International Journal of Social Robotics (2021) 13:2119–2137
https://doi.org/10.1007/s12369-020-00691-4

Empathetic Speech Synthesis and Testing for Healthcare Robots

Jesin James[1] · B. T. Balamurali[2] · Catherine I. Watson[1] · Bruce MacDonald[1]

Accepted: 10 August 2020 / Published online: 11 September 2020

Abstract
One of the major factors that affect the acceptance of robots in Human-Robot Interaction applications is the type of voice with which they interact with humans. The robot's voice can be used to express empathy, which is an affective response of the robot to the human user. In this study, the aim is to find out if social robots with empathetic voice are acceptable for users in healthcare applications. A pilot study using an empathetic voice spoken by a voice actor was conducted. Only prosody in speech is used to express empathy here, without any visual cues. Also, the emotions needed for an empathetic voice are identified. It was found that the emotions needed are not only the stronger primary emotions, but also the nuanced secondary emotions. These emotions are then synthesised using prosody modelling. A second study, replicating the pilot test is conducted using the synthesised voices to investigate if empathy is perceived from the synthetic voice as well. This paper reports the modelling and synthesises of an empathetic voice, and experimentally shows that people prefer empathetic voice for healthcare robots. The results can be further used to develop empathetic social robots, that can improve people's acceptance of social robots.

Keywords Social robots · Emotional speech synthesis · Artificial empathy · Prosody modelling · Healthcare

1 Introduction

In Human-Robot Interaction (HRI), the focus is given to make robots learn to react to users socially and engagingly [1]. Such social robots are used for various applications such as education (e.g. [2]), passenger guidance (e.g. [3]) and healthcare (e.g. [1,4,5]). Healthcare robotics is the focus of this research study. The healthcare robot (*Healthbots project* [6]) [7,8] is an application of human-robot interaction under development at the Centre for Automation and Robotic Engineering Science, the University of Auckland, New Zealand. This project aims to develop social robots that provide support and care to people living in nursing homes. The role of these *Healthbots* will be to assist the medical staff in aged-care facilities by being a companion to the aged people [9]. Currently, the technology is undergoing additional field trials in realistic environments and commercialisation [6]. This paper describes the journey towards developing an empathetic voice for *Healthbots*. The next two sections explains the motivation to develop an empathetic voice (Sect. 2) and details about empathy in social robot applications (Sect. 3). This is followed by Sect. 4 describing a pilot study conducted to understand if people prefer empathetic voice in Healthcare robots. Section 5. Further, emotional speech synthesis (Sect. 6) and another experiment (Sect. 7) to evaluate the acceptance of synthesised empathetic voice are also described in detail. Section 8 concludes the paper.

✉ Jesin James
jesin.james@auckland.ac.nz

B. T. Balamurali
balamurali_bt@sutd.edu.sg

Catherine I. Watson
c.watson@auckland.ac.nz

Bruce MacDonald
b.macdonald@auckland.ac.nz

[1] Centre for Automation and Robotic Engineering Science, Department of Electrical, Computer, and Software Engineering, The University of Auckland, Auckland, New Zealand

[2] Singapore University of Technology & Design, Singapore, Singapore

2 Motivation: Acceptance of Social Robots

In this section, the motivation for building empathetically speaking social robots is discussed. How humans interact with robot's in social situations and the impact that the robot's

voice has on their acceptance are discussed in detail. Evidence from past research is used to emphasise the importance of the robot's synthesised voice on acceptance of robots by humans. Robots that interact in social situations as companions are novel to people who use them due to the very few preconceptions about their attributes and behaviour that people have. People rationalise this novelty by projecting the familiar human-like characteristics, emotions and behaviour onto them [10]. This behaviour is called Antropomorphism[1]. First, the general factors that improve social robots' acceptance are discussed, leading to the impact of the robot's voice on acceptance.

2.1 Acceptance of Social Robots

When robots serve as companions in social situations, their acceptance among users is a primary design consideration. Some factors that enhance robots' acceptance are their appearance, humanness, personality, expressiveness and adaptability [7]. Many studies have looked into such factors that need to be taken care of, in order to improve the acceptance of social robots, with prime focus given to the elderly[2] [7,11,12]. Previous studies have observed that people anthropomorphise robots [10,13]. A study by Heerink et al. defines the social abilities that humans expect from robots [14]. The results of the study, which identifies the factors that encourage older adults to accept robots, are summarised as the Almere Model [11]. Only the results leading to the development of Almere model are discussed here, as the Almere model was primarily developed based on studies on elderly users of healthcare robots, which is the application considered in this study. Also, the results of the model adequately lead to the relevance of the voice type on the user's acceptance of social robots. The results of the study looking into the factors that encourage older adults to accept robots, provide a clear indication that humans tend to anthropomorphise robots. Humans anthropomorphise robots by expecting social abilities from them. According to the study, the social abilities that humans expect from social robots are that they should: cooperate, express empathy, show assertivity, exhibit self-control, show responsibility, competence and gain trust.

Now, the big question is - how can these social abilities be embedded in social robots? The aforementioned social abilities can be expressed during the scenarios in which a human and robot interact through some means of communication. This communication occurs in multiple modalities, principally the auditory and visual mediums [13]. Even for the *Healthbots* considered in this study, the communication is visually through the information displayed on a screen and verbally using spoken dialogues of the robot. As the verbal communication is also used by the robot to interact with humans, the synthetic voice of the robot plays a significant role in determining how people anthropomorphise robot, which in-turn impacts the robot's acceptance.

2.2 Impact of Robotic Speech on Anthropomorphism

Speech is a primary mode of communication between robots and humans. People's anthropomorphism of robots is impacted by the type of synthetic voice used by robots to converse, and this also affects the robots' acceptance. Adding literature evidence to this statement regarding the relation between synthesised speech and acceptance of the robot, a summary of various studies about this concept is presented here. Research on robot voices is based on various studies on the impact of speech from artificial agents on anthropomorphism and their acceptance [15]. Examples of social robots that use synthetic speech to interact with people are Kismet [16], a storyteller robot [17] and reception robots [18].

Experiments reported in [19] indicate that people make judgements of robots' personalities based on their voice. A study reported in 2003, [13] discusses and reviews other studies that show the impact of a robot's voice/speech on people's judgements of the robot's perceived intelligence. This study has attributed the perception of intelligence of robots by humans to be similar to the judgements humans make about other humans. Hence this can be considered as a direct impact of anthropomorphism of the social robot impacted by speech. During the same time period, Goetz et al. [20] experimented on how people's cooperation with a robot varied depending on the speaking style of the robot (synthesised speech) when it was instructing a team to complete a task. One team performed the task instructed by a playfully speaking robot. A different team performed the same task instructed by a robot with a neutral voice. Here, the performance of the team that did the task under the playful robotic voice was better than the other team. This would mean that humans get motivated by a robotic voice, even though the voice is synthesised. Another experiment conducted in 2006 investigated the difference when affect[3] was added to the robot's synthesised speech [21]. Here, a robot guided people to complete a task. In one case, the robot expressed urgency to motivate people to complete the given task. In the other case, the robot

[1] Anthropomorphism refers to the tendency of humans to see human-like characteristics, emotions, and motivations in non-human entities such as animals, gods, and objects.

[2] The focus is on elderly as the *Healthbots* - which is the application on which this study is based, is developed for aged-care facilities.

[3] Affect is a concept used in psychology to describe the experiencing of feeling or emotion. The addition of emotions/feelings into the robot's speech is explained here. This has lead to a field called affective computing, which includes developing systems that can recognise, interpret and respond to emotions, and also produce them.

spoke with a robotic voice, without motivating the people. The study arrived at the observation that the team that did the task under the expressive robotic voice performed better than the team under the neutral voice[4]. In another experiment conducted in 2012, users listened to a human-like voice type and a robot-like voice type. The two voice types were spoken by robot Flobi using its synthesised voice. The decision of the vocal cues to produce the robot-like and the human-like voice was based on pre-tests regarding human-likeness vs robot-likeness. The acceptance of the robot with the human-like voice was better than the robot-like voice [22], although specifics of the vocal cues are not reported in the paper.

Based on the studies discussed in the above paragraph, it can be seen that people anthropomorphise robotic speech, i.e., associate human attributes to the robot based on the voice, which is synthesised. This is evident in the way the robot's acceptance improved with a change in voice and how people performed better when the robot spoke expressively. Upon understanding that the synthesised voice of the social robot is a key factor in its acceptance, it is then necessary to decide what type of voice is suitable for social robots. In the next section, the type of expressive voice needed for healthcare robots is identified based on past studies and perception experiments.

3 Empathy and Emotions Needed in Healthcare Robots

Recently, it was observed that roboticists build robots in the anthropomorphic form to improve their acceptance through embodied cognition, but users are disappointed by the lack of reciprocal empathy[5] from these robots [23]. Due to the lack of definition of empathy for human-robot interactions, the tendency for humans to anthropomorphise robots is used as the key to deriving a definition in [24]. Empathy in human-human interaction is the behaviour that enables one human to experience what another human feels and respond to it. It is an emotional response that is automatically evoked by one's understanding of the other human [25]. When the companion is a robot, empathy in human-robot interaction can be defined as the programmed affective reaction of the robot to the behaviour of the human that it can sense according to the technology embedded in it. It is also called *Artificial empathy* in human-robot interaction studies [24,26].

[4] In this study, *neutral* voice is defined as voice spoken naturally (i.e. without stress). For the robot with *expressive* voice, stress was included to express urgency.

[5] Empathy is the ability to understand and share the feelings of another.

3.1 Prosody Component for Empathy Portrayal

The empathy portrayal by humans involves various communication modalities, such as facial, vocal (non-verbal and verbal) [23]. For robots, these communication modalities exist. The focus of this research study is to use speech to express empathy. Speech has two components [27] pertinent to empathy:

1. The *verbal component*, which focuses on the words alone.
2. The *prosody component*, which can be thought of as the melody and rhythm of speech. Emotions are expressed by variations in *prosody component* (like varying intonation, speech rate, stress) [27,28]. This *prosody component* refers to the affective prosody.

Empathetic behaviour via speech can be depicted by a proper choice of words, which is the *verbal component,* and the emotions portrayed by the speaker, which is the *prosody component*. The choice of words determines the lexical features, which contributes to the *verbal component*. The emotions govern the acoustic features [29] contributing to the *prosody component*. Often empathy is incorporated into synthesised speech by the inclusion of words that convey an affective response (called dialogue modelling). As stated in [30], a robot nurse assistant should be able to greet people, sound happy when informing patients of good results and express sympathy or encouragement when the test results are not satisfactory. So, a combination of speech and visual channels of the robot can be used to impart empathy. This research focuses on speech alone. Empathy is communicated more via the non-linguistic channel, as stated by [31]. The same study also cites research indicating that a speaker's emotional state can be expressed without the use of words and be understood just by listening to the speaker's voice. Hence, along with words that convey empathy, the emotions that are used in saying those words play an inevitable role in making the listener understand the speaker's empathy towards them. In this study, the aim is to express empathy in synthesised speech for healthcare robot by:

1. Using the **speech alone** as the medium for communication between the social robot and the human user.
2. Modelling the **prosody component of speech** to express empathy.

