8-14-2015

Use of Vocal Prosody to Express Emotions in Robotic Speech

Joe Crumpton

Use of vocal prosody to express emotions in robotic speech

By

Joseph John Crumpton

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Computer Science
in the Department of Computer Science and Engineering

Mississippi State, Mississippi

August 2015

Use of vocal prosody to express emotions in robotic speech

By

Joseph John Crumpton

Approved:

Cindy L. Bethel
(Major Professor)

Derek T. Anderson
(Committee Member)

J. Edward Swan II
(Committee Member)

Byron J. Williams
(Committee Member)

T. J. Jankun-Kelly
(Graduate Coordinator)

Jason M. Keith
Interim Dean
Bagley College of Engineering

Name: Joseph John Crumpton

Date of Degree: August 14, 2015

Institution: Mississippi State University

Major Field: Computer Science

Major Professor: Dr. Cindy L. Bethel

Title of Study: Use of vocal prosody to express emotions in robotic speech

Pages of Study: 138

Candidate for Degree of Doctor of Philosophy

Vocal prosody (pitch, timing, loudness, etc.) and its use to convey emotions are essential components of speech communication between humans. The objective of this dissertation research was to determine the efficacy of using varying vocal prosody in robotic speech to convey emotion. Two pilot studies and two experiments were performed to address the shortcomings of previous HRI research in this area.

The pilot studies were used to determine a set of vocal prosody modification values for a female voice model using the MARY speech synthesizer to convey the emotions: anger, fear, happiness, and sadness. Experiment 1 validated that participants perceived these emotions along with a neutral vocal prosody at rates significantly higher than chance. Four of the vocal prosodies (anger, fear, neutral, and sadness) were recognized at rates approaching the recognition rate (60%) of emotions in person to person speech.

During Experiment 2 the robot led participants through a creativity test while making statements using one of the validated emotional vocal prosodies. The ratings of the

robot's positive qualities and the creativity scores by the participant group that heard non-negative vocal prosodies (happiness, neutral) did not significantly differ from the ratings and scores of the participant group that heard the negative vocal prosodies (anger, fear, sadness). Therefore, Experiment 2 failed to show that the use of emotional vocal prosody in a robot's speech influenced the participants' appraisal of the robot or the participants' performance on this specific task.

At this time robot designers and programmers should not expect that vocal prosody alone will have a significant impact on the acceptability or the quality of human-robot interactions. Further research is required to show that multi-modal (vocal prosody along with facial expressions, body language, or linguistic content) expressions of emotions by robots will be effective at improving human-robot interactions.

DEDICATION

To Amy and Emma.

ACKNOWLEDGEMENTS

I would like to thank Dr. Cindy Bethel, my advisor and mentor, who has been most generous with her time for discussions and feedback concerning my research. The assistance of my committee members (Dr. Edward Swan II, Dr. Byron Williams, and Dr. Derek Anderson) while planning this research was invaluable.

I thank the Texas A&M Department of Computer Science and Engineering for the use of the Survivor Buddy robot in my experiments. I also thank my fellow STaRS lab members (both past and present) for their support and assistance.

Finally, I thank my best friend and wife, Amy Crumpton, for supporting all of my madcap dreams.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF TABLES vii

LIST OF FIGURES x

CHAPTER

1. INTRODUCTION 1

    1.1 Motivation 1
    1.2 Research Question 2

2. RELATED WORK 5

    2.1 Using Vocal Prosody to Communicate Emotion Between Humans 5
    2.2 Using Vocal Prosody to Communicate Emotion From Devices to Humans 10
    2.3 Using Vocal Prosody to Communicate Emotion From Robots to Humans 18
    2.4 Critique of Previous HRI Research on Robotic Vocal Prosody . . 21
        2.4.1 Embodiment 21
        2.4.2 Non-Linguistic Utterances 23
        2.4.3 Validation of Emotional Voice and its Perception 25

3. PRELIMINARY EMOTIONAL VOICE VALIDATION 26

    3.1 Experimental Design 26
    3.2 Apparatus 26
        3.2.1 Robot 27
        3.2.2 Speech Synthesizer 30
        3.2.3 Text 32
    3.3 Study Protocol 33

3.4 Voice Modification 34
3.4.1 Initial Experiment 35
3.4.2 Second Experiment 36
3.5 Results 38
3.5.1 Initial Experiment 38
3.5.2 Second Experiment 39
3.6 Discussion 41
3.6.1 Initial Experiment 41
3.6.2 Second Experiment 42
3.7 Conclusions and Future Work from Preliminary Experiments . . . 43

4. EXPERIMENT 1: VALIDATION OF EMOTIONAL VOICES 45

4.1 Experimental Design 45
4.2 Apparatus 46
4.2.1 Robot 46
4.2.2 Speech Synthesizer 47
4.2.3 Text 48
4.3 Surveys and Measures 48
4.3.1 Pre-Experiment Surveys 48
4.3.2 Experiment Measures 51
4.3.3 Post-Experiment Surveys 53
4.4 Study Protocol 53
4.5 Results 55
4.5.1 Free Choice of Emotion 55
4.5.2 Forced Choice of Emotion 57
4.5.3 Intelligibility of Emotional Robot Speech 59
4.5.4 Correlations 61
4.6 Discussion 76
4.6.1 Recognition of Emotion 76
4.6.2 Intelligibility of Emotional Robot Speech 78
4.6.3 Correlations 80
4.7 Conclusions and Future Work from Experiment 1 81

5. EXPERIMENT 2: EFFECTS OF USING EMOTIONAL ROBOTIC VOICES 83

5.1 Experimental Design 83
5.2 Apparatus 84
5.2.1 Robot 84
5.2.2 Speech Synthesizer 85
5.2.3 Text 86
5.3 Surveys and Measures 86
5.3.1 Pre-Experiment and Post-Experiment Surveys 86

5.3.2 Experiment Measures . . . 87
5.4 Study Protocol . . . 87
5.5 Results . . . 89
5.5.1 Robot Evaluation . . . 89
5.5.2 Participant Creativity . . . 98
5.5.3 Perceived Difference in Voice . . . 101
5.6 Discussion . . . 103
5.6.1 Robot Evaluation . . . 103
5.6.2 Participant Creativity . . . 104
5.7 Conclusion . . . 105

6. CONCLUSION . . . 107

6.1 Contributions . . . 107
6.2 Future Work . . . 109
6.3 Publication Plan . . . 110

REFERENCES . . . 112

APPENDIX

A. INFORMED CONSENT FORMS . . . 122

A.1 Preliminary Experiments . . . 123
A.2 Experiment 1 . . . 125
A.3 Experiment 2 . . . 126

B. ASSESSMENTS . . . 128

B.1 Preliminary Experiments . . . 129
B.2 Experiment 1 . . . 130

C. SURVEYS . . . 132

C.1 Demographics . . . 133
C.2 Mood . . . 135
C.3 Personality . . . 136
C.4 Robot Evaluation . . . 137

LIST OF TABLES

2.1 Emotions and Associated Vocal Prosody Characteristics 9

3.1 Changes Made to Standard Voice to Convey Emotions in Initial Preliminary Experiment 36

3.2 Changes Made to Standard Voice to Convey Emotions in Second Preliminary Experiment 37

3.3 Emotion Recognition Rates in Initial Experiment 39

3.4 Statistical Significance of Emotion Recognition Rates in Initial Experiment 39

3.5 Emotion Recognition Rates in Second Experiment 40

3.6 Statistical Significance of Emotion Recognition Rates in Second Experiment 40

4.1 Changes Made to Standard Voice to Convey Emotions in Experiment 1 . . 47

4.2 Big Five Inventory Scoring 50

4.3 Participant's Responses Categorized as Non-Emotion or Emotion 55

4.4 Recognition Rates for Free Choice of Emotions 56

4.5 Statistical Significance of Recognition Rates for Free Choice of Emotions . 57

4.6 Recognition Rates for Forced Choice of Emotions 57

4.7 Statistical Significance of Emotion Recognition Rates for Forced Choice of Emotions 58

4.8 Comparing for Free and Forced Choice Emotion Recognition Rates 59

4.9 Transcribed Words by Emotion 59

4.10 Terms in ω^2 for One-Way Repeated Measures ANOVA 60
4.11 Words Transcribed Incorrectly Most Often 61
4.12 Binary Participant Attributes (Experiment 1) 62
4.13 Emotion Recognition Rate by Gender 63
4.14 Significance of Emotion Recognition Rate by Gender 63
4.15 Emotion Recognition Rate by GPS Use 64
4.16 Significance of Emotion Recognition Rate by GPS Use 65
4.17 Emotion Recognition Rate by Digital Assistant Use 66
4.18 Significance of Emotion Recognition Rate by Digital Assistant Use 67
4.19 Multivalue Participant Attributes (Experiment 1) 68
4.20 Correlation of Computer Experience and Emotion Recognition Rate 69
4.21 Correlation of Robot Experience and Emotion Recognition Rate 70
4.22 Correlation of Video Game Experience and Emotion Recognition Rate . . . 70
4.23 Correlation of Positive Affect and Emotion Recognition Rate 71
4.24 Correlation of Negative Affect and Emotion Recognition Rate 72
4.25 Correlation of Extraversion and Emotion Recognition Rate 73
4.26 Correlation of Agreeableness and Emotion Recognition Rate 74
4.27 Correlation of Conscientiousness and Emotion Recognition Rate 75
4.28 Correlation of Neuroticism and Emotion Recognition Rate 75
4.29 Correlation of Openness and Emotion Recognition Rate 76
5.1 Robot Evaluation by Vocal Prosody Group 90
5.2 Significance of Robot Evaluation by Vocal Prosody Group 91

5.3 Terms in ω^2 for One-Way ANOVA 92

5.4 Results of ANOVA on Robot Qualities Grouped by Individual Vocal Prosodies 93

5.5 Pairwise Comparisons of Individual Vocal Prosodies for the Robot Quality "Happy" 97

5.6 Pairwise Comparisons of Individual Vocal Prosodies for the Robot Quality "Friendly" 97

5.7 Pairwise Comparisons of Individual Vocal Prosodies for the Robot Quality "Optimistic" 98

5.8 Creativity by Vocal Prosody Group 99

5.9 Significance of Creativity by Vocal Prosody Group 99

5.10 Results of ANOVA on Creativity Grouped by Individual Vocal Prosodies . 100

5.11 Pairwise Comparisons of Individual Vocal Prosodies for Creativity (Wooden Pencil) 101

5.12 Perceived Difference in Voice 102

LIST OF FIGURES

2.1 Sinusoid Waveform from the Recording of the First 0.015 Seconds of the Word "Hello." 6

2.2 Pitch Contour of the Statement "He is at the game." 8

2.3 Pitch Contour of the Question "He is at the game?" 8

2.4 Example of ToBI Markup 12

2.5 Screen Capture of Praat Showing ToBI Markup 14

2.6 Another Screen Capture of Praat Showing ToBI Markup 14

2.7 Specifying Vocal Prosody Using SSML 15

2.8 Pitch Contour Specified Using SSML. 16

2.9 Sample of EmotionML. 17

3.1 Image Displayed During Survivor Buddy Introduction 27

3.2 Survivor Buddy as a Stand-Alone Robot 28

3.3 Robot Operator GUI (Preliminary Experiments) 29

3.4 Image Displayed as a Face on the Survivor Buddy Monitor 30

3.5 Example of RAWMARYXML Expressing a "Sad" Vocal Prosody 32

3.6 Sentence Structures for Semantically Unpredictable Sentences 33

3.7 Emotion Choices Presented to the Participants (Preliminary Experiments) . 34

4.1 Emotion Choices Presented to the Participants (Experiment 1) 51

5.1 Robot Operator GUI (Experiment 2) 85

CHAPTER 1

INTRODUCTION

This chapter[1] presents the motivation for this line of research, the overarching research question, and the specific hypotheses that were tested.

1.1 Motivation

Robotic systems are increasingly being studied for use in social situations. Roles such as companions, tutors, and caregivers are being investigated as possible uses of robots. To reduce training needs for robot users and the robots themselves, interactions between robots and their users should be as natural as possible [14]. Given that speech is the one of the most natural ways for humans to communicate, communication between humans and robots using voice is an area that has received much attention [16, 35, 43, 56, 65, 73, 71, 78]. In situations where humans and multiple robots are working together, even communication between robots would ideally be via voice as opposed to some form of electronic networking so that the humans can understand what is being communicated between the robots [65].

One of the areas of robotic speech, and synthesized speech in general, that can be improved to make generated speech sound more authentic is the use of vocal prosody. Vocal

[1]The content of this chapter was submitted to *International Journal of Social Robotics* in a survey article on the use of vocal prosody to convey emotion in robotic speech.

prosody is the way in which words are spoken (pitch, timing, loudness, etc.) as opposed to the actual linguistic meanings of these words [36]. In human communication, vocal prosody is considered one of the paralinguistic components of speech [59]. Paralinguistic in this context refers to the features of communication that appear along side (*para* is a prefix from Greek meaning "side by side") the actual words being communicated. Other paralinguistic components of speech include voice quality, non-word utterances, pronunciation, and enunciation [59].

Vocal prosody is an essential component of speech communication between humans [88]. It has long been recognized a speaker's vocal prosody is one of the ways the emotional state of the speaker is communicated to listeners [37, 28, 80]. The emotions or mood being conveyed by speech can be crucial to interpreting the meaning of a speaker's message. For example, a tone of sarcasm can be used by a speaker to signal the listener that the message being delivered should be interpreted as literally the opposite of the actual words being spoken [103]. In addition to short term states of the speaker such as emotions or moods, vocal prosody can also be used by a listener to infer traits of the speaker such as gender and personality [59, 88].

1.2 Research Question

Does the use of vocal prosody to communicate emotions in robotic speech increase the effectiveness of a robot's spoken communication?

The inspiration for these experiments was the work of Bainbridge et al. on robot embodiment [6]. Bainbridge et al. found that study participants followed a robot's instructions

more closely than an on-screen avatar's instructions. The difference in how participants responded to the robot and the on-screen avatar demonstrated that the robot's embodiment is an important characteristic of the robot. This research intended to provide similar proof that a robot that uses vocal prosody to vary its speech and communicate emotions will be more engaging than a robot that does not utilize varying or intentional vocal prosody. In addition to surveying study participants about the acceptability of the robot and its speech, a concrete measure (the robot user's performance on a creative task) was used to judge the impact of the robot using emotional speech.

Experiment 1 (see Chapter 4) validated that the vocal prosody modifications determined by preliminary experiments (see Chapter 3) communicated the intended emotions to the robot user. Also, the intelligibility of the robot speech in view of the vocal prosody modifications made to convey emotions was investigated. The specific hypotheses used for Experiment 1 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.

Experiment 2 (see Chapter 5) intended to show that the robot's use of emotional vocal prosodies would improve the robot user's appraisal of the robot and improve objective

measures of the human-robot interaction. The specific hypotheses used for Experiment 2 were:

H_3: The robot will be rated higher on positive attributes (attractive, happy, friendly, intelligent, cooperative, etc.) by participants who hear the non-negative voices (neutral, happiness) than by participants who hear the negative voices (fear, anger, sadness).

H_4: Participants who hear the non-negative voices (neutral, happiness) will perform better on the creativity test than participants who hear the negative voices (fear, anger, sadness).

CHAPTER 2

RELATED WORK

An important aspect of using vocal prosody in synthetic speech is to understand and summarize the features that individuals use to communicate emotional intent to listeners. In order for vocal prosody to be incorporated into synthetic speech it is essential to investigate various markup languages that can be utilized by different speech synthesizers to produce changes in the vocal prosody characteristics within synthesized speech. This chapter[1] also surveys research that discusses the importance of the use of vocal prosody within synthesized speech. This chapter ends with a presentation of previous human-robot interaction (HRI) research concerning the use of vocal prosody by robots and a critique of that previous research.

2.1 Using Vocal Prosody to Communicate Emotion Between Humans

Before robots and other electronic devices with speech synthesizers can use vocal prosody to convey emotions, the features of speech that are used by humans to convey emotion must be identified and quantified. Early research consisted of listeners detecting and classifying the features of speech that accompanied emotional speech [28, 37, 80]. More recent research has applied statistical and machine learning techniques such as the Fuzzy

[1]The content of this chapter was submitted to *International Journal of Social Robotics* in a survey article on the use of vocal prosody to convey emotion in robotic speech.

Logical Model of Perception, Linear Discriminant Analysis, Support Vector Machines, and AdaBoosted Decision Trees to determine which features can be used to classify emotional speech [44, 48, 49, 58]. Typically, the features of speech that are found to correlate with the expression of emotion are pitch, timing, and loudness which are sometimes referred to as *The Big Three* of vocal prosody [98].

Pitch corresponds to the frequency at which the vocal folds of the person speaking vibrate [30]. The human voice is not a simple signal consisting of one sinusoid. The shape of the vocal tract reinforces some frequencies and dampens other frequencies. The result is a complex sinusoid as shown in Figure 2.1. Note that Figure 2.1 only shows the sinusoid for the first 0.015 seconds of the word spoken. Typically the wave form is shown for an entire word or statement (as in the top half of Figure 2.2) and the wave is too compressed to

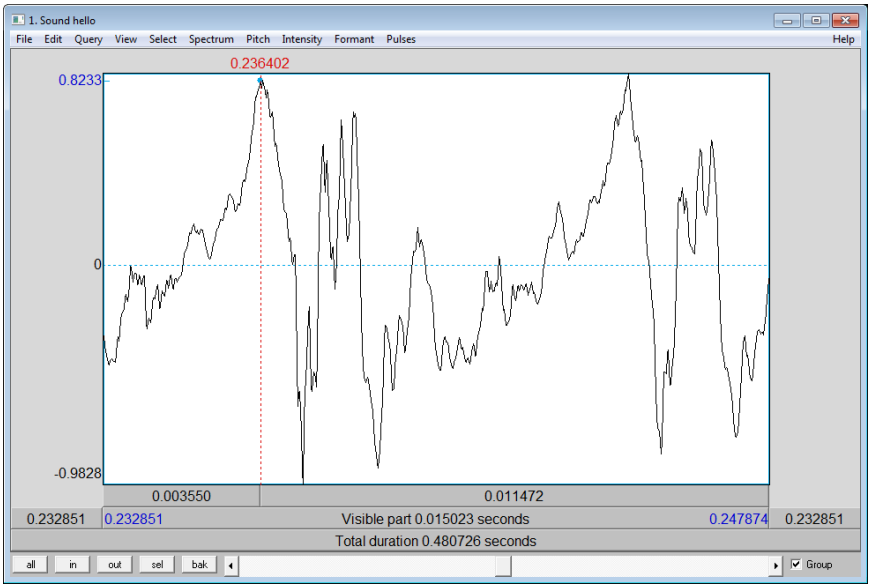


Figure 2.1

Sinusoid Waveform from the Recording of the First 0.015 Seconds of the Word “Hello.”

see the complexity and periodicity of the wave. The most prominent frequency is referred to as the *fundamental frequency* or *F0*. The other frequencies that are emphasized by the vocal tract are called *formants* [30]. A higher than normal fundamental frequency can indicate happiness and lower than normal F0 can indicate that the speaker is sad [37].

Not only is the fundamental frequency of a speech segment important, both the range of frequencies and the change in fundamental frequency during a speech segment can affect a listener's assessment of the speaker. The *pitch range* is the difference between the highest frequency and the lowest frequency during an utterance. A small pitch range usually indicates sadness while an expansive pitch range indicates happiness or perhaps anger [82]. The change in F0 during a speech segment is referred to as the *pitch contour*. The pitch contour can be critical to the meaning of an utterance. American English speakers can change a statement such as "He is at the game" to a question by raising the pitch of their voice at the end of the speech segment [103]. In Figure 2.2 and Figure 2.3 the wave form of the recorded speech is shown in the top half of the figure and the pitch contour is shown in the bottom half of the figure. For a declarative statement the speaker's pitch usually falls at the end of the statement as shown in Figure 2.2. A question is often accompanied by a rise in the pitch contour as shown in Figure 2.3.

Timing is concerned both with how fast a person is speaking and the pauses within a statement. The speed of a person speaking is typically measured in words per second while the pauses can be measured in seconds. Rapid speech can indicate the speaker is happy or angry. A slow rate of speech typically indicates the speaker is sad [37].

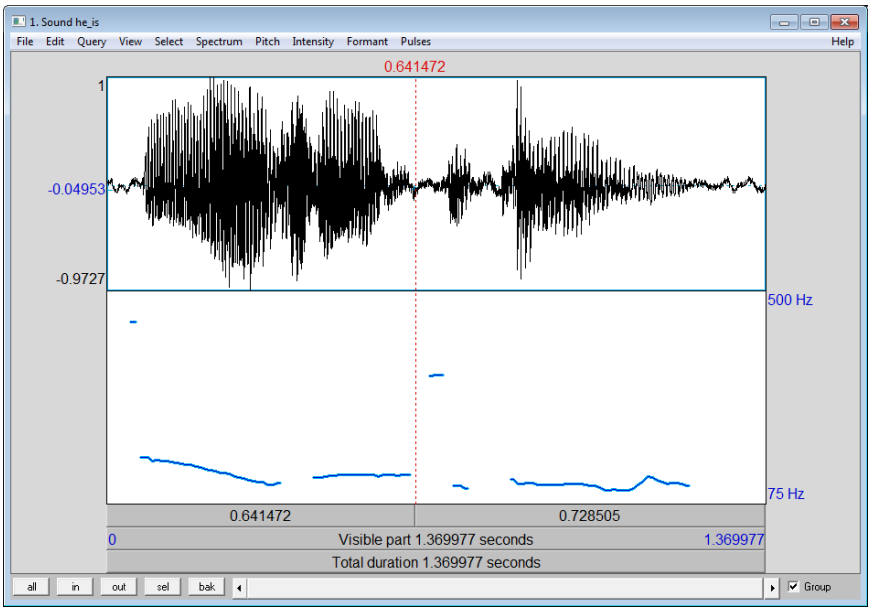


Figure 2.2

Pitch Contour of the Statement “He is at the game.”

