Expressive speech for a virtual talking head

Xingyan Li, Catherine I. Watson, Aleksandar Igic, Bruce MacDonald
Department of Electrical and Computer Engineering
University of Auckland, New Zealand
{xingyan.li, c.watson, b.macdonald}@auckland.ac.nz, aigi001@aucklanduni.ac.nz

Abstract

This paper presents our work on building an expressive facial speech synthesis system Eface, which can be used on a social or service robot. Eface aims at enabling a robot to deliver information clearly with empathetic speech and an expressive virtual face. The system is built on two open source software packages: the Festival speech synthesis system, which provides robots the capability to speak with different voices and emotions, and Xface–a 3D talking head, which enables the robot to display various human facial expressions. This paper addresses how to express different speech emotions with Festival and how to integrate the synthesized speech with Xface. We have also implemented Eface on a physical robot and tested it with some service scenarios.

1 Introduction

Not only do cute and smart robots appear in Hollywood movies, but more and more social and service robots are being introduced into peoples' everyday lives. These robots usually have direct communications with people, such as taking customers' orders and serving food, reading books and newspapers, or playing games with children. Therefore, the ability to express human emotions and be empathetic makes the robot more effective. For example, a robot nurse assistant should be able to greet people, sound happy to inform patients with good results and express sorrow or encouraging emotions when the test results are not satisfying.

To communicate with people empathetically, two aspects of the robot should be considered: an empathetic voice and expressive facial movements. Ideally, communication of emotion will happen through three channels: verbal, vocal nonverbal and facial. The verbal channel focuses on the human–robot dialogue, for example, the words "like" and "hate" can deliver different emotion information; the vocal nonverbal channel addresses on several speech features, such as voice pitch, duration and intensity; the facial channel aims at delivering vivid expressions with the facial movements. Our current work, based on two open source systems, Festival [21] and Xface [3], aims at providing robots the capability to speak with an empathetic voice and with an expressive virtual face. Voice control and ToBI [20] are used with Festival to offer more expressive, human–like voices, and Xface is used to display diverse facial expressions with different face models.

Our previous work [11] presents the expressive facial speech synthesis system with the open source robotic control interface Player [9] in Linux, and two versions of virtual faces with different polygon numbers. This paper presents improvements to automatic emotional speech synthesis and also work on integrating Xface and Festival under Windows. Related work is shown in Section 2, followed by design and implementation details in Section 3 and test results in Section 4.

2 Related work

There has been much research on social and service robots, such as CMU's nurse robots Flo [16] and Pearl [14], the interactive tour-guide robot Minerva [23] and MIT's social robot Kismet [7]. Most of these robots have the ability to speak, which is an important communication medium between robot and humans.

Nourbakhsh describes how emotions influence the synthesized speech in a tour guide robot [12]. Although the quality of synthesized speech is significantly poorer than synthesized facial expression and body language [4], it is still possible to generate empathetic speech. Another example of robot expressive speech is Kismet's vocalization system [6]. It generates expressive utterances by assembling strings of phonemes with pitch accents, which is what we have done. There are various text to speech (TTS) systems available, such as IBM's Naxpres, Microsoft's Speech SDK, AT&T's Natural Voices and OpenMARY developed by Saarland University [15]. We use the Festival speech synthesis system, a TTS re-

search framework developed by the University of Edinburgh, to generate empathetic speech for a robot. In Festival, synthesis is based on diphones while the pitch baseline, pitch range, and speech rate for the whole sentence can be controlled. It also provides ToBI for transcribing prosody and controlling the intonation contour to some extent. Eface can use both sentences marked manually with ToBI and automatically generated intonation to control the emotion of the voice.

In addition, the use of facial expressions to enhance human–computer interaction has been researched and implemented. For example Hara's 3D robotic face [10] is capable of displaying seven different facial expressions; Thalmann's virtual face [22] provides face–to–virtualface communication in a virtual world; Bouchra's synthetic faces [1] are reproduced with appearance parameters extracted from a natural image or video sequence and Sheng's 3D face [19] can be synthesized from an arbitrary head-and-shoulder image with a complex background. Currently, we use Xface [3] system to show expressive facial movements on the display screen of the robot. Xface has the capability to control several parts of the virtual face–face, eyes, lips, head and hair. It also provides a synchronization method to animate the face with speech. The original Xface uses Microsoft's Speech SDK as its text to speech engine, while we integrate it with Festival to provide an empathetic speech interface with more speech feature control fucntions to the robot.