3.2 Empathy and Emotional Expression in Robotic Speech

Empathy in social robots is a relatively new research area. To date, there are only a few published research studies, and the major findings are discussed here. One study (in 2005) in this area of empathetic social robots [19] has shown that

robots with empathy received positive ratings in the areas of likeability and trustworthiness. They were also perceived as supportive. Further, a study in 2013 shows that robots with empathy have reduced frustration and stress among users, as well as improved the users' comfort, satisfaction, and performance doing a set task [32]. Finally, in 2018, James et al. [24] reports that the positive effects of empathy are produced in users only when the robot's expressions are in congruence with the users' affective state.

Motivated by these research findings, good modelling of empathy and expression of emotion is required while building the robots. This modelling will avoid a potential mismatch between the users' expectation of the robot's emotion (based on the application) and the actual emotions expressed by the robot, which can otherwise lead to a negative effect. The studies on empathy in healthcare robots are also limited. The study reported in [24] has explored empathetic healthcare robots and people's preference to them. Also, [32] addresses the benefits of an empathetic voice (among other voice attributes like pitch and humour) improving users' ease of interaction with the robot, while stating direct advantages for healthcare robot applications.

Currently, in the *Healthbots* used in this research, a New Zealand English voice is incorporated, and pilot studies were conducted with regards to the naturalness of the voice [8,33]. The voice that has no emotional expression can be called a "neutral" voice. It was noted repeatedly that familiarity with the voice and closeness to human-like speech improves the positive attitude towards robots. This positive attitude, in turn, improves the acceptance of the robot. Also, the acceptance level can increase after meeting the robot assistant [8], and if the robot speaks with a local accent [34]. However, the age and native language of the user can impact on the perceived intelligibility of the robot voice. Watson et al. [33] found that the non-native listeners performed significantly worse than the native listeners in a synthetic speech condition. The authors report that the in-depth language model that the native speakers have, helped them parse the synthetic speech better than the non-native speakers.

From the studies discussed here, a robot that converses in a familiar language and factors such as the speaking style (including local accent), emotional expression and empathy are critical factors in improving their acceptance. As humans anthropomorphise robots, an empathetically interacting robot is expected to increase the level of acceptance of social robots based on the evidence presented in Sects. 2 and 3. To test that these findings apply to the healthcare robots that are used in this study, a large-scale perception experiment was conducted.

4 Study 1 (Pilot)

[6] This study (discussed in detail in [24]) involves a perception experiment to evaluate whether human subjects perceive empathy in robot speech. For this experiment, empathy is expressed through speech, with prosody being varied with the relevant melody and rhythm; i.e., by adding appropriate emotions to the words in the speech of the robot. A perception test was conducted to address the following *research questions*:

1. *Research question I:* Can people perceive empathetic behaviour from a robot when only the emotions in its speech are used to express empathy?
2. *Research question II:* Do people prefer empathetic voice from robots or a non-empathetic robotic voice?
3. *Research question III:* What factors of speech can be related to an empathetic voice?

The robot used for the study is the *Healthbot*. There are three different situations in which the robot speaks to the patient - (1) greeting the user, (2) providing medicine reminders, (3) guiding the user to use the touch interface.

Dialogues were framed for each of these situations, and included dialogues already used by the *Healthbots* (more details about dialogues are in Sect. 5). Each situation had 20-25 dialogues. A professional voice artist produced the dialogues in two variations.

1. One variation used a monotone voice with no variation in prosody features like intonation and intensity. This voice will be referred as *robotic voice* here.
2. The second variation was spoken like a nurse speaking empathetically to a patient, with changes in emotions. This voice will be referred as *empathetic voice* here.

A professional voice artist was used instead of synthesised voice as the current synthesised voices used in the *Healthbot* lack naturalness and quality as they are still under development. Also, robotic voices that were empathetic were not able to be created at the time of the study. Indeed, this was one of the points of the study - to ascertain what types of emotions were required for an empathetic voice. This is a pilot study to understand what type of voices are preferred by participants. If the empathetic voice is natural-sounding and the robotic voice is synthesised, it may cause the participants to be biased towards the more natural-sounding voice. This bias needed to be avoided, and hence, acted out voices were used for both the cases.

[6] The study is approved by the University of Auckland Human Participants Ethics Committee (UAHPEC) on 20/10/2017 for 3 years. Ref. No. 019845.

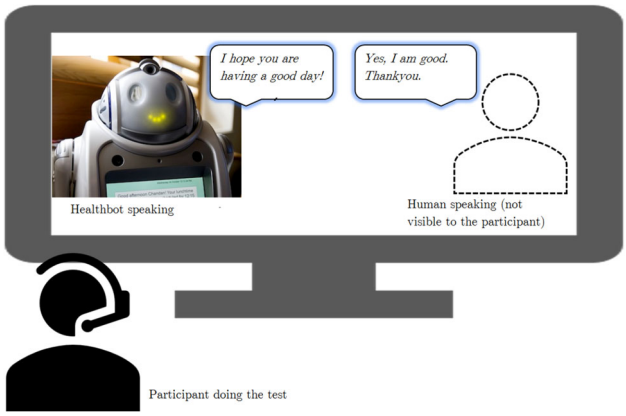


Fig. 1 The participant taking part in the study watching the Healthbot talking

An illustration of a participant taking the test watching the healthbot talking is shown in Fig. 1. A link to the online survey is provided here[7] where the video of the robot with the different voices can be seen. 120 participants, aged 16–65 (age distribution shown in Fig. 2), completed study 1. Majority of the participants were from the age group 26–35. Based on their self-reporting, all participants had above average hearing ability, with 50 participants being first language New Zealand English speakers (L1)[8] and 70 were bilingual speakers (L2). All participants completed the test. The participants could choose to use headphones or loudspeakers according to their convenience. In total, 20% of the participants used loudspeakers, and the remaining 80% used headphones. Each participant took approximately 15 minutes for the test. An online survey platform Qualtrics[9] was used. No restriction was put on recruiting participants for the test other than a minimum age of 16. Such a generalised participation was selected as the *Healthbots* will be used in applications where the users may not have any knowledge about robotics. The participants went through three parts of the survey to address each of the research questions.

4.1 Addressing Research Question I - Pilot

4.1.1 Design

For addressing *Research question I*, both the voice variations spoken by the actor used the same words with variation in only the prosody component. The robot had a neutral facial expression. The patient's dialogue was spoken by a speaker in the same manner regardless of the variation in the robot's voice. An example of a *Healthbot* dialogue was "It seems like you are taking a long time to take your medicine.", for which the patient responds "I am lately very slow in all the tasks I do!". The participants could see and hear the *Healthbot* speaking (as shown in Fig. 1), but the patient speaking to the robot was not shown to them. The patient was not shown to enable the participants to feel as if the robot was speaking to them, and hence, they could rate the robot's interaction with them. Each participant was given one scenario (greetings, reminders or instructions) with the two variations (robotic voice and empathetic voice). Both voice variations were shown to each participant one after the other, with the empathetic voice first, followed by the robotic voice. After seeing each scenario, the participants had to rate the voice based on an empathy scale.

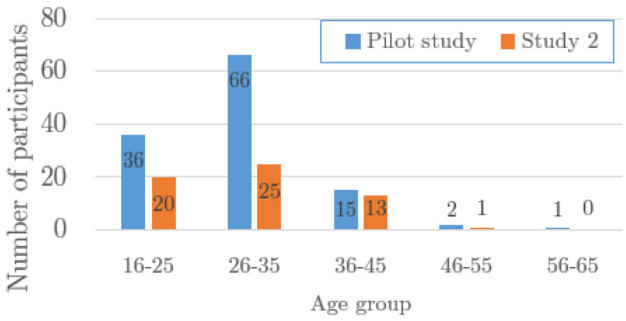


Fig. 2 The distribution of age groups in pilot study and study 2

The questions asked to the participants for *Research question I* were based on the empathy measuring scale from the Motivational Interviewing Treatment Integrity (MITI) module [35] used for human-human interaction, which was extended to human-robot interaction in [24]. MITI module defines five scales to rate a clinician's empathy. The 5 point scale used in MITI and the experiment are shown in Table 5 in the Appendix and Table 1 of [24]. A score of 1 represents the least empathy according to the MITI scale. The dialogues were not randomised as they were framed as a conversation between the robot and the patient. First, the participants saw and heard the robot speaking the empathetic voice. They were then asked to rate the voice based on the scale. They then listened to the robotic voice and rated it based on the same scale. A within-participants design is used here as the difference between the two voices types may not be captured if the same person does not hear both the voice types. This may cause the robotic voice to also be perceived as empathetic (although lower levels) if heard separately.

4.1.2 Results

Figure 3 shows the empathetic behaviour rating given by the participants to the two voices based on the empathy rating

[7] https://auckland.au1.qualtrics.com/jfe/form/SV_2hn68L1Df9lMXIh

[8] First language and second language speakers distinction is based on New Zealand English. Participants were classified as L1 if they were living in New Zealand since age seven at least.

[9] Version XM of Qualtrics. Copyright 2019 Qualtrics. Qualtrics and all other Qualtrics product or service names are registered trademarks or trademarks of Qualtrics, Provo, UT, USA. https://www.qualtrics.com.

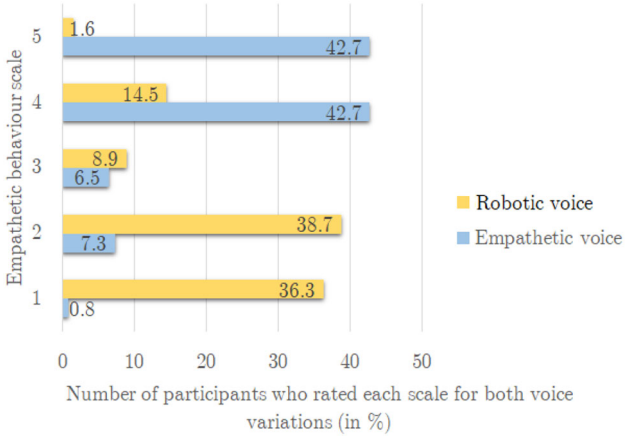


Fig. 3 Participants' rating of the two voice types. The blue bar (bottom bar of every pair) represents the empathetic voice, and the yellow bar (top bar of every pair) represents the robotic voice. (Color figure online)

scale. The bar chart shows the percentage of participants who have chosen a particular empathetic behaviour scale (1 to 5) for the robotic voice and empathetic voice.

