itself and repeated the instructions. The robot would ask the participant if they were ready to begin before starting to work through the sentences with the participant.

After the first ten sentences the robot gave the participant new instructions. The participant was asked to select the emotion conveyed by the robot's voice from a list of five emotions: angry, fear, happy, neutral, and sad. The robot stated a new answer sheet for the participant was located in the folder next to the robot. Fig. 5 shows the emotion choices for the last ten sentences as depicted on the second response sheet. The robot would ask the participant if they were ready to continue and then the robot would proceed with the last ten sentences.

Figure 5. Emotion choices presented to the participants for the last ten sentences

After all of the sentences were completed, the robot asked the participant to retrieve the researcher from the hallway. The participant would then complete the short-form PANAS survey again for thoroughness and as a follow-up measure to ensure participants were in a similar affective state to how they felt when they arrived for the study. The participant's last surveys were an evaluation of the robot and an evaluation of the study itself. The researcher then debriefed the participant and thanked the participant for his or her help.

D. Design

This experiment consisted of participants listening to sentences said by a robot which used vocal prosody modifications to convey emotions. Since each participant heard all five versions of the robot's voice, the experiment was a within-subjects design. The proposed hypotheses were:

- H_1: Participants will recognize the emotion being communicated by the robot solely based on the robot's vocal prosody (pitch, pitch range, speech rate, and volume).
- H_2: Participants will understand the robot's speech better when changes to the robot's vocal prosody (pitch, pitch range, speech rate, and volume) are small.

The approach to have participants provide their own word through free choice to describe the emotion conveyed by the robot's voice for the first ten sentences was inspired by criticisms presented in an article by Greasley *et al.* that the choices presented to the participant may influence their perception of the emotion conveyed from vocal prosody [32]. The free choice of emotion (first ten sentences) and the forced choice of emotion (last ten sentences) were not counterbalanced in this experiment. This decision was made to avoid having the list of emotions (angry, fear, happy, neutral, and sad) presented during the forced choice of emotion influence the participant's choice of words during the free choice of emotion part of the study.

The participants' free choices of emotions were categorized as one of the five expected emotions or as not an emotional word. If the participant's response was one of the five expected emotions or if the response was a word whose root was one of the five expected emotions, it was categorized as the given emotion. For example, fearful was categorized as fear. Three data sets relating affective concepts and words were consulted next: the hierarchical cluster analysis of emotions by Shaver *et al.* [33], WordNet-Affect [34], and EmoSenticNet [35]. If the word appeared in one of the three data sets, the word was categorized using the emotion specified by the data set. As an example, "indifference" is categorized as neutral-emotion in WordNet-Affect. If a word remained unclassified, the word's definition and list of similar words in WordNet [36] was consulted. If the definition or list of similar words contained

one of the five expected emotions, the word was classified as that emotion. For example, "frantic" is defined in WordNet as "distraught with fear or other violent emotion" so frantic was categorized as fear. Words that could not be categorized using the above process were labeled as "not emotion" and that response was excluded from further analysis.

IV. Results

The following results are reported for 53 participants, all of which were college students. The participants (34 females and 19 males) had an average age of 18.96 years (SD = 1.65). The students were recruited from lower-level computer science and psychology classes. The only inclusion criterion was the requirement that English be the student's first language.

A. Free Choice of Emotion

The participants' word choices were categorized as one of the expected emotions using the process described in section III-D are provided in Table III. Of the 530 participant responses, 92 could not be categorized as describing an emotion. Examples of words that were not categorized as an emotion were: alert, commanding, determined, informative, and sweet.

TABLE III
Participant's Response Categorized as Non-Emotion or Emotion

Intended Emotion	**Emotion Word**	
	No	Yes
Anger	29	77
Fear	13	93
Happiness	21	85
Neutral	16	90
Sadness	13	93
Total	92	438

Table IV shows how the participants classified the sentences when allowed to provide their own word choice to describe the emotion conveyed by the robot's voice. This free choice of emotions consisted of the first ten sentences heard by each participant.

