The 23rd IEEE International Symposium on Robot and
Human Interactive Communication
August 25-29, 2014. Edinburgh, Scotland, UK,

Conveying Emotion in Robotic Speech: Lessons Learned

Joe Crumpton[1] and Cindy Bethel[2]

Abstract—
This research explored whether robots can use modern speech synthesizers to convey emotion with their speech. We investigated the use of MARY, an open source speech synthesizer, to convey a robot's emotional intent to novice robot users. The first experiment indicated that participants were able to distinguish the intended emotions of anger, calm, fear, and sadness with success rates of 65.9%, 68.9%, 33.3%, and 49.2%, respectively. An issue was the recognition rate of the intended happiness statements, 18.2%, which was below the 20% level determined for chance. The vocal prosody modifications for the expression of happiness were adjusted and the recognition rates for happiness improved to 30.3% in a second experiment. This is an important benchmarking step in a line of research that investigates the use of emotional speech by robots to improve human-robot interaction. Recommendations and lessons learned from this research are presented.

I. INTRODUCTION

Speech is the primary way for people who are working in close proximity to communicate. If the interaction between people and robots working together is to be natural, people and robots will also communicate via speech. While we usually input information into our technological devices via keyboards and mice and receive their output via monitors or screens, the ability of robots to move make such interactions more difficult. Prasad et al. point out that even communication between robots would ideally be via voice while the robots are in the presence of people [1]. Even though robots might communicate more efficiently using a technology such as a wireless Ethernet network, the people working alongside of a robot need to understand the robot's intentions and plans. This situation is similar to a multinational team using a language understood by all the team members to avoid excluding any teammate from interactions within the team.

One important aspect of speech communication between people is the use of vocal prosody. Vocal prosody refers to the non-linguistic attributes of a person's voice while speaking. These attributes include average pitch, pitch range, volume, and speech rate [2]. A listener can use the vocal prosody produced by a speaker to infer the emotions felt by the speaker. For example, a person who is speaking slowly with a constrained pitch range is usually sad while a person who is speaking faster than usual with a high average pitch is either angry or happy. The emotion of the speaker can even cause the listeners to reverse the meaning of a spoken phrase. If the phrase *He is the best candidate* is said in a sarcastic way the listeners will understand that the person speaking does not believe the literal meaning of the phrase.

[1]Doctoral candidate at the Social, Therapeutic and Robotics Lab, Department of Computer Science and Engineering, Mississippi State University, Mississippi State, MS 39762, USA joe.crumpton@msstate.edu

[2]Director of the Social, Therapeutic and Robotics Lab and Assistant Professor in the Department of Computer Science and Engineering, Mississippi State University, Mississippi State, MS 39762, USA cbethel@cse.msstate.edu

TABLE I
EMOTIONS COMMUNICATED VIA VOCAL PROSODY [7], [8], [9]

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

The paper explores the use of a modern speech synthesizer to vary the vocal prosody characteristics of a robot's voice to express emotional intent. Section 2 outlines the communication of emotions through vocal prosody, the use of speech synthesizers to produce emotional speech, and previous research on the use of vocal prosody to communicate emotion by robots. Section 3 describes the two experiments with details on their design, the robot, and the vocal prosody modifications made to communicate emotions. Section 4 presents the results from both experiments. Section 5 contains a discussion of the data produced by the experiments. Section 6 presents lessons learned and recommendations as a result of conducting the experiments.

II. RELATED WORK

A. Recognition of Emotion through Vocal Prosody

The use of vocal prosody to recognize the emotional state of a person speaking has long been a topic of research [3], [4], [5]. The characteristics of speech determined to correlate with the communication of emotions are called the *Big Three* of vocal prosody: pitch, timing, and loudness [6]. Each of the three characteristics can be measured and quantified in many different ways. For example, pitch can be measured and reported as average pitch, the maximum and minimum pitches, the range of pitches, and the contour of the average pitch during an utterance. The many possible combinations of pitch, timing, and loudness allow for the communication of different emotions. Table I shows the vocal prosody characteristics that are typically used to express Ekman's *Big Six* emotions: happiness, surprise, sadness, anger, disgust, and fear [7], [8], [9], [10].

B. Expression of Emotions by Speech Synthesizers

The generation of emotional speech by speech synthesizers was first discussed by Cahn in the 1990s [11]. Early speech

synthesizers were limited in how users could manipulate the vocal prosody of the synthesized speech. Recent speech synthesizers are based on the use of Hidden Markov Models (HMM) of the phonemes to represent sounds to be created in order to generate a speech segment [12]. The models can be manipulated by the speech synthesizer to raise or lower the pitch and speed up or slow down the generated speech. The machine-like quality (or *buzziness*) of the synthesized speech is the tradeoff for the increased ease in speech manipulation [12].

C. Emotion in Robotic Speech

Much of the research in the Human-Robot Interaction (HRI) field has focused on robots listening to and responding to their human users [13], [14], [15], [1], [16]. Breazeal and Aryananda note that the vocal prosody used by a human when speaking to a robot can allow the robot to recognize praise, prohibition, attention, and comfort messages [13]. This paper presents research that investigates the opposite communication channel: from robot to human instead of human to robot.

