Int J of Soc Robotics
DOI 10.1007/s12369-015-0329-4

SURVEY

A Survey of Using Vocal Prosody to Convey Emotion in Robot Speech

Joe Crumpton[1] · **Cindy L. Bethel**[2]

Accepted: 18 November 2015

Abstract The use of speech for robots to communicate with their human users has been facilitated by improvements in speech synthesis technology. Now that the intelligibility of synthetic speech has advanced to the point that speech synthesizers are a widely accepted and used technology, what are other aspects of speech synthesis that can be used to improve the quality of human-robot interaction? The communication of emotions through changes in vocal prosody is one way to make synthesized speech sound more natural. This article reviews the use of vocal prosody to convey emotions between humans, the use of vocal prosody by agents and avatars to convey emotions to their human users, and previous work within the human–robot interaction (HRI) community addressing the use of vocal prosody in robot speech. The goals of this article are (1) to highlight the ability and importance of using vocal prosody to convey emotions within robot speech and (2) to identify experimental design issues when using emotional robot speech in user studies.

Keywords Synthesized speech · Emotional robot speech · Human–robot interaction · Vocal prosody

✉ Joe Crumpton
crumpton@dasi.msstate.edu

Cindy L. Bethel
cbethel@cse.msstate.edu

[1] Distributed Analytics and Security Institute, Mississippi State University, Starkville, USA

[2] Social, Therapeutic and Robotic Systems Lab, Department of Computer Science and Engineering, Mississippi State University, Starkville, USA

1 Introduction

Robot 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 [17]. 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 is receiving considerable attention [19,34,41,54,61,65,67,72]. 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 [61].

One of the areas of robot speech, and synthesized speech in general, that can be improved to make the generated speech more natural sounding is the use of vocal prosody. Vocal prosody is the way something is said (pitch, timing, loudness, etc.) as opposed to the actual linguistic meanings of the spoken words [35]. In human communication, vocal prosody is considered one of the paralinguistic components of speech [57]. 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 [57].

Vocal prosody is an essential component of speech communication between humans [81]. 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 [29,36,75]. The emotions or mood being conveyed by speech can be crucial to interpreting the meaning of a speaker's mes-

Published online: 07 December 2015

sage. For example, a tone of sarcasm can be used to signal a listener that a spoken statement should be interpreted as the opposite of the literal meaning of the words being said [94]. In addition to short term states 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 [57,81].

Section 2 presents two types of emotion models used to describe emotions and the communication of emotions via vocal prosody from person to person. Different methods of speech synthesis, different standards used to specify vocal prosody parameters to speech synthesizers, and the importance of the use of vocal prosody by devices other than robots are reviewed in Sect. 3. Section 4 summarizes human–robot interaction research on the use of vocal prosody in robot speech. Issues to avoid in human–robot interaction research on the communication of emotions through vocal prosody are described in Sect. 5.

2 Expression of Emotions Using Vocal Prosody

An important aspect of using vocal prosody in synthetic speech is to understand and summarize the features that people use to communicate emotional intent to listeners. This section starts with a review of two models that describe emotions: basic emotion models and dimensional emotion models. The section concludes with a description of the vocal prosody attributes that correspond with the communication of emotions through speech.

2.1 Emotion Models

Typically emotions are described using basic emotions models or dimensional emotion models. A basic emotions model (sometimes referred to as distinct or discrete emotions model [5,76]) ascribes each emotion "its own specific physiological, expressive, and behavioral reaction patterns" [76]. Ekman's *Big 6* basic emotions (*happiness*, *surprise*, *sadness*, *anger*, *disgust*, and *fear* [27,62]) are an example of a basic emotions model. An advantage of the basic emotions model is its universality. There is evidence that emotions are recognized pan-culturally [27,62]. Most people typically use a small set of labels (anger, fear, sad, happy, etc.) when describing their emotions [9,76]. One drawback of the basic emotions model is its complexity when describing how each emotion is expressed. Each basic emotion is characterized by its own unique neuromotor expression program [74] that produces the changes in facial expressions and vocal prosody that others use to recognize the speaker's emotional state.

Another set of models for the explanation of emotions are dimensional emotion models. In dimensional models, emotions are characterized by their locations on one or more continuous abstract scales [76]. Typically two (such as valence and activation [73]) or three (such as pleasure, arousal, and dominance [49]) dimensions are required to adequately differentiate the many possible emotions [76]. Dimensional models of emotion have several attractive features. The differing intensities of an emotion can be represented as different values along the dimensions. For example, extreme sadness can be differentiated from mild sadness by moving along the valence dimension. Another advantage of dimensional models is that the representation of two emotions expressed simultaneously (such as surprise and fear) can be accommodated by showing the expressed emotional state as a point in space between the accepted coordinates of the two emotions. The main disadvantage of the dimensional emotion models is their lack of use by the general public. It is rare for someone to describe or identify their current emotional state as *low valence and high activation*. Most people describe their emotional state using a basic emotion label such as *angry* [9,76].

In spite of the differences between the basic emotions models and the dimensional models of emotion, the models do share some similarities. Bachorowski and Owren point out that both models expect that the emotional state of a speaker will be "encoded in vocal acoustics" that allow listeners to interpret the speaker's emotions [5]. Emotions will be described in the remainder of this article using the basic emotions model for the following reasons. First, participants in human–robot interaction studies can be expected to know a set of basic emotions from their common experience. Using a dimensional emotions model when asking participants to describe an emotion would require training the participant on the meanings of the dimensions and giving examples of how common emotions (anger, happiness, etc.) can be expressed with the dimensional emotional model. Second, Laukka has found that vocal expressions of emotions are categorically perceived and there are crisp boundaries between participants' perception of different emotions in speech [42]. Laukka's findings "suggest that a discrete-emotions approach provides the best account of vocal expression" [42].