3 Design and implementation

3.1 Expressive speech synthesis

To interact well with users, the robot should have a large vocabulary and full sentence connected speech that is recognizable by any normal hearing person with a good command of the language. The robot should generate clear and expressive speech so that human users can recognize both the meaning and expression. For example, a guide robot should express a happy voice with a welcome speech such as "It is a nice day and I am very happy to meet you."

The emotion of the speech is not only affected by the words it uses, but also by the way the speech is said. The vocal nonverbal component is more efficient than the verbal content for communicating information about the speaker's state or attitude [18]. To realize a more efficient and pleasant human–robot communication, vocal cues should be included in the synthetic speech, especially for robots in social situations, such as guide robots and nursing assistant robots.

Although previous research showed that different speech features might have different emotional effects in different languages, the following specific features of speech may contribute to convey emotional information [15]:

- Pitch and duration play an important role in speech emotion. In particular, the interaction of pitch with loudness and with the grammatical features of the text seems to be critical. In some conditions, pitch and duration are sufficient to distinguish between neutral speech, joy, boredom, anger, sadness, fear and indignation.
- Loudness alone may not be important but the correct synthesis of loudness can help to deliver emotional information.
- Spectral energy distribution and spectral structure can carry much of the affective information.
- Voice quality is also significant in showing the affective information.

There is some disagreement as to how much each of the aspects above contribute to expressing emotion [15]. In addition, different sentences or different situations require different speech features. Our work focuses on the pitch mean and variation, and duration.

To create expressive speech a good intonation model is needed, and it must be possible to change the emotive state of the robot voice. Festival provides the functionality to change parameters so that the emotional speech features can be altered. Our specific focus has been generating Happy speech and Neutral speech. Happy speech (c.f. neutral speech) has a higher mean pitch and great pitch range, and is faster. Festival is able to change the pitch and duration parameters when given commands such as *f0_mean, f0_sd* and *Duration_Stretch*. Our speech synthesis system provides an interface to control the robot's speech parameters related to pitch and duration.

Although Festival's global parameters can be altered to express simple emotional speech, it turns out that changes to global parameters led to only minor improvements in the output emotion heard and are insufficient to make the speech empathetic. The intonation needs to be modeled well too. There are a few different automatic intonation generating models distributed with Festival. These include: (1) Simple down sloping pitch contour; (2) Condition and Regression Tree (CART) [8] and (3)ToBI.

ToBI is a framework for describing intonation [20] by marking the accented regions of speech as either high or low pitch, and also describing the behavior of the pitch contour at the end of the phrase. Within Festival the ToBI labels can be assigned to the text either automatically via the CART model, or manually using the ToBI intonation model.

The ToBI intonation approach involves using ToBI annotations to the text input to the TTS. With the ToBI marks, the user can control the intonation contour to

some extent. For example with the accent *High* (H) and *Low* (L), sentences can express different emotions:

- Plain accent:
  "I am very happy to meet you"
- With emotion:
  `(I ((accent H*))) (am ((accent L*)))`
  `(very ((accent H*))) (happy ((accent L*)))`
  `(to ()) (meet ((accent H*))) (you ((tone`
  `L-H%)))`
- Plain accent:
  "Now I am very sad you don't like me"
- With emotion:
  `(Now ((accent H*))) (I ((accent L*))) (am`
  `()) (very ((accent H*))) (sad ((accent`
  `H*))) (you ((accent L*))) (dont ((accent`
  `L*))) (like ((accent H*))) (me ((accent`
  `L*)(tone L-L%))`

where H(L)* describe the pitch in accented words in a phrase, H(L)– and H(L)% describe the behavior of the pitch contour in the last syllable of an end of phrase, so for L–H% the pitch contour goes down first and then finishes with a short rise. Each phrase has at least 1 pitch accent, and an end of phrase tone.

