



Perception of synthetic speech with emotion modelling delivered through a robot platform: an initial investigation with older listeners

*Aleksandar Igic¹, Catherine I. Watson¹, Rebecca Stafford²
Elizabeth Broadbent², Chandimal Jayawardena¹, Bruce MacDonald¹*

¹Department of Electrical and Computer Engineering, University of Auckland, New Zealand

²Department of Psychological Medicine, University of Auckland, New Zealand

aigi001@aucklanduni.ac.nz, c.watson@auckland.ac.nz, r.stafford@auckland.ac.nz
e.broadbent@auckland.ac.nz, c.jayawardena@auckland.ac.nz, b.macdonald@auckland.ac.nz

Abstract

In this paper we give results of an initial investigation into the perception of synthetic speech delivered through a robotic platform. The robotic speech was judged by 19 residents and 10 staff of a New Zealand retirement village. We have investigated intelligibility and quality measures on two English language diphone voices, with US and New Zealand accents. We have also looked at the effects intonation modelling has on these measures. Our results indicate that the New Zealand voice is preferred and scores higher in the quality measure, additionally we see evidence that the dialogues delivered through both voices are intelligible. We also observe a difference in opinion to the intonation modelling. Comparing the results between staff and residents, we see that residents give lower scores to intelligibility and quality measures.

Index Terms: speech synthesis, unit selection, join costs

1. Introduction

Benefiting from recent advances in robotics and the proliferation of assistive medical technologies, older people are increasingly becoming the primary users of quality of life enhancing systems, which are designed to prolong the period of independent living. Many of these technologies use a dialogue system designed to interact through artificially generated speech. Previous studies have found that older listeners, as well as listeners with hearing impairments have trouble understanding artificial speech, particularly when the dialogue contains unfamiliar words (e.g. medication names) [1]. At the University of Auckland, in a joint venture with South Korea's Electronics and Telecommunications Research Institute (ETRI), we are developing a robotic assistant to help care for older people [2]. ETRI's partner, Yujin Robot, has provided the robotic platform and the University of Auckland based group is developing older-care applications. In this paper we present the results of an investigation in voice perception. It was conducted as a part of a larger study that focused on the attitudes of retirement village staff and residents to the healthcare robot. Here we have tested the measures of: intelligibility, voice quality and the rate of speech, with two differently accented voices, each of which has been further modified by prosodic models to generate neutral and happy intonation. The voice capability is implemented in the robot using the Festival speech synthesis system [3]. We have built a male New Zealand English (NZ) accented diphone voice [4], and it is used alongside the male US accented diphone voice distributed with Festival.

2. Background

2.1. Synthetic speech perception by older listeners

The end goal of the project is to develop an assistive robot platform whose primary mode of communication is synthetic voice. From the communication standpoint, for the speech system to be useful to people across age groups, it is important for it to be both intelligible and easy to listen to. It is well documented that synthetic speech is problematic [5][6][7], when compared to natural speech, and furthermore, additional difficulties are experienced by older listeners and those with hearing difficulties [7][1]. In [5] it is reported, that the reduction of phonemic information and paralinguistic cues in synthetic speech account for additional cognitive load in processing synthetic speech signals. An investigation on the effects of hearing ability on intelligibility of synthetic speech reported in [8], found that extended high frequency hearing loss ($> 8kHz$) (which is more acute in older individuals), adversely affects the participants ability to understand synthetic speech. To combat this and improve the intelligibility of synthetic speech, dialogue needs to be designed with well placed contextual cues. When contextual cues are incorporated in the dialogue, improvements in intelligibility of synthetic speech is observed across age groups [7].

2.2. Emotive speech synthesis

Our overall aim in terms of the speech capability for the robot is to provide high quality natural sounding speech that can also be prosodically modified to emulate various emotive states. There are many studies focused on emotional characteristics of speech and ways in which emotion is communicated vocally [9]. In the current speech system on the robot, we make use of a statistical prosody model trained from the Boston University Radio Speech Corpus, which is part of the Festival standard issue. We modify the way in which this model changes the speech to simulate the emotive states of neutral and happy. We make use of the correlates of change in mean pitch and pitch variation, to the emotive states of happy and neutral. These changes are introduced through a function called SayEmotional [10] which takes in three parameters: emotion, level of activation and the text to be synthesised. SayEmotional modifies the pitch contour of the produced utterance, resulting in more expressive speech and is fully described in [10].