2.2 Vocal Prosody Use Between Humans

The features of speech that are used by humans to convey emotion must be identified and quantified before speech synthesizers can use vocal prosody to convey emotions. Early research consisted of listeners detecting and classifying the features of speech that accompanied emotional speech [29,36,75]. More recent research has applied statistical and machine learning techniques such as Linear Discriminant Analysis, Support Vector Machines, and AdaBoosted Decision Trees to determine which features can be used to classify emotional speech [43,47,48,56]. Pitch, timing, and loudness are the features of speech that are typically found to correlate with the expression of emotion through vocal prosody. Pitch,

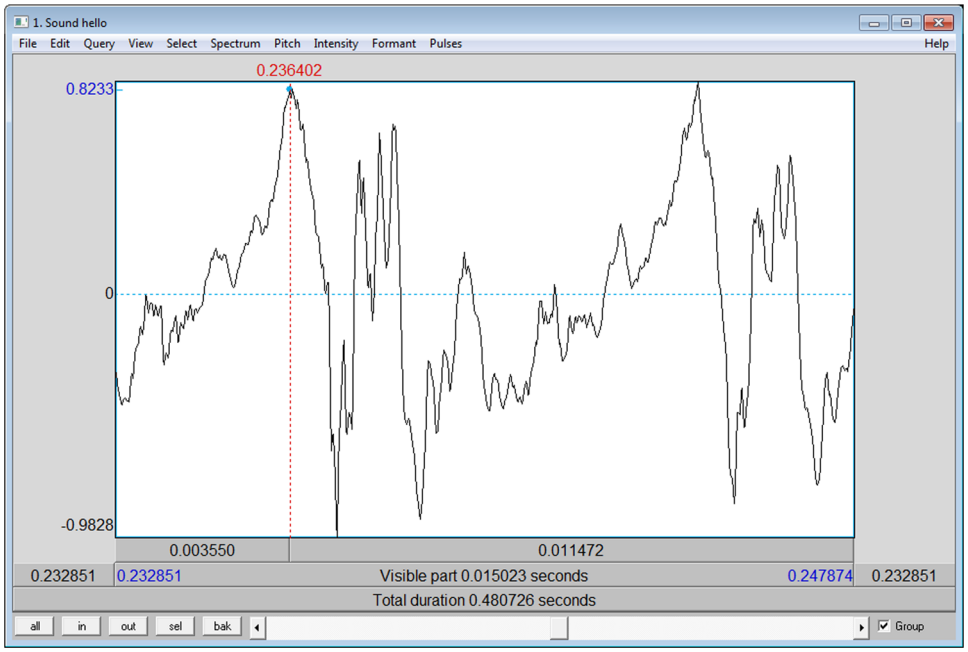

Fig. 1 Sinusoid waveform from the recording of the first 0.015 s of the word "Hello"

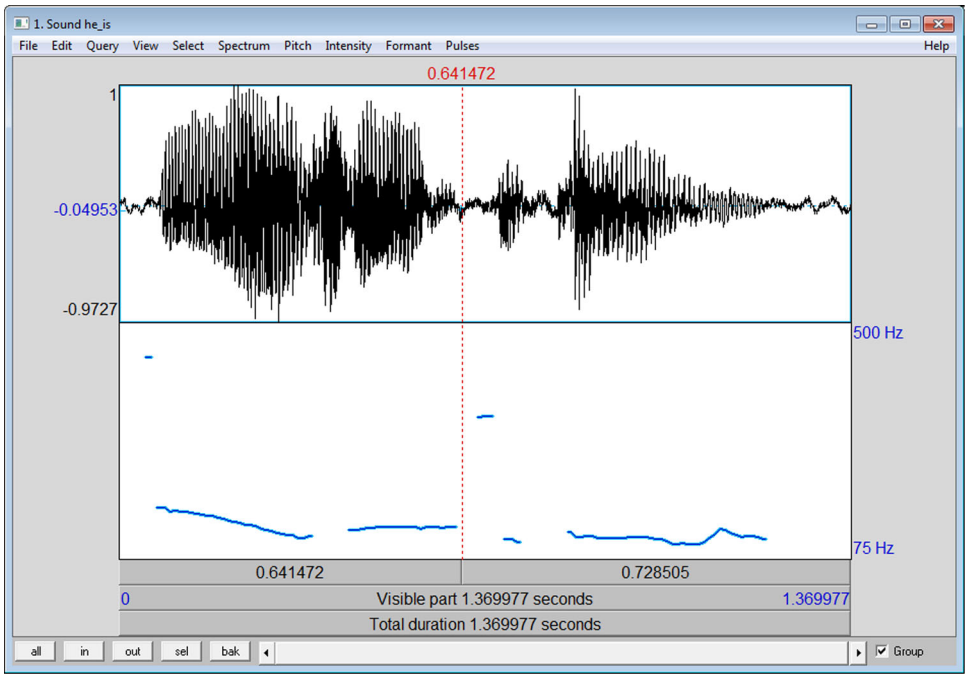

Fig. 2 Pitch contour of the statement "He is at the game"

timing, and loudness are sometimes referred to as *The Big Three* of vocal prosody [90].

Pitch corresponds to the frequency at which the vocal folds vibrate when a person is speaking [30]. The sound of the human voice is not a simple signal consisting of one sinusoid however. The shape of the speaker's vocal tract reinforces some frequencies and dampens other frequencies. The result is a complex sinusoid as shown in Fig. 1. Note that Fig. 1 shows only the first 0.015 seconds of the complex sinusoid from a recording of the word "Hello". Typically the wave form is shown for an entire word or statement (as in the top half of Fig. 2) and the wave is too compressed to 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 [36].

Not only is the fundamental frequency of a speech segment important, both the range of frequencies and the change in

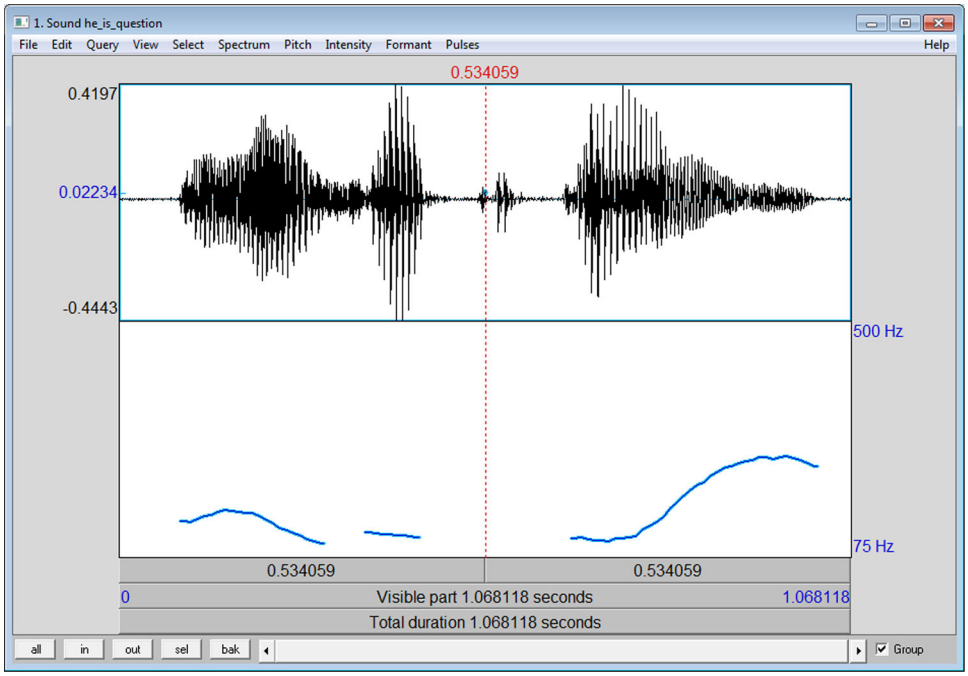

Fig. 3 Pitch contour of the question "He is at the game?"