Almost 85% (sum of scores for 4 and 5 scales − 42.7% + 42.7% - showing high levels of empathy perception) of the participants felt that the robot with the empathetic voice showed great interest in the patient and tried to engage with them. Half of this group felt that the robot responded well, while the other half felt that it could do better. The authors believe the reason people felt that the robot could do better may be related to people's inhibitions that a robot cannot feel the patient's situation; instead, it is just programmed to respond accordingly. Conversely, 75% of the participants (sum of scores for 1 and 2 scales showing low levels of empathy perception − 38.7%+36.3%) felt that the robot with the robotic voice had little interest in the patient (given a rating of 1 or 2). Curiously, two participants (1.6%) felt that the robotic voice still showed a high level of empathy. As the empathy rating scale decreases from 3 to 1, it can be seen that less than 15% of the participants have given a lower rating for the empathetic voice. At the same time, for the robotic voice, most of the participants have given a rating of 1 or 2 on the scale. This suggests that robotic speech with appropriate words alone is not sufficient for people to perceive an empathetic behaviour from the robot.

4.1.3 Statistical Analysis

Because the data is skewed for both the voice types, a Wilcoxon signed-rank test was conducted to assess the difference between the robotic voice and the empathetic voice. The empathy scale ratings 1 to 5 given by the participants was used for the analysis. The results (shown in Table 1 Row 3) indicate that the empathetic voice ratings (Median = 4) are significantly higher than the robotic voice ratings (Median = 2), $p < 0.001, r = 0.7$. An effect size $r = 0.7$ indicates that the effect is large according the Cohen's benchmark for effect sizes [36]. Hence, it can be summarised that the empathetic voice received higher ratings than the robotic voice, and the result is statistically significant.

4.2 Addressing Research Question II - Pilot

4.2.1 Design

To evaluate *Research question II* of the experiment, both voice variations were shown to each participant. Then they were asked to judge which voice they preferred. The dialogues lasted for almost 1–2 min for each of the variations.

4.2.2 Results

In total, 113 of the 120 participants (about 95%) preferred the empathetic voice over the robotic voice, which is a robust result.

4.3 Addressing Research Question III - Pilot

4.3.1 Design

For evaluating *Research question III*, participants were asked reasons (free-response and forced-response) for choosing their preferred voice from the robotic voice and the empathetic voice. The reasons given for the participants to choose from are listed on the left end of Fig. 4. Each dialogue spoken by the robot was listed, and the emotion/feeling/tone that a patient might expect from a nurse speaking was associated as a label with each dialogue. For example, a dialogue, "It seems like you are taking a long time to take your medicine", would be expected to be said with *concern* and *empathy*. The first author gave similar labels to each dialogue. The options given to the participants were based on these labels. Also, they were asked which emotions they could feel when listening to each of the voices from the options *angry, happy, sad, excited, concerned, anxious, encouraging, assertive, apologetic* and *other*. The first four emotions are primary emotions [10] (excluding neutral) and the rest are words indicating secondary emotions.

[10] Primary emotions are emotions that are innate to support reactive response behaviour (Eg. angry, happy, sad, fear). The basic/primary emotions are based on the studies by Ekman [37]. Secondary emotions arise from higher cognitive processes, based on an ability to evaluate preferences over outcomes and expectations (Eg. relief, hope) [38,39]. Various theories define primary and secondary emotions [40], but here we will be looking at emotions that were studied as part of human-robot interaction and speech synthesis studies.

Table 1 Statistical analysis results of robotic voice and empathetic voice in study 1 (Pilot) and study 2

Study type	Robotic voice Mean ± SD	Empathetic voice Mean ± SD	Robotic voice Median	Empathetic voice Median	p-value	Effect size	Confidence interval
Study 1 (pilot) N = 120	4.1 ± 0.9	2.0 ± 1.0	4	2	< 0.001	0.7	[2.0, 2.9]
Study 2 (synthesised) N = 51	3.2 ± 1.2	4.0 ± 1.1	4	3.5	0.003	0.3	[-1, 0]

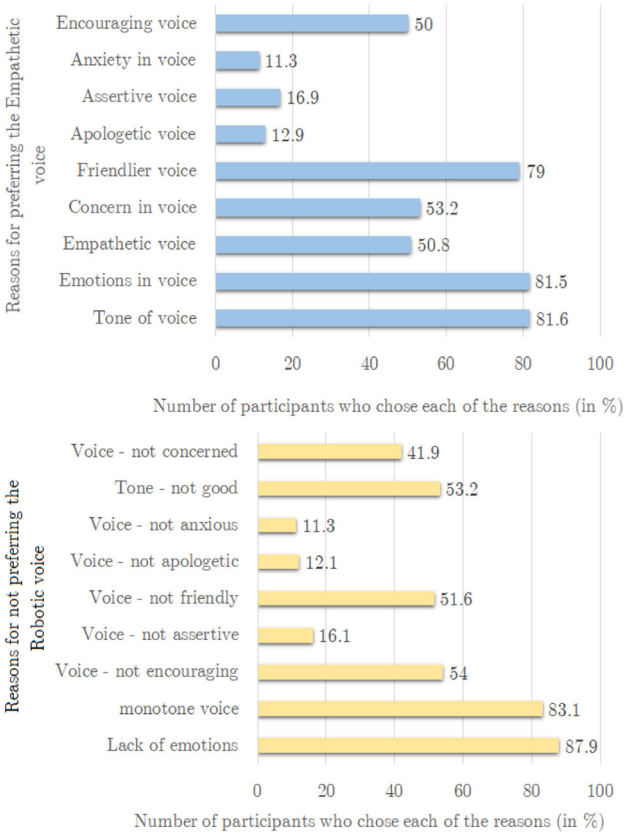


Fig. 4 Participants' reasons (forced-responses) for choosing the empathetic voice (in blue) and not preferring the robotic voice (yellow)

4.3.2 Results

The reasons for choosing the empathetic voice (blue colour) and not preferring the robotic voice (yellow colour) from the forced-responses are given in Fig. 4, along with the percentage of the participants who selected these reasons. (There were only a few free responses and no common theme evolved from them). The most influencing factors for preferring the empathetic voice was the tone and emotions in the voice, closely followed by friendliness in the voice. People could also perceive empathy, concern and encouragement in the voice, which also contributed to their choice. Looking at the reasons for not choosing the robotic voice, the lack of emotions and monotony in the voice are the most influencing factors (as the number of participants who chose these reasons is higher than the rest), followed by lack of encouragement and concern in the voice. It is important to restate here that both the voices had the same verbal content. This content contained words that portrayed encouragement or concern (For example - "Oh dear! Exercising regularly is very crucial for you" expresses concern in the words and was used for both the empathetic voice and the robotic voice). It was when these words showing active engagement were spoken expressively with the appropriate emotions that participants could perceive the empathetic behaviour of the robot. This suggests that for developing empathetic artificial agents, the interaction via speech plays a role in influencing people's perception of robots. Even though other modalities like facial expressions are under research, speech synthesis needs to be developed to express more human-like empathy. Communication via speech comprises of dialogue modelling along with the synthesis of the required emotions. From this test, it is also evident that proper dialogue modelling alone is not enough. Participants perceived higher empathetic behaviour only from the voice where the emotions matched the dialogues spoken by the robot.

Responses to the emotions that participants could perceive from the empathetic voice are consolidated in Fig. 5. The responses that came under *other* were warm, friendly and engaging. Only the participants who preferred the empathetic voice were required to provide this response, and each participant could provide multiple responses. Here, it can be seen that the emotions perceived by the participants in the empathetic voice are secondary emotions. This indicates that the synthetic voice spoken by the social robot should also be modelled to speak with a selection of secondary emotions.

The conclusions from the pilot study were:

Participants can perceive empathy from robots when empathy is portrayed by speech using variations in the prosody component. When the prosody variation is absent in speech (i.e. only the words in the sentences expressed empathy), participants perceived lower levels of empathy.

Participants prefer an empathetic voice from a robotic companion compared to a robotic voice (non-emotional) in a healthcare application.

The main factors that contributed to people's reason to prefer the empathetic voice are the emotions in the voice and the variations in prosody.

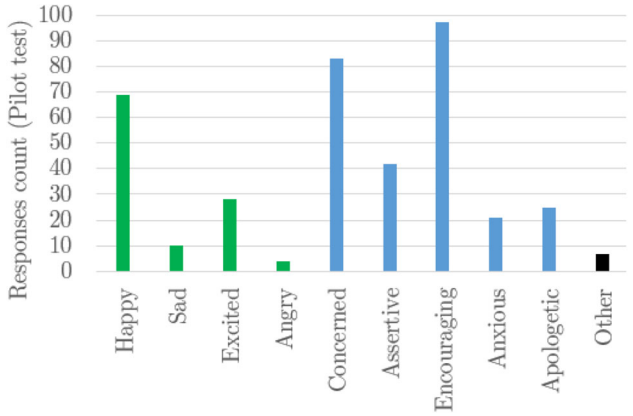


Fig. 5 Participants' responses for the emotions they could perceive from the empathetic voice (pilot study)

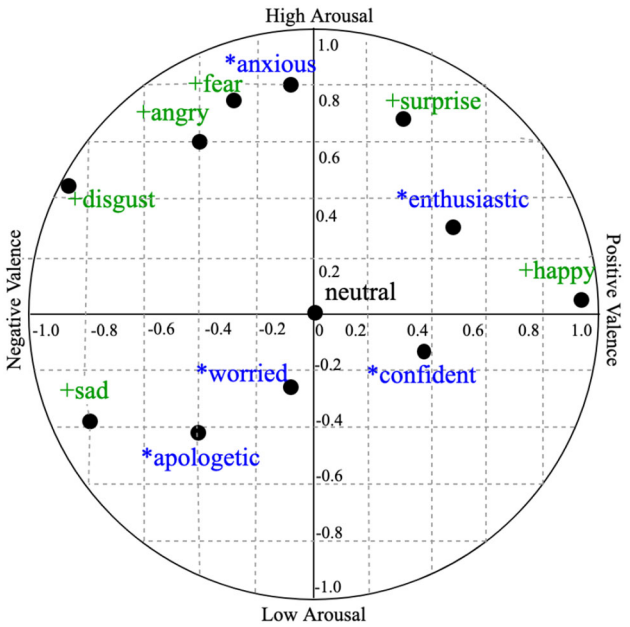


Fig. 6 Valence-arousal plane of emotions showing primary emotions defined by Ekman [37] (marked in green +) and the emotions identified for the healthcare robot based on [24] (marked in blue *). Adapted from [46]

The prosody component needs to be in alignment with the verbal component so that people can perceive empathy from a social robot. In order to correctly model the prosody component, the emotions needed for an empathetic voice and acoustic features to model an empathetic voice needs to be identified. The next section explains this in detail.

