Towards a flexible platform for voice accent and expression selection on a healthcare robot

Aleksandar Igic[1], Catherine I. Watson[1], Jonathan Teutenberg[2], Rie Tamagawa[3], Bruce MacDonald[1], Elizabeth Broadbent[3]

1 - Department of Electrical and Computer Engineering
2 - Department of Comuter Science
3 - Department of Psychological Medicine

University of Auckland, Private Bag 92019, Auckland 1142,
New Zealand

aigi001@aucklanduni.ac.nz,
jono@cs.auckland.ac.nz,
{c.watson,e.broadbent,r.tamagawa,b.macdonald}
@auckland.ac.nz

Abstract

In the application of robots in healthcare, where there is a requirement to communicate vocally with non-expert users, a capacity to generate intelligible and expressive speech is needed. The Festival Speech Synthesis System is used as a framework for speech generation on our healthcare robot. Expression is added to speech by modifying mean pitch and pitch range parameters of a statistical model distributed with Festival. US and UK English diphone voices are evaluated alongside a newly made New Zealand English accented diphone voice by human judges. Results show judges can discern different accents and correctly identify the nationality of the voice.

1. Introduction

With the rapidly ageing populations in the developed world, robots are increasingly finding a use in nursing homes in assistive medical care [1][2]. In order for such robots to facilitate the needs of older and mobility restricted users from a communication point of view, more human modes of interaction need to be implemented [1]. The most natural mode of communication for humans is speech, which for a medical robot requires both speech recognition and generation capabilities. We are currently focusing on implementing a flexible robotic speech generation framework that will provide a high standard of quality and expressiveness.

2. Healthcare Robot Background

We are currently developing a Healthcare robot to assist with elderly care. This multi-disciplined project involves personnel with backgrounds in Engineering, Health Psychology, Health Informatics, Nursing, and Gerontology. It involves academics and industry from both New Zealand and Korea [3] .The project is working closely with a retirement village in Auckland, where the healthcare robots are to be trialed. We have already evaluated a preliminary version of the robot with elderly users, in a blood pressure measurement task. The robot instructed users how to use a blood pressure measurement device, and reported back their measurements [4]. The robot (see *Figure 1*) is a mobile device with ultra sound and laser sensors for location detection. It has a screen with a talking virtual head and the face is able to convey a variety of emotions [5]. We discuss the development of the expressive face, accompanying the speech synthesis in [4].

Figure 1: *Charles with blood pressure monitor.*

The functionality of the healthcare robot is now being extended to include location monitoring, falls detection, medication management, appointment reminders, and more vital signs measurements (pulse and blood oxygenation) [1][6]. Some other non medical uses could include, delivering weather and time information and reading the news. These roles require, in most cases, human robot interaction to take place in form of speech dialogue.

The robot voice is provided by the Festival Speech Synthesis system [7]. The preliminary version of the robot used one of the default voices (KAL), which was male with an American accent. Feedback from the preliminary study [5] revealed that users found the voice the voice "too robotic". To this end we have been investigating a variety of ways to make the voice more engaging. We have considered using different accented voices and different models of intonation, and have implemented simple emotion models, and a more flexible speech synthesis system. This paper outlines our development of creating different voices for our robot, and presents evaluations of the voices to date.

3. Speech synthesis

Festival offers a robust and flexible architecture for speech and language modeling, with a powerful capability to easily integrate new speech generation modules. Scripting functionality is implemented through an internal Scheme interpreter. The standard Festival distribution contains automatic intonation and duration generating schemes, as well as a facility for manual intonation modeling through ToBI [8]. Speech synthesis methods in Festival include: Diphone concatenation [9], Multisyn unit selection [10] and HTS hidden Markov synthesis [11]. Festival is implemented on the robot in server mode, and interacts with the rest of the robot modules through a modified Player [12] framework.

In our studies we have used the three differently accented English synthetic voices generated through diphone concatenation: US, UK and NZ. US English and UK English are the two diphone voices that are part of the standard Festival distribution. The NZ voice is newly developed at the University of Auckland and contains diphones recorded by a male speaker and a New Zealand English lexicon with 500 common Maori words [13].