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 [77]. 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 [94]. In Figs. 2 and 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 such as "He is at the game", the speaker's pitch usually falls at the end of the statement as shown in Fig. 2. A question ("He is at the game?") is often accompanied by a rise in the pitch contour as shown in Fig. 3.

Timing is concerned both with how fast a person is speaking and the pauses within a statement. The speed of someone 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 [36].

Loudness is a measure of the volume at which a person is speaking. Loudness is also referred to as *intensity* in some literature [20,30,43]. Loudness is typically measured in decibels (dB), a logarithmic unit of measure that gives the ratio between two values. A soft voice (described with a low value for loudness) can indicate boredom while a loud voice indicates emotions such as happiness or anger [75].

Table 1 Emotions and associated vocal prosody characteristics [32,83,85]

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

As is evident from the previous descriptions of pitch, timing, and loudness, the expression of a single emotion can affect one or more of the measures simultaneously. Table 1 gives a summary of the changes in vocal prosody that have been found to accompany the expression of Ekman's *Big 6* basic emotions: *happiness*, *surprise*, *sadness*, *anger*, *disgust*, and *fear* [27,62]. Experiments have shown that people can recognize the emotion being communicated by another person's vocal prosody at a level much higher than chance [59, 77]. 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 approximately 60 % of the time [77]. The percent correct expected from guessing varied between 10 and 25 % based on the number of emotions 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 natural speech from common interactions [43]. 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 [22]. 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 [22].

3 Vocal Prosody in Speech From Devices to Humans

This section reviews the use of vocal prosody in speech produced by devices other than robots. First, the most popular methods of speech synthesis are explained. Second, text markup standards for specifying vocal prosody parameters are reviewed. Finally, results from two studies that demonstrate the importance of vocal prosody use in synthesized speech are presented.

3.1 Methods of Speech Synthesis

The generation of affective speech using the manipulation of vocal prosody features has been a subject of speech synthesis research for over twenty years [21,28,37,52,58,80]. Early work was hindered by the lack of capabilities of then state-of-the-art speech synthesizers to allow changes to features such as pitch range and pitch contour of the generated speech [21]. Many of the early speech synthesizers were concatenative speech synthesizers. Concatenative speech synthesizers analyze the text to be spoken for phonemes. Phonemes are the smallest distinct units of sound in a language. The concatenative speech synthesizer selects the corresponding sounds for each phoneme in a statement from a sound database. The selected sounds are concatenated together to generate the sound of speech. This process made it difficult to manipulate the pitch and speed of the sounds that comprised the synthesized speech and still maintain the intelligibility of the resulting speech. An effective but data intensive approach has been to create a database of phonemes spoken by a user while expressing each emotion.

As speech synthesizers have become more advanced, the ability to convey emotion within generated speech has improved. Hidden Markov Models (HMM) of phonemes are the basis of several speech synthesizers [87]. A series of the HMM representations of phonemes are used to generate speech. The models are easily manipulated by the speech synthesizer to alter vocal prosody parameters such as pitch and speech rate. A drawback of HMM-based speech synthesizers is a decrease in the quality of the generated speech when compared to the speech produced by concatenative speech synthesizers [87]. Zen et al. attributed the reduced quality of speech generated by HMM-based speech synthesizers to three factors: vocoding (analysis/synthesis algorithms), accuracy of acoustic models, and over-smoothing (detailed speech characteristics are not reproduced by the synthesizer) [33,96].

The difference in the size of a concatenative voice model versus a HMM voice model is significant. The difference in data size can be seen by examining the voice models based on the recordings provided by the Language Technologies Institute at Carnegie Mellon University [15]. One of the available databases of voice recordings is named `slt`. The `slt` recordings consist of 1132 phrases spoken by a US English speaking female. The concatenative voice model produced from the `slt` data is 132 megabytes of data and the HMM-based voice model is only 1.6 megabytes of data. The `slt` recordings were made by the speaker expressing a calm or neutral emotion. The speaker would have to say the 1132 phrases in each of the desired emotions to create a concatenative voice model capable of expressing emotions through vocal prosody. The resulting concatenative voice model would be much larger in size than the model that can only express the one emotion. The HMM-based voice models do not suffer from this same increase in size for each desired emotion. As long as the changes in vocal prosody can be expressed in changes to the fundamental values of pitch, timing, and volume required by the HMM-based voice model to produce speech, the one HMM-based voice model can express any of the basic emotions.

Following the success of deep learning in areas such as image and speech recognition, deep learning has been utilized in speech synthesis. Deep joint models such as deep belief networks (DBN) and deep conditional models such as deep neural networks (DNN) have been used instead of HMM to generate the acoustic models used to synthesize speech [97]. Subjective listening tests have shown that the speech produced by deep learning-based speech synthesizers is more natural sounding than the speech produced by HMM-based speech synthesizers [96]. The improvement in speech quality has been attributed to the ability of deep learning models to "better ... describe the complex and nonlinear relationships between the inputs and targets" in a speech synthesizer [97]. While impressive demonstrations[1] of deep learning-based speech synthesis research are available, the authors are not aware of any currently available commercial or open-source speech synthesizers that utilize deep learning.

The previous speech synthesis methods (concatenative, HMM-based, deep learning-based) can be adapted to directly generate emotional speech in addition to neutral sounding speech. For concatenative synthesizers, the database of phonemes is expanded to include the sound of each phoneme when expressed with each of the targeted emo-

[1] http://www.zhizheng.org/demo/dnn_tts/demo.html.

tions. For HMM and deep learning-based synthesizers, the vocal prosody parameters (pitch, timing, and loudness) can be specified in the synthesizer's input to affect the sound of the generated speech. Another method that can create synthesized speech with varying vocal prosody is voice conversion. Voice conversion is used after speech synthesis to convert neutral sounding speech to emotional speech [85]. In voice conversion, models of vocal prosody parameters learned from examples of emotional speech are used to modify the synthesized neutral sounding speech. An anticipated advantage of voice conversion over the previously described synthesis techniques is that more subtle differences in the vocal prosody of emotional speech such as voice quality can be included in the conversion process [1]. In practice, voice conversion is not always able to produce synthesized speech that is both natural sounding and contains recognizable emotional intent [88].