A drawback of the ToBI intonation model is that all utterances must be manually annotated with the labels. Further, if no ToBI labels at all are used on the text, the synthetic speech has no pitch variation at all. It is far worse than the simple intonation model.

Another option is to generate intonation automatically with the CART model. CART is a statistical model, built from a pre-labeled speech corpus, and is the automatic intonation generating method employed in Eface. A pre-built CART model is available within the Festival distribution. It is trained from a Boston University FM Radio Speech Corpus [13] and the speaker is an American female news reader (f2b).

Automatic pitch generation in Festival is a three level process. The first level predicts the ToBI labels based on input text using CART. The second level generates pitch target values based on the ToBI mark-up using linear regression (LR) as described in [5]. In the third stage a series of pitch values are interpolated based on the pitch targets. Within the CART method of generating intonation, there are two LR parameters that can be varied in order to change the expressiveness of the voice. These are the mean pitch, and the standard deviation of pitch which corresponds to the pitch range within the utterance.

In order to change the emotive state of the voice, we are currently, using a single CART tree (f2b) with different LR parameters. We have created a new Festival command *SayEmotional* to do this systematically. This command takes three variables, the emotion (happy or neutral), the text, and the degree of the emotion (currently each emotion has 3 levels).

A happy emotive state is achieved by increasing the mean pitch from 1.5 times to 2 times, and standard deviation from 2 to 4 times that of the original pre-recorded speakers mean and range. A neutral emotive state is achieved through keeping the mean pitch constant, and increasing the standard deviation from 1.5 to 2.5 times that of the original pre-recorded speakers range.

In addition, other than the typical American and British accent English voices, we have also built a New Zealand accent English voice and plan to include a lexicon of common Maori words (eg. placenames, flora, fauna, greetings) into our system.

3.2 Expressive virtual face synthesis

In addition to empathetic speech, expressive facial display is another important factor for human-robot communication. Facial expressions have the ability to portray a person's emotional state, temper and mood as well as being able to emphasize and aid in the understanding of spoken text. Our physical robot has a display to show a 3D virtual face which is capable of expressing several emotions as well as rendering the correct lip movements for speech.

Currently, an open source 3D talking head, Xface, is used. This system is based on the MPEG-4 standard [2], which has devised a coding method for graphical models and the transmission of their animation parameters that specifically allows the modeling, manipulation and animation of human facial models. The facial animation coding consists of two standardised sets of parameters: Facial Definition Parameters (FDPs) and Facial Animation Parameters (FAPs). FDPs consist of a set of 3D feature points that can be used to define the basic geometry of the face, and optionally associate a texture with these points; they make each individual facial model unique. FAPs represent a set of atomic facial movements relating to key facial features such as the eyebrows, eyes, nose, mouth, ears and tongue. By explicitly separating animation parameters from unique facial models it is possible to apply the same set of FAP movements to different models defined by a unique set of FDPs. This allows the use of one set of FAPs (expression movements) to be applied to different facial models. To display different expressions, a set of FAPs that correlate to the movement of the facial definition points are combined to form facial movements that correspond to recognizable expressions. By specifying expressions in terms of their component FAPs, it is possible to manipulate a facial model by altering the positions of the FDPs.

Xface relies on OpenGL and is optimized to achieve at least 25 frames per second with a polygon count up to 12000, using modest hardware. This virtual face can

show eight expressions: *anger, disgust, fear, rest, sad, smile(closed), smile(open) and surprise*. For each expression, each related FAP is associated with a magnitude value to specify the amount of FDP movement in the defined direction. Thus, different levels of expression can be achieved by altering the magnitude value, such as "very happy" and "a little sad." The 3D virtual face can also perform 16 face movements such as blinking, brow movement, eye squint and looking around. It also able to alter the lips and mouth to express visemes.

We have generated different face models with the FaceGen Modeller from Singular Inversions to determine the most accepted face for a user group. FaceGen can generate realistic 3D faces for any race, gender and adult age group with 36 expressions, phonemes and modifiers. The user can also control the texture color, symmetric shape and add extra parts such as eyeglasses or hats. Combining Facegen's realistic face images and Xface, we can achieve expressive facial models that look human-like. Fig. 1 shows different face models with expressions or phonemes.