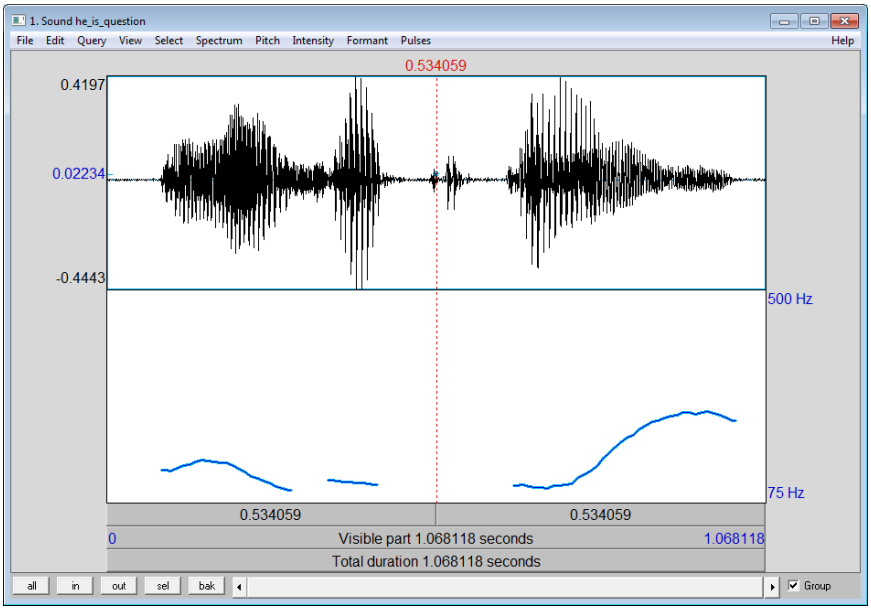


Figure 2.3

Pitch Contour of the Question “He is at the game?”

Loudness is a measure of the volume at which a person is speaking. Loudness is also referred to as *intensity* in some literature [17, 30, 44]. Loudness is typically measured in decibels (dB), a logarithmic unit of measure that gives the ratio between two values. A soft voice, a low value for loudness, can indicate boredom while a loud voice indicates emotions such as happiness or anger [80].

As is evident from the previous descriptions of pitch, timing, and loudness, a particular emotion can affect one or more of the measures simultaneously. Table 2.1 [33, 91, 93] gives a summary of the changes in vocal prosody that usually accompanies the expression of Ekman's *Big 6* basic emotions: *happiness*, *surprise*, *sadness*, *anger*, *disgust*, and *fear* [26, 66].

Table 2.1

Emotions and Associated Vocal Prosody Characteristics

Emotion	**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

Experiments have shown that people can recognize the emotion being communicated by another person's vocal prosody at a level much higher than chance [62, 82]. Scherer et al. surveyed twenty-seven previous studies of emotion recognition and reported that participants were able to identify the emotion that was meant to be conveyed about 60

percent of the time [82]. The percent correct expected from guessing varied between 10 and 25 percent based on the number of emotions that were being used in each of the individual experiments. Criticism of these experiments point out that the listeners are distinguishing between a set of listed emotions and not identifying the intended emotions with the listener's own choice of words [31]. Another criticism is the use of actors to provide speech samples instead of using recordings of more natural speech from common interactions [44]. It is less clear that a high rate of identification is possible from the spontaneous speech of non-actors recorded in more life-like situations [20]. Cowie points out that even among sound recordings chosen from TV interviews and talk shows based on their appearance of emotional content, only 34 percent of the clips were labeled by listeners as containing strong emotions [20].

2.2 Using Vocal Prosody to Communicate Emotion From Devices to Humans

The generation of affective speech using the manipulation of vocal prosody features has been a subject of speech synthesis research for over twenty years [18, 54, 38, 86, 27, 61]. Early work was hindered by the lack of capabilities of then state-of-the-art speech synthesizers to allow changes to features of the generated speech such as pitch range and pitch contour [18]. As speech synthesizers have become more advanced, the ability to convey emotion within generated speech has improved. There are now speech synthesizers, such as MARY (Modular Architecture for Research on speech sYnthesis) [87], that were designed with the generation of expressive speech as a goal. Several commercially avail-

able speech synthesizers such as Acapela Group's Acapela[2] and Cereproc's CereVoice[3] text-to-speech engines contain voices claimed to portray different emotions. However, the companies do not provide empirical evidence showing that listeners actually perceive the emotion claimed to be portrayed by the generated speech.

One of the major difficulties to overcome in the Text-to-Speech (TTS) field concerning the use of vocal prosody to convey emotions is natural language understanding. Recognition of the intended emotion from just the content of text can be a difficult problem. Early efforts for the prediction of an appropriate emotion from text focused on the identification of keywords or the use of hand-written rules to analyze text [47]. Given the increase in inexpensive computing power and the decrease in the cost of digital storage, recent research has employed machine learning techniques on a large text corpora to generate rules for the prediction of the intended emotional content of text [1, 47, 92].

Once the text to be spoken has been analyzed for emotional content, the text must be marked up with enough prosody information that the speech synthesizer can manipulate the generated speech to convey the intended emotions. There have been several efforts to create standardized markup languages that can be used to annotate text with information about how the speech synthesizer should "say" the text in order to produce more natural sounding speech.

Tone and Break Indices (ToBI) was an early standard used to mark up text with vocal prosody attributes [90]. ToBI transcriptions consist of the text being spoken along with

[2]http://www.acapela-group.com/

[3]https://www.cereproc.com/

markup in specific tiers that convey prosodic information. The most commonly used tiers are the tone tier and the break-index tier. Figure 2.4 [9] shows an example of the text of a question along with the tone and break-index tiers.

```
Will you have marmalade, or jam?
               L*       H-     L* H-H%
    1   1    1           3  1   4
```

Figure 2.4

Example of ToBI Markup

The tone tier contains information about the tone contour of the speech using symbols for low (L) and high (H) tones or frequencies. The * symbol along with a tone symbol (H or L) indicates that a syllable receives more stress than the surrounding parts of the word or phrase. The first L* in Figure 2.4 shows that the first syllable in marmalade was said with emphasis in a low tone. The – symbol along with a tone symbol marks the tone target of a phrase as opposed to a single accented syllable. The H– in Figure 2.4 shows that the phrase ends at a higher pitch than the previous accented syllable at the beginning of the word marmalade. The % symbol along with a tone symbol marks the tone target at the end of a phrase where a pause in speech occurs. The L% or H% at the end of a phrase are often combined with a L– or H– symbol. The first half of the symbol (L– or H–) represents the tone target of the phrase and the last half of the symbol represents the tone target of

the very end of the phrase. For example, the H–H% shown in Figure 2.4 indicates that the phrase *or jam* has a high phrase accent and a high tone at the ending boundary.

The numbers in the break-index tier are a scale from 0 to 4 that represent the different types of pauses within the spoken text. The typical pause between spoken words is represented by a 1 and the pause between distinct segments of speech is represented by 4. The 3 shown in Figure 2.4 corresponds to the pause that the comma represents in the question. Note that phrase accents (L– or L*) typically occur at breaks labeled with a 3. The breaks between phrases (labeled with a 4) often have the tone markup representing the phrase accent and the tone at the ending boundary such as H–H% or L–H%.

Praat is a software system that can display the waveform and pitch contour of a speech segment so that the annotator can see both the pauses in speech and movement of F0 within the pitch contour [13]. Praat also has tools for labeling the tone and break-index tiers of a speech segment. Figure 2.5 and Figure 2.6 are two examples of a speech segment and the corresponding ToBI markup displayed by Praat [97]. More information about ToBI and Praat can be found at the "Transcribing Prosodic Structure of Spoken Utterances with ToBI" course at Massachusetts Institute of Technology's OpenCourseWare website [97].

Effective speech synthesis is an important part of making the world wide web accessible to people with impaired vision. The World Wide Web Consortium (W3C) recommends the use of a text markup system for speech synthesizers that includes elements intended to affect the vocal prosody of synthesized speech. The Speech Synthesis Markup Language (SSML) [101, 99, 4] is a standard that contains elements that direct the TTS system to produce speech that will be interpreted by listeners as being from a person of a specific

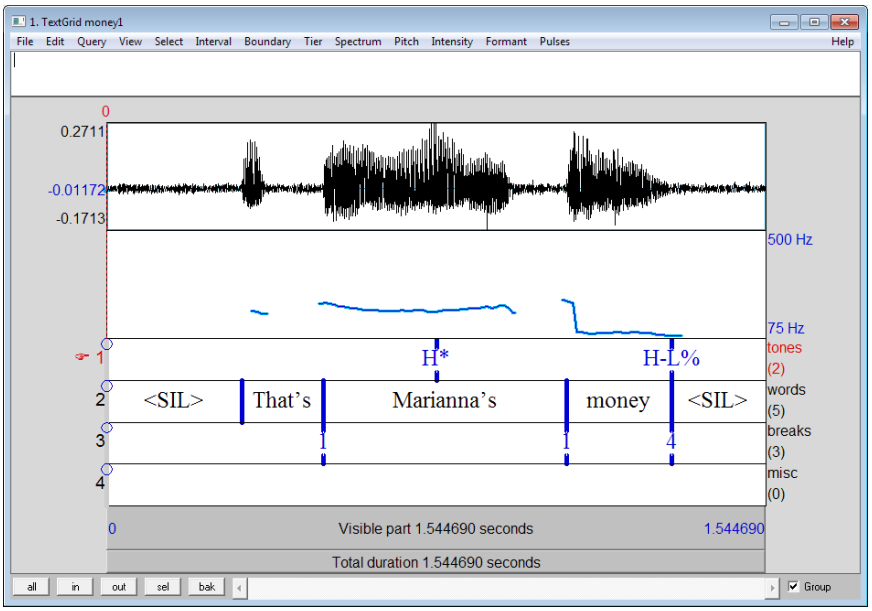


Figure 2.5

Screen Capture of Praat Showing ToBI Markup

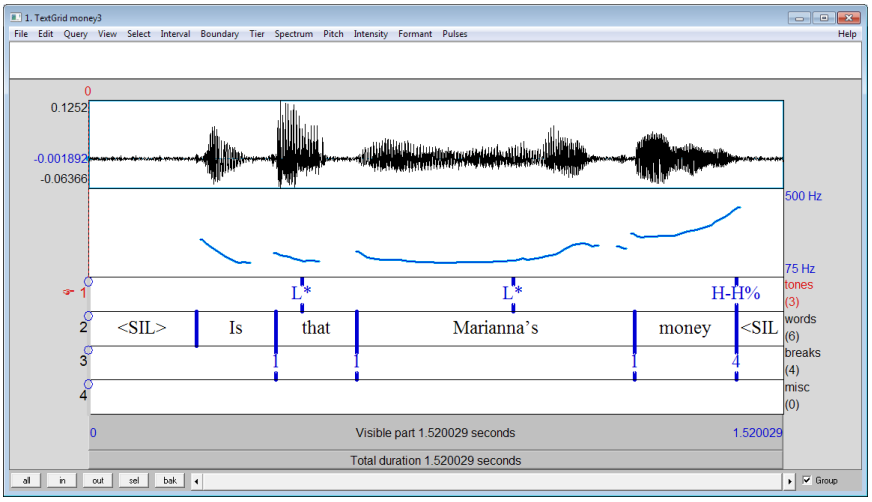


Figure 2.6

Another Screen Capture of Praat Showing ToBI Markup

gender, age, etc. Figure 2.7 shows the text “Are we there yet?” marked up using SSML so that the synthesized speech should sound like it was spoken by a young boy. Note that the speech synthesizer is responsible for the choice and implementation of the vocal prosody to meet the directives contained in the SSML markup.

```
<?xml version="1.0"?>
<speak version="1.1"
       xmlns="http://www.w3.org/2001/10/synthesis"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
             http://www.w3.org/TR/speech-synthesis11/synthesis.xsd"
       xml:lang="en-US">
    <voice gender="male" languages="en-US" age="7">
      Are we there yet?
    </voice>
</speak>
```

Figure 2.7

Specifying Vocal Prosody Using SSML

If more control over the voice output is desired, SSML also includes a `prosody` element that can specify options such as pitch, pitch contour, range, and speech rate. Figure 2.8 illustrates the SSML markup that specifies that a question should be asked at a greater than normal speed with a specific pitch contour. Currently several commercial speech synthesizers such as such as Microsoft’s .NET speech synthesizer and Nuance Communication’s Dragon Mobile claim to support SSML 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 [51, 57].

```
<?xml version="1.0"?>
<speak version="1.1"
       xmlns="http://www.w3.org/2001/10/synthesis"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
                http://www.w3.org/TR/speech-synthesis11/synthesis.xsd"
       xml:lang="en-US">
  <voice gender="male" languages="en-US" age="7">
    <prosody rate="fast" contour="(0%,+10Hz)(30%,+20Hz)(60%,+10Hz)">
      Are we there yet?
    </prosody>
  </voice>
</speak>
```

Figure 2.8

Pitch Contour Specified Using SSML.

EmotionML is another markup language proposed by the W3C to direct the expression of emotion [5]. Whereas SSML is meant to guide the output of speech synthesizers, EmotionML is also meant as the input to on-screen avatars, robots, and other electronic devices [85]. In the cases of on-screen avatars and robots, the emotions specified by EmotionML may be expressed by facial expressions or body language in addition to changes in vocal prosody. EmotionML allows for emotions to be specified using names from lists of emotions such as Ekman's *Big 6* basic emotions or using values for dimension-based definitions of emotions such as Mehrabian's Pleasure, Arousal, and Dominance model [26, 50, 3]. Note that the device that is using EmotionML for its input is responsible for translating the emotion names or dimension values into actual changes in vocal prosody, facial expressions, or body language to express the emotion. This is similar to the how the W3C's earlier SSML standard was used when specifying the speaker's age and gender. If

a speech synthesizer implements EmotionML, the synthesizer's user is no longer required to manually translate the desired emotion into changes to the SSML's `prosody` element for pitch, pitch contour, and speech rate. Figure 2.9 is an example of EmotionML markup. The question *What was that sound?* should be produced by the device in a manner meant to communicate fear. The value attribute of a named emotion is a floating point number in the closed interval [0.0, 1.0] that describes the "strength" of the emotion. The value 0.0 represents no emotion and 1.0 represents "pure uncontrolled emotion" [5].

```
<sentence id="sentence1">
   What was that sound?
</sentence>
<emotion xmlns="http://www.w3.org/2009/10/emotionml"
         category-set="http://www.w3.org/TR/emotion-voc/xml#big6">
   <category name="afraid" value="0.6"/>
   <reference role="expressedBy" uri="#sentence1"/>
</emotion>
```

Figure 2.9

Sample of EmotionML.

Since there is not just one standard for marking up text with vocal prosody information, speech synthesizers often support more than one of the standards. For example, the MARY speech synthesizer will accept GToBI (a ToBI variant for German language), SSML, and EmotionML markup elements as its input [84].

The importance of vocal prosody in synthesized speech has been shown in several experiments. Nass et al. [55] showed that an automobile driver's performance was influenced

by the emotions conveyed by a virtual passenger's speech. When the emotion conveyed by the virtual passenger's speech matched the emotion of the driver, the driver paid more attention to the road and the driver was involved in fewer accidents. D'Mello and Graesser have integrated speech feedback that contains varying vocal prosody into their Affective AutoTutor, a new version of their intelligent tutoring system AutoTutor [25]. The Affective AutoTutor system detects a student's emotion using multimodal techniques that include dialog cues, the student's posture, and the student's facial movements as inputs. The tutoring system then constructs its feedback in order to give encouragement to students that are displaying positive emotions and to reduce the continuation of negative emotions in struggling students. The system communicates affect via the facial expressions of the tutor avatar, the linguistic content of the tutor's speech, and the vocal prosody of the synthesized voice. Students, especially students with low domain knowledge, showed more learning gains when using the Affective AutoTutor system as opposed to the older AutoTutor system that did not attempt to communicate affect [25].

2.3 Using Vocal Prosody to Communicate Emotion From Robots to Humans

While the concept of robots conveying emotion through their speech might seem nonsensical given that robots do not actually feel emotions, there are several benefits from the use of vocal prosody within robotic speech. First, previous research has shown that people prefer to communicate with robots via voice and they prefer that the voice be human-like [24, 41, 70]. Second, taking advantage of the ability of humans to perceive emotions in speech might increase the effectiveness of robotic speech communication. For example, a

robot team member in an urban search and rescue situation could use its vocal prosody to convey the seriousness of a warning by sounding excited when speaking or use a calming voice to reduce the anxiety levels of a survivor once located.

There has been support for robots expressing emotion and intention in order for the robots to be perceived as "believable characters" by humans [7, 15]. Researchers have designed systems to calculate and express the emotional state of the robot in response to its environment and interactions [14, 69, 95]. For example, the Kismet robot would lower its head and/or frown when receiving negative feedback from a study participant [14]. One of the main roadblocks to the credible use of vocal prosody in text-to-speech applications has been the difficulty of determining the correct emotions to express from the content of the text. This is a case where a robotic system that has computed an appropriate emotion has an advantage. The robot has already determined its emotional state and can then use its vocal prosody to express that state.

Most of the work in the HRI field has concentrated on conveying emotions by varying the vocal prosody of simple non-linguistic utterances. Read and Belpaeme [72] found that people interpret human-like utterances made by robots as expressing emotions. Oudeyer [58] created algorithms that can modify child-like "babble" to convey emotions. Human listeners of several nationalities were able to successfully determine the communication intent of the utterances produced by Oudeyer's system. The study of non-linguistic utterances has been justified by pointing out that generating the non-linguistic sounds is computationally inexpensive [58] and the utterances should be understandable across cultures and languages [71].

As expressive speech synthesizers became more readily available, several proposals for their use in robotic systems were made [77, 104]. There has been little research into how manipulating a robot's voice would affect its users however. One study has shown that a robot learner that expresses emotion through its statements and voice causes people to provide the robot with more and better training data [45]. Leyzberg et al. asked participants to train a small robot in some simple dances. The robot would receive a score supposedly based on how well the robot performed a dance. When the robot responded to its score with appropriate emotional statements expressed through recorded speech, the participants provided more examples of the dance moves to the robot. If the robot made apathetic statements or inappropriate responses (excited by low scores or upset by high scores), the human trainer provided significantly fewer dance examples for the robot. Tielman et al. found that children showed more emotions when interacting with a robot that showed emotions through its body language and voice than a robot that did not display emotions [95]. They did note their changes to the vocal prosody of the robot's speech resulted in making the robot more difficult to understand.

Recent research has shown that people's impressions of a robot can be influenced by the pitch of the robot's voice [56]. Niculescu et al. manipulated the average fundamental frequency of a robot receptionist's voice to determine if participants would find the robot with the higher voice more attractive and more outgoing than the same robot with a lower voice [56]. The same robot was used for both high and low pitched voice conditions and it was dressed as a female in both conditions. Not only did the participants rate the robot with the higher voice as having a more attractive voice, being more aesthetically appealing, and

more outgoing, the participants responded that the robot with the higher voice exhibited better social skills. These results were expected given that both men and women find women with higher voices more attractive and ascribe more positive personality traits to women with higher voices [56].

2.4 Critique of Previous HRI Research on Robotic Vocal Prosody

This section contains a critique of previous experiments concerning the use of emotional voices by robots. First, evidence of the importance of using collocated robots in experiments is presented. Next, the use of Non-Linguistic Utterances by robots to communicate with their users is discussed. Finally, the section concludes with a discussion concerning the validation of the changes made to the vocal prosody of synthetic speech to express emotion.

2.4.1 Embodiment

A common trend in research concerning vocal prosody and robots is the use of pictures or on-screen avatars instead of actual physical robots [53, 72]. Some of the research was conducted without the mention of robots to participants, but the results were used to make recommendations about robotic voices [35]. While the use of avatars instead of physical robots is understandable in light of the lower cost of on-screen avatars and relative ease of programming as opposed to robots, it is not universally accepted that people react to images on a screen and collocated physical robots in the same way. Research has found that the use of simulated images or avatars definitely results in participants viewing the "robot" differently than collocated physical robots [40, 42, 64]. For example, Kidd and

Breazeal state that “the robot was more engaging and rated more highly on the scales of perceptions than the animated character” [42].

What is less clear is the comparison of collocated robots and remote robots that are seen and heard via video recordings or video conferencing. Early research reported that participants did not react differently to collocated and remote robots [40, 42, 64]. Kidd and Breazeal reported “it is not the presence of the robot that makes a difference, rather it is the fact that the robot is a real, physical thing, as opposed to the fictional animated character in the screen” [42]. In contrast, Bainbridge et al. [6] found that study participants obeyed unusual requests (placing new books in a trashcan) made by a collocated robot more often than unusual requests made by the same robot displayed on a monitor. The participants also respected the personal space of a collocated robot more than that of a robot presented on a monitor [6]. One explanation for the differences observed between these studies might be the difficulty of the task being performed by the study participants. The earlier experiments involved relatively simple tasks such as placing blocks on top of each other or interacting with a robotic dog [42, 40]. In the Bainbridge et al. experiment, the participants received instructions from the robot concerning moving books from one location to another location within a small office [6]. Wainer et al. [100] found that participants preferred a collocated robot over remote robots and over a robot avatar when the robot was acting as a coach while the participant solved a Towers of Hanoi puzzle. Leyzberg et al. found that participants tutored by a collocated robot became better puzzle solvers than participants who received the equivalent tutoring from a video representation of the robot [46]. One can imagine that people prefer a collocated robot when performing complex tasks that

require many interactions with the robot. While the differences in reactions to collocated and remote robots are being further studied, using collocated robots in experiments would be a wise choice especially when making recommendations about long-term human-robot interaction.