TABLE IV
Recognition Rates for Free Choice of Emotions

Intended Emotion	**Emotion Category (% correct)**				
	Anger	Fear	Happiness	Neutral	Sadness
Anger	**51.9**	3.9	14.3	15.6	14.3
Fear	0.0	**55.9**	26.9	1.1	16.1
Happiness	2.4	21.2	**38.8**	7.1	30.6
Neutral	5.6	8.9	16.7	**38.9**	30.0
Sadness	12.9	1.1	1.1	5.4	**79.6**

Table V gives the results of a one sample *t*-test (α = 0.05) for each of the emotion recognition rates during the free choice of emotion portion of the study. The participant's free choice of emotion was categorized as one of the five expected emotions using the process described in Section III-D. If the participant randomly guessed at the emotional intent of the robot's speech, we would expect the participant to be correct 20% (or 1/5) of the time. Therefore, the test value used in the one sample *t*-test was 0.2, the recognition rate that results from random guessing.

TABLE V
Statistical Significance of Recognition Rates for Free Choice of Emotions

Emotion	**Mean**	***t***	***df***	***p* (2-tailed)**	**Cohen's *d***
Anger	0.51	5.03	46	<0.001	0.73
Fear	0.59	6.69	51	<0.001	0.93
Happiness	0.40	3.60	50	0.001	0.50
Neutral	0.41	3.69	51	0.001	0.51
Sadness	0.76	11.47	50	<0.001	1.61

B. Forced Choice of Emotion

Table VI gives the recognition rates for the sentences (the last ten sentences heard by each participant) where the participant was asked to select the emotion conveyed by the robot's voice from a list consisting of angry, fear, happy, neutral, and sad (see Fig. 5).

TABLE VI
Recognition Rates for Forced Choice of Emotions

Intended Emotion	**Selected Emotion (% correct)**				
	Anger	Fear	Happiness	Neutral	Sadness
Anger	**70.8**	2.8	1.9	24.5	0.0
Fear	0.0	**62.3**	14.2	2.8	20.8
Happiness	0.0	**32.1**	31.1	11.3	25.5
Neutral	1.9	0.9	7.5	**79.2**	9.4
Sadness	20.8	1.9	0.0	35.8	**41.5**

Table VII gives the results of a one sample *t*-test (α = 0.05) for each of the emotion recognition rates during the forced choice of emotion portion of the study. The test value used in the one sample *t*-test was 0.2, the recognition rate expected from random guessing.

TABLE VII
Statistical Significance of Emotion Recognition Rates for Forced Choice of Emotions

Emotion	**Mean**	***t***	***df***	***p* (2-tailed)**	**Cohen's *d***
Anger	0.71	10.69	52	<0.001	1.47
Fear	0.62	7.87	52	<0.001	1.08
Happiness	0.31	2.36	52	0.022	0.32
Neutral	0.79	13.62	52	<0.001	1.87
Sadness	0.42	3.60	52	<0.001	0.49

Table VIII shows the results of a paired *t*-test (α = 0.05) comparing the free choice and forced choice emotion recognition rates.

C. Intelligibility of Emotional Robot Speech

A strict interpretation of correct transcription was used when compiling the following results. Other than transcribing the exact word said by the robot, only homonyms (words sharing a pronunciation) were accepted as a correct transcription.

TABLE VIII
COMPARING FOR FREE AND FORCED CHOICE EMOTION RECOGNITION RATES

Emotion	Free Choice Mean	Forced Choice Mean	*df*	*t*	*p* (2-tailed)
Anger	0.51	0.69	46	-2.63	**0.012**
Fear	0.59	0.63	51	-0.82	0.416
Happiness	0.40	0.31	50	1.27	0.211
Neutral	0.41	0.79	51	-5.48	**<0.001**
Sadness	0.76	0.40	50	5.51	**<0.001**

For example, transcribing *see* for *sea* was accepted as correct. There were 142 distinct words in the set of 40 semantically unpredictable sentences. For the 53 participants, a total of 6890 words were heard. Table IX gives the percentage of words said by the robot in each intended emotion that were transcribed correctly.