3. Robot trial in the retirement village

The robot voice evaluation was a part of a larger study that focused on the attitudes of residents and staff of the retirement village to the older care robot [11][12]. The greater study looked at how the staff and residents responded to interacting with the robot. In the study a number of tasks and interactions were performed sequentially and required the robot to self navigate to the participant, assist them to take their blood pressure and blood oxygenation measurements, remind them to hydrate, as well as providing an entertainment function by playing a song and telling a joke. All the tasks that were performed by the robot, made use of the voice system and generated speech using the NZ accented diphone voice, which was further modified with the SayEmotional method using the ‘neutral’ tone. Participants evaluated the robot by completing two paper questionnaires, one before interacting with the robot, and one after. This process took on average one hour, after which a 10 minute voice evaluation was conducted using the evaluation module implemented on the robot. The main goal of the voice evaluation was to collect the data on the speech generation system to guide further developments. We were interested in assessing whether the older participants could understand the robot, and to check measures such as the rate of speech and the overall pleasantness of the voices. In addition we were interested in comparing the differences in opinion between the ‘neutral’ and ‘happy’ toned voices generated through SayEmotional.

3.1. Robotic voice perception test

The NZ and US (KAL) voices were used in the perception study. The US (KAL) voice used the lexicon from [13], and the NZ used the NZ lexicon [4]. The two voices were additionally divided into two intonation states: ‘happy’ and ‘neutral’. The reason for choosing these states was that the robot dialogue can be divided into a greeting type, which would use the ‘happy’ intonation state, and an information/instruction type, which would use the ‘neutral’ intonation state. Ideally the ‘neutral’ and ‘happy’ renditions would be within a single robot dialogue. However at the start of the trial the robot was set up so it could be in one emotional state for each dialogue.

Overall, the participants were asked to rate four voices which were referred to throughout the study as: Voice 1, 2, 3 and 4, and are summarised in the Table 1 below. The voices were evaluated using a computerised questionnaire presented on a 10 inch touch screen on the robot. The participants were required to indicate their answers by touching the appropriate option displayed on the screen. The screenshots of the voice evaluation module are shown in Fig. 1. Voices were evaluated sequentially, and the participants were required to complete the forms in Fig. 1(b) and Fig. 1(c) for each voice. The evaluation concluded with the participant indicating their most preferred voice in the form Fig. 1(d). A 100 point scale was used to collect the rate of speech, understandability (voice intelligibility) and pleasantness (voice quality) as seen in Fig. 1(b). The reason behind using 100 point scales was to maintain consistency with the greater investigation. To explain the scales, a green triangle was displayed next to each of the scales, which when pressed would prompt the robot to ask the question corresponding to the particular measure in the currently selected voice. The questions for each of the measures are summarised below.

- Understandability - “How well can you understand me?”
- Pleasantness - “Is my voice pleasant or unpleasant?”
- Rate of Speech - “Am I talking too fast, too slow or just right?”

Because of the space constraint on the robot screen, the anchors for each scale needed to be explained by the researcher present during the trial, and these are:

- Understandability - 0 for “Not at all”, and 100 for “Very well”
- Pleasantness - 0 for “Unpleasant” and 100 for “Very pleasant”
- Rate of Speech - 0 for “Too slow”, 50 for “Just right” and 100 for “Too fast”

During the entire course of the robot interaction, a researcher was present and explained the questions to minimize any participant confusion.

Voice 1	Voice 2	Voice 3	Voice 4
NZ Neutral	NZ Happy	US Neutral	US Happy

Table 1: Diphone voice accent and tonality used in the perception study.