2.4.2 Non-Linguistic Utterances

Another trend in recent research on the use of vocal prosody in robotic speech is the use of non-linguistic utterances by the robot instead of speech consisting of words and speech. As mentioned earlier, the study of non-linguistic utterances has been justified by pointing out that generating the non-linguistic sounds is computationally inexpensive [58] and the utterances should be understandable across cultures and languages [71]. The use of non-linguistic utterances was inspired by characters such as R2-D2 and WALL-E from motion pictures [11, 71]. It is seemingly accepted that the emotional intent of the characters' utterances are interpreted correctly by other characters in the movie and by audience members. The interpretation of the communication by other characters is a non-issue, the other characters' reactions are scripted and do not require interpreting the sounds. The interpretation of the communicative intent by the audience is aided by the reaction of the other characters and the other non-verbal cues such as body language and facial expressions. It would be interesting to see how much of the intent is communicated by the utterances and how much is inferred from other cues.

Read has shown that children do assign emotional meanings to the non-linguistic utterances of a robot [71]. The children do not always agree on which emotion is expressed

by each sound. In a more recent experiment, Read and Belpaeme have shown that adults also categorize non-linguistic utterances with relation to affective content, especially when two utterances are compared [74]. In both of the mentioned experiments the participants did not differentiate between subtle changes in the level of emotion being communicated, the utterances were categorized without any acknowledgment to the degree of the emotion. Read and Belpaeme's most recent work shows that the interpretation of non-linguistic utterances is heavily influenced by what action a robot experiences [76]. For example, a sound that was previously rated as communicating a positive valence by participants was rated as communicating a negative valence if the sound was produced by the robot in response to the robot receiving a slap [76].

Even with these successful demonstrations of the limited interpretation of non-linguistic utterances by study participants, Read states that the use of non-linguistic utterances have "obvious shortcomings in comparison to natural spoken language" [71]. The amount of information that can be communicated by the non-linguistic utterances is obviously limited. In the Star Wars movies, on-screen characters may understand the exact meaning of R2-D2's chirps but the audience is limited to hearing the on-screen characters repeat the message in words before knowing the meaning. Communicating a detailed message such as "the network is down but is expected to be online in 15 minutes" via chirps and buzzes would be difficult using only non-linguistic utterances. If humans and robots are expected to work together to share information and accomplish tasks, both the humans and robots will need to use a communication medium that is able to express sometimes complicated messages. Read and Belpaeme advocate for the use of non-linguistic utterances in addi-

tion to natural language as opposed to non-linguistic utterances replacing the use of natural language by robots [75].

2.4.3 Validation of Emotional Voice and its Perception

A potential weakness to avoid in the study of emotional voices expressed by robots is the validation of the emotional voices. While the correlates of vocal prosody parameters and the intended emotion are known (see Table 2.1), researchers should verify that the manipulation of vocal prosody correctly communicates the intended emotion. Differentiating a large number of emotions based on pitch, pitch contour, volume, and timing can be challenging. Tielman et al. used arousal and valence parameters to modify a robot's speech while the robot was interacting with children [95]. But the researchers did not check that the children could correctly interpret the emotional intent of the robot's speech. Beale and Creed made a similar criticism of research on the use of emotion by agents and on-screen characters [8].

Study participants are adept at recognizing emotional intent through several modalities [2]. If a robot is using an emotional voice and also expressing its emotions through the literal meaning of its statements, its facial expression, or body language then the effect of the voice is entangled with the effects of the other modalities. While emotions are typically communicated through many modalities at once, research that makes specific claims about the communication of emotion via vocal prosody must attempt to isolate the effect of vocal prosody from the other ways to communicate emotion.

CHAPTER 3

PRELIMINARY EMOTIONAL VOICE VALIDATION

This chapter[1] presents two experiments conducted as pilot studies for the creation and validation of emotional robotic voices. The experimental design, apparatus, survey materials, and tasks performed were the same for both experiments.

3.1 Experimental Design

These experiments were within-subjects designs that evaluated the detection of emotion in robot speech. The hypothesis was:

> 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).

3.2 Apparatus

This section describes the apparatus used in the preliminary experiments: the robot, the speech synthesizer, and the text said by the robot during the experiment. Only the vocal prosody modifications made to express the five emotions (anger, calm, fear, happiness, and sadness) differed between the two preliminary experiments (see Section 3.4 for details).

[1]The content of this chapter appears in [21]

3.2.1 Robot

The Survivor Buddy robot was used for the robot interactions with participants in these experiments. The Survivor Buddy robot was developed at Texas A&M University and is usually mounted to a mobile robotic platform (see Figure 3.1) [34]. The robot was designed to aid in research that investigates how a robot can be used to communicate with and comfort disaster survivors. For this study the Survivor Buddy robot was not mounted to a mobile base, rather it was placed on a table facing the study participant (see Figure 3.2). The Survivor Buddy robot consists of a small monitor manufactured by Mimo Monitors, Inc.[2] mounted to the end of an arm. The arm contains four Robotis Dynamixel[3] actuators. One actuator raises and lowers the arm while the remaining three actuators allow the monitor to raise and lower, turn to the left and right, and tilt to the left and right.

Figure 3.1

Image Displayed During Survivor Buddy Introduction

[2]http://www.mimomonitors.com/

[3]http://www.robotis.com/xe/dynamixel_en/

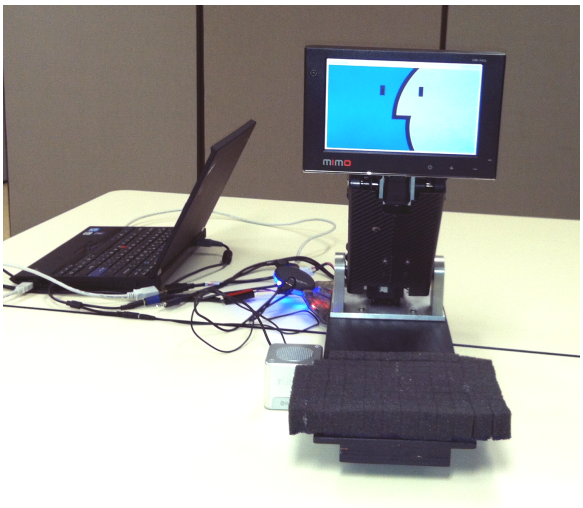


Figure 3.2

Survivor Buddy as a Stand-Alone Robot

The experiments were conducted using the *Wizard-of-Oz* technique [23]. The sound and video from the robot's microphone and camera were streamed to the robot operator's PC which was located in another room. The robot operator could use pre-programmed functionality to perform routine actions such as raising the robot's monitor from a resting position and having the robot give instructions to the participant (see Figure 3.3). Less routine tasks such as turning the robot's head to face a participant or asking the participant to speak more loudly were also possible to accomplish with the manual controls available to the robot operator through the graphical user interface.

To avoid the implication of emotion from the robot's "face", static images were shown on the Survivor Buddy's monitor. For most of the experiment an image derived from Apple's Finder icon was used as the Survivor Buddy's face. As Figure 3.4 illustrates, the smile was removed to avoid a bias toward "happy" emotions. An image (shown in Figure 3.1) of

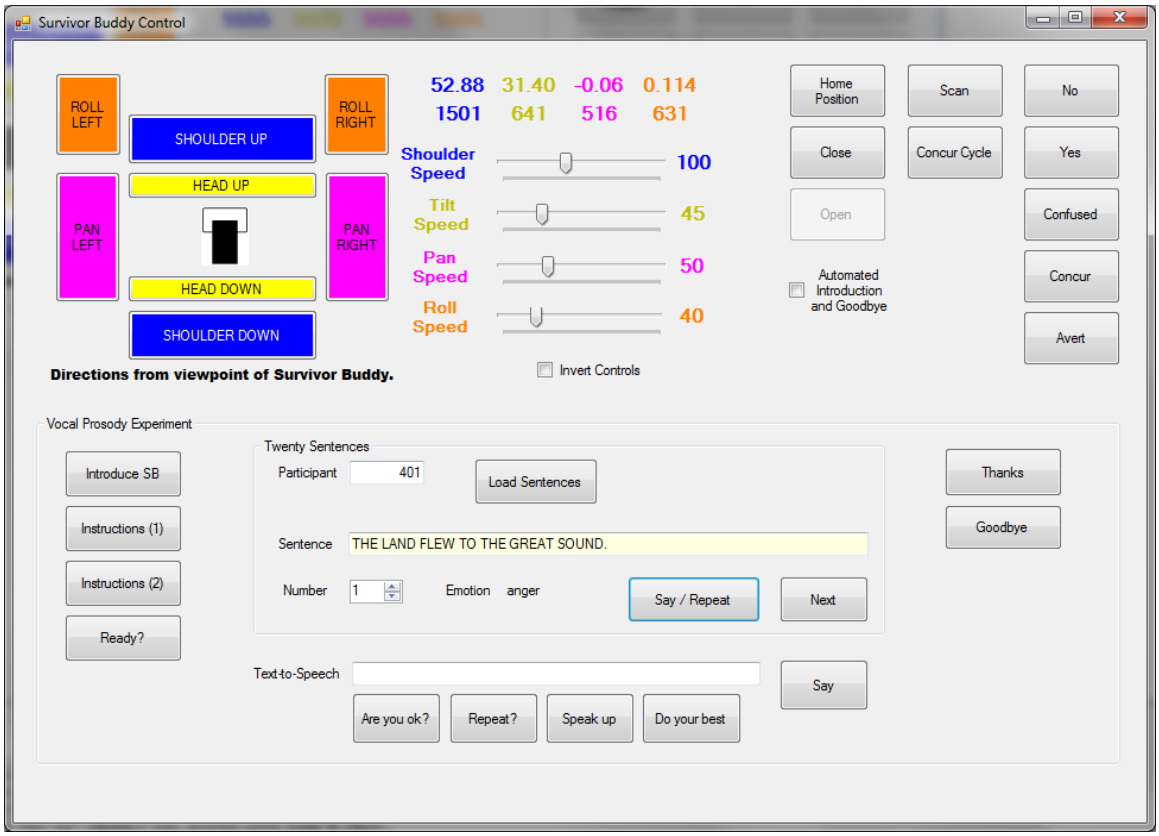


Figure 3.3

Robot Operator GUI (Preliminary Experiments)

the Survivor Buddy mounted to a mobile robot base was shown to the participants while the Survivor Buddy robot introduced itself to the participants. During its introduction, the robot explained that it was meant to be used to communicate with people trapped by rubble at disaster scenes. One robot operator noted that after seeing the image of the Survivor Buddy robot in the rubble the participants appeared much more interested in the robot.

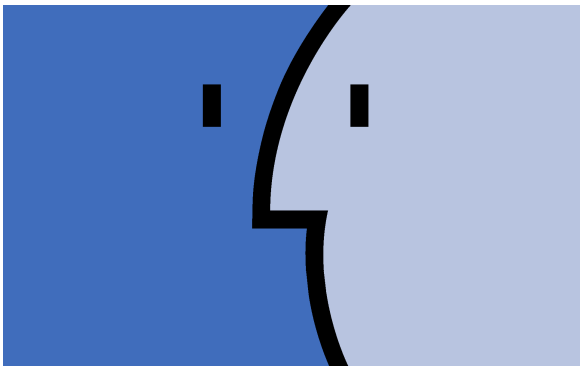


Figure 3.4

Image Displayed as a Face on the Survivor Buddy Monitor

3.2.2 Speech Synthesizer

The original software supplied with the Survivor Buddy robot uses Microsoft's text-to-speech system to produce speech. For these experiments the MARY (Modular Architecture for Research on speech sYnthesis) speech synthesizer was used. MARY was designed with the generation of expressive speech as a goal [87]. While MARY supports many input formats such as ToBI and EmotionML (see Section 2.2) to specify vocal prosody modifications or what emotion to convey in its speech, voice model training issues prohibited their use for this research. As reported in [21], informal listening tests found that the out-

put of the MARY speech synthesizer did not appear to change the pitch of statements in response to the ToBI markup. The Praat software system [13] was used to verify that the ToBI commands were not affecting the generated speech. After much investigation, the root cause was found. The `slt` voice data from the Language Technologies Institute at Carnegie Mellon University [12] contained sentences without ending punctuation. Therefore, the voice model was never trained to respond to frequency changes such as the rising pitch at the end of a question. The resulting voice model did not contain the information needed to respond correctly to the ToBI markup language.