3.2 Specification of Vocal Prosody

One of the major difficulties to overcome in the Text-to-Speech (TTS) field related to 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 [46]. Given the increase in computing power and the decrease in the cost of digital storage, recent research has employed machine learning techniques on large text corpora to generate models for the prediction of the intended emotional content of text [2,46,84].

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 [82]. ToBI transcriptions consist of the text being spoken along with markup in specific tiers that convey prosodic information. The most commonly used tiers are the tone tier and the break-index tier. Fig. 4 shows an example of the text of a question along with the tone and break-index tiers [12].

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 Fig. 4 shows that the first syllable in marmalade was

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

Fig. 4 Example of ToBI markup

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 Fig. 4 shows that the phrase *will you have marmalade* 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 Fig. 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 Fig. 4 corresponds to the pause that the comma represents in the question. Note that phrase accents (L- or H-) 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 software 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 [16]. Praat also has tools for labeling the tone and break-index tiers of a speech segment. Figures 5 and 6 are two examples of speech segments and the corresponding ToBI markup displayed by Praat [89].

High quality speech synthesis is an important part of making the world wide web accessible to people with impaired vision. The World Wide Web Consortium 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) [6,91,93] 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 gender, age, etc. Fig. 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 system is responsible for the choosing how the vocal prosody of the generated speech should be changed to meet the directives contained in the SSML markup.

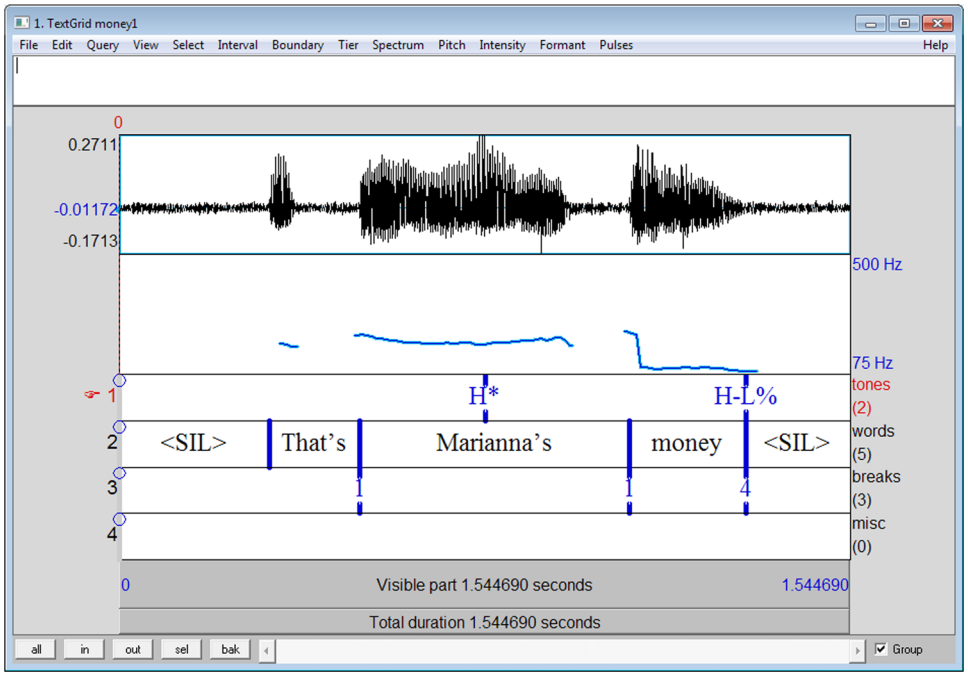

Fig. 5 Screen capture of Praat Showing ToBI Markup of a Statement

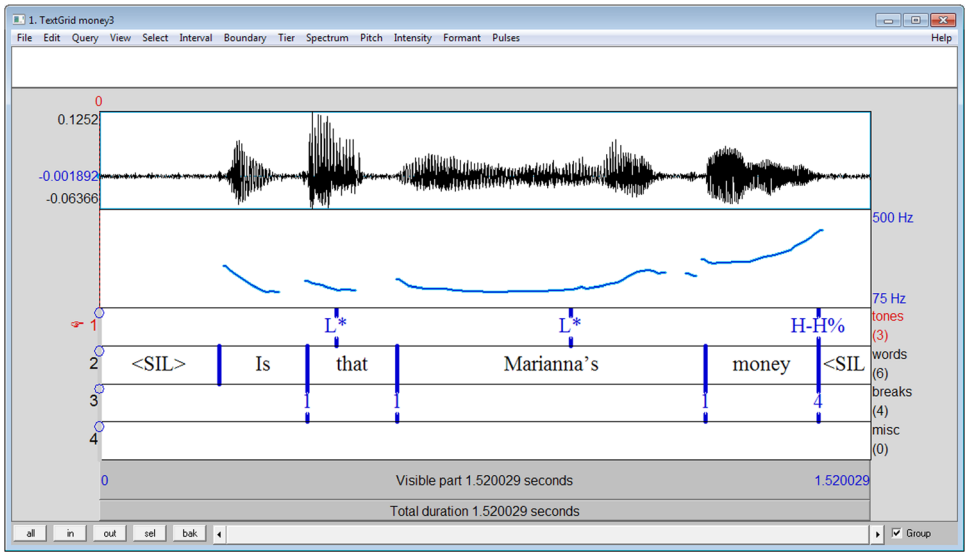

Fig. 6 Screen Capture of Praat showing ToBI markup of a question

```
<?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>
```

Fig. 7 Specifying vocal prosody using SSML

```
<?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>
```

Fig. 8 Pitch contour specified using SSML

```
<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>
```

Fig. 9 Sample of EmotionML

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 8 illustrates the SSML markup that specifies that a question should be asked at a greater than normal speed with a specific pitch contour.

EmotionML is another markup language proposed by the World Wide Web Consortium to direct the expression of emotion [7]. 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 [79]. 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 [4,27,49]. 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 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 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" [7].

Since there is not just one standard for marking up text with vocal prosody information, speech synthesizers often support input expressed in 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 [78].

3.3 Importance of Vocal Prosody in Synthesized Speech

The importance of appropriate vocal prosody in synthesized speech has been shown in several experiments. Nass et al. [53] 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 [26]. 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 [26].

4 Vocal Prosody in Speech 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 robot speech. First, previous research has shown that people prefer to communicate with robots via voice and they prefer that the voice be human-like [25,39,64]. Second, taking advantage of the ability of humans to perceive emotions in speech may increase the effectiveness of robot 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. The use of vocal prosody to communicate emotion would be most applicable to social and anthropomorphic robots, but even non-anthropomorphic robots that utilize synthesized speech could use vocal prosody when communicating with human users.