TABLE IX
TRANSCRIBED WORDS BY EMOTION

Intended Emotion	Number of Words	Transcribed Correctly (%)
Anger	1373	76.6
Fear	1379	68.2
Happiness	1372	80.4
Neutral	1381	83.5
Sadness	1385	85.1

A one-way repeated measures ANOVA was conducted to see if the correct transcription rates of words spoken in the different vocal prosody modifications were significantly different. Mauchly's test indicated that the assumption of sphericity had not been violated, $\chi^2(9) = 7.579$, $p = 0.58$. The ANOVA results show that there was a significant difference in the correct transcription rates for words said using the different emotions' vocal prosody modifications, $F(4,208) = 25.633$, $p < 0.001$.

Just fourteen words (shown in Table X) accounted for 21% of the total number of transcription errors (1463). Each of the words in Table X were transcribed incorrectly at least 75% of the total times that the words appeared in the sentences heard by the participants. Table XI lists the thirteen words always transcribed correctly by the participants.

TABLE X
WORDS TRANSCRIBED INCORRECTLY MOST OFTEN (BY PERCENTAGE)

Word	Appeared	Transcribed Incorrectly	Most Common Incorrect Answer
end	33	33	pen (17)
looped	33	33	moved (20)
law	20	20	lost (10)
snored	20	20	more (8)
owned	53	47	on (20)
year	20	17	mirror (5)
posed	33	28	post (18)
cook	20	16	put (7)
fact	33	26	fat (10)
week	33	25	wheat (9)
helped	20	15	held (13)
sport	20	15	port (5)
staff	20	15	stuff (14)

TABLE XI
WORDS ALWAYS TRANSCRIBED CORRECTLY

Word	Transcribed Correctly
held	40
box	33
fish	33
girl	33
grab	33
great	33
plan	33
trip	33
fresh	20
road	20
teach	20
tree	20
wife	20

V. DISCUSSION

The results concerning the recognition rates of intended emotion for both free choice and forced choice of emotions are discussed in this section. Also discussed are the correct transcription rates for the words spoken with the different vocal prosody modifications performed to convey the emotions anger, fear, happiness, neutral, and sadness.

A. Recognition of Emotion

Table IV shows the recognition rates of the intended emotion conveyed by vocal prosody when the participant choose their own word to describe the emotion. The recognition rates of anger (51.9%), fear (55.0%), and sadness (79.6%) were comparable to the successful emotion recognition rate (60%) of people listening to human speakers [37]. The recognition rates of happiness and neutral were lower than the recognition rates of the three previously mentioned emotions. The recognition rate expected from random guessing would have been 20% (1/5). The happiness and neutral recognition rates were still significantly higher than 20% as shown in Table V. The fact that happiness had a lower recognition rate than the other emotions was anticipated since previous research on vocal prosody and the recognition of emotions has shown that happiness is difficult to recognize from vocal prosody alone [27]. The null hypothesis for H_1 would be that the participants would make emotion choices randomly regardless of vocal prosody. The null hypothesis is rejected because all of the emotion recognition rates during the free choice of emotions were significantly higher than chance.

The criticism by Greasley *et al.* that the emotion options presented to participants would influence the participants' choices is supported. The results of paired sample *t*-tests comparing the participants' emotion recognition rates for the free choice of emotion sentences and the forced choice of emotion sentences are shown in Table VIII. Only two (fear and happiness) of the forced choice emotion recognition rates are not significantly different than the free choice emotion recognition rates even though the vocal prosody modifications

were exactly the same between the two conditions. Of the three remaining emotions, two emotion recognition rates were significantly higher (anger: free choice = 0.51, forced choice = 0.69, $t(46) = -2.63$, $p = 0.012$; neutral: free choice = 0.41, forced choice = 0.79, $t(51) = -5.48$, $p < 0.001$) for forced choice of emotions. One emotion recognition rate was significantly lower (sadness: free choice = 0.76, forced choice = 0.40, $t(50) = 5.51$, $p < 0.001$) for forced choice of emotions.