Prior investigations that involved the communication of emotion by robots through the use of vocal prosody have focused on non-linguistic utterances [17], [18]. Read and Belpaeme state that non-linguistic utterances are not computationally expensive to generate and modify to fit a particular emotion. They also assert that people who speak different languages will correctly interpret the utterances. However they recognize "the shortcomings in comparison to natural language are obvious" [18], referring to the small amount of information conveyed by non-linguistic utterances as opposed to statements spoken in a natural language. While the use of non-linguistic utterances to communicate emotion are useful in some situations; robots that perform complex tasks in cooperation with people need the expressiveness of natural language to communicate their action plans and intentions.

III. Experiments

This section details two experiments conducted as part of this research. The experimental design, robot, survey materials, and tasks performed were the same for both experiments.

A. Design

These experiments were within-subjects designs that evaluated the detection of emotion in semantically unpredictable sentences (SUS). The proposed hypothesis was:
H_1: Participants will recognize the emotional intent of a statement based on vocal prosody alone.

B. 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 Fig. 1) [20]. 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. The Survivor Buddy robot consists of a small monitor manufactured by Mimo Monitors, Inc. mounted to the end of an arm. The arm contains four Robotis Dynamixel 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.

Fig. 1. Image displayed during Survivor Buddy introduction [19]

The experiments were conducted using the *Wizard-of-Oz* technique [21]. 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 routines to perform routine functions 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 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.

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 Fig. 2 illustrates, the smile was removed to avoid a bias toward "happy" emotions. An image (shown in Fig. 1) of the Survivor Buddy mounted to a mobile robot base was shown to the participants while the Survivor Buddy robot introduced itself to the participants. The robot explained that it was meant to be used to communicate with people trapped by rubble. 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.

C. Tasks

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

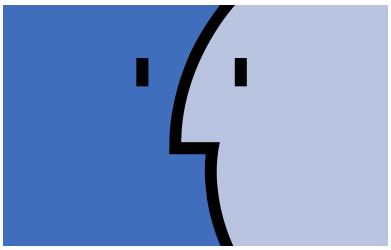

Fig. 2. Image displayed as a face on the Survivor Buddy monitor

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)

Fig. 3. Sentence structures for semantically unpredictable sentences [22]

sadness. 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."

Semantically unpredictable sentences (SUS) [22] 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 Fig. 3). 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.

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.

D. Data Collection

Each participant completed an informed consent form, a demographics questionnaire, 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. 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.

E. Voice Modification for the Initial Experiment

These experiments utilized MARY (Modular Architecture for Research on speech sYnthesis), an open source speech synthesizer designed to produce expressive speech [23]. The pitch and volume of the synthesized speech are 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 [24]. 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 is not conveying 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 uses a speech rate of 75% and a volume of 60%. These values allow 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 [7], [8], [9]. For example, Tao et al. 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 II shows the modifications made to the standard voice that were used in the first experiment .

TABLE II
CHANGES MADE TO STANDARD VOICE TO CONVEY EMOTIONS IN INITIAL EXPERIMENT

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

TABLE III
CHANGES MADE TO STANDARD VOICE TO CONVEY EMOTIONS IN SECOND EXPERIMENT

Emotion	Pitch	Pitch Range	Pitch Contour	Speech Rate	Volume
Anger	-50Hz	120%	each word has 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%	varies between -5% and +25%	varies between 70% and 90%	80%
Sadness	-30Hz	70%	falling	50%	40%

F. Voice Modification for the Second Experiment

The program that specified the speech synthesizer's vocal prosody was modified and a second experiment was conducted. The essential 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. Table III shows the vocal prosody changes made to the default voice to express each of the five emotions.

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 has 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 is calculated over the entire sentence so that pitch rises and falls in a smooth pattern as recommended by Burkhardt and Sendlmeier [25]. The new "happiness" vocal prosody's speech rate also varies 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.

IV. RESULTS

A. Initial Experiment

Thirty-three university students (17 females and 16 males) participated in this experiment. Their average age was 19.7 years old (SD = 2.18). Table IV 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 V gives the results of a one sample *t*-test (α=0.05) for each of the emotion recognition rates. The test value used in the one sample *t*-test was 0.2, the recognition rate that results from random guessing.

TABLE IV
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	19.7	18.2	**36.4**
Sadness	29.5	19.7	0.8	0	**49.2**

TABLE V
STATISTICAL SIGNIFICANCE OF EMOTION RECOGNITION RATES IN INITIAL EXPERIMENT

Emotion	t	df	p (2-tailed)	Cohen's *d*
Anger	9.727	32	<0.001	1.69
Calm	12.969	32	<0.001	2.26
Fear	2.844	32	0.008	0.49
Happiness	-0.521	32	0.606	-0.09
Sadness	4.790	32	<0.001	0.83

B. Second Experiment

Nineteen university students (11 females and 8 males) participated in the second experiment. Data collection was stopped short of the goal of 30 participants due to a lack of availability of participants. The participants' average age was 18.7 years old (SD = 1.06). Table VI is a confusion matrix that displays the classification of sentences said for the intended emotions across all participants. Table VII gives the results of a one sample *t*-test (α=0.05) for each of the emotion recognition rates. The test value used in the one sample *t*-test was 0.2, the recognition rate that results from random guessing.