4. Adding Expression to Speech

We have subdivided the robot dialogue into five different types: greeting, instruction (eg. instructing a patient to put a cuff for blood pressure measurements), information (eg. delivering measurement results), question, social (eg. reading news, telling jokes). Our goal is to ensure that each of these dialogue types has the appropriate tone. This means we need to be able to adjust the both the intonation and emotion quotient of the voice. At present we are only employing very simple techniques, but coupled with the virtual robot head we can convey different emotional states.

Generating expressive intonation is a multi-tier process within Festival. The text to be spoken is first ToBI labeled [8] manually, or automatically through a CART tree model [14] [15]. These labels are then converted to pitch targets using linear regression [16]. Interpolation is done between target points to generate a pitch contour for the utterance. Two parameters, mean speaker pitch and speaker pitch standard deviation, allow control over the average value and the range of the final pitch contour.

4.1 A New Method of Changing Intonation in Festival

To allow for utterances to be synthesized with different levels of expression, a function 'SayEmotional' was written in Scheme which takes three parameters: input text to be synthesized, one of two emotions 'Happy' or 'Neutral', and the level of 'emotional intensity'. *Figure 2* illustrates this functionality by comparing plots of pitch contours of the utterance "I am very happy to meet you" generated with four different methods: *a)* with no intonation, *b)* with manually labeled text, *c)* and *d)* through 'SayEmotional' utilizing 'Neutral' and 'Happy' parameters respectfully.

These plots show the value of the fundamental frequency (f0) of voiced speech as it changes throughout the duration of the utterance. All are generated with the New Zealand English voice, and are of the same duration.

The 'Happy' utterance differs from the 'Neutral' through increased pitch mean and range. This follows findings in psychological studies of acoustic properties of emotion as reviewed in [17].

The case of no intonation being applied the contour is flat, and in the manually labeled text case:

(I((accent H*))) (am((accent L*))) (very((accent H*)))(happy((accent L*))) (to()) (meet((accent H*))) (you((tone L-H%)))

The contour is dynamic, with f0 rises occurring at (H*) labeled words, and f0 falls occurring at (L*) labeled words.

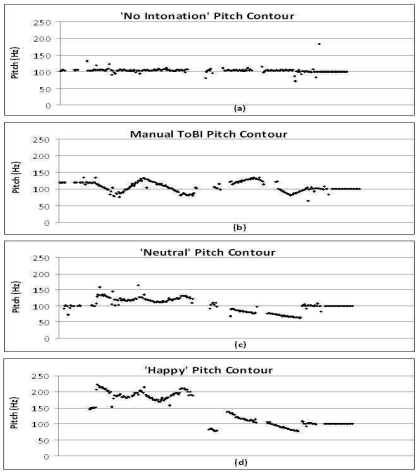

Figure 2: *Pitch contours of the phrase "I am very happy to meet you" with: a) no intonation, b) manual ToBI intonation, c) automatically generated 'Neutral' intonation and d) automatically generated 'Happy' intonation*

4.2 Changing Emotion

The 'SayEmotional' method makes use of the automatic intonation generation and manipulates the two baseline linear regression parameters to generate emotional speech. The baseline parameters are mean pitch and pitch standard deviation, and they are calculated from original recorded diphones. These parameters are dependent on the vocal characteristics of the speaker and are different for each diphone voice. As the diphones are context neutral, baseline parameters for mean pitch and range are used for generating 'Neutral' utterances.

In order to vary the emotive state of the generated speech, we are systematically changing the mean and the standard deviation parameters of the CART model. To move from 'Neutral' to lowest intensity 'Happy' we are increasing the mean pitch to 1.5 times and the standard deviation to 2 times that of the original. High intensity 'Happy' is achieved by increasing the mean to 2.5 and standard deviation to 4 times that of the original.

5. Diphone Voice Evaluations

A study was conducted to evaluate the human perception on the three English accented voices: US, UK and NZ. All voices are synthesized using diphone concatenation using context neutral diphones. The study group comprised of 20 participants, 6 males and 14 females with a mean age of 31.95 and standard deviation of 11.65. Participants had lived in NZ for average of 20.87 years with standard deviation of 12.26.