RAWMARYXML is one way the vocal prosody modifications necessary to convey emotion were specified in this experiment. RAWMARYXML features tags to markup changes in the synthesized speech's speech rate and pitch contour [84]. Figure 3.5 shows the RAWMARYXML markup to convey a "sad" vocal prosody through a slow speech rate and a falling pitch contour. The MARY speech synthesizer also allowed modification of the produced speech's volume, average pitch, and amount of pitch variation through request parameters sent to the speech synthesizer as part of speech synthesis request [83].

```
<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 3.5

Example of RAWMARYXML Expressing a “Sad” Vocal Prosody

3.2.3 Text

Semantically unpredictable sentences (SUS) [10] were used in the listening task to ensure that the participants were choosing an emotion based on the robot’s vocal prosody, not the linguistic content of the sentence. Sets of semantically unpredictable sentences are typically used to test the intelligibility of speech synthesizers. The sentences are generated by first compiling a list of the most commonly used words for several parts of speech (noun, verb, adjective, and determiner). Then the words are placed into one of five sentence structures (shown in Figure 3.6). The resulting sentences contain real words in structurally acceptable arrangements. However, the linguistic content of each sentence is meaningless. 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.

Determiner + Noun + Verb (intransitive) + Preposition + Determiner + Adjective + Noun
Determiner + Adjective + Noun + Verb (transitive) + Determiner + Noun
Verb (transitive) + Determiner + Noun + Conjunction + Determiner + Noun
Question Adverb + Verb (auxiliary) + Determiner + Noun + Verb (transitive) + Determiner + Noun
Determiner + Noun + Verb (transitive) + Determiner + Noun + Relative Pronoun + Verb (intransitive)

Figure 3.6

Sentence Structures for Semantically Unpredictable Sentences

A set of fifty semantically unpredictable sentences was generated. A subset of twenty sentences was selected for each participant. For each subset, each participant listened to four sentences for each of the five emotions (anger, calm, fear, happiness, sadness) in a random order of presentation.

3.3 Study Protocol

Each participant completed an informed consent form (see Appendix A.1), a demographics questionnaire (see Appendix C.1), and a mood survey before the researcher gave instructions for the tasks of choosing emotions and transcribing the sentences. The researcher then left the room. The robot operator remotely controlled the Survivor Buddy robot while the robot introduced itself and repeated the instructions to the participant. The robot operator then used the robot to lead the participant through twenty sentences.

The tasks completed by the participants related to listening to sentences said by the Survivor Buddy robot. When a sentence was first said by the robot the participant would select the emotion being conveyed by the vocal prosody of the robot's speech. The list

of emotions that the participant could choose from was anger, calm, fear, happiness, and sadness. See Figure 3.7 for how the emotions were depicted on the participants' assessment form (Appendix B.1). The participant would then transcribe the sentence. The robot would automatically repeat the sentence once while the participant wrote the sentence. The participant could ask the robot to repeat the sentence by saying "repeat." The participant would signal the robot to move to the next sentence by saying "next."

Figure 3.7

Emotion Choices Presented to the Participants (Preliminary Experiments)

Once the participant completed the twenty sentences the robot asked the participant to retrieve the researcher from the hallway. The participant finished the study by completing a questionnaire evaluating the robot, a second mood survey, a personality survey, and a short survey on the participant's experience during the study.

3.4 Voice Modification

This sections describes the changes in vocal prosody that were made to the standard voice model in both the initial and second preliminary experiments. The changes in vocal prosody were intended to convey emotions in the generated speech.

3.4.1 Initial Experiment

These experiments utilized MARY (Modular Architecture for Research on speech sYnthesis), an open source speech synthesizer designed to produce expressive speech [87]. The pitch and volume of the synthesized speech were specified in speech synthesis requests sent to the MARY server. MaryXML, one of the input languages for MARY, contains tags and elements that allow for the modification of the pitch contour and speech rate of synthesized statements [84]. These modifications to the standard voice were intended to convey four of Ekman's *Big Six* emotions: anger, fear, happiness, and sadness. Disgust and surprise were omitted to reduce the number of emotion choices. Disgust and surprise were chosen because the authors did not envision a scenario where the quality of human-robot interaction would depend on the communication of those two emotions.

The vocal prosody modification labeled *calm* was used to represent a normal vocal prosody that did not convey an emotion. The calm vocal prosody was used as a baseline for the pitch, speech rate, and volume modifications made to express the other four emotions. The calm vocal prosody used a speech rate of 75% and a volume of 60%. These values allowed for changing both parameters higher and lower without making the produced speech difficult to understand. For example, anger is expressed by a faster than normal speech rate and sadness is expressed by a slower than normal speech rate. For the initial experiment, sentences said with the anger vocal prosody used a speech rate of 95% (faster than the calm's vocal prosody speech rate of 75%) and sentences said with the vocal prosody intended to convey sadness used a speech rate of 50% (slower than the calm's vocal prosody speech rate of 75%). The values used for the vocal prosody modification

parameters were initially chosen based on literature reporting the prosody characteristics of emotional speech [33, 91, 93]. For example, Tao et al. [93] report that one of the transformations required to change a neutral vocal prosody to a strong happiness vocal prosody is a rise in the F0 pitch by 37.2%. The particular voice model used for these experiments had an average pitch of 180 Hz. An increase of 37.2% would be an increase of 67 Hz over the normal voice. Pilot testing of the vocal prosody modifications was performed and the parameter values were adjusted based on feedback from the listeners. Table 3.1 shows the modifications made to the standard voice that were used in the first experiment.

Table 3.1

Changes Made to Standard Voice to Convey Emotions in Initial Preliminary Experiment

Emotion	**Pitch**	**Pitch Range**	**Pitch Contour**	**Speech Rate**	**Volume**
Anger	-50Hz	120%	each word had a falling contour	95%	100%
Calm	unchanged	unchanged	unchanged	75%	60%
Fear	+40Hz	30%	rising	90%	80%
Happiness	+50Hz	150%	each word had a rising contour	85%	60%
Sadness	-30Hz	70%	falling	50%	40%

3.4.2 Second Experiment

The program that specified the speech synthesizer's vocal prosody was modified and a second experiment was conducted. The essential study design features of the first experiment were repeated in this second experiment. The only difference was the vocal prosody

instructions given to the MARY speech synthesizer were modified. Table 3.2 shows the vocal prosody changes made to the default voice to express each of the five emotions.

Table 3.2

Changes Made to Standard Voice to Convey Emotions in Second Preliminary Experiment

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

The "fear" and "happiness" vocal prosodies changed the most from the initial experiment. The pitch of the new "fear" vocal prosody was raised an additional 30Hz, the pitch range was decreased 10%, and the volume was decreased by 10%. The speech rate of the new "fear" vocal prosody was increased by 10% but random pauses were inserted between words to mimic a halting speech pattern.

The new "happiness" vocal prosody had an increased pitch range (200% as opposed to 150%) and an increased volume (80% instead of 60%). The pitch contour of the new "happiness" vocal prosody was calculated over the entire sentence so that the pitch rose and fell in a smooth pattern as recommended by Burkhardt and Sendlmeier [17]. The new "happiness" vocal prosody's speech rate also varied between 70% and 90% over the

entire sentence. These last two changes were made to give the sentences said with the "happiness" vocal prosody a melodic quality.

3.5 Results

This section reports the results of both the initial and second preliminary experiments. The purpose of these experiments was to validate the changes in vocal prosody of the robot's speech made to express emotion. The surveys administered before and after listening task (see Section 3.3) in the preliminary experiments were being evaluated for use in the following larger scale experiments. The survey results were not analyzed and the survey results are not reported.

3.5.1 Initial Experiment

Thirty-six university students participated in this experiment. Program malfunctions did not permit three students to finish the experiment so the following results are for 33 participants. The 33 participants included 17 females and 16 males. Their average age was 19.7 years old (SD = 2.18). Table 3.3 is a confusion matrix that displays the classification of sentences said with the intended emotions across all participants. For example, the first row of the table shows that sentences spoken with the "anger" vocal prosody were recognized correctly 65.9% of the time while 18.2% were classified as "calm", 7.6% as "fear", 5.3% as "happiness", and 3.0% as "sadness."

Table 3.4 gives the results of a one sample *t*-test (α = 0.05) for each of the emotion recognition rates including effect size using Cohen's *d* value. Cohen's *d* is used to express and categorize an effect size as small (*d* = 0.2), medium (*d* = 0.4), or large (*d* = 0.8) [19].

Table 3.3

Emotion Recognition Rates in Initial Experiment

Intended Emotion	**Selected Emotion (% correct)**				
	Anger	Calm	Fear	Happiness	Sadness
Anger	**65.9**	18.2	7.6	5.3	3.0
Calm	4.5	**68.9**	4.5	2.3	18.9
Fear	0	11.4	**37.9**	33.3	17.4
Happiness	0	25.0	19.7	18.2	**36.4**
Sadness	29.5	19.7	0.8	0	**49.2**

For a one sample *t*-test Cohen's *d* is calculated as

$$d = \frac{\overline{X} - \mu}{\text{SD}} \tag{3.1}$$

where μ is the test value being compared to the mean. The test value used in this one sample *t*-test was 0.2, the recognition rate that results from random guessing.

Table 3.4

Statistical Significance of Emotion Recognition Rates in Initial Experiment

Emotion	**Mean**	**SD**	***t***	***df***	***p* (2-tailed)**	**Cohen's *d***
Anger	0.659	0.27	9.727	32	<0.001	1.69
Calm	0.689	0.22	12.969	32	<0.001	2.26
Fear	0.379	0.34	2.994	32	0.005	0.52
Happiness	0.182	0.20	-0.521	32	0.606	
Sadness	0.492	0.35	4.790	32	<0.001	0.83

3.5.2 Second Experiment

Nineteen university students participated in the second experiment. There were eleven female and eight male participants. Their average age was 18.7 years old (SD = 1.06).

Table 3.5 is a confusion matrix that displays the classification of sentences said for the intended emotions across all participants. Table 3.6 gives the results of a one sample *t*-test (α = 0.05) for each of the emotion recognition rates including effect size. The test value used in the one sample *t*-test was 0.2, the recognition rate that results from random guessing.

Table 3.5

Emotion Recognition Rates in Second Experiment

Intended Emotion	**Selected Emotion (% correct)**				
	Anger	Calm	Fear	Happiness	Sadness
Anger	**76.3**	9.2	5.3	6.6	2.6
Calm	7.9	**76.3**	2.6	6.6	6.6
Fear	3.9	1.3	**46.1**	14.5	31.6
Happiness	7.9	15.8	18.4	**30.3**	26.3
Sadness	23.7	**40.8**	3.9	0	30.3

Table 3.6

Statistical Significance of Emotion Recognition Rates in Second Experiment

Emotion	**Mean**	**SD**	***t***	***df***	***p* (2-tailed)**	**Cohen's *d***
Anger	0.763	0.26	9.570	18	<0.001	2.20
Calm	0.763	0.23	10.775	18	<0.001	2.47
Fear	0.461	0.39	2.888	18	0.01	0.66
Happiness	0.303	0.21	2.093	18	0.05	0.48
Sadness	0.303	0.28	1.578	18	0.13	

3.6 Discussion

The emotional vocal prosody recognition rates from the initial and second preliminary experiments are discussed in this section.

3.6.1 Initial Experiment

The recognition rate for each emotion was initially compared to the recognition rate of random guessing by the participant. Since there were five choices of emotion for each sentence, the probability of correctly guessing the intended emotion was 20% (1/5). The recognition rates for the intended emotion of anger (65.9%) and calm (68.9%) were both well above chance (see Table 3.4). These rates were comparable to the successful emotion recognition rate (60%) of people listening to human speakers [82]. The recognition rates for fear (37.9%) and sadness (49.2%) were significantly higher than chance but are lower than the recognition rates of anger and calm.

The most surprising result was the recognition rate for happiness (18.2%). Not only was this rate below the level of chance, sentences said with a "happy" vocal prosody were more likely rated as fear, calm, or sadness than rated as conveying happiness. This finding should have been anticipated given that previous research has shown that happiness is difficult to recognize from vocal prosody alone [60].

This result lead to more research into the expression of happiness through vocal prosody. Frick [30] noted that speech expressing happiness "is often described as containing gentle contours in pitch." This idea was repeated by Burkhardt and Sendlmeier [17] who used a "wave pitch contour model" to express joy / happiness. These findings were the basis of

the change to the "happiness" vocal prosody for the second experiment. Instead of applying pitch contour changes to individual words (as in the "anger" vocal prosody), the pitch contour of the entire sentence was modified to produce a gentle rising and falling contour. A similar modification to the speech rate (speeding up and slowing down) was made for the entire sentence as well. These two changes produced a melodic "sing-song" quality in the sentences synthesized with the "happiness" vocal prosody.

3.6.2 Second Experiment

After changes were made to the vocal prosody characteristics for calm, fear, and happiness (see Table 3.1 and Table 3.2), the recognition rates for four of the five intended emotions increased from the rates observed in the initial experiment. Four of the five intended emotion recognition rates were significantly higher than chance (20%) as shown in Table 3.6. Sentences said with the modified "happiness" vocal prosody were correctly classified 30.3% of the time, an improvement over the 18.2% recognition rate in the previous experiment.

The recognition rate for sadness fell from 49.2% in the first experiment to 30.3% in this experiment even though no changes were made to the vocal prosody used to express sadness. More importantly, the statements said in a "sad" vocal prosody were classified as calm more often than they were classified as sadness.

The null hypothesis for H_1 was that participants would not be able to recognize the emotional intent of a statement based on vocal prosody alone. The null hypothesis was rejected after the intended emotion recognition rates for four of the five emotions were

significantly higher than chance. However, there is much room for improvement in the recognition rates of the "sadness" and "happiness" vocal prosody modifications. For statements voiced using the "happiness" vocal prosody, almost as many (26.3%) statements were labeled "sad" as were labeled "happy." The misclassification of "sad" statements as "calm" statements might not have a serious impact on interactions between a robot and a person. The misclassification of "happy" statements as "sad" statements was a more troubling mistake and could lead to many misunderstandings.

3.7 Conclusions and Future Work from Preliminary Experiments

Semantically unpredictable sentences were used in these experiments to ensure that the linguistic content of the robot's speech would not affect the listener's choice of emotion. While the content of the sentences themselves did not communicate emotion, individual words in the sentences may have influenced the participant's choice of emotion for individual statements. One example is the word *cried* in the sentence *The dream cried by the great way*. The negative connotations of the word *cry* might have lead participants to label this sentence as "sad" no matter what vocal prosody was used while the sentence was spoken.

Based on informal discussions with STaRS lab members who listened to the vocal prosody modifications, the description of the vocal prosody that should not communicate a particular emotion should be changed from *calm* to *neutral*. The term calm might imply a slowness of speech that leads to the confusion of sadness and calm.

The second experiment did not involve as many participants as the initial experiment. The second experiment was repeated using a larger number of participants with changes

addressing the above concerns (see Chapter 4) in order to increase confidence in the vocal prosody modifications before proceeding with experiments concerning the efficacy of emotional robotic voices.

CHAPTER 4

EXPERIMENT 1: VALIDATION OF EMOTIONAL VOICES

This chapter[1] presents an experiment conducted to validate the vocal prosody modifications made to portray emotions with a robot voice.

4.1 Experimental Design

Previous HRI research related to the use of emotional voices by robots often omitted the validation of the vocal prosody modifications used to communicate emotions during experiments. While the vocal prosody modifications were based on the research on how people communicate emotions through vocal prosody (see Section 2.1), the researchers did not validate that the people interacting with a robot perceived the emotion that the robot was attempting to communicate through its vocal prosody. The purposes of this experiment was to validate that the modifications to the robot's vocal prosody communicated the intended emotion and that the robot's speech remained intelligible. The 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).

[1] The content of this chapter appears in [22]

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.

Note that the first hypothesis (H_1) was the same hypothesis used in the preliminary experiments (see Chapter 3). Each participant heard the vocal prosody modifications for all five emotions: anger, fear, happiness, neutral, and sadness. Therefore this was a within-subjects design.

4.2 Apparatus

This section describes the apparatus used in Experiment 1: the robot, the speech synthesizer, and the text said by the robot during the experiment.

4.2.1 Robot

The robot (Survivor Buddy) used in this experiment was the same robot used in the preliminary experiments (see Section 3.2.1). To address concerns that Survivor Robot might not be seen as a robot by some of the participants, the monitor representing Survivor Buddy's face was probabilistically moved after Survivor Buddy said a sentence during the experiment portion of the study. Forty percent of the time Survivor Buddy would perform an avert motion that moved its face a small amount in a random direction. Twenty percent of the time Survivor Buddy would start a "small scan" motion that turned its head left and right. The remaining forty percent of the time the Survivor Buddy robot would not make a movement. Survivor Buddy would continue these movements while the participant asked it to repeat the same sentence. Once the participant asked Survivor Buddy to move to the

next sentence, Survivor Buddy moved its head back to its "home" (or normal position) before saying a new sentence.

This experiments was conducted using the *Wizard-of-Oz* technique [23]. The sound and video from the robot's microphone and camera were streamed to the robot operator's PC which was located in another room. This experiment used the same graphical user interface program written for robot operator as the preliminary experiments (see Section 3.2.1).

4.2.2 Speech Synthesizer

MARY, the open source speech synthesizer used in the preliminary experiments (see Chapter 3) was used again in this experiment. Table 4.1 lists the changes to the normal voice generated by the MARY speech synthesizer to express the five emotions used in the experiment. The changes were the result of the preliminary emotional voice validation experiments reported in Chapter 3. Note that the term neutral was substituted for the term calm used in preliminary experiments as recommended in Section 3.7.

Table 4.1

Changes Made to Standard Voice to Convey Emotions in Experiment 1

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

4.2.3 Text

Semantically unpredictable sentences (see Section 3.2.3 for a detailed explanation) [10] were used in the listening task to ensure that the participants chose an emotion based on the robot's vocal prosody, not the linguistic content of the sentence. During the preliminary experiments (see Section 3.7) 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 semantically unpredictable sentences were selected to avoid symbolism and the unintentional portrayal of emotion.

A set of forty semantically unpredictable sentences was generated. A subset of twenty sentences was selected for each participant. For each subset, each participant listened to four sentences for each of the five emotions (anger, fear, happiness, neutral, sadness) in a random order of presentation.

4.3 Surveys and Measures

This section describes the pre-experiment surveys, experiment measures, and post-experiment surveys used for Experiment 1.

4.3.1 Pre-Experiment Surveys

Participants completed several pre-experiment surveys. Each participant completed a demographics survey (see Appendix C.1) asking for their gender, age, occupation, highest level of education, ethnicity, race, prior computer experience, and prior robot experience. In addition to previous demographic questions that were asked of all participants in Social, Therapeutic & Robotic Systems (STaRS) lab experiments, a set of questions concerning

previous experience with synthesized speech were also asked. These questions included asking the participants for their experience with video games, their use of Global Positioning System (GPS) units that gave directions by speech, and their use of digital personal assistants such as Siri or Google Now that announce answers by speech.

The short-form of the International Positive and Negative Affect Schedule (I-PANAS-SF) [102, 94] (see Appendix C.2) was used to measure the participant's mood before the experiment. Given that the participants were attempting to recognize the emotion conveyed by the Survivor Buddy's voice, it was important to determine the participant's current mood to investigate if the participant's mood affected their judgment of the emotions presented. The participant was instructed to read each term on the I-PANAS-SF and mark to what extent the term described how they felt at the moment. The scale for each item ranged from 1 (not at all) to 5 (extremely). The I-PANAS-SF was scored by summing the responses for the positive affect items (active, alert, attentive, determined, and inspired) and summing the responses for the negative affect items (afraid, ashamed, hostile, nervous, and upset). Therefore the participant received two affect scores (positive and negative) that ranged between 5 and 25.

The shortened Big Five Inventory (BFI-10) [39, 68] (see Appendix C.3) was administered to measure the participant's personality in terms of the five dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. The participant was asked to mark how well the statements (see Table 4.2) described their personality on a scale of 1 (disagree strongly) to 5 (agree strongly). The participant received a score for each of the 5 dimensions that ranged from 2 to 10.

Table 4.2

Big Five Inventory Scoring

Dimension	**I see myself as someone who . . .**
Extraversion	. . . is reserved *(reverse scored)* . . . is outgoing, sociable
Agreeableness	. . . is generally trusting . . . tends to find fault with others *(reverse scored)*
Conscientiousness	. . . tends to be lazy *(reverse scored)* . . . does a thorough job