B. Intelligibility of Emotional Robot Speech

The highest correct transcription rate of words spoken in the different vocal prosody modifications was 85.1% for the "sadness" vocal prosody (see Table IX). It was expected that the "sadness" vocal prosody correct transcription rate would be higher than the correct transcription rates for anger, fear, and happiness because the vocal prosody modifications made to convey "sadness" were relatively small. For the "sadness" vocal prosody, the average pitch was lowered by 30Hz (the smallest change for any emotion) and the pitch range reduced by 30% (the smallest change for any emotion). A pairwise comparison of correct transcription rates for the five vocal prosody modifications showed that of words said in the "fear" vocal prosody were transcribed correctly at a significantly lower rate than words said in the other vocal prosodies. The low correct transcription rate for "fear" can be explained by the fact that the vocal prosody modifications performed to express fear were among the largest modifications made for any of the emotions. The average pitch was raised by 70Hz (the largest change for any emotion), the pitch range was only 20% (the smallest range for any emotion), and the speech rate was 100% (the fastest speech rate for any emotion). The second lowest correct transcription rate (76.7%) was for words said in the "anger" vocal prosody. The vocal prosody modifications made to to express fear were also quite large. The average pitch was lowered by 50Hz (the second largest absolute change for any emotion), the pitch range was 120%, and the speech rate was 95% (the second fastest speech rate for any emotion). The null hypothesis for H_2 would be that the correct transcription rate would be uniform for all of the vocal prosody modifications. The null hypothesis is rejected based on the results of the one-way repeated measures ANOVA and the pairwise comparisons of the correct transcription rates.

The only unexpected result in the correct transcription rates was the fact that the words said in the "sadness" vocal prosody were transcribed correctly more often than the words said in the "neutral" vocal prosody. The "neutral" vocal prosody was the baseline for the vocal prosody modifications to convey the four emotions. It was expected that words said with the "neutral" vocal prosody would have the highest correct transcription rate. The difference might be due to the speech rate used to express the different emotions. The "sadness" vocal prosody used the slowest speech rate, 50% of the normal speech rate for the `slt` voice model, while the "neutral" vocal prosody used a speech rate of 85%. This implies that speech rate might be a more important factor influencing the correct transcription rate than pitch, pitch range, and volume.

Some of the transcription errors were obviously due to words sounding similar to each other. For example, *posed* was often transcribed incorrectly as *post*. On the other hand, some of the transcription errors appear to be the result of the participants using the context of a sentence to provide a word. The word *looped* was transcribed in all cases as *moved* even though the words do not sound similar to each other. The sentence heard by the participants was *The site placed the arm that looped*. The word *moved* does seem to make more sense in that sentence than the word *looped*.

Relaxing the standard of only counting exact matches and homonyms as correct transcriptions would have slightly increased the correct transcription rate by the participants. Using the strict definition of correct transcription resulted in a 78.8% correct transcription rate over the 6890 words heard by participants. Allowing different word endings to count as correct (allowing *places* or *place* to count as *placed*) would raise the overall correct transcription rate to 80.4%.

VI. Conclusion and Future Work

This experiment has shown that a person interacting with a robot will utilize a robot's varying vocal prosody to determine what emotion a robot is attempting to convey. Robot speech can now be used alongside body language and facial expressions (when possible) to present multimodal expressions of emotion that should improve the naturalness of human-robot interactions.