5 Emotion Analysis of Social Robot

From the pilot study conducted, it was found that the addition of emotions to the verbal component is essential for people's perception of empathy in robotic speech. To synthesise empathetic speech, proper modelling of emotions is essential to enhance the verbal component. To identify the emotions associated with an empathetic voice, an emotion analysis of the *Healthbot* was done as described in [24]. Defining an emotional range that can be called as "empathetic" was the focus of the study, and also a pre-requisite for synthesising empathetic voice. Each dialogue spoken by the robot in the empathetic voice was perceptually analysed and marked on the valence-arousal plane to identify the emotional range (details are provided in [24]) This analysis was independent of the responses provided by the participants in the pilot study. Based on the analysis, the emotions needed for an empathetic healthcare robot were identified as: *anxious, apologetic, confident, enthusiastic*, and *worried*.

These dialogues that are designed for the Healthbot require emotions that do not fall under the primary emotion categories (marked as green "+" in Fig. 6) but are rather variants of them which are the secondary emotions (marked as blue "*"). Many studies, including [30,41–45] have focused on social robots which are capable of synthesising speech with the primary emotions. As important, these primary emotions are in real life; this study of the dialogues suggests that synthesising these nuanced secondary emotions are for an empathetic robot voice.

5.1 Discussion Based on the Pilot Study and Emotions for Empathetic Robot

We believe that to improve HRI interactions, synthetic speech needs to have the capacity to emulate secondary emotions, in addition to primary emotion. This requires us to have knowledge of the acoustic features of these emotions. However, in contrast to the primary emotions, there are very few studies on the acoustics of these emotions. Further, with regards to the specific set of secondary emotions required for our Healthbots, there were no resources to get the acoustic features. To that end, an emotional corpus which includes these secondary emotions needs to be developed.

It is not possible to use existing speech corpora of the primary emotions to determine the acoustic features of the secondary emotions. The primary emotions are well apart in the valence-arousal plane. An inspection of the position of the secondary emotions in the valence-arousal plane (Fig. 6) shows that they are not well separated on the valence-arousal plane. This will be a significant challenge when trying to model and synthesise these secondary emotions; hence there is a need for a purpose-built speech corpora.

6 Emotional Speech Synthesis

6.1 Corpus Development

The emotions needed for the social robot are identified as - *anxious, apologetic, confident, enthusiastic* and *worried*. To study the secondary emotions, a New Zealand English speech corpus with strictly-guided simulated emotions was developed [47]. This corpus, called JLcorpus [11] contains five primary (*angry*, *excited*, *happy*, *neutral*, *sad*) and five secondary emotions (*anxious*, *confident*, *worried*, *apologetic*, *enthusiastic*). The JLcorpus has an equal number of four English long vowels-/a:/, /o:/, /i:/ and /u:/, to facilitate emotion-related formant and glottal source features comparison across vowel types. The corpus contains 2400 sentences spoken by two male and two female professional actors. The semantic context of all the sentences in the corpus was kept the same for all primary emotions, while the secondary emotions have 13 emotion-incongruent sentences and two emotion-congruent sentences. The inclusion of emotion-congruent sentences allows analysis of the effect of semantic influence on emotion portrayal and acoustic features, as seen in [48]. The emotion quality of the JLcorpus was evaluated by a large scale perception test of 120 participants, where the participants evaluated the emotions portrayed in the corpus. The corpus was labelled at the word and phonetic levels by webMAUS [49] with hand correction for wrongly marked boundaries.

6.2 Features Modelled

Modelling and synthesising the secondary emotions was done using three prosody features - fundamental frequency (f_0), speech rate, and mean intensity. A preliminary analysis of the emotions in the JLCorpus is reported in [47]. Detailed analysis of the f_0 contour based on the Fujisaki model (a method to parameterise the f_0 contour, more details to follow) is reported in [50]. The decision to model f_0 contour and speech rate is based on these analyses.

The Fujisaki model [51] parameterises the f_0 contour superimposing (all parameters marked for a sentence in Fig. 7: (1) the base frequency F_b (indicated by the horizontal line at the floor of the f_0 pattern), (2) the phrase component - declining phrasal contours accompanying each prosodic phrase, and (3) the accent component - reflecting fast f_0 movements on accented syllables and boundary tones. These commands are specified by the following parameters:

1. *Phrase command onset time* (T_0): Onset time of the phrasal contour, typically before the segmental onset of the phrase of the ensuing prosodic phrase. (Phrase command duration *Dur_phr* = *End of phrase time* $- T_0$)
2. *Phrase command amplitude* (A_p): Magnitude of the phrase command that precedes each new prosodic phrase, quantifying the reset of the declining phrase component.
3. *Accent command Amplitude* (A_a): Amplitude of accent command associated with every pitch accent.
4. *Accent command onset time* (T_1) *and offset time* (T_2): The timing of the accent command that can be related to the timing of the underlying segments. (Accent command duration *Dur_acc* = $T_2 - T_1$)

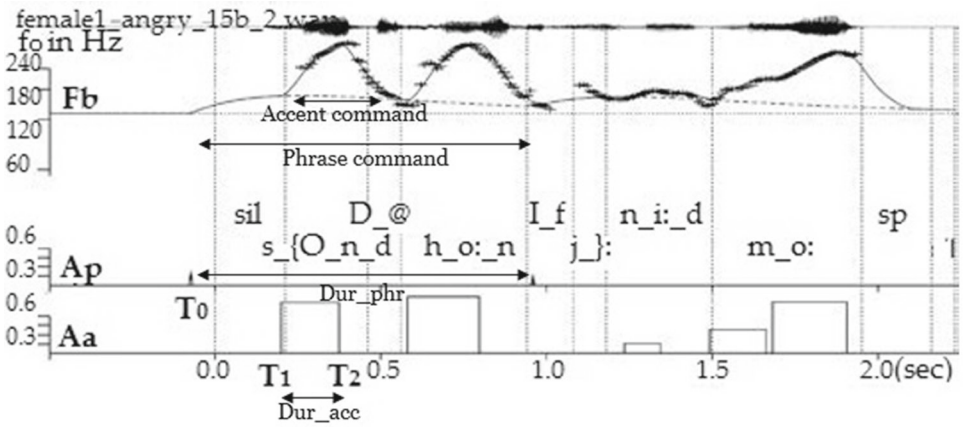
A_a, A_p, T_0, T_1, T_2, F_b are referred to as the Fujisaki parameters. *Dur_phr* and *Dur_acc* are derived parameters from the Fujisaki parameters. The Fujisaki parameters for each utterance was extracted using *AutoFuji extractor* [52]. Checking was done so that potential errors in f_0 tracking did not affect the parameters. Analysis of the effect of emotions on the Fujisaki parameters [50] showed that they were affected by the emotions, with accent command parameters (smaller units - A_a and Accent command duration $T_2 - T_1$) and F_b having the most significant effect.

Mean values were obtained for the speech rate (in syllables/s) [47] and intensity (in dB) of the sentences for rule-based modelling of these prosody features for each emotion.

6.3 Emotional Text-to-Speech Synthesis System

The inputs to an emotional text-to-speech synthesis system are the text to be converted to speech and the emotion to be produced. To facilitate real-time implementation, all the features used for prosody modelling here are based on these two inputs only. Fig. 8 shows the proposed system for emotional text-to-speech synthesis system. The input text is analysed linguistically to extract context features. A text-to-speech synthesis system for New Zealand English based on MaryTTS [53] has been built [54,55] . This New Zealand English text-to-speech synthesis system produces speech without any emotion and will be referred as *non-emotional speech* here. The input text is passed through the text-to-speech synthesis system to obtain non-emotional speech. The pitch is extracted from the non-emotional speech (by *Praat* Auto Correlation Function [56]) and label files are obtained from input text and non-emotional speech using the New Zealand English option of the Munich Automatic

[11] JLCorpus contains five primary and five secondary emotions. *"Assertive"* was one of the secondary emotions. The actors of the corpus were instructed to speak seriously and confidently while recording this emotion. In a previous paper, the reviewers strongly criticised the use of *"assertive"* as an emotion, and asked to reconsider it. From the existing list of emotions in Russel's circumplex model of emotions (Fig. 6), *"confident"* was the best match. This journey is a clear indication of the difficulty in analysing and classifying secondary emotions as they can be difficult to define. The corpus is available at: github.com/tli725/JL-Corpus.



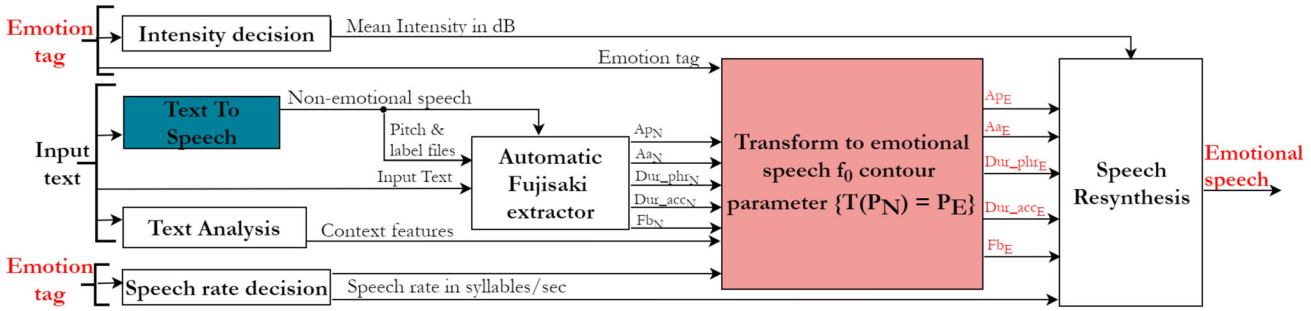


Fig. 7 Fujisaki parameters for *'Sound the horn if you need more'* (SAMPA phonetic symbols). T_0, T_1, T_2 marked for first phrase and accent commands

Fig. 8 Emotion-based f_0 contour transformation implementation in emotional text-to-speech synthesis system

Web Segmentation System [49]. The pitch and label files are passed on to *AutoFuji extractor* to obtain the 5 derived Fujisaki parameters of non-emotional speech (Ap_N, Aa_N, Dur_phr_N, Dur_acc_N, Fb_N - subscript *"N"* added to denote "Non-emotional"). The parameters are then time-aligned to the input text at the phonetic level. The decision of speech rate and intensity are made based on the emotion tag. With the context features, non-emotional speech Fujisaki parameters, emotion and speech rate as features, a transformation is applied on each of the non-emotional speech Fujisaki parameters to obtain the emotional speech parameters (the feature list given in Table 2). The context features and the non-emotional speech Fujisaki parameters are extracted by automatic algorithms, while the emotion, speaker are tags assigned depending on the emotion and speaker to which the conversion needs to be done. Feature extraction is done at the phonetic level as the transformation is phone-based.