Neuroticism	. . . is relaxed, handles stress well *(reverse scored)* . . . gets nervous easily
Openness	. . . has few artistic interests *(reverse scored)* . . . has an active imagination

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.

4.3.2 Experiment Measures

Each participant heard a set of 20 semantically unpredictable sentences (four for each emotion: anger, fear, neutral, sadness, and happiness) said by the robot. The order of presentation was balanced so that two sentences expressed in each emotion were contained in the first ten sentences. For the first ten sentences, the participant wrote a word describing the robot's intended emotion and then transcribed the sentence (see Appendix B.2). For the last ten sentences, the participant chose one of the five emotions as depicted in Figure 4.1 and then transcribed the sentence. This was a change from the preliminary experiments reported in Chapter 3. The preliminary experiments presented the participant with the five emotion choices for all 20 sentences. The assessment completed by the participant was used to calculate the recognition percentage for each of the five intended emotions and the number of words misunderstood when said using each of the five vocal prosody modifications.

Figure 4.1

Emotion Choices Presented to the Participants (Experiment 1)

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 [31]. 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. [89], WordNet-Affect [96], and EmoSenticNet [63]. 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 [52] 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.

4.3.3 Post-Experiment Surveys

The participant completed the short-form of the International Positive and Negative Affect Schedule (I-PANAS-SF) [102, 94] (see Appendix C.2) again to measure the participant's mood after the experiment. An evaluation of the robot (see Appendix C.4) was completed by the participant. The participant finished the study by completing a short survey on the participant's experience during the study.

4.4 Study Protocol

Each participant completed an informed consent form (see Appendix A.2) and a demographics survey. Next, short form versions of the International Positive and Negative Affect Schedule (I-PANAS-SF) and Big Five Inventory (BFI-10) were completed by the participant. 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 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 3.2.3. 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. Appendix B.2 shows the emotion choices for the last ten sentences as depicted on the second response sheet. The robot asked the participant if they were ready to continue and then the robot would proceed with 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 I-PANAS-SF survey again 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.

4.5 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.

4.5.1 Free Choice of Emotion

The participants' word choices were categorized as one of the five expected emotions using the process described in section 4.3.2. The results are provided in Table 4.3. Of the 530 participant responses, 92 could not be categorized to one of the five emotion categories. Examples of words that were not categorized as an emotion were: alert, commanding, determined, informative, and sweet.

Table 4.3

Participant's Responses 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 4.4 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 4.4

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**

The participant's free choice of emotion was categorized as one of the five expected emotions using the process described in Section 4.3.2. Table 4.5 gives the results and effect sizes 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. Cohen's *d* is used to express and categorize an effect size as small (*d* = 0.2), medium (*d* = 0.4), or large (*d* = 0.8) [19]. For a one sample *t*-test Cohen's *d* is calculated as

$$d = \frac{\overline{X} - \mu}{\text{SD}} \tag{4.1}$$

where μ is the test value being compared to the mean. If the participant randomly guessed the emotional intent of the robot's speech, the participant would be expected to guess correctly 20% (or 1/5) of the time. Therefore, the test value used in the one sample *t*-test

and when calculating Cohen's *d* was 0.2, the recognition rate that results from random guessing.

Table 4.5

Statistical Significance of Recognition Rates for Free Choice of Emotions

Emotion	**Mean**	**SD**	***t***	***df***	***p* (2-tailed)**	**Cohen's *d***
Anger	0.51	0.42	5.03	46	<0.001	0.73
Fear	0.59	0.42	6.69	51	<0.001	0.93
Happiness	0.40	0.40	3.60	50	0.001	0.50
Neutral	0.41	0.42	3.69	51	0.001	0.51
Sadness	0.76	0.35	11.47	50	<0.001	1.61

4.5.2 Forced Choice of Emotion

Table 4.6 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 Figure 3.7).

Table 4.6

Recognition Rates for Forced Choice of Emotions

Intended	**Selected Emotion (% correct)**				
Emotion	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 4.7 gives the results and effect sizes 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 4.7

Statistical Significance of Emotion Recognition Rates for Forced Choice of Emotions

Emotion	**Mean**	**SD**	***t***	***df***	***p* (2-tailed)**	**Cohen's *d***
Anger	0.71	0.35	10.69	52	<0.001	1.47
Fear	0.62	0.39	7.87	52	<0.001	1.08
Happiness	0.31	0.34	2.36	52	0.022	0.32
Neutral	0.79	0.32	13.62	52	<0.001	1.87
Sadness	0.42	0.44	3.60	52	<0.001	0.49

Table 4.8 shows the results of a paired sample *t*-test (α = 0.05) comparing the free choice and forced choice emotion recognition rates. The effect sizes are given in Cohen's *d* which is calculated for a paired sample *t*-test as

$$d = \frac{\overline{X_D}}{\mathrm{SD}_D} \tag{4.2}$$

where $\overline{X_D}$ is the mean of the paired differences and SD_D is the standard deviation of the paired differences. Cohen's *d* categorizes an effect size as small (*d* = 0.2), medium (*d* = 0.4), or large (*d* = 0.8) [19].

Table 4.8

Comparing for Free and Forced Choice Emotion Recognition Rates

Emotion	Paired Differences Mean	SD	*df*	*t*	*p* (2-tailed)	Cohen's *d*
Anger	−0.18	0.47	46	−2.63	0.012	−0.38
Fear	−0.05	0.42	51	−0.82	0.416	
Happiness	0.09	0.50	50	1.27	0.211	
Neutral	−0.38	0.49	51	−5.48	<0.001	−0.76
Sadness	0.36	0.47	50	5.51	<0.001	0.77

4.5.3 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. 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 4.9 gives the percentage of words said by the robot in each intended emotion that were transcribed correctly.

Table 4.9

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 effect size (ω^2) for a one-way repeated measures ANOVA is calculated as

$$\omega^2 = \frac{\left[\frac{k-1}{nk}(\mathrm{MS_M} - \mathrm{MS_R})\right]}{\mathrm{MS_R} + \frac{\mathrm{MS_B} - \mathrm{MS_R}}{k} + \left[\frac{k-1}{nk}(\mathrm{MS_M} - \mathrm{MS_R})\right]} \tag{4.3}$$

Table 4.10 lists definitions for the terms used in the ω^2 definition. For effect sizes expressed as ω^2, 0.01 is a small effect size, 0.06 is a medium effect size, and 0.14 is a large effect size [29]. The ANOVA results [$F(4,208) = 25.633$, $p < 0.001$, $\omega^2 = 0.23$] show that there was a significant difference in the correct transcription rates for words said using the different vocal prosody modifications for each of the five emotions.

Table 4.10

Terms in ω^2 for One-Way Repeated Measures ANOVA

Symbol	**Definition**
k	number of conditions
n	number of participants
$\mathrm{MS_M}$	mean square of the model sum of squares
$\mathrm{MS_R}$	mean square of the residual sum of squares
$\mathrm{MS_B}$	mean square of the between-participants sum of squares

Just fourteen words (shown in Table 4.11) accounted for 24.5% of the total number of transcription errors. Each of the words in Table 4.11 were transcribed incorrectly at least 72.7% of the total times that the words appeared in the sentences heard by the participants.

Thirteen words were correctly transcribed by all of the participants: box, fish, fresh, girl, grab, great, held, plan, road, teach, tree, trip, wife.

Table 4.11

Words Transcribed Incorrectly Most Often

Word	Incorrectly Transcribed	Most Common Incorrect Transcription
end	100%	pen
looped	100%	moved
law	100%	lost
snored	100%	more
owned	88.7%	on
year	85.0%	mirror
posed	84.8%	post
cook	80.0%	put
fact	78.8%	fat
week	75.8%	wheat
helped	75.0%	held
sport	75.0%	port
staff	75.0%	stuff
bag	72.7%	back

4.5.4 Correlations

The results of the demographic, mood, and personality surveys were used to investigate the possibility of correlations between participant attributes and their ability to successfully distinguish between the emotional vocal prosodies. Table 4.12 lists the counts of participants in groups defined by binary-valued demographic attributes. The binary-valued attribute "Owns Robot" was not evaluated for correlations because only one participant answered that they owned a robot. The remaining binary-valued attributes ("Gender", "Uses

GPS with Voice", and "Uses Digital Assistant with Voice") were evaluated with independent samples *t*-tests.

Table 4.12

Binary Participant Attributes (Experiment 1)

Attribute	**Value**	***n***	**Value**	***n***
Gender	Female	34	Male	19
Owns Robot	Yes	1	No	51
Uses GPS with Voice	Yes	48	No	5
Uses Digital Assistant with Voice	Yes	38	No	15

Table 4.13 lists descriptive statistics (mean, SD, and *n*) for the emotion recognition rates grouped by gender. Note that the emotion recognition rates are reported for both the free choice and forced choice portions of the experiment. Table 4.14 reports the results of an independent-samples *t*-test (α = 0.05) which compared the emotion recognition rates of males and females. The effect size was calculated using the equation for Cohen's *d* when the groups being compared are not equal sizes.

$$d = \frac{t(n_1 + n_2)}{\sqrt{df}\sqrt{n_1 n_2}} \tag{4.4}$$

Females recognized the "anger" vocal prosody (mean = 0.625, SD = 0.402, *n* = 33) during the free choice of emotion portion of the experiment better [$t(45) = 2.917$, $p = 0.005$, Cohen's $d = 0.93$] than males (mean = 0.267, SD = 0.372, *n* = 18) .

Table 4.13

Emotion Recognition Rate by Gender

		Female			Male		
	Emotion	**Mean**	**SD**	***n***	**Mean**	**SD**	***n***
Free Choice	Anger	0.625	0.402	32	0.267	0.372	15
	Fear	0.515	0.417	34	0.722	0.392	18
	Happiness	0.455	0.402	33	0.306	0.389	18
	Neutral	0.333	0.389	33	0.553	0.438	19
	Sadness	0.758	0.356	33	0.778	0.352	18
Forced Choice	Anger	0.706	0.351	34	0.711	0.346	19
	Fear	0.588	0.379	34	0.684	0.415	19
	Happiness	0.324	0.346	34	0.290	0.346	19
	Neutral	0.824	0.299	34	0.737	0.348	19
	Sadness	0.353	0.418	34	0.526	0.456	19

Table 4.14

Significance of Emotion Recognition Rate by Gender

	Emotion	**Mean Difference**	***t***	***df***	***p***	**Cohen's *d***
Free Choice	Anger	0.358	2.917	45	0.005	0.93
	Fear	-0.208	-1.742	50	0.088	
	Happiness	0.149	1.278	49	0.207	
	Neutral	-0.219	-1.871	50	0.067	
	Sadness	-0.020	-0.194	49	0.847	
Forced Choice	Anger	-0.005	-0.046	51	0.963	
	Fear	-0.096	-0.855	51	0.397	
	Happiness	0.034	0.344	51	0.732	
	Neutral	0.087	0.955	51	0.344	
	Sadness	-0.173	-1.402	51	0.167	

Forty-eight of the participants answered that they have used a GPS device that gives directions via voice. Table 4.15 lists descriptive statistics for the emotion recognition rates when grouped by GPS use. Table 4.16 the results of an independent-samples *t*-test (α = 0.05) which compared the emotion recognition rates of the two groups. There were no significant differences in the emotion recognition rates between participant groups based on who had or had not used a GPS device that used speech to give directions.

Table 4.15

Emotion Recognition Rate by GPS Use

		No			Yes		
	Emotion	**Mean**	**SD**	***n***	**Mean**	**SD**	***n***
Free Choice	Anger	0.625	0.479	4	0.500	0.423	43
	Fear	0.375	0.479	4	0.604	0.412	48
	Happiness	0.500	0.500	5	0.391	0.393	46
	Neutral	0.500	0.500	5	0.404	0.412	47
	Sadness	0.900	0.224	5	0.750	0.361	46
Forced Choice	Anger	0.800	0.274	5	0.698	0.353	48
	Fear	0.500	0.500	5	0.635	0.382	48
	Happiness	0.400	0.418	5	0.302	0.338	48
	Neutral	0.900	0.224	5	0.781	0.325	48
	Sadness	0.500	0.500	5	0.406	0.433	48

Table 4.16

Significance of Emotion Recognition Rate by GPS Use

	Emotion	**Mean Difference**	***t***	***df***	***p* (2-tailed)**	**Cohen's *d***
Free Choice	Anger	0.125	0.561	45	0.578	
	Fear	–0.229	–1.058	50	0.295	
	Happiness	0.109	0.573	49	0.569	
	Neutral	0.096	0.485	50	0.630	
	Sadness	0.150	0.905	49	0.370	
Forced Choice	Anger	0.102	0.625	51	0.535	
	Fear	–0.135	–0.734	51	0.467	
	Happiness	0.098	0.604	51	0.549	
	Neutral	0.119	0.795	51	0.430	
	Sadness	0.094	0.454	51	0.651	

Thirty-eight of the participants answered they had used a digital personal assistant such as Apple Siri or Google Now that communicates via voice. Table 4.17 lists descriptive statistics for the emotion recognition rates when grouped by digital personal assistant use. Table 4.18 presents the results of an independent-samples *t*-test ($\alpha = 0.05$) which compared the emotion recognition rates of the two groups. Effect sizes for significant results are reported in Cohen's *d* (for unequal size groups). The group of participants who had used a digital personal assistant (mean = 0.737, SD = 0.344, $n = 38$) successfully identified the "neutral" vocal prosody at a lower rate [$t(47.3) = 2.733$, $p = 0.009$, $d = 1.22$] than the group of participants who had not used a digital personal assistant (mean = 0.933, SD = 0.176, $n = 15$) during the forced choice of emotion section of the experiment.

Table 4.17

Emotion Recognition Rate by Digital Assistant Use

		No			**Yes**		
	Emotion	**Mean**	**SD**	***n***	**Mean**	**SD**	***n***
Free Choice	Anger	0.462	0.431	13	0.529	0.425	34
	Fear	0.679	0.421	14	0.553	0.416	38
	Happiness	0.433	0.417	15	0.389	0.398	36
	Neutral	0.429	0.475	14	0.408	0.400	38
	Sadness	0.767	0.320	15	0.764	0.368	36
Forced Choice	Anger	0.700	0.316	15	0.711	0.361	38
	Fear	0.633	0.442	15	0.618	0.376	38
	Happiness	0.367	0.352	15	0.290	0.342	38
	Neutral	0.933	0.176	15	0.737	0.344	38
	Sadness	0.333	0.450	15	0.447	0.433	38

Table 4.18

Significance of Emotion Recognition Rate by Digital Assistant Use

	Emotion	**Mean Difference**	***t***	***df***	***p* (2-tailed)**	**Cohen's *d***
Free Choice	Anger	−0.068	−0.488	45	0.628	
	Fear	0.126	0.966	50	0.339	
	Happiness	0.044	0.358	49	0.722	
	Neutral	0.021	0.157	50	0.876	
	Sadness	0.003	0.025	49	0.980	
Forced Choice	Anger	−0.011	−0.099	51	0.922	
	Fear	0.015	0.124	51	0.902	
	Happiness	0.077	0.735	51	0.466	
	Neutral	0.196	2.733	47.3	0.009	1.22
	Sadness	−0.114	−0.856	51	0.396	

Table 4.19 lists descriptive statistics (minimum, maximum, mean, SD, and *n*) for the multi-valued participant attributes (demographics, mood, and personality).

Table 4.19

Multivalue Participant Attributes (Experiment 1)

Attribute	**Min**	**Max**	**Mean**	**SD**	***n***
Computer Experience	0	5	2.36	1.06	53
Robot Experience	0	4	0.58	0.95	53
Video Gaming Experience	0	5	2.04	1.36	53
Positive Affect	5	23	16.23	3.87	53
Negative Affect	5	8	5.68	0.83	53
Extraversion	2	10	6.57	1.64	53
Agreeableness	5	10	8.02	1.43	53
Conscientiousness	4	10	7.91	1.56	53
Neuroticism	2	10	5.25	1.78	53
Openness	4	10	6.89	1.60	53

The following tables give the results of Kendall's tau-*b* (τ_b) correlation tests (α = 0.5) for the multi-valued participant attributes and the emotion recognition rates. Kendall's tau-*b* is a non-parametric correlation that performs well in the presence of tied ranks [29]. Values for the correlation coefficient (τ_b) range from -1.0 for a perfect negative correlation to 1.0 for a perfect positive correlation. When (τ_b) is interpreted as an effect size, values of $\pm$0.1 represent a small effect, $\pm$0.3 is a medium effect, and $\pm$0.5 is a large effect. Table 4.20 reports the results of a Kendall's tau-*b* (τ_b) correlation test (α = 0.5) between the participants' self-reported level of experience with computers and emotion recognition rates. There were no statistically significant correlations between computer experience and emotion recognition rates.

Table 4.20

Correlation of Computer Experience and Emotion Recognition Rate

	Emotion	τ_b	***p* (2-tailed)**	***n***
Free Choice	Anger	–0.082	0.524	47
	Fear	0.113	0.355	52
	Happiness	–0.038	0.756	51
	Neutral	0.029	0.811	52
	Sadness	0.238	0.057	51
Forced Choice	Anger	0.014	0.907	53
	Fear	–0.122	0.310	53
	Happiness	–0.013	0.915	53
	Neutral	–0.037	0.764	53
	Sadness	0.069	0.566	53

Table 4.21 reports the results of a Kendall's tau-*b* (τ_b) correlation test (α = 0.5) between the participants' self-reported level of experience with robots and emotion recognition rates. There was a significant positive correlation [τ_b = 0.255, p = 0.049, n = 51] between level of experience with robots (mean = 0.58, SD = 0.95, n = 53) and the recognition of the "sad" vocal prosody during the free choice of emotion word portion of the experiment.

Table 4.22 reports the results of a Kendall's tau-*b* (τ_b) correlation test (α = 0.5) between the participants' self-reported level of experience with video gaming and emotion recognition rates. There were no statistically significant correlations between video gaming experience and emotion recognition rates.

Table 4.21

Correlation of Robot Experience and Emotion Recognition Rate

	Emotion	τ_b	***p* (2-tailed)**	***n***
Free Choice	Anger	0.056	0.675	47
	Fear	0.174	0.167	52
	Happiness	–0.134	0.291	51
	Neutral	0.125	0.322	52
	Sadness	0.255	0.049	51
Forced Choice	Anger	–0.052	0.680	53
	Fear	–0.002	0.984	53
	Happiness	0.084	0.510	53
	Neutral	–0.031	0.811	53
	Sadness	0.219	0.080	53

Table 4.22

Correlation of Video Game Experience and Emotion Recognition Rate

	Emotion	τ_b	***p* (2-tailed)**	***n***
Free Choice	Anger	–0.240	0.057	47
	Fear	0.149	0.213	52
	Happiness	–0.043	0.724	51
	Neutral	0.107	0.373	52
	Sadness	–0.084	0.496	51
Forced Choice	Anger	–0.010	0.936	53
	Fear	0.231	0.052	53
	Happiness	–0.051	0.671	53
	Neutral	–0.052	0.673	53
	Sadness	0.004	0.972	53

Each participant's positive and negative affect (mood) was measured before their interaction with the robot. The correlations between affect (both positive and negative) and emotion recognition rates were investigated using Kendall's tau-*b* test (α = 0.5). Positive affect (mean = 16.23, SD = 3.87, n = 53) was negatively correlated [τ_b = -0.237, p = 0.042, n = 51] with the sad emotion recognition rate in the free choice of emotional word portion of the study (see Table 4.23). Positive affect was positively correlated [τ_b = 0.261, p = 0.023, n = 53] with the neutral emotion recognition rate in the forced choice portion of the study.

Table 4.23

Correlation of Positive Affect and Emotion Recognition Rate

	Emotion	τ_b	***p* (2-tailed)**	***n***
Free Choice	Anger	0.105	0.377	47
	Fear	0.184	0.103	52
	Happiness	0.211	0.064	51
	Neutral	0.106	0.348	52
	Sadness	–0.237	0.042	51
Forced Choice	Anger	–0.036	0.750	53
	Fear	0.087	0.438	53
	Happiness	–0.083	0.467	53
	Neutral	0.261	0.023	53
	Sadness	0.057	0.610	53

Table 4.24 reports the results of a Kendall's tau-*b* (τ_b) correlation test (α = 0.5) between the participants' negative affect and emotion recognition rates. There were no statistically significant correlations between video gaming experience and emotion recognition rates.

Table 4.24

Correlation of Negative Affect and Emotion Recognition Rate

	Emotion	τ_b	***p* (2-tailed)**	***n***
Free Choice	Anger	–0.080	0.546	47
	Fear	0.022	0.859	52
	Happiness	0.052	0.684	51
	Neutral	0.001	0.993	52
	Sadness	–0.039	0.765	51
Forced Choice	Anger	0.172	0.174	53
	Fear	0.061	0.628	53
	Happiness	0.110	0.384	53
	Neutral	0.074	0.562	53
	Sadness	0.095	0.448	53

Each participant's personality was also measured before their interaction with the robot. Kendall's tau-*b* test (α = 0.5) was used to discover correlations between scores on the five personality dimensions (extraversion, agreeableness, conscientiousness, neuroticism, and openness) and emotion recognition rates. Table 4.25 reports the results of a Kendall's tau-*b* (τ_b) correlation test (α = 0.5) between extraversion and emotion recognition rates. There were no statistically significant correlations between extraversion and emotion recognition rates.

Table 4.25

Correlation of Extraversion and Emotion Recognition Rate

	Emotion	τ_b	***p* (2-tailed)**	***n***
Free Choice	Anger	0.075	0.548	47
	Fear	–0.129	0.277	52
	Happiness	–0.216	0.071	51
	Neutral	0.146	0.215	52
	Sadness	0.029	0.813	51
Forced Choice	Anger	–0.061	0.611	53
	Fear	–0.019	0.872	53
	Happiness	0.042	0.721	53
	Neutral	–0.094	0.432	53
	Sadness	0.118	0.315	53

Table 4.26 reports the results of a Kendall's tau-*b* (τ_b) correlation test (α = 0.5) between agreeableness and emotion recognition rates. There were no statistically significant correlations between agreeableness and emotion recognition rates. Agreeableness (mean = 8.02, SD = 1.43, *n* = 53) was negatively correlated [τ_b = -0.256, *p* = 0.030, *n* = 52] with the

recognition rate of the "fear" vocal prososdy in the free choice of emotional word portion of the study.

Table 4.26

Correlation of Agreeableness and Emotion Recognition Rate

	Emotion	τ_b	***p* (2-tailed)**	***n***
Free Choice	Anger	–0.073	0.561	47
	Fear	–0.256	0.030	52
	Happiness	–0.063	0.598	51
	Neutral	0.064	0.591	52
	Sadness	–0.104	0.391	51
Forced Choice	Anger	0.027	0.820	53
	Fear	0.078	0.508	53
	Happiness	0.027	0.821	53
	Neutral	0.014	0.910	53
	Sadness	0.095	0.420	53

Table 4.27 reports the results of a Kendall's tau-*b* (τ_b) correlation test ($\alpha = 0.5$) between conscientiousness and emotion recognition rates. There were no statistically significant correlations between conscientiousness and emotion recognition rates.

Table 4.28 reports the results of a Kendall's tau-*b* (τ_b) correlation test ($\alpha = 0.5$) between neuroticism and emotion recognition rates. Neuroticism (mean = 5.25, SD = 1.78, $n = 53$) was negatively correlated with the sad emotion recognition rates in both the free choice of emotional word [τ_b = -0.288, $p = 0.016$, $n = 51$] and the forced choice [τ_b = -0.399, $p = 0.001$, $n = 53$] portions of the study.

Table 4.27

Correlation of Conscientiousness and Emotion Recognition Rate

	Emotion	τ_b	***p* (2-tailed)**	***n***
Free Choice	Anger	0.081	0.514	47
	Fear	–0.130	0.272	52
	Happiness	0.093	0.435	51
	Neutral	0.065	0.583	52
	Sadness	–0.046	0.705	51
Forced Choice	Anger	–0.058	0.627	53
	Fear	–0.014	0.905	53
	Happiness	0.078	0.512	53
	Neutral	0.222	0.065	53
	Sadness	0.206	0.079	53

Table 4.28

Correlation of Neuroticism and Emotion Recognition Rate

	Emotion	τ_b	***p* (2-tailed)**	***n***
Free Choice	Anger	–0.073	0.552	47
	Fear	0.087	0.454	52
	Happiness	0.034	0.770	51
	Neutral	–0.087	0.458	52
	Sadness	–0.288	0.016	51
Forced Choice	Anger	–0.016	0.889	53
	Fear	–0.119	0.304	53
	Happiness	–0.152	0.195	53
	Neutral	0.132	0.264	53
	Sadness	–0.399	0.001	53

Table 4.29 reports the results of a Kendall's tau-b (τ_b) correlation test ($\alpha = 0.5$) between openness and emotion recognition rates. There were no statistically significant correlations between openness and emotion recognition rates.

Table 4.29

Correlation of Openness and Emotion Recognition Rate

	Emotion	τ_b	p **(2-tailed)**	n
Free Choice	Anger	0.025	0.840	47
	Fear	0.003	0.979	52
	Happiness	0.000	1.000	51
	Neutral	–0.048	0.682	52
	Sadness	–0.028	0.820	51
Forced Choice	Anger	0.129	0.279	53
	Fear	0.125	0.287	53
	Happiness	0.168	0.157	53
	Neutral	–0.047	0.697	53
	Sadness	0.156	0.183	53

4.6 Discussion

This section is a discussion of the results from Experiment 1: the emotion recognition rates during the free and forced choice of emotion words portions of the experiment, the intelligibility of robot speech when using vocal prosody to communicate emotions, and correlations between participant attributes and the recognition of emotion in robot speech.

4.6.1 Recognition of Emotion

Table 4.4 shows the recognition rates of the intended emotion conveyed by vocal prosody when the participant choose their own word to describe the emotion. The recog-

nition 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 [82]. 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 4.5. 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 [60]. The null hypothesis for H_1 was that the participants would make emotion choices randomly regardless of vocal prosody. The null hypothesis was 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 4.8. 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.

4.6.2 Intelligibility of Emotional Robot Speech