In the procedure each participant was asked to listen to a minute long sentence synthesized by one of the three English voices. Two sentences were synthesized per voice; one with manually and the other with the automatically ToBI annotated text, comprising in total of 6 different sentences being evaluated by each participant. Three measures were investigated: the quality of the voice, the nationality of the voice and the 'roboticness' of the voice, lastly participants were asked to indicate which voice was the most preferred and which was the least.

ANOVA analysis was performed on all the results and showed no significant differences in the quality score among the voices regardless of whether the intonation of the speech was generated from ToBI labels, automatically generated or labeled by hand., $F(5, 114) = 1.75$, $p = .128$.

When participants were asked to rate the roboticness of the voice, the results of ANOVA showed that the rating was significantly different between the 6 voices, $F(5, 114) = 2.31$, $p = .048$. The US original voice was rated as the most robotic while NZ original was rated as the most human-like. There was no significant difference in roboticness between intonation from automatically generated ToBI labels or labeled by hand. Since there was no difference in quality and roboticness between the two intonation methods, we will focus on the results of the intonation from the automatically generated ToBI labels from now on.

We tested to see whether the participants could identify the accent type of the voices. Each was given 9 options (New Zealand, Australian, South African, British, Asian, Canadian, American, Irish, Other (non-definable)). The majority of participants guessed the correct nationality of the given voice although the recognition rate for US voices was lower than for NZ and UK voices.

The New Zealand accent was correctly identified by 65% of the participants, the US accent was correctly identified by 45 % of the participants, and the UK accent by 50 % of the participants.

Preferred		Non preferred	
	% recalled		% recalled
New Zealand	35	New Zealand	35
American	10	American	55
British	55	British	10

Table 1: *Preferred and non preferred accent*

We also asked the participants what accent they preferred the most and the least (see table 1). Results of Chi-square were significant for both Preferred $\chi^2(2, N = 20)= 6.10$, $p = .047$, and Non-preferred answers, $\chi^2(2, N = 20)= 6.10$, $p = .047$, indicating that participants had significantly varied opinions about which voice they prefer and do not prefer. The British accent was preferred by more participants, while American accent was least preferred.

The main outcome of the study shows that there is no statistical difference between the effects of manual and automatic intonation schemes on the perception of quality and the roboticness of synthetic voices. Due to these results we have decided to move away from manual ToBI labeling and focus solely on automatic intonation schemes. This realization in turn prompted the development of the 'SayEmotional' method described in Section 4, which was based solely on adapting the automated intonation scheme. These results also indicate that there is a personal preference element in voice accent. This is suggests users should have a choice about the voice nationality on the robot.

6. Improving the speech synthesizer

Manipulating the pre-recorded diphone speech waveforms through intonation modeling as with 'SayEmotional' introduces audible artifacts that reduce the quality of the generated speech. Currently a harmonic plus noise synthesis model (HNM) is being added into Festival which allows for waveform manipulation to be achieved with a lower loss of quality compared to other systems [18]. We have further improved the original HNM system described in [18] by using continuous sinusoids to synthesize speech [19] which further improves the quality of generated speech and gives a two fold increase in computational efficiency of generating speech. The initial focus of the work is to allow New Zealand diphone voice synthesis to work with the HNM system. Eventually we aim to incorporate it into other synthesis methods within festival.

7. Conclusion

We are working on a healthcare robot for nursing homes, with a flexible speech synthesis system as a means of human robot interaction. In the final stage, we intend to have a speech framework with the ability to automatically generate emotive, high quality speech with the capacity to change the nationality of the voice dependant on user preference. The speech system, based on Festival, makes use of differently accented voices including a newly created New Zealand English voice. It is able to change its speech emotive state depending on the context. We are in the process of implementing an improved harmonic plus noise model of speech synthesis.

Throughout the development, there will be usability trials. Next trials, scheduled for October, will focus on the interactions of older people with the robot system in a nursing home.

Acknowledgement

This work was supported by the R&D program of the Korea Ministry of Knowledge and Economy (MKE) and the Korea Evaluation Institute of Industrial Technology (KEIT). [2008-F039-01, Development of Mediated Interface Technology for HRI].

This work is supported by a grant from the NZ government Foundation for Research, Science and Technology for robotics to help care for older people and by a University of Auckland New Staff Grant.

Authors would like to acknowledge the work of Xingyan Li and Tony Kuo, for their work on the development of the Healthcare Robot.