Context features, speaker and emotion tag are categorical, while the non-emotional speech features and speech rate are continuous. Hand-corrected Fujisaki parameters obtained from the natural emotional speech is the target value to be transformed. The transformation is applied to the Fujisaki parameters of non-emotional speech to convert them to Fujisaki parameters of emotional speech. Ensemble learning using two regressors - *Random Forests* [57] and *Adaboost* [58] is employed. The average of the predicted values from these regressors would be the final transformed value. The hyperparameters of these regressors are tuned via grid search cross-validation. An emotion-dependent model is built for each of the Fujisaki parameters. All the emotions-dependent models are combined to form the f_0 contour transformation model. The database for modelling contains 7413 phones with their corresponding Fujisaki parameters. 80% of the database is used for training and rest for testing using random selection.

Once the non-emotional Fujisaki parameters are transformed to that of emotional speech, the resynthesis is done (last block in Fig. 8). Predicted Fujisaki parameters are time-aligned to the phones in the sentence. Fujisaki parameters are then used to reconstruct the f_0 contour by superimposing the F_b, accent commands and phrase commands based on the Fujisaki model. Once the f_0 contour is reconstructed, the speech is re-synthesised by pitch-synchronous overlap and add using *Praat*. The re-synthesised speech has f_0 contour obtained from the transformation model developed. Intensity and speech rate rules are assigned by emotion-based mean values.

6.4 Performance Analysis of Synthesised Speech

Performance analysis of the synthesised speech produced using the method described above was conducted using a perception test. In this perception test, sentences from the JLCorpus like "Tom beats that farmer" are used, and not

Table 2 Features used for f_0 contour transformation

Feature	Description	Extraction method
Context Features	Count = 102, Eg. accented/unaccented, vowel/ consonant	Text analysis at the phonetic level using MaryTTS.
Emotion tag	5 primary & 5 secondary emotions	Each emotion tag is assigned to the sentence
Speaker	2 male speakers	Speaker tag is assigned
Non-emotional f_0 features	5 Fujisaki parameters - Ap_N, Aa_N, Dur_phr_N, Dur_acc_N, Fb_N	Passing non-emotional speech to *AutoFuji extractor*.
Speech rate	Speech rate of the sentence	Based on emotion type.

Table 3 Confusion matrix showing hit rates from pair-wise subjective test

Actual emotions	Perceived emotions				
	Apologetic	*Anxious*		*Apologetic*	*Enthusiastic*
Apologetic	**97.9%**	2.1%	*Apologetic*	**100%**	0%
Anxious	0%	**100%**	*Enthusiastic*	1.4%	**98.6%**
	Confident	*anxious*		*Apologetic*	*Worried*
Confident	**88.3%**	11.7%	*Apologetic*	**64.3%**	35.2%
Anxious	12.4%	**87.6%**	*Worried*	32.4%	**67.6%**
	Enthusiastic	*Anxious*		*Confident*	*Enthusiastic*
Enthusiastic	**78.6%**	21.4%	*Confident*	**69%**	31%
Anxious	24.8%	**75.2%**	*Enthusiastic*	30.3%	**69.7%**
	Worried	*Anxious*		*Confident*	*Worried*
Worried	**97.9%**	2.1%	*Confident*	**95.2%**	4.8%
Anxious	4.19%	**95.9%**	*Worried*	22.8%	**77.2%**
	Apologetic	*Confident*		*Worried*	*Enthusiastic*
Apologetic	**94.5%**	5.5%	*Worried*	**97.9%**	2.1%
Confident	9.7%	**90.3%**	*Enthusiastic*	0.7%	**99.3%**

the robot dialogues as the Pilot study. Hence, this test is independent of the healthcare robot application, and tests only the quality of the emotions in the synthesised voice and thereby the machine learning approach. The synthesised emotional speech was evaluated by a perception test with 29 participants, where the participants evaluated emotions in the synthesised emotional speech. The participants had almost 50% distribution of first and second-language speakers of English (all variants of English were included). The majority of the participants were from the age group 16–35 (82%) and the remaining distributed over 36–65. All the participants had an average, above average or excellent (self-reported) hearing ability. In a forced response emotion classification task, the participants had to choose which emotion they perceived and group the sentences into the two emotion pairs provided. In total, 100 sentences were evaluated by 29 participants, giving 2900 evaluations.

Table 3 shows the confusion matrix obtained from the perception test for each emotion pair. The most confused emotion pairs were *enthusiastic* vs *anxious*, *confident* vs *enthusiastic* and *worried* vs *apologetic* (expected due to their closeness in the valence arousal levels). The confusion between *anxious* vs *enthusiastic* is the only problematic pair that can be an inappropriate reaction from the robot to the human user. On average, the perception accuracy was 87% to differentiate between the emotion pairs. The results obtained here are comparable to other emotional synthesis studies that used different techniques to model f_0 contour. For instance [59] reported 50% perception accuracy for expressions of good, bad news, question, [60] reported 75% perception accuracy for happy, angry, neutral and [61] reported 65% perception accuracy for joy, sadness, anger, fear. However, no past studies did contour modelling on the secondary emotions we studied; hence direct comparison will not be possible. It is of note that the accuracy rate for the secondary emotions in the perception test in [47] for the JLCorpus was 40%, which is considerably lower than the 87% obtained here. However, this was quite a different test where participants had five emotions to select from, rather than two. The secondary emotions are not as well separated on the valence-arousal place, and giving participants a choice of five emotions, will lead to confusions between emotions close to each other on the valence-arousal plane, as can be seen in Table 4 (Comparison between apologetic and anxious vs confident and enthusiastic).

7 Study 2 - Perceived Empathy from Synthesised Emotional Voice

[12] Based on the results obtained from the pilot study (Sect. 4), the emotions required for an empathetic voice were identified (Sect. 5), then modelled and finally synthesised (Sect. 6. The next step is to find out if humans can actually perceive empathy from the developed voice. A second perception test with the synthesised voice being spoken by a robot was conducted. This perception test is a replica of the test conducted in Sect. 4, except that the voices used here are synthesised. The testing setup provided to the participants were also the same as the pilot test, as shown in Fig. 1. A link to the survey is provided here[13] where the robot speaking to the patient can be seen. A total of 51 participants aged 16–55 (age distribution is shown in Fig. 2) with average or above-average hearing ability took part in this experiment. The aim is to address the three research questions listed in Sect. 4 with the synthesised voice.

7.1 Addressing Research Question I

7.1.1 Design

The participants saw a video and listened to two sets of dialogues between a *Healthbot* and a patient. The text associated with both the dialogue sets were the same. As before, there are two voices:

1. *Synthesised robotic voice* - This is the robotic voice synthesised without any emotions. This voice is the output from the New Zealand English text-to-speech synthesis system. This voice is rendered in a neutral tone without any emotions associated with it.
2. *Synthesised empathetic voice* - This is the emotional voice that is produced by the emotion transformation model. All the dialogs spoken by this voice contain one of the five secondary emotions - *anxious*, *apologetic*, *confident*, *enthusiastic*, *worried*.

The same voice talent was used to create both the synthesised voices. The participants listened to each of these voices separately (the empathetic voice first, followed by the robotic voice) and rated the voices on a five-point empathy scale, which was also used in the pilot experiment.

[12] Approved by the University of Auckland Human Participants Ethics Committee (UAHPEC) on 20/10/2017 for 3 years. Ref. No. 019845.

[13] https://auckland.au1.qualtrics.com/jfe/form/SV_9tvDP800i4oLmXX

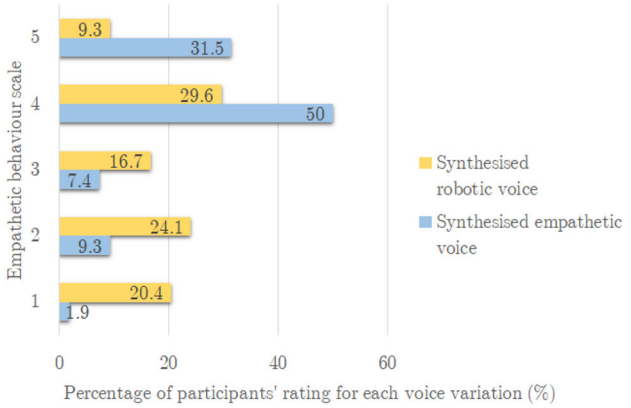


Fig. 9 Participants' rating of the two voice types

7.1.2 Aim

From the pilot study, the participants could perceive empathy from the voice of the robot when the empathy was expressed using changes in prosody only. However, this was done using natural speech. The synthesised voice will not be as perfect or human-like as natural voice. Hence, this test is essential to understand if the expression of empathy in the synthesised voices is being perceived by participants.

7.1.3 Results

Figure 9 summarises the rating given by the participants for the two voice types in %. It can be seen that 81.5% of the participants rated 4 or 5 for the empathetic voice. This means that the participants could perceive higher levels of empathy from the synthesised emotional voice. For the robotic voice, 44.5% of participants rated it on scale 1 or 2. The participants perceived only lower levels of empathy from the synthetic robotic voice. Both the voices spoke the same text. The only difference was in the prosody modelling in synthesised empathetic voice to produce emotional speech. This suggests that the addition of prosody modelling contributed to the perception of an empathetic voice from the robot.

7.1.4 Statistical Analysis

Because the data is skewed for both the voice types, a Wilcoxon signed-rank test was conducted to assess the difference between the ratings received for the synthesised robotic voice and the synthesised empathetic voice. The empathy scale ratings 1 to 5 given by the participants was used for the analysis. Some interesting findings were obtained from the statistical analysis, and they are:

(a) Difference in empathy ratings for study 1 and 2: The results (shown in Table 1 Row 4) indicates that the empathetic voice ranks (Median = 4) are higher than the robotic voice

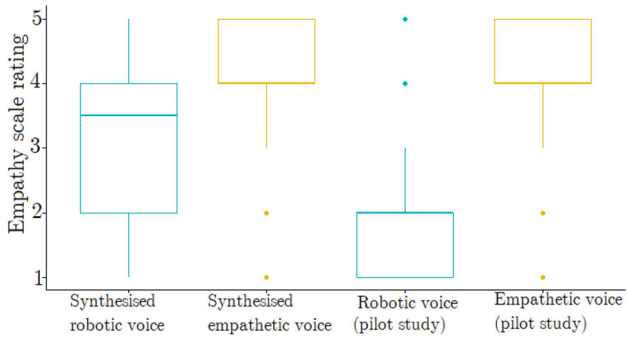


Fig. 10 Boxplot showing participants' rating of the two voice type for study 1 and study 2

ranks (Median = 3.5), $p = 0.003$, $r = 0.3$). Effect size $r = 0.3$ indicates that the effect is small to medium according to the Cohen's benchmark for effect sizes.