The results of this experiment suggest several avenues for further investigation. An obvious start is more research on the vocal prosody attributes' correlates of happiness. The vocal prosody modifications intended to communicate happiness resulted in the lowest recognition rates for both free choice and forced choice of emotions. Expanding the vocal prosody modifications to include more subtle attributes such as voice quality and articulation accuracy (once supported by speech synthesizers) could possibly increase the recognition rates of happiness and the other emotions.

Validating that these vocal prosody modifications could be applied to a male voice in order to convey emotions is another extension of this experiment. While the absolute values of the parameter changes might be affected by the lower average pitch of a male voice, it is expected that the direction and relative changes in the pitch and pitch range would be a suitable starting point for modifying a male voice model's vocal prosody to communicate emotion.

Although there is some support for the universal interpretation of emotions via vocal prosody [27], [38], the results from this experiment currently apply only to native English speakers. An obvious follow-on study would test these vocal prosody modifications for emotion recognition by participants from other cultures or who speak other languages. Another planned extension of this work is investigating the effects of robot body shape on the interpretation of emotional intent by human listeners.

Finally, looking at emotion recognition rates when these vocal prosody modifications are made to meaningful sentences

would be quite interesting. Would applying the modifications intended to convey sadness increase the level of emotion conveyed by a sad sentence such as "*I miss the time we spent together*" [39]. What would happen if the emotion conveyed by the vocal prosody attributes were mismatched with a sentence's linguistic content? Would using a "happy" vocal prosody with a sad sentence result in the listener recognizing sarcasm?

It is expected that this experiment's results along with the above suggested research extensions will eventually result in robots that naturally interact with their human users via voice in various domains such entertainment, education, and personal assistants.

REFERENCES

[1] R. Prasad, H. Saruwatari, and K. Shikano, "Robots that can hear, understand and talk," *Advanced Robotics*, vol. 18, no. 5, pp. 533–564, 2004.

[2] J. Crumpton and C. L. Bethel, "Conveying emotion in robotic speech: Lessons learned," in *Proceedings: 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Edinburgh, Scotland: IEEE, 2014.

[3] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. New Jersey: Prentice Hall PTR, 2001.

[4] G. Fairbanks and W. Pronovost, "An experimental study of the pitch characteristics of the voice during the expression of emotion," *Speech Monographs*, vol. 6, no. 1, p. 87, 1939.

[5] G. L. Huttar, "Relations between prosodic variables and emotions in normal American English utterances," *Journal of Speech and Hearing Research*, vol. 11, no. 3, pp. 481–487, 1968.

[6] K. R. Scherer, "Vocal affect expression: a review and a model for future research," *Psychological Bulletin*, vol. 99, no. 2, pp. 143–165, 1986.

[7] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[8] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signal processing: State-of-the-art and future perspectives of an emerging domain," in *Proceedings: 16th ACM International Conference on Multimedia*. Vancouver, British Columbia, Canada: ACM, 2008, pp. 1061–1070.

[9] K. Hammerschmidt and U. Jürgens, "Acoustical correlates of affective prosody," *Journal of Voice*, vol. 21, no. 5, pp. 531–540, 2007.

[10] C. Sobin and M. Alpert, "Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy," *Journal of Psycholinguistic Research*, vol. 28, no. 4, pp. 347–365, 1999.

[11] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1145–1154, 2006.

[12] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.

[13] J. Cahn, "The generation of affect in synthesized speech," *Journal of the American Voice Input / Output Society*, vol. 8, pp. 1–19, 1990.

[14] M. Schröder. MaryXML. http://mary.dfki.de/documentation/maryxml (current May 22, 2013).

[15] M. Tielman, M. Neerincx, J.-J. Meyer, and R. Looije, "Adaptive emotional expression in robot-child interaction," in *Proceedings: 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Bielefeld, Germany: ACM, 2014, pp. 407–414.

[16] R. Beale and C. Creed, "Affective interaction: How emotional agents affect users," *International Journal of Human-Computer Studies*, vol. 67, no. 9, pp. 755–776, 2009.