References

[1] I. H. Kuo, E. Broadbent, B. MacDonald. *"Designing a robotic assistant for healthcare applications"*, in the 7th conference of Health Informatics New Zealand, Rotorua, Oct 2008.

[2] J.C. Bauer *"Service robots in health care: The evolution of mechanical solutions to human resource problems"*, Bon Secours Health System, Inc. Technology Early Warning System – White Paper. 2003. Available from: http://bshsi.com/tews/docs/TEWS%20Service%20Robots.pdf.

[3] B. MacDonald, W. Abdulla, E. Broadbent, M. Connolly, K. Day, N. Kerse, M. Neve, J. Warren, C. I. Watson *"Robot assistant for care of older people"*, 5th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI 2008) November 20-22, 2008.

[4] X. Li, B. MacDonald, C. I. Watson *"Expressive Facial Speech Synthesis on a Robotic Platform"*, 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems: IROS 2009, St Loius, October 11 to 15.

[5] I. H. Kuo, J. M. Rabindran, E. Broadbent, Y. I. Lee, N. Kerse, R. M. Q. Stafford, B. MacDonald, *"Age and gender factors in user acceptance of healthcare robots"*. In Proceedings of 18th IEEE International Symposium on Robot and Human Interactive Communication, Toyama, Japan, September 27 – October 2 2009, 214-219.

[6] E. Broadbent, R. Tamagawa, N. Kerse, B. Knock, A. Patience, B. MacDonald "*Retirement home staff and residents' preferences for healthcare robots"*, 18th IEEE International Symposium on Robot and Human Interactive Communication, Toyama, Japan, September 27-October 2, 2009, 645 - 650.

[7] A. W. Black, P. Taylor, R. Caley, *"The Festival Speech Synthesis System"*. In http://www.cstr.ed.ac.uk/projects/festival/, 1999

[8] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg, *"TOBI: a standard for labeling English prosody"*. In Second International Conference on Spoken Language Processing, October 13 - 16, 1992, 867-870.

[9] F. Charpentier, M. Stella, *"Diphone synthesis using an overlap-add technique for speech waveforms concatenation"*. In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP, April 1986, 2015-2018.

[10] R. A. J. Clark, K. Richmond, S. King, *"Multisyn: Open-domain unit selection for the Festival speech synthesis system"*. In Speech Communication, Volume 49 Issue 4, 2007, 317-330.

[11] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, K. Tokuda *"The HMM-basedSpeech SynthesisSystem (HTS) Version 2.0"*. In Proc. of Sixth ISCA Workshop on Speech Synthesis, 2007

[12] B. P. Gerkey, R. T. Vaughan, A. Howard *"The player/stage project: Tools for multi-robot and distributed sensor systems"* In Proceedings of the International Conference on Advanced Robotics, June 30 –July 3, 2003, 317-323.

[13] C. I. Watson, J. Teutenberg, L. Thompson, S. Roehling, A. Igic, *"How to build a New Zealand voice"*, (Submitted), NZ Linguistic Society Conference, Palmerston North, November 30 – December 1, 2009.

[14] A. K. Syrdal, J. Hirschberg, J. McGory, M. Beckman, *"Automatic ToBI prediction and alignment to speed manual labeling of prosody"* In Speech Communication, Volume 33, Issues 1 - 2, 2001. 135-151.

[15] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, *"Classification and Regression Trees"* Chapman & Hall (Wadsworth, Inc.): New York, 1984.

[16] A. W. Black, A. J. Hunt, *"Generating F0 contours from ToBI labels using linear regression"*. In Fourth International Conference on Spoken Language Processing, October 3 – 6, 1996, 1385-1388.

[17] K. R. Scherer, *"Vocal communication of emotion: A review of research paradigms"*. In Speech Communication, Volume 40, Issue 3, April 2003, 227 – 256.

[18] Y. Stylianou, *"On the implementation of the harmonic plus noise model for concatenative speech synthesis"*, In Third ESCA/COCOSDA Workshop on Speech Synthesis, November 26-29, 1996, 261-266.

[19] J. Teutenberg, C. I. Watson, "*Flexible and efficient harmonic resynthesis by modulated sinusoids*". In the Proceedings of the 17th European Signal Processing Conference, Glasgow, August, 2009.