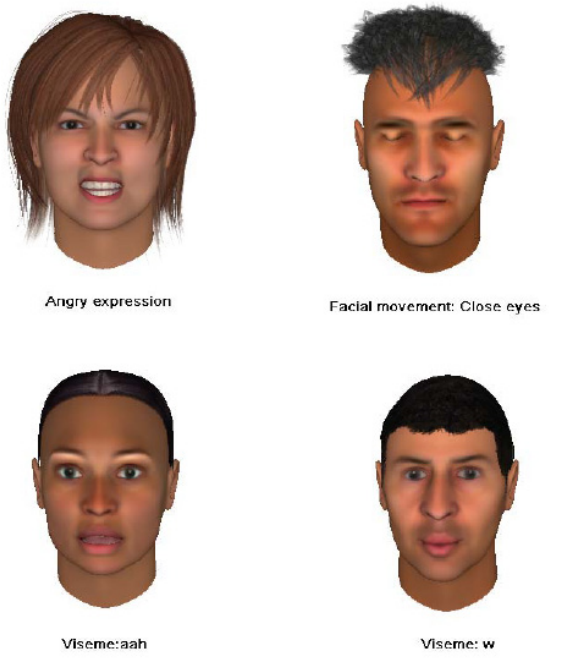

Figure 1: 3D face models with different expressions, visemes and facial movements.

3.3 Integrate speech and face animation

Although Xface uses Microsoft's Speech SDK as its text to speech engine, it also allows the user to use other TTS software. Fig. 2 shows how the Festival speech synthesis system is integrated with Xface.

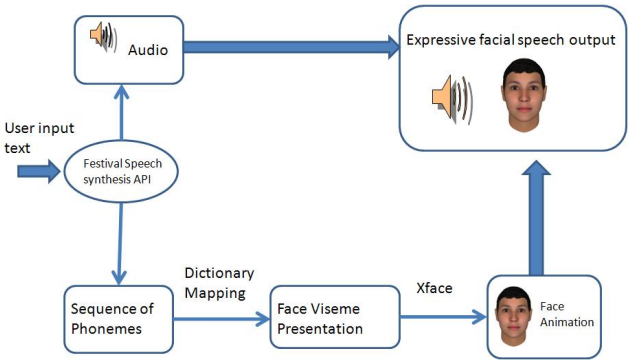

Figure 2: The structure of Festival Speech and Xface integration

The key point is to generate each speech audio file according to the user input text and requirement (specific expressions or ToBI marks) and the phoneme file which records the related sequence of phonemes. Festival is able to output a text-based list of phonemes for any given block of text spoken, using the utterance command *utt.segs*. The list contains both the specific phonemes spoken and the spoken duration time for each phoneme. The phonemes are associated with their respective viseme output using a dictionary file. The file contains all 42 phonemes in the English language (with another phoneme standing for silence) and relates them to their corresponding viseme values that represent a set of FAPs to move all the required definition points around the mouth area into the correct position. The time to display each viseme is taken directly from the phoneme duration time in Festival so the total time to display the entire viseme animation should be the same as the total time of the audio file being played. Therefore, Xface can instruct the face animation and the audio file is played at the same time.

In addition, since the 3D face model is composed with separated parts — head, eyes, hair and face — Xface can perform more complicated facial movements. For example, the robot can speak while smiling, and at the same time, blink the eyes and turn the head. With the expressive facial expression and emotional speech, the robot can deliver information naturally, clearly and empathically.

4 Experimental results

The expressive speech animation synthesis system has been implemented and tested with a robot nursing assistant. In this scenario, an iRobiQ robot with an onboard Intel Pentium 800 MHz processor is used to act the part of a helper for a human nurse. The robot is equipped with a speaker to talk, a display to show its face and a blood pressure monitor with a cuff, connected to the

robot via USB to measure the patient's blood pressure (shown in Fig. 3). Dialogues are designed for the robot to instruct the patient to use the blood pressure monitor and the result is then recorded and communicated to the patient.