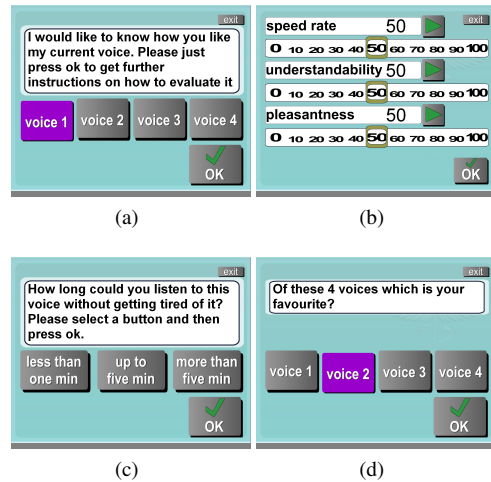


Figure 1: Screen-shots of the voice evaluation module implemented on the robot.

3.2. Robotic voice test results

The group of participants involved in gathering the voice perception data composed of 19 residents with a mean age of 78.84 years, SD = 5.8, and 10 staff with a mean age of 44.6 years, SD = 10.99. Our particular interest was to investigate if the older participants understood the synthetic speech - as there is evidence from previous research that older individuals have difficulties in doing so [6]. We saw this as a critical design question; if the robot speech is not understandable then we must rethink the speech approach. The results are given in Table 2 and indicate that most participants rated the NZ accented voice higher than the US voice in intelligibility.

In Table 3 we compare the intelligibility between the four voices using the Bonferroni corrected p -values obtained through non-parametric Wilcoxon signed-rank tests. These results indicate that there is no statistical difference in intelligibility scores between voices of the same accent, generated using ‘neutral’ and ‘happy’ intonation. A difference ($v = 182$, $p = 0.02$) is observed when cross accent comparison is made between US and NZ accents generated with ‘happy’ intonation.

Participant Group	NZ Neutral		NZ Happy	
	μ	σ	μ	σ
All	56.8	24.50	53.1	14.82
Residents	52.95	21.82	49.42	8.87
Staff	64	28.75	60	21.08

(a)

Participant Group	US Neutral		US Happy	
	μ	σ	μ	σ
All	47.45	17.24	42.72	16.57
Residents	43.47	13.74	38.89	13.50
Staff	55	21.21	50	20

(b)

Table 2: Summary of the intelligibility (understandability) scores showing the mean (μ) and the standard deviation (σ) for each of the voices judged by the two participant groups

Comparison	V	p - Value
NZ Neutral and NZ Happy	149.5	0.96
US Neutral and US Happy	105	0.23
NZ Neutral and US Neutral	165	0.35
NZ Happy and US Happy	182	0.02*

Table 3: Summary of the four, non-parametric two sample Wilcoxon signed-rank tests, comparing the intelligibility scores between the two voices (NZ vs. US) and intonation tones ('Neutral' vs. 'Happy').(* $p < 0.05$)

In a similar manner to intelligibility, the measure for voice quality was investigated. The results are given in Table 4, and show that NZ voice scored higher in the quality measure too.

Participant Group	NZ Neutral		NZ Happy	
	μ	σ	μ	σ
All	55.24	17.19	47.03	12.60
Residents	55.89	13.36	44.95	9.76
Staff	54	23.66	51	16.63

(a)

Participant Group	US Neutral		US Happy	
	μ	σ	μ	σ
All	31.69	18.82	35.79	19.15
Residents	33.11	17.83	33.05	14.46
Staff	29	21.32	41	26.01

(b)

Table 4: Summary of the quality (pleasantness) scores showing the mean (μ) and the standard deviation (σ) for each of the voices judged by the two participant groups

We have compared the quality scores using the same test as the intelligibility measure, using p -values obtained through non-parametric two sample Wilcoxon signed-rank tests, given in Table 5. Results indicate that there is a statistical difference between the quality scores when comparing the two accents with the NZ accent scoring higher than the US voice. There was a significant difference between the NZ voices generated with the two different intonation models ($v = 105$, $p = 0.04$), though this is not observed for the US accented voice.

For the extended duration question the three options show

Comparison	V	p - Value
NZ Neutral and NZ Happy	105.5	0.04*
US Neutral and US Happy	57.5	0.91
NZ Neutral and US Neutral	290	0.0003***
NZ Happy and US Happy	229.5	0.003**

Table 5: Summary of the four, non-parametric two sample Wilcoxon signed-rank tests, comparing the quality scores between the two voices (NZ vs. US) and intonation tones (Neutral vs. Happy).(* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

in Fig. 1 were: "under 1 minute," "up to 5 minutes" and "over 5 minutes." The question attempted to determine the appropriate duration times for instruction dialogues, if they were to use the two voices installed on the robot. The results are summarised in the Table 6, and indicate that the instructions generated with the NZ accented voice with both intonation models, can be of at least 5 minutes in length.