The effect size is lower for the synthesised speech ($r = 0.3$) compared to the acted-out speech ($r = 0.7$). To ascertain what is happening, consider the boxplot in Fig. 10 with the participants' ratings of the robotic voice and the empathetic voice from both study 1 and 2. From Fig. 10, it can be seen that, for the pilot study there is no overlap between the two voice type ratings (which also reflects in the effect size $r = 0.7$ from Table 1). However, in the second study using the synthesised voice, it can be seen that there is some overlap between the ratings for robotic voice and empathetic voice.

(b) Empathetic voice rating from both studies: The empathetic voice has higher empathy rating compared to the robotic voice for both study 1 and 2 (see boxplots in Fig. 10). The ratings received for robotic voice and empathetic voice in both study 1 and 2 are significantly different (From Table 1 Columns 6 and 7)) from each other. Table 4 shows pair-wise Wilcoxon signed-rank test results for both voice types in study 1 and 2. This comparison helps to understand if peoples' responses are statistically different for the two studies. It can be seen that the difference in ratings received for the empathetic voice for study 1 and 2 is not statistically significant ($p = 0.176$, $r = 0.1$[14]). This suggests that participants felt that the robot was both interested in what the patient was saying, and was trying to engage with the patient, regardless of whether the empathetic voice was acted or synthesised. Hence, the aim of this study - which is to develop a synthesised empathetic voice is successful.

*(c) Difference in robotic voice rating from study 1 and 2:*For the robotic voice in the two studies, there is a statistically significant difference ($p < 0.001$, $r = 0.4$[14] - From Table 4). This difference can be visually observed from the box plots in Fig. 10. Here, the synthesised robotic voice has higher ratings compared to the robotic voice produced by an actor. This could be because people give more allowances to the synthesised voice as it is not human. So, they tune their mind to actively listen to the voice and the words (which portrayed empathy), knowing very well that it is synthesised. However, when the participants know that the voice is produced by a human, then people probably expect more empathy in the voice. This could be a reason why the acted-out robotic speech was poorly rated on the empathy scale.

7.2 Addressing Research Question II

7.2.1 Design

In this part of the test, the participants were asked which of the two voices they preferred if they were the patient, and they were talking to a healthcare robot. In this stage, they could see the video of the two voices any number of times to make their decision.

7.2.2 Aim

This test is to understand which voice the participants prefer in the actual application of the robot.

7.2.3 Results

83% of the participants preferred the synthesised empathetic voice and 17% preferred synthesised robotic voice. In the pilot experiment with natural speech (described in Sect. 4.2.2) similar findings were observed with the majority of the participants (95%) preferring the empathetic voice over the robotic voice. It is clear that in both studies, the empathetic voice was preferred over the robotic one. However, the participants' reasons for making a choice may not necessarily be the same. A robotic voice spoken by a human could be perceived as creepy, wheres a synthetic voice with modelled empathy might be considered acceptable for the task. However, without further study, we can only speculate the reason. We also need to consider the impact of participant numbers. The pilot experiment was done by 120 participants, while 51 participants did this second experiment. The larger number of participants in the initial test may also be a reason for the stronger results. Additionally, some participants found empathy in the synthetic voice to be "not real", which made them choose robotic voice instead.

7.3 Addressing Research Question III

7.3.1 Design

In this part of the experiment, the participants were asked the reason for preferring of the two voice types. The participants were provided with a series of options to justify their choice

[14] r indicates the effect size.

Table 4 Comparison between robotic voice and empathetic voice in study 1 (Pilot) and study 2

	Robotic voice	Empathetic voice
Synthesised Robotic voice	$p < 0.001, r = 0.4$	$p < 0.001, r = 0.4$
Synthesised empathetic voice	$p < 0.001, r = 0.6$	$p = 0.176, r = 0.1$

of the voice they preferred (same options as the pilot study - Forced-response). Also, there was a free-response section, where the participants could write what they wanted. The participants were also asked which all emotions they could perceive from the voice. They could see the video of the two voices any number of times to make their decision.

7.3.2 Aim

This was designed to understand why people chose either of the two voices.

7.3.3 Results

The forced-response reasons for choosing synthesised empathetic voice and for not choosing synthesised robotic voice are given in Fig. 11. The tone of the voice, the emotions, the empathy, the friendliness and encouragement in the voice are the most frequent reasons for participants choice of the synthesised empathetic voice. The lack of emotions, the tone being not appropriate and the lack of friendliness in the voice are the key reasons for not preferring synthesised robotic voice.

The participants also had a free-response section where they could make comments on their choices, other than the ones already listed in Fig. 11. Figure 13 presents a mind map of the reasons the participants provided for preferring the synthesised empathetic voice. The key ideas that came up from the thematic analysis[15] are the preference due to suitability for the application (Social robots, specifically healthcare robots), the influence of the changes introduced to the affective prosody in the voice and the properties of the voice. Reasons that were most quoted were the naturalness perceived in the empathetic voice and the emotions in the voice. Participants also commented that the voice is suitable and appropriate for a healthcare application, and they felt that the robot was engaging and interested in the patient. Another important reason the participants mentioned was the tone of the voice, which is a direct reflection of the f_0 contour modelling done to synthesise the emotions in the empathetic speech. Figure 14 illustrates a mind map summarising participants' responses for not preferring the synthesised robotic voice. All the responses were related to the major themes - voice suitability for a social robot application and the property of the voice. The most quoted reasons for not preferring

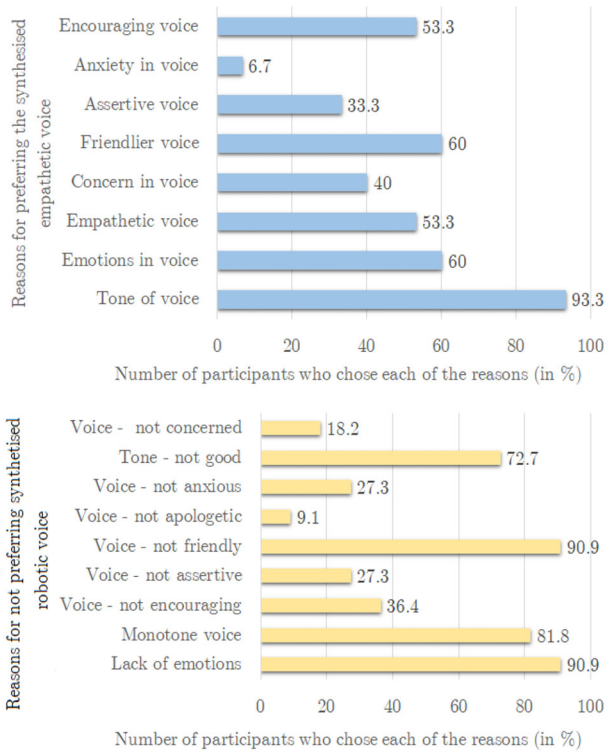


Fig. 11 Participants' reasons for choosing the synthesised empathetic voice (in blue) and not preferring the synthesised robotic voice (yellow)

the robotic voice were that the voice sounded unnatural and the lack of emotions. Also, many participants found the lack of engagement from the robot speaking with the robotic voice as a reason to not prefer it.

The participants were also asked to choose the emotions they could perceive from the two voice types from a set *angry, happy, sad, excited, concerned, anxious, encouraging, assertive, apologetic, other*. The results are summarised in Fig. 12. Similar to the results in the pilot study (Fig. 5, most of the emotions that the participants could perceive were words indicating secondary emotions. Empathy and confidence were the responses that came under the *other* option.

The major takeaways from study 2 are:

The reasons for choosing the synthesised empathetic voice are that the participants preferred to have a *healthbot* which portrays emotions while expressing empathy. Even though the text content in the dialogues was the same, higher empathy was perceived only when the acoustics of emotions were added to the voice in congruence with the text. This strongly

[15] Using Taguette [62]; mind map plot using Miro [63]

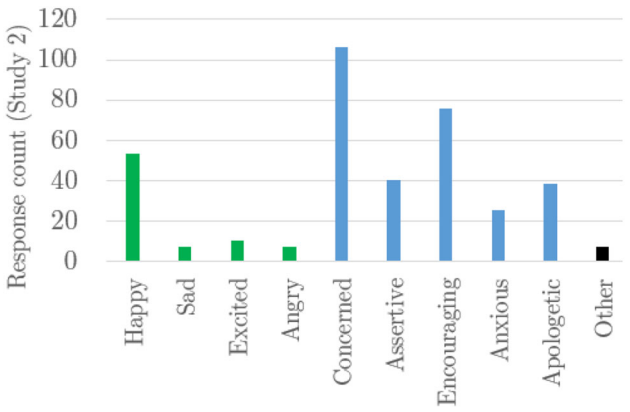


Fig. 12 Participants' responses for the emotions they could perceive from the empathetic voice (study 2)

suggests that the prosody modelling produced by the speech synthesis system must be in alignment with the linguistic content of the sentence, for empathy to be correctly perceived by the users.

The synthesised robotic voice was perceived as being uninterested or rude, which made participants not like the voice in the healthcare scenario. Without proper prosody modelling, there is the possibility that robotic dialogues may sound like the robot is uninterested in the patient. This could reduce the acceptance of social robots.

The participants could perceive many secondary emotions from synthesised empathetic voice - like *apologetic*, concerned (*worried*), encouraging (*enthusiastic*), assertive (*confident*). These are the emotions that were modelled by the emotion transformation, and the perception test has shown that the participants can perceive the same emotions in the synthesised voice.