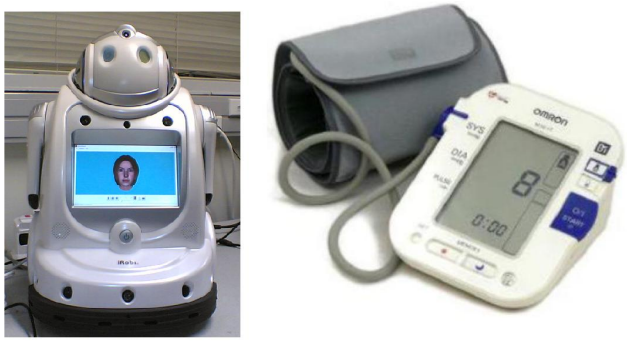

Figure 3: Left: iRobiQ; Right: blood pressure monitor

We have tested different empathetic speech for a nursing assistant. It is relatively difficult to express different emotions just by altering the global speech features, which can cause side effects. For example, increasing the pitch and decreasing the duration can express a "happy" emotion but it is hard for people to hear the command clearly.

We notice that by improving the intonation model, the speech becomes more empathetic, and the differences in the emotions are more apparent. Therefore, we implement an empathetic voice with two methods for the user experiments: manually annotated with ToBI markers and automatically annotated with the CART model. Two expressions are used with the CART model: Happy and Neutral. In Fig. 4 we compare the pitch contours under different conditions with the New Zealand voice for the sentence: "I am very happy to meet you."

The pitch contour is a graphical representation of the fundamental frequency of voiced speech (Hz), as it changes throughout the duration (seconds) of the utterance. The top-most plot (Fig. 4a) shows the pitch contour of a Festival utterance without any intonation modelling. The next plot (Fig. 4b) shows the pitch contour of the utterance with manual ToBI annotations (I ((accent H*))) (am ((accent L*))) (very ((accent H*))) (happy ((accent L*))) (to ()) (meet ((accent H*))) (you ((tone L-H%))).

Fig. 4c shows the pitch contour of the utterance with "Neutral" expression and the bottom-most plot (Fig. 4d) shows the pitch contour of the utterance with "Happy" expression. Both were created by applying the modified CART intonation model.

The no-intonation modelling case is characterised by the flatness of the pitch contour, producing a very monotonic utterance. Manually applying the ToBI annotations produces a more dynamic contour, in which the pitch rises occur during the words marked with (H*), and the falls occur during (L*) marked words.

In the two automatic intonation generating cases: "Neutral" and "Happy"; the pitch contours are similar in the shape (the time in the utterance that the rises and falls occur), but differ in the mean value of the pitch, and in the amount of pitch deviation around the mean. The "Happy" utterance has a larger pitch range and an increased pitch mean, compared to the "Neutral" case. This follows the findings in psychological studies focusing on acoustic properties, in the vocal portrayal of emotion, of which [17] is a good review. Note both Fig. 4b and d are pitch contours for "Happy" speech, but in the case of Fig. 4d this contour will create more empathetic speech than the simple intonation model, without the overhead of hand annotation required when implementing the ToBI model.

Eleven people are asked to listen to the sample speeches in Fig. 4, all of them can tell the difference between the "Neutral" expression and "Happy" expression. However, we need further study to figure out the proper speech voice in a healthcare scenario.

The virtual face can display nine different expressions smoothly, however, due to insufficient CPU speed, there exist a small delay in lip synchronization with the speech. For each of the sentences, there is a time difference (average 0.15 seconds, maximum 0.5 seconds) between the audio voice and the visual display. For most users, this is acceptable. The same application on a desktop with Intel Pentium 2.3 GHz processor runs very smoothly and the lip synchronization is clear and satisfying. Since the CPU speed is comparable to a normal PC machine for most service and social robots, we believe that our expressive facial speech synthesis system can be used smoothly. Fig. 5 shows the virtual face with Neutral and Happy expression during speech. Similarly, eleven people are asked to look to the sample videos and all of them can tell the difference between the "Neutral" facial expression and "Happy" facial expression.