Accent	Tone	$t < 1\text{min}$	$t \leq 5\text{min}$	$t > 5\text{min}$
NZ	Neutral	2	11	16
NZ	Happy	5	10	14
US	Neutral	18	10	1
US	Happy	11	16	2

Table 6: Results showing the number of participants indicating how long they could listen to a given voice before tiring.

In the rate of speech measure, the majority of participants indicated they found the speech speed 'just right' by scoring it at 50 points. To illustrate the differences in score we have marked the voices that scored less than 50 as too slow, and the ones that scored over 50 points as too fast. The results are summarised in the Table 7. Results indicate that the generated voice speed is judged as being appropriate. There are concerns as to the validity of these speech rate results, primarily because the scale used could easily be misinterpreted. Unlike the quality and intelligibility scales, the rate of speech had three anchors that ranged in interpretation from 'bad(too slow) — 0', to 'good (just right) — 50' to 'bad (too fast) — 100' again.

Accent	Tone	Too Slow	Just Right	Too Fast
NZ	Neutral	5	22	2
NZ	Happy	6	15	8
US	Neutral	8	18	3
US	Happy	8	18	3

Table 7: Summary of the results relating to the rate of speech. Participants were asked to indicate if a particular voice speed was 'too slow', 'just right', or 'too fast'.

In the last question the participants were asked to indicate their preferred voice out of the possible four. The majority preferred the NZ accented voice that used the 'neutral' intonation, followed by the NZ 'happy' voice. Pushing the voice button on the screen at this time would play a speech sample of the word "Hello" generated with the selected voice. The overall responses are given in Table 8.

3.3. Discussion

There seems to be very little intelligibility difference between the NZ and US voices, with the only significant difference be-

Participant Group	NZ		US	
	Neutral	Happy	Neutral	Happy
All	22	4	1	2
Residents	15	2	1	1
Staff	7	2	0	1

Table 8: Results showing the voice preference by the number of participants indicating a voice as their most liked.

tween the NZ happy and US happy. This result follows findings in [5] where it was found that prosody in synthetic speech does not contribute to its intelligibility. The participants gave middle of the road scores in intelligibility and combining these with the results asking the length of time participants could listen to the voices before tiring, leads us to infer that participants could understand the synthetic speech. In the quality measure, the NZ voice is clearly more preferred. We see some differences in scores given to the two voice tones, showing evidence that participants could distinguish between the neutral and happy models. Looking into the differences in intonation modelling scores, indicate that happy intonation has a negative effect on the voice quality. Intonation for diphone voices is modified using time domain overlap and add method, which is distributed with festival. Manipulating intonation introduces audible artefacts into the generated speech. The extent to which the original diphones are modified in happy intonation is greater than neutral which results in greater signal degradation, and could explain the negative effect of happy intonation on quality scores. To address this problem we are implementing a new synthesis method based on harmonic and noise modelling of speech through which better quality prosodic manipulation is possible [14]. When we compare the scores between the two participant groups we see a difference in opinion to the voices between the staff and residents. The residents on average give lower scores and are more consistent in their evaluations than the staff, a result which is supported by wider research that older people perceive synthetic speech differently. When the participants were asked to indicate how long they could see themselves listening to each voice before tiring, the NZ voices performed better than the US ones. The NZ neutral voice had the largest number of participants indicating that they could listen to the voice up to and longer than five minutes before tiring. Again we see a difference in opinion regarding the preference in the two voices with the NZ scoring higher than the US.