TABLE VI
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 VII
STATISTICAL SIGNIFICANCE OF EMOTION RECOGNITION RATES IN SECOND EXPERIMENT

Emotion	t	df	p (2-tailed)	Cohen's *d*
Anger	9.939	18	<0.001	1.90
Calm	8.767	18	<0.001	2.01
Fear	3.531	18	0.002	0.81
Happiness	3.970	18	0.001	0.91
Sadness	2.014	18	0.059	0.46

V. DISCUSSION

A. 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 V). These rates were comparable to the successful emotion recognition rate (60%) of people listening to human speakers [26]. The recognition rates for fear (37.8%) and sadness (49.2%) were significantly higher than chance but are lower than the recognition rates of anger and calm.

The most surprising result is the recognition rate for happiness (18.2%). Not only is 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 [27].

This result lead to more research into the expression of happiness through vocal prosody. Frick [28] notes that speech expressing happiness "is often described as containing gentle contours in pitch." This idea was repeated by Burkhardt and Sendlmeier [25] 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.

B. Second Experiment

After changes were made to the vocal prosody characteristics for calm, fear, and happiness (see Tables II and III), 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 VII. 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 H_0 was that participants would not be able to recognize the emotional intent of a statement based on vocal prosody alone. H_0 was rejected after the intended emotion recognition rate for four of the five emotions was 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 is a more troubling mistake and could lead to many misunderstandings.

VI. CONCLUSIONS

The intersection of human-robot interaction and emotional speech synthesis is an exciting, but often difficult, area of research. This section details some of the lessons learned while preparing and conducting these experiments to aid and further this research effort by others considering similar lines of inquiry.

First, carefully check that commands controlling the speech synthesizer are affecting the output of the speech synthesizer. Initially, TOne and Break Indices (ToBI) [29], [30] was considered as the markup language to specify the changes in vocal prosody in the speech synthesizer's input. In informal listening tests, the output of the MARY speech synthesizer did not appear to change the pitch of statements in response to the ToBI markup. The Praat software system [31] 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 [32] 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.

Second, the use of vocal prosody by people to communicate emotions is a heavily researched field. The features of vocal prosody correlated with different emotions have been identified by many researchers. However, the literature often reports the changes in vocal prosody without magnitudes or units. For example, Scherer reports that pitch variability is "$\leq$" for sadness and "$\geq\geq$" for fear [5]. Articles such Hammerschmidt and Jürgens [7] provide detailed information (including measurement units, means, and standard deviation) about the vocal prosody correlates for different emotions. This information was invaluable for the manipulation of these characteristics in synthesized speech.

Finally, semantically unpredictable sentences were used in this experiment 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. Additional data analysis is required to determine with certainty that the semantically unpredictable sentences were neutral in their linguistic content.

ACKNOWLEDGMENTS

The authors would like to thank the STaRS lab members who assisted with data collection: John Kelly, Kayla Huddleston, Malcolm McCullum, and Kaleb Stuart. We also thank the Texas A&M Department of Computer Science and Engineering for the use of the Survivor Buddy robot.

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

[12] K. Tokuda, Z. Heiga, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proceedings: 2002 IEEE Workshop on Speech Synthesis*, 2002, pp. 227–230.

[13] C. Breazeal and L. Aryananda, "Recognition of affective communicative intent in robot-directed speech," *Autonomous Robots*, vol. 12, no. 1, pp. 83–104, 2002.

[14] 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," in *Proceedings: Workshops at Twenty-Sixth AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence, 2012.

[15] E. Kim, D. Leyzberg, K. Tsui, and B. Scassellati, "How people talk when teaching a robot," in *Proceedings: 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, 2009, pp. 23–30.

[16] O. Rogalla, M. Ehrenmann, R. Zöllner, R. Becher, and R. Dillmann, "Using gesture and speech control for commanding a robot assistant," in *Proceedings: 11th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2002, pp. 454–459.

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

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

[19] M. Cimons. (2011) Behind the scenes: Robots to the rescue. [Online]. Available: http://www.livescience.com/13089-scenes-robots-rescue-bts-110304.html

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

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

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

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

[24] M. Schröder. MaryXML. [Online]. Available: http://mary.dfki.de/documentation/maryxml

[25] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *Proceedings: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. International Speech Communication Association, 2000.

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

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

[28] R. W. Frick, "Communicating emotion: The role of prosodic features," *Psychological Bulletin*, vol. 97, no. 3, pp. 412–429, 1985.

[29] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *Proceedings: 2nd International Conference on Spoken Language Processing (ICSLP)*, vol. 2. International Speech Communication Association, 1992, pp. 867–870.

[30] M. E. Beckman and G. M. Ayers. Guidelines for ToBI labelling. [Online]. Available: http://www.speech.cs.cmu.edu/tobi/

[31] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[32] A. W. Black. CMU_ARCTIC speech synthesis databases. [Online]. Available: http://festvox.org/cmu_arctic/