Now that the expressive talking face model has been developed and tested, a professional study will be conducted under the direction of a psychologist in the near future. That test will involve different voice parameters, ToBI marks, different speech emotions, face models with different age, sex and race group. Through the test, we will discover what kind of voice/face model is most acceptable for our target user group in a healthcare scenario.

5 Conclusion

This paper introduces our expressive facial speech synthesis system which can be used by social and service robots. The focus is on the robot's ability to produce

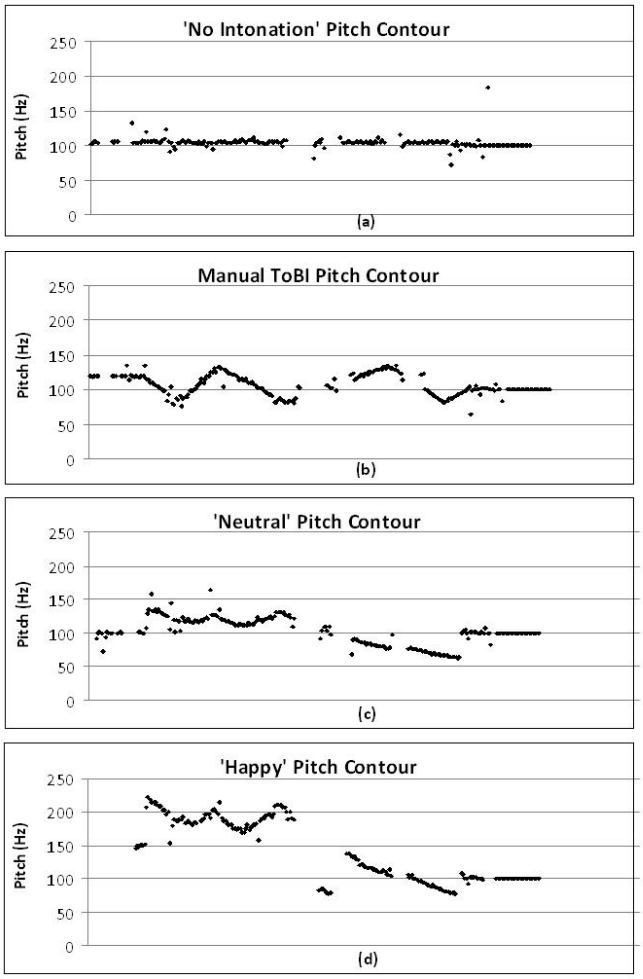

Figure 4: The Pitch contour obtained from four different intonation models: (a) Pitch contour of the utterance without any intonation modelling, (b) Pitch contour of the utterance with ToBI annotations, (c) Pitch contour of the utterance with Neutral expression (CART model), (d) Pitch contour of the utterance with Happy expression (CART model).

expressive facial speech so that clear help and commands will be delivered to users. The empathetic speech, expressive virtual face and the association between them are presented; in particular how to synthesize emotional speech with the Festival Speech Synthesis system and integrate the speech with face animation provided by Xface. The system is implemented and tested on a physical robot with experimental scenario, we believe that it can be used in a general service or social robot, or other applications that require human robot interaction.

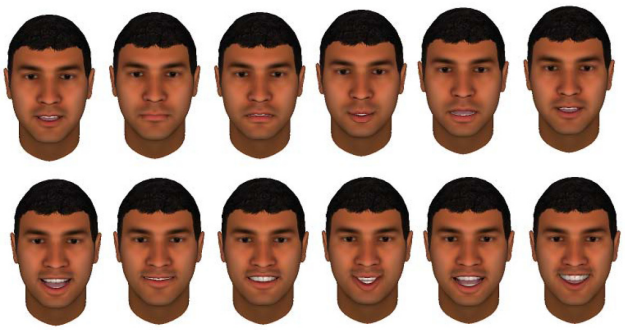

Figure 5: Top: The virtual face with Neutral expression while speaking ; bottom: The virtual face with Happy expression while speaking

References

[1] Bouchra Abboud, Franck Davoine, and Mo Dang. Expressive face recognition and synthesis. *Computer Vision and Pattern Recognition Workshop*, 5:54, 2003.