The highest correct transcription rate of words spoken using the different vocal prosody modifications was 85.1% for the "sadness" vocal prosody (see Table 4.9). 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 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 express anger 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 was rejected based on the results of the one-way repeated measures ANOVA and the pairwise comparisons of the correct transcription rates as there was a statistically significant difference.

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 which was constructed from voice data from the Language Technologies Institute at Carnegie Mellon University (see Section 3.2.2). The "neutral" vocal prosody used a speech rate of 85% of the normal speech rate for the `slt` voice model. 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%.

4.6.3 Correlations

A post hoc analysis of the pilot tests for this experiment suggested that females identified the vocal prosodies more accurately than males. That finding was not supported by this experiment except for one case: free choice of emotion word for the anger vocal prosody. The difference in emotion recognition rates of males (mean = 0.267, SD = 0.372, n = 15) and females (mean = 0.625, SD = 0.402, n = 32) while hearing the "anger" vocal prosody during the free choice of emotion portion of the experiment was the only statistically significant result [$t(45) = 2.917$, $p = 0.005$, Cohen's $d = 0.93$]. The difference might be due to females recognizing the "anger" vocal prosody at a higher rate or it might be due to the females having a better vocabulary to describe the "anger" vocal prosody.

Neuroticism (mean = 5.25, SD = 1.775, n = 53) was negatively correlated with the sad emotion recognition rates in both the free choice of emotional word (τ_b = -0.288, p = 0.016, n = 51) and the forced choice (τ_b = -0.399, p = 0.001, n = 53) portions of the

study. Neuroticism is a measure of a participant's emotional stability with high scores representing "negative emotionality, such as feeling anxious, nervous, sad, and tense" [39]. The negative correlation means that participants who were more emotionally stable (low neuroticism scores) recognized the robot's portrayal of sadness at a higher rate. The fact that the correlation was present in both the free choice of emotion word and the forced choice of emotion portions of the study lends credence to the acceptance of this correlation not being due to chance.

Most of the other correlations (see Section 4.5.4) found after Experiment 1 are most likely due to chance. The large number of tests run on the demographic categories, affect measures, and personality measures make finding a few significant correlations likely.

4.7 Conclusions and Future Work from Experiment 1

This experiment has shown that a person interacting with a robot can utilize a robot's varying vocal prosody to determine what emotion a robot is attempting to convey. The intelligibility of the robot speech was high even when using the vocal prosody modifications necessary to portray emotions. 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.

Although there is some support for the universal interpretation of emotions via vocal prosody [81, 60], 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.

A planned extension of this work is investigating the effects of robot body shape on the interpretation of emotional intent by human listeners.

The results currently apply only to the MARY speech synthesizer with the HMM-based `slt` female voice model. The validation of these vocal prosody modifications when applied in other speech synthesizers is another extension of this work that could be done once other speech synthesizers allow the modification of vocal prosody parameters such as pitch, pitch range, and pitch contour. Also, applying the vocal prosody modifications to a male voice in order to convey emotions is another extension of this research. 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 suitable starting points for modifying a male voice model's vocal prosody to communicate emotion.

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.

CHAPTER 5

EXPERIMENT 2: EFFECTS OF USING EMOTIONAL ROBOTIC VOICES

This chapter describes an experiment that investigates the effects of an emotional robotic voice on novice robot users. Experiment 2 uses the robot voice and vocal prosody modifications validated in Experiment 1.

5.1 Experimental Design

Each participant will be assigned to one of five groups. Members of each group will experience one of the five emotional vocal prosodies (anger, fear, happiness, neutral, and sadness) during certain interactions with the robot. Therefore, this is a between-subjects design. The hypotheses were:

> H_3: The robot will be rated higher on positive attributes (attractive, happy, friendly, intelligent, cooperative, etc.) by participants who hear the non-negative voices (neutral, happiness) than by participants who hear the negative voices (fear, anger, sadness).
>
> H_4: Participants who hear the non-negative voices (neutral, happiness) will perform better on the creativity test than participants who hear the negative voices (fear, anger, sadness).

5.2 Apparatus

This section describes the apparatus used in Experiment 2: the robot, the speech synthesizer, and the text spoken by the robot during the experiment.

5.2.1 Robot

The same robot and speech synthesizer was used in this experiment as the emotional voice validation experiments described in Chapters 3 and 4. Experiment 2 used a different experiment task and required a new robot operator graphical user interface (see Figure 5.1). The top half of the robot operator interface allowed control of the robot's movement. The bottom half of the interface allowed the robot operator to control the robot while the robot led the participant through the experiment. The robot operator could use pre-programmed functionality to perform routine actions such as raising the robot's monitor from a resting position and having the robot give instructions to the participant. Less routine tasks such as asking the participant to speak more loudly were also possible to accomplish with the manual controls available to the robot operator through the graphical user interface.

To address concerns that Survivor Buddy robot might not be seen as a robot by some of the participants, the monitor representing Survivor Buddy's face was probabilistically moved while the robot listened to the participant during the creativity task. Once a second the robot would select a small movement or choose to remain still. Ten percent of the time the robot would make a concur motion that consisted of shaking its head up and down. Twenty percent of the time the robot would make an avert motion that moved its face in a random direction for a small amount. The remaining seventy percent of the time the

robot remained still. Before speaking to the participant, the robot would return to its home position which was facing the participant.

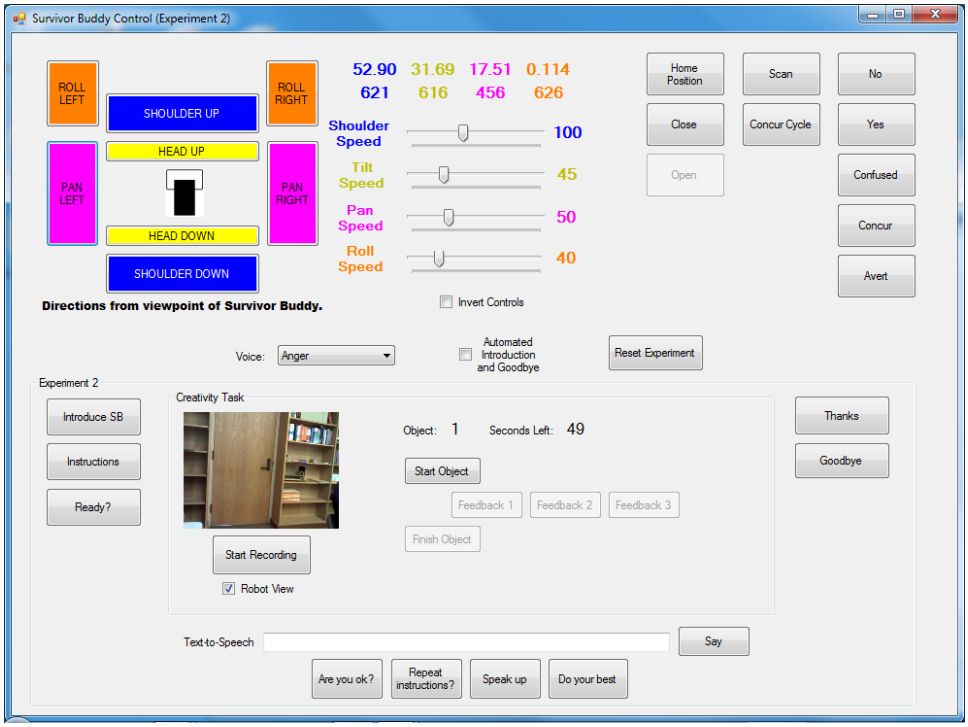


Figure 5.1

Robot Operator GUI (Experiment 2)

5.2.2 Speech Synthesizer

MARY, the open source speech synthesizer used in the Preliminary Experiments and Experiment 1 (see Chapters 3 and 4) was used again in this experiment. The modifications to the standard voice model to express emotions were the same modifications that were previously validated in Experiment 1 (see Table 4.1).

5.2.3 Text

The robot led the participant through a creativity task (see Section 5.3.2 for details). The robot introduced itself and gave instructions for the task in its neutral vocal prosody. During the creativity task, the robot made three statements in one of the five vocal prosodies (anger, fear, happiness, neutral, or sadness):

- Good. Keep going.
- You have *x* seconds left.
- Can you think of any more?

Each participant heard the three statements for all three items in the creativity task in the same vocal prosody for their assigned group.

5.3 Surveys and Measures

The pre-experiment surveys, post-experiment surveys, and experiment measures for Experiment 2 are described in this section.

5.3.1 Pre-Experiment and Post-Experiment Surveys

Experiment 2 used the same pre-experiment (demographics, mood, and personality) and post-experiment (mood, robot evaluation, and study evaluation) surveys as Experiment 1. See Sections 4.3.1 and 4.3.3 for more details. Before the debriefing of the participant by the researcher, the participant was asked if they could tell the difference in the robot's voice when it was giving instructions and giving feedback during the experiment task. The question was asked to determine if they could perceive a difference in the vocal prosody presented in the creativity tasks, if it was different. For the group that heard the neutral

vocal prosody during the task, the instructions and feedback would have been said using the same vocal prosody.

5.3.2 Experiment Measures

Guilford's Alternative Uses Task [32, 67] was used to generate objective measures of the quality of the human-robot interaction. The Alternative Uses Task is a commonly used test of creativity (divergent thinking). The Alternative Uses Task consists of asking a participant to name possible uses for common objects such as brick or paperclip. A timed version of the Alternative Uses Task was used in this experiment. Participants were asked to name as many possible uses of an object as they could think of in 45 seconds. Subscores for fluency, flexibility, and elaboration were assigned based on the quantity and content of the participant's answers. The three subscores were summed to create a participant's creativity score for a single item. Each participant was asked to complete the creativity task for three different objects: (chair, button, and wooden pencil). A participant's final creativity score was the sum of their creativity scores for the three items.

5.4 Study Protocol

Each participant completed an informed consent form (see Appendix A.3). Once informed consent was given by the participant, the participant completed the pre-experiment surveys (demographics, mood, and personality). The researcher then gave instructions for the Alternative Uses Task. The researcher left the room in order to not influence the participant during the experiment.

The robot operator remotely controlled the Survivor Buddy robot while the robot introduced itself and repeated the instructions to the participant. While the robot introduced itself and gave instructions to the participant, the robot used its neutral vocal prosody. The robot operator used the robot to lead the participant through three iterations of the Alternative Uses Task for distinct objects. While the participant was completing the task the robot would speak three times for each of the objects. After the participant stated one or two alternative uses for an object, the robot would say "Good. Keep going." When the participant paused, the robot would announce the how many seconds were remaining. With approximately 5 seconds remaining for an object, the robot would ask the participant "Can you think of any more?" After the completion of the task for each object, the robot would announce that time was up and ask the participant to wait while the robot thought of another object. Once the participant completed the Alternative Uses Task for all three objects, the robot asked (in its neutral vocal prosody) the participant to retrieve the researcher from the hallway.

The participant finished the study by completing a second mood survey, an evaluation of the robot, and a short survey on the participant's experience during the study. The researcher asked the participant if they could tell a difference between Survivor Buddy's voice when it was giving instructions and when it was giving feedback for the task. If the participant stated that they could tell a difference, they were asked to describe the difference. Finally, the participant was debriefed and given a chance to ask any questions about the study and equipment.

Note that each participant heard the robot's neutral vocal prosody during the robot's introduction and general instructions. Each participant heard one of the five emotional vocal prosodies (anger, fear, happiness, neutral, or sadness) while the participant was completing the Alternative Uses Task. The encouraging statement, time remaining announcement, and prompt for more uses were exactly the same across conditions, only the vocal prosody used while speaking the statements was different.

5.5 Results

The following results are reported for 110 participants, all of which were college students. The participants (58 females and 52 males) had an average age of 20.15 years (SD = 2.42). 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.

5.5.1 Robot Evaluation

Eighteen of the twenty-five robot evaluation questions asked the participant to rate the robot on positive qualities. Table 5.1 presents descriptive statistics for the robot ratings when grouped by participants who heard negative (fear, anger, sadness) or non-negative (neutral, happiness) vocal prosodies during the creativity task. Table 5.2 gives the results of independent-samples *t*-test ($\alpha = 0.05$) that compared the mean ratings by participants who heard negative (fear, anger, sadness) or non-negative (neutral, happiness) vocal prosodies during the creativity task. The effect size, Cohen's *d* for groups with unequal sizes (see Equation 4.4), is shown for statistically significant results. Cohen's *d* categorizes an effect

size as small (d = 0.2), medium (d = 0.4), or large (d = 0.8) [19]. Participants in the non-negative voice group rated the robot [$t(108)$ = -2.234, p = 0.028, d = -0.44] as more "Warm" than the participants in the negative voice group.

Table 5.1

Robot Evaluation by Vocal Prosody Group

	Negative			**Non-Negative**		
	Mean	**SD**	***n***	**Mean**	**SD**	***n***
Looks Human-Like	1.61	1.19	66	1.66	0.96	44
Behaves Human-Like	3.68	1.43	66	3.75	1.54	44
Attractive	3.20	1.44	66	3.18	1.78	44
Happy	4.17	1.21	66	4.14	1.25	44
Friendly	5.18	1.11	66	5.34	1.29	44
Optimistic	4.74	1.32	66	5.00	1.48	44
Warm	3.73	1.13	66	4.23	1.18	44
Believable	4.52	1.48	66	4.39	1.42	44
Knowledgeable	5.39	1.05	66	5.25	1.24	44
Responsible	4.92	1.27	66	5.30	1.30	43
Intelligent	5.32	1.18	66	5.75	1.16	44
You Like Robot	5.52	1.26	66	5.61	1.40	44
Robot Likes You	4.26	1.21	66	4.18	1.32	44
Honest	5.91	1.20	66	5.84	1.45	44
Cooperative	5.48	1.24	66	5.84	1.10	44
Attentive To You	5.82	1.23	66	6.16	1.14	44
You Trust Robot	5.18	1.19	66	5.41	1.28	44
You Engaged	5.45	1.15	66	5.43	1.23	44

Table 5.2

Significance of Robot Evaluation by Vocal Prosody Group

	Mean Difference	***t***	***df***	***p* (2-tailed)**	**Cohen's *d***
Looks Human-Like	–0.053	–0.247	108	0.806	
Behaves Human-Like	–0.068	–0.238	108	0.813	
Attractive	0.015	0.049	108	0.961	
Happy	0.030	0.127	108	0.899	
Friendly	–0.159	–0.690	108	0.492	
Optimistic	–0.258	–0.957	108	0.341	
Warm	–0.500	–2.234	108	0.028	–0.44
Believable	0.129	0.455	108	0.650	
Knowledgeable	0.144	0.654	108	0.514	
Responsible	–0.378	–1.506	107	0.135	
Intelligent	–0.432	–1.892	108	0.061	
You Like Robot	–0.098	–0.385	108	0.701	
Robot Likes You	0.076	0.311	108	0.756	
Honest	0.068	0.269	108	0.789	
Cooperative	–0.356	–1.540	108	0.126	
Attentive To You	–0.341	–1.468	108	0.145	
You Trust Robot	–0.227	–0.952	108	0.343	
You Engaged	0.023	0.099	108	0.922	

A one-way ANOVA test was conducted to investigate if the positive quality ratings differed among the five individual vocal prosodies. Table 5.4 gives the result of the one-way ANOVA test (α = 0.05). The effect size of significant differences in the quality ratings found during the one-way ANOVA test is expressed in ω^2. For a one-way ANOVA test, ω^2 is calculated as

$$\omega^2 = \frac{\mathrm{SS_M} - (df_\mathrm{M})\mathrm{MS_R}}{\mathrm{SS_T} + \mathrm{MS_R}} \tag{5.1}$$

Table 5.3 defines the terms used in the ω^2 equation. For effect sizes reported in ω^2, 0.01 is a small effect size, 0.06 is a medium effect size, and 0.14 is a large effect size [29].

Table 5.3

Terms in ω^2 for One-Way ANOVA

Symbol	**Definition**
df_M	degrees of freedom for the effect
$\mathrm{SS_M}$	model sum of squares
$\mathrm{SS_T}$	total sum of squares
$\mathrm{MS_R}$	mean square of the residual sum of squares

Three of the robot qualities (Happy, Friendly, and Optimistic) showed a significant difference in ratings among the five groups based on the vocal prosodies (anger, fear, happiness, neutral, and sadness). The effects of the different vocal prosodies were categorized as medium size effects based on ω^2 for the three robot qualities (Happy, Friendly, and Optimistic) that displayed significant differences in robot ratings.

Table 5.4

Results of ANOVA on Robot Qualities Grouped by Individual Vocal Prosodies

Attribute	Vocal Prosody	Mean	SD	$F(4, 105)$	p	ω^2
Looks Human-Like	Anger	1.55	1.011	0.946	0.441	
	Fear	1.91	1.630			
	Happiness	1.82	1.181			
	Neutral	1.50	0.673			
	Sadness	1.36	0.727			
Behaves Human-Like	Anger	3.82	1.332	0.610	0.657	
	Fear	3.86	1.612			
	Happiness	3.95	1.588			
	Neutral	3.55	1.503			
	Sadness	3.36	1.329			
Attractive	Anger	3.14	1.521	1.366	0.251	
	Fear	3.45	1.438			
	Happiness	3.68	2.079			
	Neutral	2.68	1.287			
	Sadness	3.00	1.380			
Happy	Anger	3.82	0.907	4.996	0.001	0.13
	Fear	5.05	1.046			
	Happiness	4.18	1.140			
	Neutral	4.09	1.377			
	Sadness	3.64	1.177			
Friendly	Anger	5.00	1.113	2.878	0.026	0.06
	Fear	5.82	0.958			
	Happiness	5.41	1.260			
	Neutral	5.27	1.352			
	Sadness	4.73	0.985			
Optimistic	Anger	4.68	0.894	2.698	0.035	0.06
	Fear	5.32	1.427			
	Happiness	5.32	1.041			
	Neutral	4.68	1.783			
	Sadness	4.23	1.378			
Warm	Anger	3.68	0.995	1.514	0.204	
	Fear	3.91	1.151			
	Happiness	4.32	1.129			
	Neutral	4.14	1.246			
	Sadness	3.59	1.260			

Table 5.4

(continued)

Attribute	Vocal Prosody	Mean	SD	$F(4, 105)$	p	ω^2
Believable	Anger	4.27	1.609	0.546	0.702	
	Fear	4.86	1.457			
	Happiness	4.36	1.293			
	Neutral	4.41	1.563			
	Sadness	4.41	1.368			
Knowledgeable	Anger	5.55	0.963	1.427	0.230	
	Fear	5.68	1.086			
	Happiness	5.27	1.241			
	Neutral	5.23	1.270			
	Sadness	4.95	0.999			
Responsible	Anger	5.23	1.232	1.172	0.327	
	Fear	4.91	1.444			
	Happiness	5.36	1.217			
	Neutral	5.24	1.411			
	Sadness	4.64	1.093			
Intelligent	Anger	5.32	1.129	0.924	0.453	
	Fear	5.36	1.399			
	Happiness	5.68	1.211			
	Neutral	5.82	1.140			
	Sadness	5.27	1.032			
You Like Robot	Anger	5.55	1.143	0.069	0.991	
	Fear	5.50	1.144			
	Happiness	5.55	1.405			
	Neutral	5.68	1.427			
	Sadness	5.50	1.504			
Robot Likes You	Anger	4.18	1.181	0.724	0.577	
	Fear	4.59	1.098			
	Happiness	4.27	1.386			
	Neutral	4.09	1.269			
	Sadness	4.00	1.309			
Honest	Anger	6.36	1.002	2.316	0.062	
	Fear	6.09	1.065			
	Happiness	5.95	1.527			
	Neutral	5.73	1.386			
	Sadness	5.27	1.279			

Table 5.4

(continued)

Attribute	**Vocal Prosody**	**Mean**	**SD**	$F(4, 105)$	p	ω^2
Cooperative	Anger	5.45	1.224	1.497	0.208	
	Fear	5.82	1.220			
	Happiness	5.73	1.032			
	Neutral	5.95	1.174			
	Sadness	5.18	1.259			
Attentive To You	Anger	5.77	1.232	2.051	0.092	
	Fear	6.27	0.935			
	Happiness	6.18	1.140			
	Neutral	6.14	1.167			
	Sadness	5.41	1.368			
You Trust Robot	Anger	4.95	1.214	1.380	0.246	
	Fear	5.50	1.144			
	Happiness	5.68	1.211			
	Neutral	5.14	1.320			
	Sadness	5.09	1.192			
You Engaged	Anger	5.50	1.263	0.086	0.987	
	Fear	5.50	1.058			
	Happiness	5.50	1.012			
	Neutral	5.36	1.432			
	Sadness	5.36	1.177			

Tables 5.5, 5.6, and 5.7 are the results of pairwise comparisons of the vocal prosody groups for the three robot qualities (Happy, Friendly, and Optimistic) that showed significant differences. The pairwise comparisons for "Happy" and "Friendly" were conducted using Tukey's HSD test. The pairwise comparisons for "Optimistic" were conducted using the Games-Howell procedure because the population variances were found to be unequal [Levene Statistic = 2.836, p = 0.028]. The effect size for all significant differences in the pairwise comparisons is given in Cohen's d calculated as

$$d = \frac{\overline{X_1} - \overline{X_2}}{\mathrm{SD}_{pooled}} \tag{5.2}$$

Cohen's d categorizes an effect size as small (d = 0.2), medium (d = 0.4), or large (d = 0.8) [19].

The pairwise comparisons of vocal prosody groups for the "Happy" robot rating (see Table 5.5) resulted in three significant differences all involving the "fear" vocal prosody. The participants in the group who heard the "fear" vocal prosody rated the robot as more "Happy" than the participants in the "anger" (p = 0.005, d = 1.28), "neutral" (p = 0.050, d = 0.64), or "sadness" (p = 0.001, d = 1.14) vocal prosody groups. The effect size for the differences in the "Happy" rating between the "fear", "anger", and "sadness" groups is large while effect size between the "fear" and "neutral" groups was medium.

The pairwise comparisons of vocal prosody groups for the "Friendly" robot rating (see Table 5.6) revealed one significant difference involving the "fear" and "sadness" vocal prosodies. The participants in the group who heard the "fear" vocal prosody rated the robot as more "Friendly" than the participants in the "sadness" (p = 0.017, d = 1.16) group.

Table 5.5

Pairwise Comparisons of Individual Vocal Prosodies for the Robot Quality "Happy"

Vocal Prosody 1	**Vocal Prosody 2**	**Mean Difference**	***p***	**Cohen's *d***
Anger	Fear	−1.227	0.005	−1.28
	Happiness	−0.364	0.827	
	Neutral	−0.273	0.932	
	Sadness	0.182	0.984	
Fear	Happiness	0.864	0.095	
	Neutral	0.955	0.050	0.64
	Sadness	1.409	0.001	1.14
Happiness	Neutral	0.091	0.999	
	Sadness	0.545	0.509	
Neutral	Sadness	0.455	0.678	

Table 5.6

Pairwise Comparisons of Individual Vocal Prosodies for the Robot Quality "Friendly"

Vocal Prosody 1	**Vocal Prosody 2**	**Mean Difference**	***p***	**Cohen's *d***
Anger	Fear	−0.818	0.131	
	Happiness	−0.409	0.759	
	Neutral	−0.273	0.933	
	Sadness	0.273	0.933	
Fear	Happiness	0.409	0.759	
	Neutral	0.545	0.512	
	Sadness	1.091	0.017	1.16
Happiness	Neutral	0.136	0.995	
	Sadness	0.682	0.284	
Neutral	Sadness	0.545	0.512	

The pairwise comparisons of vocal prosody groups for the "Optimistic" robot rating (see Table 5.7) revealed one significant difference involving the "happiness" and "sadness" vocal prosodies. The participants in the group who heard the "happiness" vocal prosody rated the robot as more "Friendly" than the participants in the "sadness" (p = 0.039, d = 0.73) group.

Table 5.7

Pairwise Comparisons of Individual Vocal Prosodies for the Robot Quality "Optimistic"

Vocal Prosody 1	**Vocal Prosody 2**	**Mean Difference**	***p***	**Cohen's *d***
Anger	Fear	–0.636	0.405	
	Happiness	–0.636	0.209	
	Neutral	0.000	1.000	
	Sadness	0.455	0.694	
Fear	Happiness	0.000	1.000	
	Neutral	0.636	0.689	
	Sadness	1.091	0.093	
Happiness	Neutral	0.636	0.604	
	Sadness	1.091	0.039	0.73
Neutral	Sadness	0.455	0.877	

5.5.2 Participant Creativity

Table 5.9 reports the results of an independent-samples t-test (α = 0.05) comparing creativity scores grouped by the vocal prosody (negative or non-negative, see Table 5.8) heard during the creativity task. The effect size of significant results are given in Cohen's d for groups with unequal sizes (see Equation 4.4). The only significant difference for creativity scores was for the third item: wooden pencil. Participants in the negative vocal

prosody group (mean = 6.79, SD = 3.64, n = 66) were more creative (t(108) = -1.341, p = 0.037, d = 0.42) than the participants in the non-negative vocal prosody group (mean = 5.39, SD = 3.01, n = 44).

Table 5.8

Creativity by Vocal Prosody Group

	Negative			**Non-Negative**		
	Mean	**SD**	***n***	**Mean**	**SD**	***n***
Chair	7.32	3.43	66	8.20	3.35	44
Button	6.65	3.29	66	6.77	3.48	44
Wooden Pencil	6.79	3.64	66	5.39	3.01	44
Total	20.76	8.13	66	20.36	7.34	44

Table 5.9

Significance of Creativity by Vocal Prosody Group

	Mean Difference	***t***	***df***	***p* (2-tailed)**	**Cohen's *d***
Chair	−0.886	−1.341	108	0.183	
Button	−0.121	−0.185	108	0.854	
Wooden Pencil	1.402	2.115	108	0.037	0.42
Total	0.394	0.259	108	0.796	

Table 5.10 gives the result of the one-way ANOVA test (α = 0.05) of the creativity scores among the five vocal prosody groups. The effect size of significant differences is expressed in ω^2 (see Equation 5.3). The only significant difference [$F(4, 105) = 4.60$, $p = 0.002$, $\omega^2 = 0.12$] in creativity scores between the individual vocal prosody groups was found again for the third item, wooden pencil. The effect size was classified as medium.

Table 5.10

Results of ANOVA on Creativity Grouped by Individual Vocal Prosodies

Item	**Vocal Prosody**	**Mean**	**SD**	$F(4, 105)$	p	ω^2
Chair	Anger	6.55	1.993	1.54	0.195	
	Fear	6.95	3.579			
	Happiness	7.86	3.212			
	Neutral	8.55	3.515			
	Sadness	8.45	4.183			
Button	Anger	6.55	2.345	0.85	0.494	
	Fear	5.95	3.697			
	Happiness	7.32	4.052			
	Neutral	6.23	2.793			
	Sadness	7.45	3.622			