8 Discussions and Conclusion

This paper presents two studies done in a symmetric fashion to develop acceptable synthetic voices for healthcare robots. Study 1 starts with trying to identify what type of voice is acceptable for healthcare robots. This is done by conducting a perception test using voices spoken by a professional voice artist. Once the type of voice needed was identified to be empathetic, the emotions needed for an empathetic voice were then found out based on the application - healthcare robots. The emotions needed for the social robot was found to be secondary emotions- *anxious, apologetic, concerned, enthusiastic, worried.* A corpus containing these emotions were then developed, and model to synthesise these emotions was formed using ensemble regressors. The emotional speech model was perceptually evaluated. Further, to complete the process, a second study was conducted using the synthesised voice as the voice of the healthcare robot. This study was a replica of the pilot study conducted initially, the only difference being the use of synthesised voice as the voice of the robot. The major contributions of this paper are: (a) the development of an emotional speech model for the secondary emotions that were identified to be needed for a healthcare robot (based on the pilot study 1), (b) synthesising the emotional speech based on the model, and (c) conducting a similar study to that done in the pilot study, but using synthesised speech for the healthcare robot. This study tested the acceptability of a healthcare robot speaking empathetically using synthesised speech with five secondary emotions.

Major findings of the pilot study are that the emotions needed for healthcare robots are not only the well-researched primary emotions, but also nuanced secondary emotions. There is a lot of resource development and research needed to understand the acoustics of these secondary emotions.

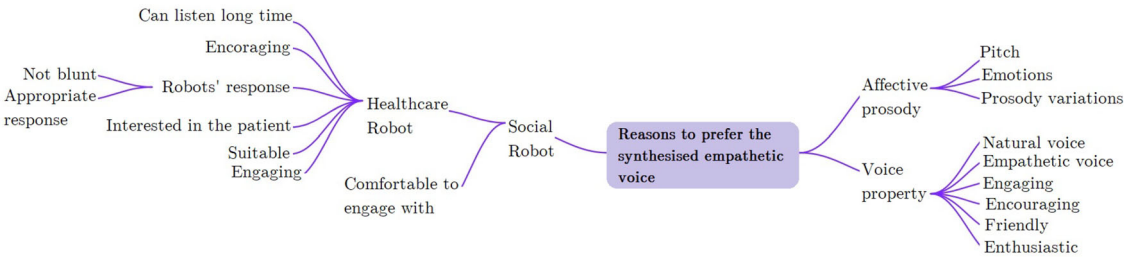


Fig. 13 Mind map showing free responses from participants for preferring the Synthesised empathetic voice

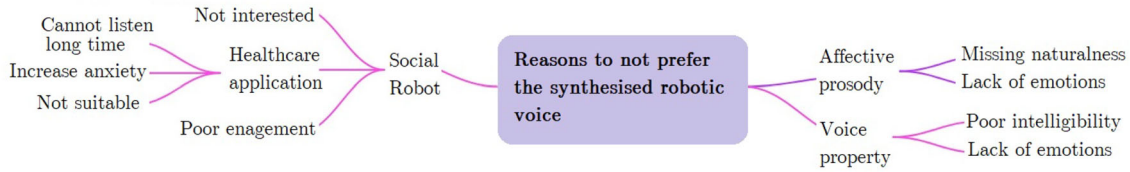


Fig. 14 Mind map showing free responses from participants for not choosing the Synthesised robotic voice

These secondary emotions were synthesised via ensemble regression modelling applied to the output of a New Zealand English text-to-speech synthesis system. Participants in study 2 found the emotional speech containing the five secondary emotions to be more empathetic than a robotic voice saying the same textual information. This result further strengthens the motivation to study nuanced emotions and include these emotions in the voice of social robots, along with the primary emotions that are already well-researched.

The participants preferred the empathetic voice over the robotic voice for a healthcare application. Hence the second important finding is that people can perceive empathy from the healthcare robot's voice when empathy was expressed only by the prosody component of speech. The text of the dialogues spoken by the synthesised robotic voice and the synthesised empathetic voice are both the same. But only when the emotions (prosody component) matched the text did people perceive empathy in the voice. And this could be perceived even when the voice was synthesised. This result emphasises the importance of the emotions in the speech being congruent with the textual content of what is being said. This congruence is essential for participants to perceive empathy from the robots, and this perception of empathy also improves the acceptance of social robots.

This study was based on the *Healthbots* application and using the dialogue set of the *Healthbots* for all the analysis. More nuanced emotions may be identified to be needed when the application is different. Such an application-oriented analysis should be done to identify the emotions that are needed. The ensemble regression-based model can be extended for more emotions.

For the experiments reported in this paper, the participants were shown videos of the *Healthbot* talking with different voice types. But we cannot extend these results to scenarios when people directly interact with a physical robot. The reaction of people when directly interacting with robots can be affected by the presence of the robot near them [64], perceived age and gender [65], engagement techniques (like gaze, nodding) [66,67], accent of the robot's voice [68] and other factors. Hence, direct interaction of people with a robot will have to be studied with the same voice types (as used in this study) to evaluate the empathy perceived from the robot. The authors will be conducting such a study (similar to [68]) in the near future.

This study focused on the prosody component of speech. The verbal component also impacts empathy portrayal by speech. Hence dialogue modelling also needs to be conducted to develop empathetic voices. Along with speech, the other communication channels like facial and para-linguistic channels also contribute to the perception of empathy. These are areas where more research needs to be done to develop social robots that express empathy.

Acknowledgements This research was supported by the Centre for Automation and Robotic Engineering Science, University of Auckland Seed funding. The authors would like to thank the professional actors who recorded their voices for the JLCorpus and the perception test participants for their time and effort.

Compliance with Ethical Standards

Conflicts of interest The authors declare that they have no conflict of interest.

Ethical Standard The authors also thank the very detailed review provided by the reviewers of the journal that helped improve this paper.

Appendix-Scale of Empathy Questionnaire

Table 5 Empathy scale in MITI and its extension to HRI

Scale	Human-human interaction	Human–robot interaction
1	Clinician has no apparent interest in client's worldview. Gives little or no attention to the client's perspective.	The robot has no interest in the patient.
2	Clinician makes sporadic efforts to explore the client's perspective. Clinicians' understanding may be inaccurate or may detract from the client's true meaning.	The robot shows some interest in what the patient is saying, but makes no efforts to engage the patient.
3	Clinician is actively trying to understand the client's perspective, with modest success.	The robot shows great interest in what the patient is saying, but makes no efforts to engage the patient.
4	Clinician shows evidence of an accurate understanding of the client's worldview. Makes active and repeated efforts to understand the client's point of view. Understanding is mostly limited to explicit content.	The robot shows great interest in what the patient is saying and tries to engage the patient, but the robot's response could be better.
5	Clinician shows evidence of deep understanding of the client's point of view, not just for what has been explicitly stated but what the client means but has not yet said.	The robot shows great interest in what the patient is saying, engages the patient well and responds appropriately to the patient.

References

1. Tapus A (2009) Assistive robotics for healthcare and rehabilitation. In: Int. conf. on control systems and computer science, Romania, pp 1–7
2. Toh LPE, Causo A, Tzuo P, Chen I, Yeo SH (2016) A review on the use of robots in education and young children. Educ Technol Soc 19:148–163
3. Triebel R, Arras K, Alami R, Beyer L, Breuers S, Chatila R, Chetouani M, Cremers D, Evers V, Fiore M (2016) Spencer: a socially aware service robot for passenger guidance and help in busy airports. In: Field and service robotics, pp 607–622
4. Pineau Joelle, Montemerlo Michael, Pollack Martha, Roy Nicholas, Thrun Sebastian (2002) Towards robotic assistants in nursing homes: challenges and results. Robot Auton Syst 42(271–281):6
5. Chu M, Khosla R, Khaksar SMS, Nguyen K (2017) Service innovation through social robot engagement to improve dementia care quality. Assist Technol 29(1):8–18
6. Centre for automation and robotic engineering science-Healthbots. https://cares.blogs.auckland.ac.nz/research/healthcare-assistive-technologies/healthbots/. Accessed 29 Oct 2019
7. Broadbent E, Stafford R, MacDonald B (2009) Acceptance of healthcare robots for the older population: review and future directions. Int J Social Robot 1(4):319
8. Igic A, Watson CI, Stafford RQ, Broadbent E, Jayawardena C, MacDonald BA (2010) Perception of synthetic speech with emotion modelling delivered through a robot platform: an initial investigation with older listeners. In: Australasian int. conf. on speech science and technology, Australia, pp 189–192
9. Igic A (2010) Synthetic speech for a healthcare robot: investigation, issues and implementation. Master's thesis, The University of Auckland, New Zealand
10. Fussell SR, Kiesler S, Setlock LD, Victoria Y (2008) How people anthropomorphize robots. In: ACM/IEEE int. conf. on human-robot interaction, Netherlands, pp 145–152
11. Heerink Marcel, Kröse Ben, Evers Vanessa, Wielinga Bob (2010) Assessing acceptance of assistive social agent technology by older adults: the Almere model. Int J Social Robot 2(4):361–375
12. Heerink M (2011) Exploring the influence of age, gender, education and computer experience on robot acceptance by older adults. In: Int. conf. on Human-robot interaction, Switzerland, pp 147–148
13. Duffy Brian R (2003) Anthropomorphism and the social robot. Robot Auton Syst 42(3–4):177–190
14. Marcel Heerink, Ben Krose, Vanessa Evers, Bob Wielinga (2006) The influence of a robot's social abilities on acceptance by elderly users. IEEE Int. Symposium on Robot and Human Interactive Communication, UK, pp 521–526
15. Markowitz J (2017) Speech and language for acceptance of social robots: an overview. Voice Interact Design 2:1–11
16. Breazeal C, Scassellati B (1999) A context-dependent attention system for a social robot. In: Int. joint conf.s on artificial intelligence, USA, pp 1146–1151
17. Chella A, Barone RE, Pilato G, Sorbello R (2008) An emotional storyteller robot. Emotion, personality, and social behavior, USA. In: AAAI spring symposium, pp 17–22
18. Mavridis Nikolaos (2015) A review of verbal and non-verbal human-robot interactive communication. Robot Auton Syst 63:22–35
19. Ivar Nass Clifford, Brave Scott (2005) Wired for speech: how voice activates and advances the human–computer relationship. MIT press, Cambridge
20. Goetz J, Kiesler S, Powers A (2003) Matching robot appearance and behavior to tasks to improve human-robot cooperation. In: IEEE int. workshop on robot and human interactive communication, USA, pp 55–60
21. Scheutz M, Schermerhorn P, Kramer J, Middendorff C (2006) The utility of affect expression in natural language interactions in joint human–robot tasks. In: ACM conf. on human–robot interaction. USA 2:226–233
22. Eyssel F, Ruiter L, Kuchenbrandt D, Bobinger S, Hegel F (2012) If you sound like me, you must be more human: on the interplay of robot and user features on human-robot acceptance and anthropomorphism. In: ACM/IEEE int. conf. on human–robot interaction, USA, pp 125–126
23. Fung P, Bertero D, Wan Y, Dey A, Chan RHY, Siddique F, Yang Y, Wu C, Lin R (2016) Towards empathetic human-robot interactions. In: Int. conf. on intelligent text processing & computational linguistics, Turkey, pp 173–193
24. James J, Watson CI, MacDonald B (2018) Artificial empathy in social robots: an analysis of emotions in speech. In: IEEE int. symposium on robot and human interactive communication, China, pp 632–637
25. Cuff Benjamin MP, Brown Sarah J, Taylor Laura, Howat Douglas J (2016) Empathy: a review of the concept. Emot Rev 8(2):144–153
26. Asada Minoru (2015) Towards artificial empathy. Int J Social Robot 7(1):19–33
27. Taylor P (2009) Text-to-speech synthesis. Cambridge university press, Cambridge
28. Crumpton J, Bethel CL (2015) Validation of vocal prosody modifications to communicate emotion in robot speech. In: Int. conf. on collaboration technologies and systems, USA, pp 39–46
29. Alam Firoj, Danieli Morena, Riccardi Giuseppe (2018) Annotating and modeling empathy in spoken conversations. Computer Speech Lang 50:40–61
30. Li X, Watson CI, Igic A, MacDonald B (2009) Expressive speech for a virtual talking head. In: Australasian conf. on robotics and automation, Australia, pp 5009–5014
31. Moore Lisa A (2006) Empathy: a clinician's perspective. ASHA Leader 11(10):16–35
32. Niculescu Andreea, van Dijk Betsy, Nijholt Anton, Li Haizhou, See Swee Lan (2013) Making social robots more attractive: the effects of voice pitch, humor and empathy. Int J Social Robot 5(2):171–191
33. Watson C, Liu W, MacDonald B (2013) The effect of age and native speaker status on synthetic speech intelligibility. In: ISCA workshop on speech synthesis, Spain, pp 195–200
34. Broadbent E, Tamagawa R, Kerse N, Knock B, Patience A, MacDonald B (2009) Retirement home staff and residents preferences for healthcare robots. In: IEEE int. symposium on robot and human interactive communication, Japan, pp 645–650
35. Moyers TB, Martin T, Manuel JK, Miller WR, Ernst D (2003) The motivational interviewing treatment integrity (miti) code: Version 2.0. http://casaa.unm.edu/download/miti.pdf. Accessed 29 Oct 2019
36. Field A, Miles J, Field Z (2012) Discovering statistics using R. Sage, Thousand Oaks, pp 666–673
37. Ekman Paul (1992) An argument for basic emotions. Cogn Emotion 6(3–4):169–200
38. Damasio A (1994) Descartes error, emotion reason and the human brain. Avon books, New York
39. Becker-Asano Christian, Wachsmuth Ipke (2010) Affective computing with primary and secondary emotions in a virtual human. Auton Agent Multi-Agent Syst 20(1):32
40. Kemper Theodore D (1987) How many emotions are there? wedding the social and the autonomic components. Am J Sociol 93(2):263–289
41. Ochs Magalie, Sadek David, Pelachaud Catherine (2012) A formal model of emotions for an empathic rational dialog agent. Auton Agent Multi-Agent Syst 24(3):410–440