[2] Gabriel A. Abrantes and Fernando Pereira. Mpeg-4 facial animation technology: survey, implementation, and results. *IEEE Transactions on Circuits and Systems for Video Technology*, 9:290–305, 1999.

[3] Koray Balci. Xface: Mpeg-4 based open source toolkit for 3d facial animation. In *Proceedings of the working conference on Advanced visual interfaces*, pages 399–402, 2004.

[4] Christoph Bartneck. Interacting with an embodied emotional character. In *Proceedings of the International Conference on Designing pleasurable products and interfaces*, pages 55–60, 2003.

[5] Alan W. Black and Andrew J. Hunt. Generating f0 contours from tobi labels using linear regression. In *International Conference on Spoken Language Processing*, pages 1385–1388, 1996.

[6] Cynthia Breazeal. Designing sociable robots. *Robotics and Autonomous Systems*, 2002.

[7] Cynthia Breazeal and Brian Scassellati. How to build robots that make friends and influence people. In *Proceedings of IEEE International Conference on Intelligent Robots and Systems*, pages 858–863, 1999.

[8] Leo Breiman, Jerome Friedman, and Charles J. Stone amd R.A. Olshen. *Classification and Regression Trees*. Chapman Hall, 1984.

[9] Brian P. Gerkey, Richard T. Vaughan, and Andrew Howard. The player/stage project: Tools for multi-robot and distributed sensor systems. In *Proceedings of the International Conference on Advanced Robotics*, 2003.

[10] F. Hara and H. Kobayashi. Use of face robot for human-computer communication. In *IEEE Inter-*

national Conference on Systems, Man and Cybernetics, pages 1515–1520, 1995.

[11] Xingyan Li, Bruce MacDonald, and Catherine I. Watson. Expressive facial speech synthesis on a robotic platform. In *Proceedings of IEEE International Conference on Intelligent Robots and Systems*, 2009.

[12] llah Nourbakhsh, Judith Bobenageb, and Sebastien Grangec. An affective mobile robot educator with a full-time job. *Artificial Intelligence*, 114:95–124, 1999.

[13] Mari Ostendorf, Patti Price, and Stefanie Shattuck-Hufnagel. *Boston University Radio Speech Corpus*. Linguistic Data Consortium, 1996.

[14] Joelle Pineau, Michael Montemerlo, and M. Pollack. Towards robotic assistants in nursing homes: Challenges and results. *Special issue on Socially Interactive Robots, Robotics and Autonomous Systems*, 42:271 – 281, 2003.

[15] Sigrid Roehling, Catherine Watson, and Bruce MacDonald. Towards expressive speech synthesis in english on a robotic platform. In *Proceedings of the Australasian International Conference on Speech Science and Technology*, pages 130–135, 2006.

[16] Nicholas Roy, Gregory Baltus, and Dieter Fox. Towards personal service robots for the elderly. In *Workshop on Interactive Robots and Entertainment*, 2000.

[17] Klaus Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40:227–256, 2003.

[18] Klaus Scherer, Robert Ladd, and Kim Silverman. Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America*, pages 1346–1356, 1984.

[19] Yun Sheng, Abdul H. Sadka, and Ahmet M. Kondoz. Automatic single view-based 3-d face synthesis for unsupervised multimedia applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 18:961–974, 2008.

[20] Kim Silverman, Mary Beckman, and John Pitrelli. Tobi: a standard for labeling english prosody. In *Proceedings of International Conference on Spoken Language*, pages 867–870, 1992.

[21] Paul Taylor, Alan Black, and Richard Caley. The architecture of the festival speech synthesis system. In *Proceedings of the ESCA Workshop in Speech Synthesis*, pages 147–151, 1998.

[22] Magnenat Thalmann, Kalra P, and Escher M. Face to virtual face. In *Proceedings of IEEE special issue on multimedia*, 1998.

[23] Sebastian Thrun, Maren Bennewitz, and Wolfram Burgard. Minerva: A second generation mobile tour-guide robot. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 1999.