4. Conclusion

In this paper we have presented the voice perception evaluation, which formed a smaller part of a greater robot acceptability study conducted in a retirement village. The study participants included both staff and residents from the village. We have asked the participants to evaluate the artificial speech in an attempt to quantify intelligibility, quality and speech rate. The results of the study support earlier findings that speech quality needs to be improved, and that different voices are perceived differently with a clear preference for the NZ voice versus the US. We note that in order for the intonation modelling to be a more flexible feature, better technique needs to be implemented through which signal degradation can be reduced. A future stand alone study will need to be conducted to take a more detailed look into the differences between intonation models and the differences in perception between older and younger listeners.

5. Acknowledgements

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 the R&D program of the Korea Ministry of Knowledge and Economy (MKE) and the Korea Evaluation Institute of Industrial Technology (KEIT). [KI001836, Development of Mediated Interface Technology for HRI]. Authors would like to acknowledge the work of Tony Kuo, Richie Wong, Ulrike Unger for their work in getting the robot ready for the trial.

6. References

- [1] M. Wolters, P. Campbell, C. DePlacido, A. Liddell, and D. Owens, "The effect of hearing loss on the intelligibility of synthetic speech," in *Proc. Intl. Conf. Phon. Sci.*, 2007.
- [2] B. MacDonald, W. Abdulla, E. Broadbent, M. Connolly, K. Day, N. Kerse, M. Neve, J. Warren, and C. I. Watson, "Robot assistant for care of older people," in *5th International Conference on Ubiquitous Robots and Ambient Intelligence*, November 20-22 2008.
- [3] A. Black, P. Taylor, and R. Caley, "The festival speech synthesis system," 1998. [Online]. Available: <http://www.cstr.ed.ac.uk/projects/festival.html>
- [4] C. I. Watson, J. Teutenberg, L. Thompson, S. Roehling, and A. Igic, "How to build a new zealand voice," in *NZ Linguistic Society Conference, Palmerston North*, November 30 - December 1 2009.
- [5] C. Paris, M. Thomas, R. Gilson, and J. Kincaid, "Linguistic cues and memory for synthetic and natural speech," *Human Factors*, vol. 42, no. 3, p. 421, 2000.
- [6] M. Wolters, P. Campbell, C. DePlacido, A. Liddell, and D. Owens, "Making speech synthesis more accessible to older people," in *Proc. 6th ISCA Speech Synthesis Workshop*, August 2007.
- [7] R. Roring, F. Hines, and N. Charness, "Age differences in identifying words in synthetic speech," *Human factors*, vol. 49, no. 1, p. 25, 2007.
- [8] M. Wolters, P. Campbell, C. DePlacido, A. Liddell, and D. Owens, "Making synthetic speech accessible to older people," in *Proc. Sixth ISCA Workshop on Speech Synthesis, Bonn, Germany*, 2007.
- [9] M. Schroder, "Emotional speech synthesis: A review," in *7th European Conference on Speech Communication and Technology*, September 3-7 2001, pp. 561-564.
- [10] A. Igic, C. Watson, J. Teutenberg, E. Broadbent, R. Tamagawa, and B. MacDonald, "Towards a flexible platform for voice accent and expression selection on a healthcare robot," in *In proceedings of the 7th Australasian Language Technology Association Workshop*, December 3-4 2009.
- [11] R. Q. Stafford, E. Broadbent, C. Jayawardena, U. Unger, I. H. Kuo, A. Igic, R. Wong, N. Kerse, C. I. Watson, and B. MacDonald, "Improved robot attitudes and emotions at a retirement home after meeting a robot," in *19th IEEE International Symposium on Robot and Human Interactive Communication [In press]*, 2010.
- [12] C. Jayawardena, I. H. Kuo, U. Unger, A. Igic, R. Wong, C. I. Watson, R. Q. Stafford, E. Broadbent, P. Tiwari, J. Warren, B. MacDonald, and J. Sohn, "Deployment of a Service Robot to Help Older People," in *IEEE/RSJ International Conference on Intelligent Robots and Systems [In press]*, 2010.
- [13] R. Weide, "The Carnegie Mellon pronouncing dictionary: cmudict. 0.4.[On-line]," 1995.
- [14] A. Syrdal, Y. Srylianou, L. Garrison, A. Conkie, and J. Schroeter, "TD-PSOLA versus Harmonic plus Noise Model in diphone based speech synthesis," in *IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING*, vol. 1. Citeseer, 1998.