There has been support for robots expressing emotion and intention in order for the robots to be perceived as "believable characters" by humans [10,18]. Researchers have designed systems to calculate and express the emotional state of the robot in response to its environment and interactions [17,63,86]. For example, the Kismet robot would lower its head and/or frown when receiving negative feedback from a study participant [17]. 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 robot 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.

Much of the work in the HRI field concerning vocal prosody use by robots has concentrated on conveying emotions by varying the vocal prosody of simple non-linguistic utterances. Read and Belpaeme [66] found that people interpret human-like utterances made by robots as expressing emotions. Oudeyer [56] created algorithms that could 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 [56] and the utterances should be understandable across cultures and languages [65].

Motion picture characters such as R2-D2 and WALL-E use non-linguistic utterances. The popularity of these characters inspired the use of non-linguistic utterances in HRI studies [14,65]. 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 [65]. 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 [68]. 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 [70]. 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 [70].

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" [65]. 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 audi-

ence 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 min" 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 addition to natural language as opposed to non-linguistic utterances replacing the use of natural language by robots [69].

As expressive speech synthesizers became more readily available, several proposals for the use of emotional natural language speech by robot systems were made [71,95]. 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 [44]. 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 [86]. Note that the referenced studies are not attempting to convince the study participants that the robots actually have emotions, the studies are investigating how the use of emotional speech by a robot changes the actions of people who are interacting with the robots.

Recent research has shown that people's impressions of a robot can be influenced by the pitch of the robot's voice [54]. 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 [54]. 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 [54].

5 Discussion

This section contains a discussion of issues that researchers should consider when designing studies concerning the use of vocal prosody by robots to communicate emotions. The issues were identified during the review of previous human–robot interaction studies (see Sect. 4) and during the authors' design of a study to verify that study participants would correctly recognize the emotional intent of robot speech [23,24]. While the recommendations concerning the issues discussed below are most applicable to social and anthropomorphic robots, the recommendations should also be applicable to non-anthropomorphic robots that utilize synthesized speech to communicate with human users.

5.1 Embodiment Effects of Robots

A common trend in research concerning vocal prosody and robots is the use of pictures or on-screen avatars instead of actual physical robots [51,66]. Some of the research was conducted without the mention of robots to participants while the results were used to make recommendations about robot voices [34]. The use of avatars instead of physical robots is understandable. On-screen avatars are less expensive and easier to program than actual robots. However, 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 participants react to an image represnting a robot differently than collocated physical robots [38,40,60]. For example, Kidd and Breazeal state that a collocated "robot was more engaging and rated more highly on the scales of perceptions than the animated character" [40].

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 [38,40,60]. 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" [40]. In contrast, Bainbridge et al. [8] 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 [8]. One explanation for the differences observed between these studies might be the difficulty of task being performed by study participants. The earlier experiments involved relatively simple tasks such as placing blocks on top of each other or interacting with a robotic dog [38,40]. In the Bainbridge et al. experiment, the participants received instructions from the robot concerning moving books from one location

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. 10 Sentence structures for semantically unpredictable sentences [13]

to another location within a small office [8]. Wainer et al. [92] 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 [45]. 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.

5.2 Avoid Confounding Factors when Portraying Emotions

Study participants are adept at recognizing emotional intent through several modalities [3]. If a robot is using an emotional voice and also expressing its emotional intent through the literal meaning of its statements, its facial expression, or body language then the effect of the emotional voice is entangled with the effects of the other modalities. If a study is intended to investigate the effects of vocal prosody to communicate emotions, the effects of the other modalities to communicate emotion must be controlled. For example, the robot's facial expression should be chosen to not communicate a specific emotion.

Avoiding the communication of emotions through the linguistic meaning of a robot's speech can be difficult especially when a robot is communicating via natural language. One solution is to use sets of Semantically Unpredictable Sentences (SUS) [13] as the text to be spoken by the robot during an experiment. The SUS Test was proposed Benoit et al. as a way to measure the intelligibility of text-to-speech systems [13]. Sets of semantically unpredictable sentences are useful in an intelligibility test because the listener can not predict the words appearing in a sentence from any of the previous words. The property of being semantically unpredictable is also useful for experiments concerning the communication of emotion via vocal prosody. If the individual words in a semantically unpredictable sentence does not imply an emotion, the linguistic content of the sentence should not imply a particular emotion.

Semantically unpredictable sentences are created by first choosing short, commonly used words for several parts of speech: noun, verb, adjective, and determinative. As mentioned above, words such as *cry* or *shrieked* should be excluded because the individual word implies an emotion. The words are then placed into one of five sentence structures (see Fig. 10). Examples of semantically unpredictable sentences are:

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

5.3 Generation and Validation of Emotional Robot Speech

Researchers can generate emotional speech for a robot by either using an "emotional" voice model provided with a speech synthesizer or by using markup such as SSML (see Sect. 3.2) to specify vocal prosody parameters that will express the intended emotion. Several commercially available speech synthesizers such as Acapela Group's Acapela[2] and Cereproc's CereVoice[3] contain voices claimed to portray different emotions. However, the companies do not provide empirical evidence showing that listeners actually perceive the emotion meant to be portrayed by the generated speech. It would be up to a researcher using a provided "emotional" voice to validate that participants perceive the emotion that the voice model is intended to convey.

While the correlates of vocal prosody parameters and the intended emotion are known (see Table 1), using markup such

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

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

as SSML to vary vocal prosody parameters to express emotions in synthesized speech is problematic. Currently several commercial speech synthesizers such as such as Microsoft's .NET speech synthesizer[4] and Nuance Communication's Dragon Mobile[5] 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 [50,55]. If a researcher is not sure if input markup is being utilized by a speech synthesizer, a software tool such as Praat [16] can be used to verify that changes in the input markup are reflected in the generated speech. Praat, used for many of the earlier figures in this article, can display the pitch contour of a sound file and makes objectively comparing two different segments of synthesized speech possible.

A potential weakness to avoid in studies of robots using emotional voices is the failure to validate the emotional voices. Tielman et al. used arousal and valence parameters to modify a robot's speech while the robot was interacting with children [86]. But the researchers did not check that the children could correctly interpret the emotional intent of the robot's speech. Before claiming that specific emotions expressed by the robot changed how participants interacted with the robot, the researcher must validate that the participants are interpreting the robot's emotional intent correctly. Beale and Creed made a similar criticism of research on the use of emotion by agents and on-screen characters [11].

6 Conclusion

In the previous sections, the use of vocal prosody to communicate emotion by humans was examined, the ability of speech synthesizers to produce varying vocal prosody was reviewed, and the previous work within the HRI community on the use of vocal prosody to convey emotions in robot speech was discussed. Most importantly, several issues that researchers should consider when planning experiments involving emotional voices by robots and human–robot interaction were identified:

- Embodiment effects of robots
- Avoid confounding factors when portraying emotions
- Generation and validation of emotional robot speech

The use of vocal prosody to convey emotions in robot speech is another tool that robot designers and programmers can utilize to increase the quality and naturalness of human–robot interaction.