Wooden Pencil	Anger	7.45	3.158	4.60	0.002	0.12
	Fear	4.82	3.607			
	Happiness	5.86	2.965			
	Neutral	4.91	3.054			
	Sadness	8.09	3.421			
Total	Anger	20.55	6.231	1.96	0.107	
	Fear	17.73	7.869			
	Happiness	21.05	7.644			
	Neutral	19.68	7.127			
	Sadness	24.00	9.129			

Table 5.11 is the result of a pairwise comparison of the vocal prosody groups for creativity scores for the third object (wooden pencil). The pairwise comparisons were conducted using Tukey's HSD test. The effect size for all significant differences in the pairwise comparisons is given in Cohen's d (see Equation 5.2). The creativity of the group who heard the "sadness" vocal prosody (mean = 8.09, SD = 3.421, n = 22) was significantly higher (p = 0.01, $d > 0.26$) than the groups who heard the "fear" (mean = 4.82, SD = 3.607, n = 22) or "neutral" (mean = 4.91, SD = 3.054, n = 22) vocal prosodies.

Table 5.11

Pairwise Comparisons of Individual Vocal Prosodies for Creativity (Wooden Pencil)

Vocal Prosody 1	**Vocal Prosody 2**	**Mean Difference**	***p***	**Cohen's *d***
Anger	Fear	2.636	0.06	
	Happiness	1.591	0.49	
	Neutral	2.545	0.08	
	Sadness	–0.636	0.97	
Fear	Happiness	–1.045	0.82	
	Neutral	–0.091	1.00	
	Sadness	–3.273	0.01	–0.26
Happiness	Neutral	0.955	0.87	
	Sadness	–2.227	0.16	
Neutral	Sadness	–3.182	0.01	–0.30

5.5.3 Perceived Difference in Voice

Before a participant was debriefed and learned the purpose of the study, they were asked "Could you tell a difference between Survivor Buddy's voice when it was giving

instructions and when it was giving you feedback during the task?" Table 5.12 reports a summary of the participants' responses.

Table 5.12

Perceived Difference in Voice

Vocal Prosody	**Difference in Voice**	
	No	**Yes**
Anger	9	13
Fear	7	15
Happiness	15	7
Neutral	17	5
Sadness	11	11

A chi-square analysis was conducted to determine if participant groups perceived the voice differences at the same rate. The result [$\chi^2(4)$ = 12.576, p = 0.014, V = 0.11] indicates that there is a significant difference in the number of participants that correctly perceived the voice difference among the groups. The effect size for the chi-square analysis is expressed in Cramér's V:

$$V = \frac{\chi^2}{n \cdot df^*} \tag{5.3}$$

where df^* is the minimum of (rows - 1) and (columns - 1). Effect sizes expressed in Cramér's V for df^* = 1 can be interpreted as large (V = 0.5), medium (V = 0.3), or small (V = 0.1).

5.6 Discussion

This section provides a discussion of the robot evaluations and participant creativity when compared based on vocal prosody groups (negative and non-negative) or individual vocal prosody groups (anger, fear, happiness, neutral, and sadness).

5.6.1 Robot Evaluation

Table 5.1 shows the average participant rating of the robot's positive qualities grouped by vocal prosody type (negative or non-negative). Of the 18 positive qualities, participants who heard non-negative vocal prosodies rated 11 qualities higher than the participants who heard negative vocal prosodies. However, only one attribute's mean rating was significantly different ("Warm": $p = 0.028$, $d = 0.44$).

Seven of the robot's positive qualities ("Attractive", "Happy", "Believable", "Knowledgeable", "Robot Likes You", "Honest", and "You Engaged") of the 18 qualities were rated more highly when the robot used a negative vocal prosody than used a non-negative vocal prosody. None of the these differences were significant.

The pairwise comparisons of the individual vocal prosody groups (see Table 5.4) shows that the group who heard the "fear" vocal prosody rated the robot highest on the "Happy". "Friendly", and "Optimistic" qualities. The hypotheses for Experiment 2 partition the individual vocal prosodies into two groups: negative (anger, fear, sadness) and non-negative (happiness and neutral). For those three robot qualities, it appears that the two vocal prosody groups would be fear, happiness, and neutral versus anger and sadness. The "fear" and "happiness" vocal prosodies were often confused by participants in the initial prelim-

inary experiment (see Section 3.5.1). These two vocal prosodies may still be confused by some participants in spite of the changes made to differentiate the two vocal prosodies (see Section 3.4.2).

The null hypothesis for H_3 was the robot would be rated the same on positive attributes by participants who heard the non-negative vocal prosodies (neutral, happiness) compared to participants who hear the negative vocal prosodies (fear, anger, sadness). The null hypothesis cannot be rejected because only one of the 18 positive attributes received significantly different ratings from the participants who heard non-negative vocal prosodies and the participants who heard negative vocal prosodies.

5.6.2 Participant Creativity

Table 5.8 reports the average creativity score for the participants grouped by vocal prosody type (negative or non-negative). Of the three items (chair, button and wooden pencil) used in the creativity task, the average creativity score was only significantly different for wooden pencil (p = 0.037, d = 0.42). The difference was in the unexpected direction and the participants who heard the negative vocal prosodies during the creativity task scored higher on the third item than the participants who heard the non-negative vocal prosodies. For the first two items (chair and button), the participants who heard the non-negative vocal prosodies scored higher than the participants who heard the negative vocal prosodies but the difference was not significant. The total creativity score means were higher for the participants who heard the negative vocal prosodies but the difference was again not significant.

The pairwise comparisons of the individual vocal prosody groups (see Table 5.11) on the third item in the creativity task shows that the group who heard the "sadness" vocal prosody were more creative (p = 0.01, $d <= 0.30$) than the "fear" or "neutral" vocal prosody groups. The difference in creativity was in the unexpected direction, the "neutral" vocal prosody group was expected to be more creative than the "sadness" vocal prosody group.

The null hypothesis for H_4 was the participants who heard the non-negative vocal prosodies (neutral, happiness) would perform the same on the creativity task as the participants who heard the negative vocal prosodies (fear, anger, sadness). The null hypothesis cannot be rejected because the creativity scores for the two groups of participants were not significantly different.

5.7 Conclusion

The null hypotheses for both H_3 and H_4 could not be rejected. One interpretation of these results is that the use of vocal prosody by the robot to communicate emotional intent had no effect on the creativity of the participants. If that interpretation is true, time and effort spent on improving the communication of emotional intent of robots through varying vocal prosody might be better spent on improving the communication of emotional intent through other modalities such as the linguistic content of speech, body language, and facial expressions (when possible). Another explanation of the results might be that the task was not difficult enough or involved adequate time to manifest the differences expected from the robot's use of different emotional voices. This experiment did not meet its goal of

providing proof that human-robot interaction would be demonstrably improved when the robot used vocal prosody to communicate emotional intent.

CHAPTER 6

CONCLUSION

6.1 Contributions

The design of Experiment 1 included several features to address the shortcomings of previous human-robot interaction (HRI) research concerning the use of emotional voices by robots. Many previous research studies did not use robots in their experiments while claiming that their results were applicable to the interaction between robots and humans. Until the effects of robot embodiment are shown to not exist, the use of actual robots in experiments concerning HRI is a necessity. The use of semantically unpredictable sentences (SUS) allowed the robot to speak syntactically correct sentences comprised of real words while avoiding the confounding factor of linguistic content. Much of the earlier research had used child-like babble or other nonlinguistic utterances for the robot's speech while the robot was portraying emotions. The use of nonlinguistic utterances obviously limits the amount of information that a robot can successfully communicate through speech. Finally, the transcription task performed by the participants showed that the robot's speech was still understandable even when the robot's vocal prosody was manipulated to portray emotions. The goal of Experiment 1 was to show that the emotional intent of a robot could be communicated solely through the robot's use of varying vocal prosody even when the robot was speaking actual sentences. The results reported in Chapter 4 show that this goal

was achieved. These voice validation steps (both the participants' perceptions of the robot's emotional intent and the intelligibility of the robot speech) are often skipped by researchers when investigating how a robot's portrayal of emotion might improve human-robot interactions. The vocal prosody modifications used to convey emotions in Experiment 1 have been validated and can now be used by other researchers when using the MARY speech synthesizer with the `slt` voice model. The vocal prosody modifications reported in Chapter 4 are also appropriate starting points for creating vocal prosody modifications to convey emotions with other speech synthesizers or voice models.

The goals of Experiment 2 were to show that a robot's use of emotional speech would improve the robot user's appraisal of the robot and improve tangible measures of the human-robot interaction. Specifically, it was expected that a person performing a task while being guided by the robot using a non-negative emotional voice would perform the task better than if the robot had spoken with a negative emotional voice. The results reported in Chapter 5 did not meet these goals. The robot's use of vocal prosody to communicate emotional intent did not significantly affect the robot user's appraisal of the robot or influence the user working with the robot to perform better on this specific creativity task. Further research is required to assess whether the participant's lack of improvement on the task is due to the vocal prosody being used without other modalities (facial expressions, body language, and linguistic content) typically used to convey emotions or if the task itself was not suitable to show the effects of the robot's use of vocal prosody to convey emotions.

6.2 Future Work

Three suggestions for future work were already described in Chapter 4:

- investigate the communication of emotion through vocal prosody in languages other than English
- validate the vocal prosody modifications to communicate emotion with speech synthesizers other than MARY
- validate the vocal prosody modifications to communicate emotion for a male voice

In addition to the extensions of Experiment 1 listed above, there are several more avenues for future research on the communication of emotional intent by robots. 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*" [79]? 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 coupled with a sad sentence result in the listener recognizing sarcasm? The design of Experiment 2 purposely constrained the communication of emotions to the robot's varying vocal prosody. The use of multi-modal expressions (vocal prosody along with facial expressions, body language, or linguistic content) of emotion might have influenced the robot users more and resulted in larger differences between the groups who experienced non-negative and negative emotion portrayals by the robot.

The current experiments address communication from the robot to a novice robot user. Obviously the novice robot user does not have much experience with a specific robot and might fail to recognize subtle changes in the robot's voice. A different use case would be

robots serving as long term companions to their users. Presumably the robot user's ability to recognize the robot's emotional intent from the robot's vocal prosody would improve as the user and robot spent more time together. Another possibility is for the robot to learn how its specific user expresses emotions and utilize similar vocal prosody modifications when the robot is attempting to communicate an emotion. This would allow the robot to adapt to a specific user over time and match the user's expectations of how specific emotions are portrayed.

Speech is one of the most natural ways for mobile robots and their users to communicate. It is expected that these experiments and results along with the above suggested research extensions will provide information and techniques that will assist robot designers and programmers to improve the acceptability and quality of human-robot interactions.

6.3 Publication Plan

The following conference papers related to this research have been published:

- J. Crumpton, "Use of Vocal Prosody to Communicate Emotion in Robot Speech", *Proceedings: 2014 Human Robot Interaction Pioneers Workshop*, Bielefeld, Germany, 2014.
- J. Crumpton and C. L. Bethel, "Conveying Emotion in Robotic Speech: Lessons Learned," *Proceedings: 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Edinburgh, Scotland, 2014, IEEE.
- J. Crumpton and C. L. Bethel, "Validation of Vocal Prosody Modifications to Communicate Emotion in Robot Speech," *Proceedings: 2015 International Conference on Collaboration Technologies and Systems (CTS 2015)*, Atlanta, GA, 2015, IEEE. (Nominated for Best Paper Award, decision pending)

The following publication related to this research is currently in review:

- J. Crumpton and C. L. Bethel, "A Survey of Using Vocal Prosody to Convey Emotion in Robot Speech," *International Journal of Social Robotics*

The following publication related to this research is planned:

- “Effects of Using an Emotional Robot Voice on Human Robot Interaction”

REFERENCES

[1] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," *Proceedings: Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, 2005, pp. 579–586, Association for Computational Linguistics.

[2] N. Amir, A. Weiss, and R. Hadad, "Is there a dominant channel in perception of emotions?," *Proceedings: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII)*, Amsterdam, The Netherlands, 2009, pp. 1–6, Association for the Advancement of Artificial Intelligence.

[3] K. Ashimura, P. Baggia, F. Burkhardt, A. Oltramari, C. Peter, and E. Zovato, "Vocabularies for EmotionML," W3C, 2013, http://www.w3.org/TR/2012/NOTE-emotion-voc-20120510/ (current Dec. 20, 2013).

[4] P. Baggia, P. Bagshaw, M. Bodell, D. Z. Huang, L. Xiaoyan, S. McGlashan, J. Tao, Y. Jun, H. Fang, Y. Kang, H. Meng, W. Xia, X. Hairong, and Z. Wu, "Speech Synthesis Markup Language (SSML) Version 1.1," W3C, 2010, http://www.w3.org/TR/2010/REC-speech-synthesis11-20100907/ (current April 25, 2014).

[5] P. Baggia, C. Pelachaud, C. Peter, and E. Zovato, "Emotion Markup Language (EmotionML) 1.0," W3C, 2013, http://www.w3.org/TR/2013/PR-emotionml-20130416/ (current April 25, 2014).

[6] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati, "The Benefits of Interactions with Physically Present Robots over Video-Displayed Agents," *International Journal of Social Robotics*, vol. 3, no. 1, 2011, pp. 41–52.

[7] J. Bates, "The Role of Emotion in Believable Agents," *Communications of the ACM*, vol. 37, no. 7, 1994, pp. 122–125.

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

[9] M. E. Beckman and G. M. Ayers, "Guidelines for ToBI Labelling," 1994, http://www.speech.cs.cmu.edu/tobi/ (current Dec. 20, 2013).

[10] 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, 1996, pp. 381–392.

[11] C. L. Bethel and R. R. Murphy, "Auditory and Other Non-verbal Expressions of Affect for Robots," *Proceedings: 2006 AAAI Fall Symposium Series, Aurally Informed Performance: Integrating Machine Listening and Auditory Presentation in Robotic Systems*, Washington, DC, 2006, AAAI.

[12] A. W. Black, "CMU_ARCTIC speech synthesis databases," *festvox*, Carnegie Mellon University's Speech Group, http://festvox.org/cmu_arctic/ (current Feb. 8, 2014).

[13] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, 2001, pp. 341–345.

[14] C. Breazeal and L. Aryananda, "Recognition of Affective Communicative Intent In Robot-Directed Speech," *Autonomous Robots*, vol. 12, no. 1, 2002, pp. 83–104.

[15] C. Breazeal and B. Scassellati, "How to build robots that make friends and influence people," *Proceedings: 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Kyongju, Korea, 1999, pp. 858–863, IEEE.

[16] D. J. Brooks, C. Lignos, C. Finucane, M. S. Medvedev, I. Perera, V. Raman, H. Kress-Gazit, M. Marcus, and H. A. Yanco, "Make It So: Continuous, Flexible Natural Language Interaction with an Autonomous Robot," *Proceedings: Workshops at Twenty-Sixth AAAI Conference on Artificial Intelligence*, Toronto, Ontario, Canada, 2012, Association for the Advancement of Artificial Intelligence.

[17] F. Burkhardt and W. F. Sendlmeier, "Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis," *Proceedings: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, 2000, International Speech Communication Association.

[18] J. Cahn, "The Generation of Affect in Synthesized Speech," *Journal of the American Voice Input / Output Society*, vol. 8, 1990, pp. 1–19.

[19] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd edition, Hillsdale, N.J. : L. Erlbaum Associates, 1988., 1988.

[20] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1, 2003, pp. 5–32.

[21] J. Crumpton and C. L. Bethel, "Conveying Emotion in Robotic Speech: Lessons Learned," *Proceedings: 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Edinburgh, Scotland, 2014, IEEE.

[22] J. Crumpton and C. L. Bethel, "Validation of Vocal Prosody Modifications to Communicate Emotion in Robot Speech," *Proceedings: 2015 International Conference on Collaboration Technologies and Systems (CTS 2015)*, Atlanta, GA, 2015, IEEE.

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

[24] K. Dautenhahn, S. Woods, C. Kaouri, M. L. Walters, K. L. Koay, and I. Werry, "What is a Robot Companion - Friend, Assistant or Butler?," *Proceedings: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Alberta, Canada, 2005, pp. 1192–1197, IEEE.

[25] S. D'Mello and A. Graesser, "AutoTutor and Affective AutoTutor: Learning by Talking with Cognitively and Emotionally Intelligent Computers that Talk Back," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 4, 2013, pp. 1–39.

[26] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-Cultural Elements in Facial Displays of Emotion," *Science*, vol. 164, no. 3875, 1969, pp. 86–88.

[27] D. Erickson, "Expressive speech: Production, perception and application to speech synthesis," *Acoustical Science and Technology*, vol. 26, no. 4, 2005, pp. 317–325.

[28] 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, 1939, p. 87.

[29] A. P. Field, *Discovering Statistics Using IBM SPSS Statistics*, 4th edition, Sage, Los Angeles, CA, 2013.

[30] R. W. Frick, "Communicating Emotion: The Role of Prosodic Features," *Psychological Bulletin*, vol. 97, no. 3, 1985, pp. 412–429.

[31] P. Greasley, C. Sherrard, and M. Waterman, "Emotion in Language and Speech: Methodological Issues in Naturalistic Approaches," *Language and Speech*, vol. 43, no. 4, 2000, pp. 355–375.

[32] J. P. Guilford, *The nature of human intelligence*, McGraw-Hill series in psychology. New York, McGraw-Hill [1967], 1967.

[33] K. Hammerschmidt and U. Jürgens, "Acoustical Correlates of Affective Prosody," *Journal of Voice*, vol. 21, no. 5, 2007, pp. 531–540.

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

[35] S. Hennig and R. Chellali, "Expressive Synthetic Voices: Considerations for Human Robot Interaction," *Proceedings: 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Paris, France, 2012, pp. 589–595, IEEE.

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

[37] G. L. Huttar, "Relations Between Prosodic Variables and Emotions in Normal American English Utterances," *Journal of Speech and Hearing Research*, vol. 11, no. 3, 1968, pp. 481–487.

[38] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, "A Speech Synthesis System with Emotion for Assisting Communication," *Proceedings: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, 2000, pp. 167–172.

[39] O. P. John and S. Srivastava, "The Big Five Trait taxonomy: History, measurement, and theoretical perspectives," *Handbook of Personality: Theory and Research*, L. A. Pervin and O. P. John, eds., 2nd edition, Guilford Press, New York, NY US, 1999, chapter 4, pp. 102–138.

[40] Y. Jung and K. M. Lee, "Effects of Physical Embodiment on Social Presence of Social Robots," *Proceedings: 7th Annual International Workshop on Presence*, Valencia, Spain, 2004, pp. 80–87, International Society for Presence Research.

[41] Z. Khan, *Attitudes towards intelligent service robots*, Tech. Rep. TRITA-NA-P9821, IPLab-154, Royal Institure of Technology (KTH), 1998.

[42] C. D. Kidd and C. Breazeal, "Effect of a Robot on User Perceptions," *Proceedings: 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sendai, Japan, 2004, vol. 4, pp. 3559–3564.

[43] E. Kim, D. Leyzberg, K. Tsui, and B. Scassellati, "How People Talk When Teaching a Robot," *Proceedings: 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, San Diego, CA, 2009, pp. 23–30, ACM.

[44] P. Laukka, D. Neiberg, M. Forsell, I. Karlsson, and K. Elenius, "Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation," *Computer Speech & Language*, vol. 25, no. 1, 2011, pp. 84–104.

[45] D. Leyzberg, E. Avrunin, J. Liu, and B. Scassellati, "Robots That Express Emotion Elicit Better Human Teaching," *Proceedings: 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Lausanne, Switzerland, 2011, pp. 347–354.

[46] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati, "The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains," *Proceedings: 34th Annual Conference of the Cognitive Science Society (CogSci)*, Sapporo, Japan, 2012, Cognitive Science Society.

[47] H. Liu, H. Lieberman, and T. Selker, "A Model of Textual Affect Sensing using Real-World Knowledge," *Proceedings: 2003 International Conference on Intelligent User Interfaces (IUI)*, Miami, Florida, 2003, pp. 125–132, ACM.

[48] D. W. Massaro, "The logic of the fuzzy logical model of perception," *Behavioral and Brain Sciences*, vol. 12, no. 04, 1989, pp. 778–794.

[49] D. W. Massaro and P. B. Egan, "Perceiving affect from the voice and the face," *Psychonomic Bulletin & Review*, vol. 3, no. 2, 1996, pp. 215–221.

[50] A. Mehrabian, "Pleasure-Arousal-Dominance: A General Framework for Describing and Measuring Individual Differences in Temperament," *Current Psychology*, vol. 14, no. 4, 1996, p. 261.

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

[52] 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, 1990, pp. 235–244.

[53] W. J. Mitchell, K. A. Szerszen Sr, A. S. Lu, P. W. Schermerhorn, M. Scheutz, and K. F. MacDorman, "A mismatch in the human realism of face and voice produces an uncanny valley," *i-Perception*, vol. 2, no. 1, 2011, pp. 10–12.

[54] I. R. Murray, J. L. Arnott, and E. A. Rohwer, "Emotional stress in synthetic speech: Progress and future directions," *Speech Communication*, vol. 20, no. 12, 1996, pp. 85–91.

[55] C. Nass, I. Jonsson, H. Harris, B. Reaves, J. Endo, S. Brave, and L. Takayama, "Improving Automotive Safety by Pairing Driver Emotion and Car Voice Emotion," *Proceedings: CHI '05 Extended Abstracts on Human Factors in Computing Systems*, Portland, Oregon, USA, 2005, pp. 1973–1976, ACM.

[56] A. Niculescu, B. Dijk, A. Nijholt, H. Li, and S. L. See, "Making Social Robots More Attractive: The Effects of Voice Pitch, Humor and Empathy," *International Journal of Social Robotics*, vol. 5, no. 2, 2013, pp. 171–191.

[57] Nuance Communications, "SSML compliance," *Dragon Mobile SDK Reference*, http://dragonmobile.nuancemobiledeveloper.com/public/Help/DragonMobileSDK Reference_Android/SpeechKit_Guide/SpeakingText.html (current Feb. 25, 2015).

[58] P.-Y. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, no. 12, 2003, pp. 157–183.

[59] J. C. Pearson and P. E. Nelson, *An Introduction to Human Communication: Understanding and Sharing*, 8th edition, McGraw-Hill Higher Education, Boston, MA, 2000.

[60] 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, 2009, pp. 417–435.