42. Boukricha H, Wachsmuth I, Carminati MN, Knoeferle P (2013) A computational model of empathy: empirical evaluation. In: Humaine association conf. on affective computing and intelligent interaction, USA, pp 1–6
43. Schröder M (2001) Emotional speech synthesis: a review. In: Eurospeech, Scandinavia, pp 561–64
44. Breazeal C (2001) Emotive qualities in robot speech. In: IEEE/RSJ IROS, USA, pp 1389–1394. IEEE
45. Crumpton Joe, Bethel Cindy L (2016) A survey of using vocal prosody to convey emotion in robot speech. Int J Social Robot 8(2):271–285
46. Paltoglou Georgios, Thelwall Michael (2012) Seeing stars of valence and arousal in blog posts. IEEE Trans Affect Comput 4(1):116–123
47. James J, Tian L, Watson CI (2018) An open source emotional speech corpus for human robot interaction applications. In: Interspeech, India, pp 2768–2772
48. James J, Watson CI, Stoakes H(2019) Influence of prosodic features and semantics on secondary emotion production and perception. In: Int. congress of phonetic sciences, Australia, pp 1779–1782
49. Kisler T, Schiel F, Sloetjes H (2012) Signal processing via web services: the use case webmaus. In: Digital humanities conf, Germany, pp 30–34
50. James J, Mixdorff H, Watson CI (2019) Quantitative model-based analysis of f_0 contours of emotional speech. In: Int. congress of phonetic sciences, Australia, pp 72–76
51. Mixdorff H, Cossio-Mercado C, Hönemann A, Gurlekian J, Evin D, Torres H(2015) Acoustic correlates of perceived syllable prominence in German. In: Annual conf. of the int. speech communication association, Germany, pp 51–55
52. Mixdorff H (2000) A novel approach to the fully automatic extraction of fujisaki model parameters. In: IEEE int. conf. on acoustics, speech, and signal processing. Proceedings, Turkey, pages 1281–1284
53. Schröder Marc, Trouvain J'urgen (2003) The German text-to-speech synthesis system MARY: a tool for research, development and teaching. Int J Speech Technol 6(4):365–377
54. Watson CI, Marchi A (2014) Resources created for building New Zealand english voices. In: Australasian int. conf. of speech science and technology, New Zealand, pp 92–95
55. Jain S (2015) Towards the creation of customised synthetic voices using Hidden Markov Models on a Healthcare Robot. Master's thesis, The University of Auckland, New Zealand
56. Paul Boersma (1993) Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Inst Phonetic Sci 17:97–110
57. Liaw Andy, Wiener Matthew (2002) Classification and regression by Random Forest. R news 2.3 23:18–22
58. Yoav Freund, Schapire Robert E (1996) Experiments with a new boosting algorithm. In: Int. conf. on machine learning, Italy, pp 148–156
59. Eide E, Aaron A, Bakis R, Hamza W, Picheny M, Pitrelli J (2004) A corpus-based approach to expressive speech synthesis. In: ISCA ITRW on speech synthesis, USA, pp 79–84
60. Ming H, Huang D, Dong M, Li H, Xie L, Zhang S (2015) Fundamental frequency modeling using Wavelets for emotional voice conversion. In: Int. conf. on affective computing and intelligent interaction, China, pp 804–809
61. Robinson C, Obin N, Roebel A (2019) Sequence-to-sequence modelling of F_0 for speech emotion conversion. In: Int. conf. on acoustics, speech, and signal processing, UK, pp 6830–6834
62. Taguette version: 0.9. https://www.taguette.org. Publisher: Zenodo
63. Miro. https://miro.com/app/
64. Powers A, Kiesler S, Fussell S, Torrey C (2007) Comparing a computer agent with a humanoid robot. In: Proceedings of the ACM/IEEE int. conf. on human-robot interaction, pp 145–152
65. McGinn C, Torre I (2019) Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots. In: 2019 14th ACM/IEEE int. conf. on human–robot interaction (HRI), pp 211–221. IEEE
66. Anzalone Salvatore M, Boucenna Sofiane, Ivaldi Serena, Chetouani Mohamed (2015) Evaluating the engagement with social robots. Int J Social Robot 7(4):465–478
67. Leite Iolanda, Castellano Ginevra, Pereira André, Martinho Carlos, Paiva Ana (2014) Empathic robots for long-term interaction. Int J Social Robot 6(3):329–341
68. Tamagawa Rie, Watson Catherine I, Han Kuo I, MacDonald Bruce A, Broadbent Elizabeth (2011) The effects of synthesized voice accents on user perceptions of robots. Int J Social Robot 3(3):253–262

Jesin James is a Lecturer in the Department of Electrical, Computer, and Software Engineering at the University of Auckland, New Zealand. Jesin did her Ph.D at the University of Auckland on the topic of developing synthetic voices for social robots. Jesin's main research areas are speech signal processing, under-resourced languages, machine learning and human-robot interaction. Jesin has experience in developing speech technology resources for languages like Malayalam, New Zealand English and te reo Maori.

B. T. Balamurali is a postdoctoral research fellow affiliated to the Audio Research Group (ARG) in Singapore University of Technology and Design (SUTD). He got his Ph.D from the University of Auckland, New Zealand in 2015. During his past position as researcher at Auckland Bio-Engineering Institute (ABI), and the current one at SUTD, he has worked in solving many interdisciplinary research problems using the state-of-the-art statistical modelling, AI algorithms. He has published and presented the findings in reputed journals and conferences. The problems range from bio-signal classification, speech/speaker recognition, spoofing signal detection, music type/musical instrument prediction/classification, fluid flow pattern prediction/classification, vocal tract impedance analysis etc. He has a number of project collaboration with major hospitals in Singapore including SGH, CGH, KKH and NUH and also collaborated with industrial partners such as Panasonic, Singapore. Prior to PhD, he was also a design engineer developing signal processing algorithms ported onto embedded processors which are integral to the many commercial entertainment devices available in today's market.

Catherine I. Watson completed her BE (Hons) and PhD from the University of Canterbury, New Zealand. After that, Catherine was a postdoctoral fellow at Macquarie University, Sydney, Australia for 8 years. Then she joined the Department of Electrical and Computing Engineering at The University of Auckland. Currently, Dr. Catherine I. Watson is an Associate Professor at the University of Auckland. Catherine's research interest is speech production for both humans and machines. Her research includes building models of speech articulators, speech synthesis, robot speech, and acoustic phonetics. The two languages she mainly focusses on are New Zealand English and te reo Maori, and her research has impact in Engineering, Speech Science and Phonetics.

Bruce MacDonald completed a BE (1st class) and PhD in the Electrical Engineering Department of the University of Canterbury. After working with NZ Electricity for three years and a year at the DSIR in Wellington, he moved to Canada and spent ten years in the Computer Science Department of the University of Calgary. Returning to New Zealand in 1995, he joined the Department of Electrical and Computer Engineering at the University of Auckland. He helped set up a new programme in computer systems engineering. Bruce also started the Robotics Laboratory. His long term goal is to design intelligent robotic assistants that improve the quality of people's lives, with primary research interests in human robot interaction and robot programming systems, and applications in areas such as healthcare and agriculture. He is the director of the department's robotics group and the leader for the multidisciplinary CARES robotics team at the University of Auckland. He is vice-Chairman for NZ's robotics, automation and sensing association. For NZ's national science challenge Science for Technological Innovation, he is the theme leader for Sensors, Robotics and Automation and deputy director. One of his current research programmes is to develop robots to help care for people, which is a multidisciplinary project undertaken jointly with Korean researchers and companies. Another current research project is for orchard robotics, a joint project with NZ researchers and companies.