[4] https://msdn.microsoft.com/.

[5] http://dragonmobile.nuancemobiledeveloper.com/.

References

1. Aihara R, Takashima R, Takiguchi T, Ariki Y (2012) GMM-based emotional voice conversion using spectrum and prosody features. Am J Signal Process 2(5):134–138
2. Alm CO, Roth D, Sproat R (2005) Emotions from text: machine learning for text-based emotion prediction. In: Proceedings of the human language technology conference and conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 579–586
3. Amir N, Weiss A, Hadad R (2009) Is there a dominant channel in perception of emotions? In: Proceedings of the 3rd international conference on affective computing and intelligent interaction and workshops (ACII). Association for the Advancement of Artificial Intelligence, pp 1–6
4. Ashimura K, Baggia P, Burkhardt F, Oltramari A, Peter C, Zovato E. (2013) Vocabularies for EmotionML http://www.w3.org/TR/2012/NOTE-emotion-voc-20120510/
5. Bachorowski JA, Owren MJ (2008) Vocal expressions of emotion. In: Lewis M, Haviland-Jones JM, Barrett LF (eds) Handbook of emotions, 3rd edn. The Guilford Press, New York, pp 196–210
6. Baggia P, Bagshaw P, Bodell M, Huang DZ, Xiaoyan L, McGlashan S, Tao J, Jun Y, Fang H, Kang Y, Meng H, Xia W, Hairong X, Wu Z (2010) Speech synthesis markup language (SSML) version 1.1. http://www.w3.org/TR/2010/REC-speech-synthesis11-20100907/
7. Baggia P, Pelachaud C, Peter C, Zovato E (2013) Emotion markup language (EmotionML) 1.0. http://www.w3.org/TR/2013/PR-emotionml-20130416/
8. Bainbridge WA, Hart JW, Kim ES, Scassellati B (2011) The benefits of interactions with physically present robots overvideo-displayed agents. Int J Soc Robot 3(1):41–52
9. Barrett LF (2006) Solving the emotion paradox: categorization and the experience of emotion. Pers Soc Psychol Rev 10(1):20–46
10. Bates J (1994) The role of emotion in believable agents. Commun ACM 37(7):122–125
11. Beale R, Creed C (2009) Affective interaction: How emotional agents affect users. Int J Hum-Comput Stud 67(9):755–776
12. Beckman ME, Ayers GM (1994) Guidelines for ToBI labelling. http://www.speech.cs.cmu.edu/tobi/
13. Benoît C, Grice M, Hazan V (1996) The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. Speech Commun 18(4):381–392
14. Bethel CL, Murphy RR (2006) Auditory and other non-verbal expressions of affect for robots. In: Proceedings of the 2006 AAAI fall symposium series, aurally informed performance: integrating machine listening and auditory presentation in robotic systems. AAAI
15. Black, A.W.: CMU_ARCTIC speech synthesis databases. http://festvox.org/cmu_arctic/
16. Boersma P (2001) Praat, a system for doing phonetics by computer. Glot Int 5(9/10):341–345
17. Breazeal C, Aryananda L (2002) Recognition of affective communicative intent in robot-directed speech. Auton Robots 12(1):83–104
18. Breazeal C, Scassellati B (1999) How to build robots that make friends and influence people. In: Proceedings of the 1999 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 858–863
19. Brooks DJ, Lignos C, Finucane C, Medvedev MS, Perera I, Raman V, Kress-Gazit H, Marcus M, Yanco HA (2012) Make it so: continuous, flexible natural language interaction with an autonomous robot. In: Proceedings of the workshops at twenty-sixth AAAI conference on artificial intelligence. Association for the Advancement of Artificial Intelligence, Palo Alto