[61] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The IBM expressive text-to-speech synthesis system for American English," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, 2006, pp. 1099–1108.

[62] J. Pittam and K. R. Scherer, "Vocal Expression and Communication of Emotion," *Handbook of Emotions*, The Guilford Press, New York, 1993, chapter 13, pp. 185–197.

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

[64] A. Powers, S. Kiesler, S. Fussell, and C. Torrey, "Comparing a Computer Agent with a Humanoid Robot," *Proceedings: 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Washington DC, USA, 2007, pp. 145–152, ACM.

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

[66] J. Prinz, "Which Emotions are Basic?," *Emotion, Evolution, and Rationality*, Oxford University Press, Oxford, UK, 2004, chapter 4, pp. 69–87.

[67] Psychology Career Center, "Guilford's Test of Divergent Thinking," *Measuring Creativity: Learn how experts are measuring creativity*, http://www.allpsychology careers.com/topics/measuring-creativity.html (current May 1, 2014).

[68] 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, 2007, pp. 203–212.

[69] P. Rani and N. Sarkar, "Emotion-Sensitive Robots - A New Paradigm for Human-Robot Interaction," *Proceedings: 4th IEEE/RAS International Conference on Humanoid Robots*, Los Angeles, CA, 2004, vol. 1, pp. 149–167, IEEE.

[70] C. Ray, F. Mondada, and R. Siegwart, "What do people expect from robots?," *Proceedings: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nice, France, 2008, pp. 3816–3821, IEEE.

[71] R. Read, "Speaking Without Words: Affective Displays in Social Robots through Non-Linguistic Utterances," *Proceedings: 2012 HRI Pioneers Workshop*, Boston, Massachusetts, 2012, ACM.

[72] R. Read and T. Belpaeme, "Interpreting Non-Linguistic Utterances by Robots: Studying the Influence of Physical Appearance," *Proceedings: 3rd International Workshop on Affective Interaction in Natural Environments (AFFINE)*, Firenze, Italy, 2010, pp. 65–70, ACM.

[73] R. Read and T. Belpaeme, "How to Use Non-linguistic Utterances to Convey Emotion in Child-Robot Interaction," *Proceedings: 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Boston, Massachusetts, 2012, pp. 219–220, ACM.

[74] R. Read and T. Belpaeme, "People Interpret Robotic Non-Linguistic Utterances Categorically," *Proceedings: 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Tokyo, Japan, 2013, pp. 209–210, ACM.

[75] R. Read and T. Belpaeme, "Non-Linguistic Utterances Should be Used Alongside Language, Rather than on their Own or as a Replacement," *Proceedings: 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Bielefeld, Germany, 2014, pp. 276–277, ACM.

[76] R. Read and T. Belpaeme, "Situational Context Directs How People Affectively Interpret Robotic Non-Linguistic Utterances," *Proceedings: 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Bielefeld, Germany, 2014, pp. 41–48, ACM.

[77] S. Roehling, B. MacDonald, and C. Watson, "Towards Expressive Speech Synthesis in English on a Robotic Platform," *Proceedings: 11th Australian International Conference on Speech Science & Technology*, University of Auckland, New Zealand, 2006, Australian Speech Science & Technology Association.

[78] O. Rogalla, M. Ehrenmann, R. Zöllner, R. Becher, and R. Dillmann, "Using Gesture and Speech Control for Commanding a Robot Assistant," *Proceedings: 11th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN)*, Berlin, Germany, 2002, pp. 454–459, IEEE.

[79] 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, 2008, pp. 935–939.

[80] K. R. Scherer, "Vocal Affect Expression: a Review and a Model for Future Research," *Psychological Bulletin*, vol. 99, no. 2, 1986, pp. 143–165.

[81] 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, 2001, pp. 76–92.

[82] 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, 1991, pp. 123–148.

[83] M. Schröder, "MARY HTTP Interface: Documentation by example," *MARY Text To Speech*, http://mary.dfki.de:59125/documentation.html (current May 2, 2014).

[84] M. Schröder, "MaryXML," *MARY Text To Speech*, http://mary.dfki.de/documentation/maryxml (current May 22, 2013).

[85] M. Schröder, P. Baggia, F. Burkhardt, C. Pelachaud, C. Peter, and E. Zovato, "EmotionML - An Upcoming Standard for Representing Emotions and Related States," *Affective Computing and Intelligent Interaction*, vol. 6974, Springer Berlin Heidelberg, 2011, chapter 35, pp. 316–325.

[86] M. Schröder, R. Cowie, and E. Douglas-Cowie, "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis," *Proceedings: 7th European Conference on Speech Communication and Technology (EUROSPEECH)*, Scandinavia, Aalborg, Denmark, 2001, pp. 87–90.

[87] 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, 2003, pp. 365–377.

[88] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language - State-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, 2013, pp. 4–39.

[89] 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, 1987, p. 1061.

[90] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A Standard for Labeling English Prosody," *Proceedings: 2nd International Conference on Spoken Language Processing (ICSLP)*, Banff, Alberta, Canada, 1992, vol. 2, pp. 867–870, International Speech Communication Association.

[91] 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, 1999, pp. 347–365.

[92] H. Tang, X. Zhou, M. Odisio, M. Hasegawa-Johnson, and T. S. Huang, "Two-Stage Prosody Prediction for Emotional Text-to-Speech Synthesis," *Proceedings: 9th Annual Conference of the International Speech Communication Association*, Brisbane, QLD, Australia, 2008, pp. 2138–2141, International Speech Communication Association.

[93] 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, 2006, pp. 1145–1154.

[94] 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, 2007, pp. 227–242.

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

[96] R. Valitutti, "WordNet-Affect: an Affective Extension of WordNet," *Proceedings: 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, 2004, pp. 1083–1086, European Language Resources Association.

[97] N. Veilleux, S. Shattuck-Hufnagel, and A. Brugos, "6.911 Transcribing Prosodic Structure of Spoken Utterances with ToBI," *MIT OpenCourseWare*, Massachusetts Institute of Technology, 2006, http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-911-transcribing-prosodic-structure-of-spoken-utterances-with-tobi-january-iap-2006 (current Dec. 20, 2013).

[98] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social Signal Processing: State-of-the-Art and Future Perspectives of an Emerging Domain," *Proceedings: 16th ACM International Conference on Multimedia*, Vancouver, British Columbia, Canada, 2008, pp. 1061–1070, ACM.

[99] W3C, "Speech Synthesis Markup Language (SSML) Version 1.0," W3C, 2004, http://www.w3.org/TR/2004/REC-speech-synthesis-20040907/ (current Dec. 20, 2013).

[100] J. Wainer, D. J. Feil-Seifer, D. A. Shell, and M. J. Mataric, "Embodiment and Human-Robot Interaction: A Task-Based Perspective," *Proceedings: 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Jeju, Korea, 2007, pp. 872–877.

[101] M. R. Walker, J. Larson, and A. Hunt, "A New W3C Markup Standard for Text-to-Speech Synthesis," *Proceedings: 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, Utah, 2001, vol. 2, pp. 965–968, IEEE.

[102] 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, 1988, pp. 1063–1070.

[103] C. H. Weaver and W. L. Strausbaugh, "Hearing the Vocal Cues," *Fundamentals of Speech Communication*, American Book Company, New York, 1964, chapter 11, pp. 283–303.

[104] L. Xingyan, B. MacDonald, and C. I. Watson, "Expressive facial speech synthesis on a robotic platform," *Proceedings: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, St. Louis, MO, USA, 2009, pp. 5009–5014, IEEE.

APPENDIX A

INFORMED CONSENT FORMS

A.1 Preliminary Experiments

Mississippi State University
Informed Consent Form for Participation in Research

Title of Research Study: Vocal Prosody of Robots

Investigator: Joseph Crumpton

Faculty Advisor: Cindy L. Bethel, Ph.D.

Study Site: Mississippi State University Campus

Researchers:

Joe Crumpton, MSU	Dr. Cindy L. Bethel, MSU
John Kelly, MSU	Richard Sween, MSU
James Kaleb Stuart, MSU	Kayla Huddleston, MSU
Amy Crumpton, MSU	Andrew LaFrance, MSU
Malcolm McCullum, MSU	

Purpose
The purpose of this research is to determine how people interact with a robot that speaks English. Additionally, there will be an evaluation of your opinions of the robot.

Procedures
If you decide to participate in this study, you will be asked to complete a demographics questionnaire and surveys concerning your personality and mood. You will speak to and listen to a robot. You will be asked to write down what the robot says. You will also answer a series of questions about the interview process.

Risks or Discomforts
There are no major physical discomforts involved in this study. Risks are minimal and do not exceed those of normal office work. Please tell us if you are having trouble with any task or if you need additional rest and the investigator will be happy to accommodate you in any way possible. If you feel any discomfort, please tell the person assisting you immediately.

Benefits
Although there is no benefit directly to you for participating, this study will help to further our understanding of human-robot interaction. As a participant, you will interact with and learn more about robots and you will have the knowledge that you are helping to make robots and their use more effective.

Confidentiality
All of your responses will be kept strictly confidential. To protect the confidentiality of this information, we will assign your data a code number that will only be known to the members of the research project. All of the information which you provide us today will be marked with the code number, not your name. All information will be stored in a computer for analysis using only your code number for identification. The information collected during the study will be used solely for the purposes of understanding and evaluating human-robot interaction. No indication of your individual answers to questions will be given to anyone. We want you to be completely

MSU IRB #13-211
Approved: 7/18/13
Expires: 7/18/18

Page 1 of 2
Version: 07/18/2013

confident that you may feel free to answer all questions without concern that it may affect you in any way.

Please note that these records will be held by a state entity and therefore are subject to disclosure if required by law. Research information may be shared with the MSU Institutional Review Board (IRB) and the Office for Human Research Protections (OHRP).

Questions

If you have any questions about this research project, please feel free to contact the investigator, Joseph (Joe) Crumpton, (662) [redacted], jjc52@msstate.edu, Department of Computer Science and Engineering, Mississippi State University, 665 George Perry St., P.O. Box 9637, Butler Hall, Room 300, Mississippi State, MS 39762. You may also contact the Faculty Advisor, Cindy L. Bethel, Ph.D., (662) 325-2757, cbethel@cse.msstate.edu, Butler Hall, Room 313.

For questions regarding your rights as a research participant, or to discuss problems, express concerns or complaints, request information, or offer input, please feel free to contact the MSU Regulatory Compliance Office by phone at 662-325-3994, by e-mail at irb@research.msstate.edu, or on the web at http://orc.msstate.edu/humansubjects/participant/.

Voluntary Participation

Please understand that your **participation is voluntary.** Your **refusal to participate will involve no penalty or loss** of benefits to which you are otherwise entitled. You **may discontinue your participation** at any time without penalty or loss of benefits.

Statement of Consent

We ask that you read all of the above information and that you ask any questions you may have about this study before continuing. After reading the above information, you should understand that you do not have to fill out this form and that you do not have to participate in this study. If you do voluntarily consent to participate in this study, you may express your consent by signing and dating below. You will be given a copy of this form for your records.

_____	_____
Participant Name (printed)	Subject ID
_____	_____
Participant Signature	Date
_____	_____
Investigator Signature	Date

A.2 Experiment 1

Mississippi State University
Informed Consent Form for Participation in Research

Title of Research Study: Vocal Prosody of Robots

Researchers: Joe Crumpton, MSU Dr. Cindy L. Bethel, MSU

Purpose
The purpose of this research is to determine how people interact with a robot that speaks English. Additionally, there will be an evaluation of your opinions of the robot.

Procedures
If you decide to participate in this study, you will be asked to complete a demographics questionnaire and surveys concerning your personality and mood. You will speak to and listen to a robot. You will be asked to write down what the robot says. You will also answer a series of questions about the interview process.

Questions
If you have any questions about this research project, please feel free to contact the investigator, Joseph (Joe) Crumpton, (662) [redacted], jjc52@msstate.edu, Butler Hall, Room 300. You may also contact the Faculty Advisor, Cindy L. Bethel, Ph.D., (662) 325-2757, cbethel@cse.msstate.edu, Butler Hall, Room 313.

For questions regarding your rights as a research participant, or to discuss problems, express concerns or complaints, request information, or offer input, please feel free to contact the MSU Regulatory Compliance Office by phone at 662-325-3994, by e-mail at irb@research.msstate.edu, or on the web at http://orc.msstate.edu/humansubjects/participant/.

Voluntary Participation
Please understand that your **participation is voluntary.** Your **refusal to participate will involve no penalty or loss** of benefits to which you are otherwise entitled. You **may discontinue your participation** at any time without penalty or loss of benefits.

Statement of Consent
Please take all the time you need to read through this document and decide whether you would like to participate in this research study.

If you agree to participate in this research study, please sign below. You will be given a copy of this form for your records.

Participant Signature _____ Subject ID _____

Investigator Signature _____ Date _____

MSU HRPP Approved: 6/16/14 Expires: 7/18/18 IRB # 13-211

Version: 06/09/2014

A.3 Experiment 2

Mississippi State University
Informed Consent Form for Participation in Research

Title of Research Study: Vocal Prosody of Robots

Researchers: Joe Crumpton, MSU Dr. Cindy L. Bethel, MSU

Purpose
The purpose of this research is to determine how people interact with a robot that speaks English**.** Additionally, there will be an evaluation of your opinions of the robot.

Procedures
If you decide to participate in this study, you will be asked to complete a demographics questionnaire and surveys concerning your personality and mood. You will speak to and listen to a robot. You will also answer a series of questions about the robot and the study process.

Questions
If you have any questions about this research project, please feel free to contact the investigator, Joseph (Joe) Crumpton, (662) 325-7915, jjc52@msstate.edu, High Performance Computing Collaboratory, Room 219. You may also contact the Faculty Advisor, Cindy L. Bethel, Ph.D., (662) 325-2757, cbethel@cse.msstate.edu, Butler Hall, Room 313.

For questions regarding your rights as a research participant, or to discuss problems, express concerns or complaints, request information, or offer input, please feel free to contact the MSU Regulatory Compliance Office by phone at 662-325-3994, by e-mail at irb@research.msstate.edu, or on the web at http://orc.msstate.edu/humansubjects/participant/.

Confidentiality
Your name and identifying information will not be connected in any way to your responses in this study. The online PRP Sona system will grant you credit based on your PRP Identity Code. Your responses and audiovisual recordings are saved using a code number that will only be known to the members of the research project. All of the information which you provide us today will be marked with the code number, not your name. The information collected during the study will be used solely for evaluating human-robot interaction. No indication of your individual answers to questions will be given to anyone.

Please note that these records will be held by a state entity and therefore are subject to disclosure if required by law. Research information may be shared with the MSU Institutional Review Board (IRB) and the Office for Human Research Protections (OHRP).

Voluntary Participation
Please understand that your **participation is voluntary.** Your **refusal to participate will involve no penalty or loss** of benefits to which you are otherwise entitled. You **may discontinue your participation** at any time without penalty or loss of benefits.

MSU HRPP Full Accreditation

Approved:	Expires:
11/6/14	7/8/18

IRB # 13-211

Version: 10/20/2014

Statement of Consent
Please take all the time you need to read through this document and decide whether you would like to participate in this research study.

If you agree to participate in this research study, please sign below. You will be given a copy of this form for your records.

_____ _____
Participant Signature Date

_____
Investigator Signature

Consent for Video and Audio Recording

I, _____, agree to be video and audio recorded as part of this research study on human-robot interaction. The recordings will be used to evaluate my interactions with a robot. I have been informed that the video and audio recordings may be played to other professionals at research meetings or made available to other professionals.

In addition, I give my consent to Joe Crumpton and Dr. Cindy L. Bethel, Ph.D., of the Social, Therapeutic, and Robotic Systems (STaRS) laboratory to use photographic images, video-recorded stills, or video / audio recordings for:

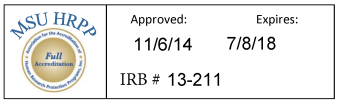
☐ Educational and training purposes _____ (initial)

This would include but is not limited to classroom settings, workshops, and training sessions.

☐ Publication purposes _____ (initial)

This would include but is not limited to textbooks, journal articles, conference papers and other publication venues. Recordings will not be published or made publicly available on social media or video sharing venues (e.g., Vine, YouTube, Facebook)

_____
Participant Signature



Version: 10/20/2014

APPENDIX B

ASSESSMENTS

B.1 Preliminary Experiments

Subject ID _____

Assessment

Select the emotion conveyed by the robot when you first hear a sentence. The robot will then repeat the sentence so that you can write the sentence.

1. Emotion: 

   Sentence:

2. Emotion: 

   Sentence:

3. Emotion: 

   Sentence:

4. Emotion: 

   Sentence:

5. Emotion: 

   Sentence:

Version: 07/10/2013

B.2 Experiment 1

Subject ID _____

Assessment

Write a word that names the emotion conveyed by the robot when you first hear a sentence. The robot will then repeat the sentence so that you can write the sentence.

1. Emotion:

   Sentence:

2. Emotion:

   Sentence:

3. Emotion:

   Sentence:

4. Emotion:

   Sentence:

5. Emotion:

   Sentence:

Version: 07/11/2014

Subject ID _____

Select the emotion conveyed by the robot when you first hear a sentence. The robot will then repeat the sentence so that you can write the sentence.

---

11. Emotion: 

Sentence:

---

12. Emotion: 

Sentence:

---

13. Emotion: 

Sentence:

---

14. Emotion: 

Sentence:

---

15. Emotion: 

Sentence:

Version: 07/11/2014

APPENDIX C

SURVEYS

C.1 Demographics

Subject ID _____

Demographic Information

(Complete each question and mark the circle that most appropriately matches where applicable)

1. **What is your gender?** ○ Male ○ Female

2. **What is your age?** _____

3. **What is your occupation?** _____

4. **What is the highest level of education you have completed?**

Some High School ○ | High School Graduate ○ | Some College ○ | AA/AS ○

Bachelor ○ | Some Graduate School ○ | Masters ○ | Ph.D./M.D./J.D. ○ | Postdoctorate ○

5. **What is your major field of study?** *(Where applicable)*

_____

6. **What is your ethnicity?** *(Select one or both)*

Hispanic/Latino ○ | Not Hispanic/Latino ○

7. **What is your race?** *(Select all that apply)*

American Indian or Alaska Native ○ | Asian ○ | Black or African American ○

Native Hawaiian or Other Pacific Islander ○ | White ○

8. What is your level of prior computer experience?

No Experience	Novice				Expert
0	1	2	3	4	5
○	○	○	○	○	○

9. What is your level of prior robot experience?

No Experience	Novice				Expert
0	1	2	3	4	5
○	○	○	○	○	○

10. Do you own a robot? ○No ○Yes What kind? _____

11. What is your level of prior video gaming experience?

No Experience	Novice				Expert
0	1	2	3	4	5
○	○	○	○	○	○

12. Do you use a GPS system that gives directions via voice? ○No ○Yes

13. Do you use a digital personal assistant such as Apple Siri or Google Now that speaks its answers? ○No ○Yes

C.2 Mood

Subject ID _____ Pre-Survey or Post-Survey

International Positive and Negative Affect Schedule Short Form

Thinking about yourself, to what extent do you feel this way right now, that is, at the present moment? Circle your answers on the scale from 1 to 5.

		not at all	a little	moderately	quite a bit	extremely
1.	Upset	1	2	3	4	5
2.	Hostile	1	2	3	4	5
3.	Alert	1	2	3	4	5
4.	Ashamed	1	2	3	4	5
5.	Inspired	1	2	3	4	5
6.	Nervous	1	2	3	4	5
7.	Determined	1	2	3	4	5
8.	Attentive	1	2	3	4	5
9.	Afraid	1	2	3	4	5
10.	Active	1	2	3	4	5

Version: 06/05/2014

C.3 Personality

Subject ID _____

Big Five Inventory-10

How well do the following statements describe your personality? Circle your answers on the scale from 1 to 5.

I see myself as someone who …	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
1. … is reserved.	1	2	3	4	5
2. … is generally trusting	1	2	3	4	5
3. … tends to be lazy	1	2	3	4	5
4. … is relaxed, handles stress well	1	2	3	4	5
5. … has few artistic interests	1	2	3	4	5
6. … is outgoing, sociable	1	2	3	4	5
7. … tends to find fault with others	1	2	3	4	5
8. … does a thorough job	1	2	3	4	5
9. … gets nervous easily	1	2	3	4	5
10. … has an active imagination	1	2	3	4	5

Version: 06/05/2014

C.4 Robot Evaluation

Subject ID _____

Evaluation of the Robot

*The following are questions regarding your **opinions about the robot.** Read each question and **circle** your answer on the scale from 1 to 7.*

1. How human-like did the robot look?

Very little	1	2	3	4	5	6	7	Very much

2. How human-like did the robot behave?

Very little	1	2	3	4	5	6	7	Very much

3. How unattractive/attractive was the robot?

Unattractive	1	2	3	4	5	6	7	Attractive

4. How unhappy/happy was the robot?

Unhappy	1	2	3	4	5	6	7	Happy

5. How unfriendly/friendly was the robot?

Unfriendly	1	2	3	4	5	6	7	Friendly

6. How pessimistic/optimistic was the robot?

Pessimistic	1	2	3	4	5	6	7	Optimistic

7. How cool/warm was the robot?

Cool	1	2	3	4	5	6	7	Warm

8. How believable was the robot?

Very little	1	2	3	4	5	6	7	Very much

9. How ignorant/knowledgeable was the robot?

Ignorant	1	2	3	4	5	6	7	Knowledgeable

10. How irresponsible/responsible was the robot?

Irresponsible	1	2	3	4	5	6	7	Responsible

11. How unintelligent/intelligent was the robot?

Unintelligent	1	2	3	4	5	6	7	Intelligent

12. How much did you like the robot?

Very little	1	2	3	4	5	6	7	Very much

Version: 07/10/2014

13. How much did the robot like you?

Very little	1	2	3	4	5	6	7	Very much

14. How comfortable/uncomfortable were you with the robot?

Comfortable	1	2	3	4	5	6	7	Uncomfortable

15. How dishonest/honest was the robot?

Dishonest	1	2	3	4	5	6	7	Honest

16. How trustworthy/untrustworthy was the robot?

Trustworthy	1	2	3	4	5	6	7	Untrustworthy

17. How competitive/cooperative was the robot?

Competitive	1	2	3	4	5	6	7	Cooperative

18. How inattentive/attentive was the robot to you?

Inattentive	1	2	3	4	5	6	7	Attentive

19. How much did you trust the robot?

Very little	1	2	3	4	5	6	7	Very much

20. How attentive/inattentive were you to the robot?

Attentive	1	2	3	4	5	6	7	Inattentive

21. How easy/difficult was it to answer the questions you were asked by the robot?

Easy	1	2	3	4	5	6	7	Difficult

22. To what extent did you feel stressed by the robot?

Very little	1	2	3	4	5	6	7	Very much

23. How hard was it to understand the robot?

Very little	1	2	3	4	5	6	7	Very much

24. How engaged were you with the robot?

Very little	1	2	3	4	5	6	7	Very much

25. How much pressure did you feel from the robot?

Very little	1	2	3	4	5	6	7	Very much

Version: 07/10/2014