[17] R. Read and T. Belpaeme, "How to use non-linguistic utterances to convey emotion in child-robot interaction," in *Proceedings: 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Boston, Massachusetts: ACM, 2012, pp. 219–220.

[18] ——, "People interpret robotic non-linguistic utterances categorically," in *Proceedings: 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Tokyo, Japan: ACM, 2013, pp. 209–210.

[19] ——, "Situational context directs how people affectively interpret robotic non-linguistic utterances," in *Proceedings: 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Bielefeld, Germany: ACM, 2014, pp. 41–48.

[20] ——, "Non-linguistic utterances should be used alongside language, rather than on their own or as a replacement," in *Proceedings: 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Bielefeld, Germany: ACM, 2014, pp. 276–277.

[21] Z. Henkel, N. Rashidi, A. Rice, and R. Murphy, "Survivor buddy: A social medium robot," in *Proceedings: 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Lausanne, Switzerland: ACM, 2011, pp. 387–387.

[22] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of oz studies - why and how," *Knowledge-Based Systems*, vol. 6, no. 4, pp. 258–266, 1993.

[23] C. Benoît, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," *Speech Communication*, vol. 18, no. 4, pp. 381–392, 1996.

[24] Microsoft. prosody element. https://msdn.microsoft.com/en-us/library/hh361583(v=office.14).aspx (current Feb. 25, 2015).

[25] Nuance Communications. SSML compliance. http://dragonmobile.nuancemobiledeveloper.com/public/Help/DragonMobileSDKReference_Android/SpeechKit_Guide/SpeakingText.html (current Feb. 25, 2015).

[26] A. W. Black. CMU_ARCTIC speech synthesis databases. http://festvox.org/cmu_arctic/ (current Feb. 8, 2014).

[27] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, "Factors in the recognition of vocally expressed emotions: A comparison of four languages," *Journal of Phonetics*, vol. 37, no. 4, pp. 417–435, 2009.

[28] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," *Journal of Personality and Social Psychology*, vol. 54, no. 6, pp. 1063–1070, 1988.

[29] E. R. Thompson, "Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS)," *Journal of Cross-Cultural Psychology*, vol. 38, no. 2, pp. 227–242, 2007.

[30] O. P. John and S. Srivastava, *The Big Five Trait taxonomy: History, measurement, and theoretical perspectives*. New York, NY US: Guilford Press, 1999, ch. 4, pp. 102–138.

[31] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German," *Journal of Research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[32] P. Greasley, C. Sherrard, and M. Waterman, "Emotion in language and speech: Methodological issues in naturalistic approaches," *Language and Speech*, vol. 43, no. 4, pp. 355–375, 2000.

[33] P. Shaver, J. Schwartz, D. Kirson, and C. O'Connor, "Emotion knowledge: further exploration of a prototype approach." *Journal of Personality and Social Psychology*, vol. 52, no. 6, p. 1061, 1987.

[34] R. Valitutti, "Wordnet-affect: an affective extension of wordnet," in *Proceedings: 4th International Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal: European Language Resources Association, 2004, pp. 1083–1086.

[35] S. Poria, A. Gelbukh, E. Cambria, P. Yang, A. Hussain, and T. Durrani, "Merging senticnet and wordnet-affect emotion lists for sentiment analysis," in *Proceedings: International Conference on Signal Processing (ICSP)*, vol. 2. Beijing, China: IEEE, 2012, pp. 1251–1255.

[36] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[37] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, "Vocal cues in emotion encoding and decoding," *Motivation and Emotion*, vol. 15, no. 2, pp. 123–148, 1991.

[38] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-Cultural Psychology*, vol. 32, no. 1, pp. 76–92, 2001.

[39] J. B. Russ, R. C. Gur, and W. B. Bilker, "Validation of affective and neutral sentence content for prosodic testing," *Behavior Research Methods*, vol. 40, no. 4, pp. 935–939, 2008.