20. Burkhardt F, Sendlmeier WF (2000) Verification of acoustical correlates of emotional speech using formant-synthesis. In: Proceedings of the ISCA tutorial and research workshop (ITRW) on speech and emotion. International Speech Communication Association, Singapore
21. Cahn J (1990) The generation of affect in synthesized speech. J Am Voice Input/Output Soc 8:1–19
22. Cowie R, Cornelius RR (2003) Describing the emotional states that are expressed in speech. Speech Commun 40(1):5–32
23. Crumpton J, Bethel CL (2014) Conveying emotion in robotic speech: Lessons learned. In: Proceedings of the 23rd IEEE international symposium on robot and human interactive communication (RO-MAN)
24. Crumpton J, Bethel CL (2015) Validation of vocal prosody modifications to communicate emotion in robot speech. In: Proceedings of the 2015 international conference on collaboration technologies and systems (CTS 2015)
25. Dautenhahn K, Woods S, Kaouri C, Walters ML, Koay KL, Werry I (2005) What is a robot companion - friend, assistant or butler? In: Proceedings of the 2005 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp. 1192–1197
26. D'Mello S, Graesser A (2013) Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Trans Interact Intell Syst 2(4):1–39
27. Ekman P, Sorenson ER, Friesen WV (1969) Pan-cultural elements in facial displays of emotion. Science 164(3875):86–88
28. Erickson D (2005) Expressive speech: production, perception and application to speech synthesis. Acoust Sci Technol 26(4):317–325
29. Fairbanks G, Pronovost W (1939) An experimental study of the pitch characteristics of the voice during the expression of emotion. Speech Monogr 6(1):87
30. Frick RW (1985) Communicating emotion: the role of prosodic features. Psychol Bull 97(3):412–429
31. Greasley P, Sherrard C, Waterman M (2000) Emotion in language and speech: methodological issues in naturalistic approaches. Lang Speech 43(4):355–375
32. Hammerschmidt K, Jürgens U (2007) Acoustical correlates of affective prosody. J Voice 21(5):531–540
33. Heiga Z, Senior A, Schuster M (2013) Statistical parametric speech synthesis using deep neural networks. In: Proceedings: 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 7962–7966
34. Hennig S, Chellali R (2012) Expressive synthetic voices: Considerations for human robot interaction. In: Proceedings of the 21st IEEE international symposium on robot and human interactive communication (RO-MAN), pp 589–595
35. Huang X, Acero A, Hon HW (2001) Spoken language processing: a guide to theory, algorithm, and system development. Prentice Hall PTR, Upper Saddle River
36. Huttar GL (1968) Relations between prosodic variables and emotions in normal American English utterances. J Speech Hear Res 11(3):481–487
37. Iida A, Campbell N, Iga S, Higuchi F, Yasumura M (2000) A speech synthesis system with emotion for assisting communication. In: Proceedings of the ISCA tutorial and research workshop (ITRW) on speech and emotion, pp 167–172
38. Jung Y, Lee KM (2004) Effects of physical embodiment on social presence of social robots. In: Proceedings of the 7th annual international workshop on presence. International Society for Presence Research, pp 80–87
39. Khan Z (1998) Attitudes towards intelligent service robots. Technical Report TRITA-NA-P9821, IPLab-154, Royal Institure of Technology (KTH)
40. Kidd CD, Breazeal C (2004) Effect of a robot on user perceptions. In: Proceedings of the 2004 IEEE/RSJ international conference on intelligent robots and systems (IROS), vol 4, pp 3559–3564
41. Kim E, Leyzberg D, Tsui K, Scassellati B (2009) How people talk when teaching a robot. In: Proceedings of the 4th ACM/IEEE international conference on human-robot interaction (HRI). ACM, New York, pp 23–30
42. Laukka P (2004) Vocal expression of emotion: discrete-emotions and dimensional accounts. Dissertation, Uppsala University
43. Laukka P, Neiberg D, Forsell M, Karlsson I, Elenius K (2011) Expression of affect in spontaneous speech: acoustic correlates and automatic detection of irritation and resignation. Comput Speech Lang 25(1):84–104
44. Leyzberg D, Avrunin E, Liu J, Scassellati B (2011) Robots that express emotion elicit better human teaching. In: Proceedings of the 6th ACM/IEEE international conference on human–robot interaction (HRI), pp 347–354
45. Leyzberg D, Spaulding S, Toneva M, Scassellati B (2012) The physical presence of a robot tutor increases cognitive learning gains. In: Proceedings of the 34th annual conference of the Cognitive Science Society (CogSci). Cognitive Science Society
46. Liu H, Lieberman H, Selker T (2003) A model of textual affect sensing using real-world knowledge. In: Proceedings of the 2003 international conference on intelligent user interfaces (IUI). ACM, New York, pp 125–132
47. Massaro DW (1989) The logic of the fuzzy logical model of perception. Behav Brain Sci 12(04):778–794
48. Massaro DW, Egan PB (1996) Perceiving affect from the voice and the face. Psychon Bull Rev 3(2):215–221
49. Mehrabian A (1996) Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. Curr Psychol 14(4):261
50. Microsoft (2015) Prosody element. Microsoft. https://msdn.microsoft.com/en-us/library/hh361583(v=office.14).aspx
51. Mitchell WJ, Szerszen Sr KA, Lu AS, Schermerhorn PW, Scheutz M, MacDorman KF (2011) A mismatch in the human realism of face and voice produces an uncanny valley. i-Percept 2(1):10–12
52. Murray IR, Arnott JL, Rohwer EA (1996) Emotional stress in synthetic speech: progress and future directions. Speech Commun 20(12):85–91
53. Nass C, Jonsson I, Harris H, Reaves B, Endo J, Brave S, Takayama L (2005) Improving automotive safety by pairing driver emotion and car voice emotion. In: Proceedings of the CHI '05 extended abstracts on human factors in computing systems. ACM, New York, pp 1973–1976
54. Niculescu A, Dijk B, Nijholt A, Li H, See SL (2013) Making social robots more attractive: the effects of voice pitch, humor and empathy. Int J Soc Robot 5(2):171–191
55. Nuance Communications: (2015) SSML compliance. In: Dragon Mobile SDK Reference. http://dragonmobile.nuancemobiledeveloper.com/public/Help/DragonMobileSDKReference_Android/SpeechKit_Guide/SpeakingText.html
56. Oudeyer PY (2003) The production and recognition of emotions in speech: features and algorithms. Int J Hum-Comput Stud 59(12):157–183
57. Pearson JC, Nelson PE (2000) An introduction to human communication: understanding and sharing, 8th edn. McGraw-Hill Higher Education, Boston
58. Pitrelli JF, Bakis R, Eide EM, Fernandez R, Hamza W, Picheny MA (2006) The IBM expressive text-to-speech synthesis system for American English. IEEE Trans Audio Speech Lang Process 14(4):1099–1108
59. Pittam J, Scherer KR (1993) Vocal expression and communication of emotion. In: Lewis M, Haviland JM (eds) Handbook of emotions, chap. 13. The Guilford Press, New York, pp 185–197

60. Powers A, Kiesler S, Fussell S, Torrey C (2007) Comparing a computer agent with a humanoid robot. In: Proceedings of the 2nd ACM/IEEE international conference on human–robot interaction (HRI). ACM, New York, pp 145–152
61. Prasad R, Saruwatari H, Shikano K (2004) Robots that can hear, understand and talk. Adv Robot 18(5):533–564
62. Prinz J (2004) Which emotions are basic? In: Evans D, Cruse P (eds) Emotion, evolution, and rationality, chap. 4. Oxford University Press, Oxford, pp 69–87
63. Rani P, Sarkar N (2004) Emotion-sensitive robots - a new paradigm for human-robot interaction. In: Proceedings of the 4th IEEE/RAS international conference on humanoid robots, vol 1, pp 149–167
64. Ray C, Mondada F, Siegwart R (2008) What do people expect from robots? In: Proceedings of the 2008 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 3816–3821
65. Read R (2012) Speaking without words: Affective displays in social robots through non-linguistic utterances. In: Proceedings of the 2012 HRI pioneers workshop. ACM, New York
66. Read R, Belpaeme T (2010) Interpreting non-linguistic utterances by robots: Studying the influence of physical appearance. In: Proceedings of the 3rd international workshop on affective interaction in natural environments (AFFINE). ACM, New York, pp 65–70
67. Read R, Belpaeme T (2012) How to use non-linguistic utterances to convey emotion in child-robot interaction. In: Proceedings of the 7th ACM/IEEE international conference on human–robot interaction (HRI). ACM, New York, pp 219–220
68. Read R, Belpaeme T (2013) People interpret robotic non-linguistic utterances categorically. In: Proceedings of the 8th ACM/IEEE international conference on human–robot interaction (HRI). ACM, New York, pp 209–210
69. Read R, Belpaeme T (2014) Non-linguistic utterances should be used alongside language, rather than on their own or as a replacement. In: Proceedings of the 9th ACM/IEEE international conference on human–robot interaction (HRI). ACM, New York, pp 276–277
70. Read R, Belpaeme T (2014) Situational context directs how people affectively interpret robotic non-linguistic utterances. In: Proceedings of the 9th ACM/IEEE international conference on human–robot interaction (HRI). ACM, New York, pp 41–48
71. Roehling S, MacDonald B, Watson C (2006) Towards expressive speech synthesis in English on a robotic platform. In: Proceedings of the 11th Australian international conference on speech science & technology. Australian Speech Science & Technology Association, Gold Coast
72. Rogalla O, Ehrenmann M, Zöllner R, Becher R, Dillmann R (2002) Using gesture and speech control for commanding a robot assistant. In: Proceedings of the 11th IEEE international workshop on robot and human interactive communication (RO-MAN), pp 454–459
73. Russell JA, Barrett LF (1999) Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. J Pers and Soc Psychol 76(5):805–819
74. Scherer K (2009) Emotion theories and concepts (psychological perspectives). In: Sander D, Scherer K (eds) Oxford companion to emotion and the affective sciences. Oxford University Press, Oxford, pp 145–149
75. Scherer KR (1986) Vocal affect expression: a review and a model for future research. Psychol Bull 99(2):143–165
76. Scherer KR (2000) Psychological models of emotion. In: Borod J (ed) The neuropsychology of emotion, chap. 6. Oxford University Press, Oxford, pp 138–162
77. Scherer KR, Banse R, Wallbott HG, Goldbeck T (1991) Vocal cues in emotion encoding and decoding. Motiv Emot 15(2):123–148
78. Schröder, M.: MaryXML. http://mary.dfki.de/documentation/maryxml
79. Schröder M, Baggia P, Burkhardt F, Pelachaud C, Peter C, Zovato E (2011) EmotionML—an upcoming standard for representing emotions and related states. In: D'Mello S, Graesser A, Schuller B, Martin JC (eds) Affective computing and intelligent interaction, vol 6974. Springer, Berlin, pp 316–325
80. Schröder M, Cowie R, Douglas-Cowie E (2001) Acoustic correlates of emotion dimensions in view of speech synthesis. In: Proceedings of the 7th European conference on speech communication and technology (EUROSPEECH), pp 87–90
81. Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller C, Narayanan S (2013) Paralinguistics in speech and language—state-of-the-art and the challenge. Comput Speech Lang 27(1):4–39
82. Silverman K, Beckman M, Pitrelli J, Ostendorf M, Wightman C, Price P, Pierrehumbert J, Hirschberg J (1992) ToBI: A standard for labeling English prosody. In: Proceedings of the 2nd international conference on spoken language processing (ICSLP), vol 2. International Speech Communication Association, Singapore, pp 867–870
83. Sobin C, Alpert M (1999) Emotion in speech: the acoustic attributes of fear, anger, sadness, and joy. J Psycholinguist Res 28(4):347–365
84. Tang H, Zhou X, Odisio M, Hasegawa-Johnson M, Huang TS (2008) Two-stage prosody prediction for emotional text-to-speech synthesis. In: Proceedings of the 9th annual conference of the International Speech Communication Association. International Speech Communication Association, Singapore, pp 2138–2141
85. Tao J, Kang Y, Li A (2006) Prosody conversion from neutral speech to emotional speech. IEEE Trans Audio Speech Lang Process 14(4):1145–1154
86. Tielman M, Neerincx M, Meyer JJ, Looije R (2014) Adaptive emotional expression in robot-child interaction. In: Proceedings of the 9th ACM/IEEE international conference on human–robot interaction (HRI). ACM, New York, pp 407–414
87. Tokuda K, Heiga Z, Black AW (2002) An HMM-based speech synthesis system applied to English. In: Proceedings of the 2002 IEEE workshop on speech synthesis, pp 227–230
88. Türk O, Schröder M (2010) Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques. IEEE Trans Audio Speech Lang Process 18(5):965–973
89. Veilleux N, Shattuck-Hufnagel S, Brugos A (2006) 6.911 transcribing prosodic structure of spoken utterances with ToBI. http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-911-transcribing-prosodic-structure-of-spoken-utterances-with-tobi-january-iap-2006
90. Vinciarelli A, Pantic M, Bourlard H, Pentland A (2008) Social signal processing: State-of-the-art and future perspectives of an emerging domain. In: Proceedings of the 16th ACM international conference on multimedia. ACM, New York, pp 1061–1070
91. W3C (2004) Speech synthesis markup language (SSML) version 1.0. http://www.w3.org/TR/2004/REC-speech-synthesis-20040907/
92. Wainer J, Feil-Seifer DJ, Shell DA, Mataric MJ (2007) Embodiment and human-robot interaction: A task-based perspective. In: Proceedings of the 16th IEEE international symposium on robot and human interactive communication (RO-MAN), pp 872–877
93. Walker MR, Larson J, Hunt A (2001) A new W3C markup standard for text-to-speech synthesis. In: Proceedings of the 2011 IEEE international conference on acoustics, speech, and signal processing (ICASSP), vol 2, pp 965–968
94. Weaver CH, Strausbaugh WL (1964) Hearing the vocal cues. In: Fundamentals of speech communication, chap. 11. American Book Company, New York, pp 283–303
95. Xingyan L, MacDonald B, Watson CI (2009) Expressive facial speech synthesis on a robotic platform. In: Proceedings of the 2009 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 5009–5014

96. Zen H, Tokuda K, Black AW (2009) Statistical parametric speech synthesis. Speech Commun 51(11):1039–1064
97. Zhen-Hua L, Shi-Yin K, Heiga Z, Senior A, Schuster M, Xiao-Jun Q, Meng HM, Li D (2015) Deep learning for acoustic modeling in parametric speech generation: a systematic review of existing techniques and future trends. IEEE Signal Process Mag 32(3):35–52

Joe Crumpton is an Assistant Research Professor at the Distributed Analytics and Security Institute (DASI) in the High Performance Computing Collaboratory at Mississippi State University. He graduated in August 2015 with his Ph.D. in Computer Science from Mississippi State University. He received the BS degree in Computer Engineering from Mississippi State University and the MS degree in Computer Science from the University of Tennessee, Knoxville. He attended the 2014 ACM/IEEE International Conference on Human-Robot Interaction (HRI) as a HRI Pioneer. His research interests include human-robot interaction and software engineering.

Cindy L. Bethel is an Associate Professor in the Computer Science and Engineering Department. She is the Director of the Social, Therapeutic, and Robotic Systems (STaRS) lab. She was a NSF/CRA/CCC Computing Innovation Postdoctoral Fellow in the Social Robotics Laboratory at Yale University. In 2008, Dr. Bethel was a National Science Foundation Graduate Research Fellow and the recipient of the 2008 IEEE Robotics and Automation Society Graduate Fellowship. She graduated in August 2009 with her Ph.D. in Computer Science and Engineering from the University of South Florida. Her research interests are in human-robot interaction, affective computing, robotics, human-computer interaction and interface design, artificial intelligence, psychology, experimental design, and statistical analysis. Her research focuses on applications associated with the use of robots for therapeutic support, information gathering, law enforcement, search and rescue, and military.