Modelling prosodic features for Empathetic speech of a Healthcare Robot

Jesin James

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

jjam194@aucklanduni.ac.nz

Abstract

Artificial speech developed using speech synthesizers has been used as the voice for robots in Human Robot Interaction. As humans anthropomorphize robots, an empathetically interacting robot is expected to increase the level of acceptance of social robots. But how empathy can be conveyed via synthesised speech is the research question addressed here. This research looks into understanding what type of voice can be perceived as empathetic by human users. This involves analyzing the emotional ranges to be covered by the voice of the robot, and analyzing which ranges would be suitable for an empathetic companion. It has been identified from this study that rather than focusing on synthesizing primary emotions in speech, it is essential to model secondary emotions to synthesize empathetic speech. Modelling of secondary emotions are restricted by the lack of resources and the difficulty in separating them on a valence-arousal plane. This research addresses these needs, and attempts to model and synthesize the secondary emotions in order to develop empathetic speech for social robots.

Index Terms: Empathetic speech, primary and secondary emotions, prosody modelling, Human Robot Interaction

1. Introduction & Motivation

A simple one-channel speech producing system that transmits explicit verbal messages is not sufficient for modern Human-Robot Interaction (HRI) applications. The interacting robots at the receiving end should be able to react to the human users in a way that encourages users to continue interacting with the social robots. The *Healthbots* (Healthcare Robots) developed at the University of Auckland to assist people in old age homes is a HRI technology which uses a synthesized voice. The robot is designed to be a companion to the elderly, which requires regular use of the synthesised voice as a communication medium. When developing robots that interact with humans, their acceptance is a primary concern [1]. Robots that interact in social situations are novel to people who use them due to their limited experience in actually interacting with robots [13]. People rationalize this novelty by anthropomorphism, i.e. projecting familiar human-like characteristics, emotions and behavior onto robots. Speech is a primary mode of interaction between robots and humans. Currently, roboticists build robots that look like humans to improve their acceptance. However, users are disappointed by the lack of reciprocal empathy from robots [3]. People's anthropomorphism of robots is impacted by the type of voice used by robots to converse, and this also affects the robots' acceptance. Due to this gap, a concept called Artificial Empathy (AE) has been introduced in HRI [4, 5, 6]. AE is the affective response portrayed by the artificial intelligence in companion robots. There have been attempts to model empathetic behavior in social robots using modalities such as facial expressions, gestures and speech [5-9]. But the focus on this work is to use the vocal channel (speech) to express empathy. Speech has two components [10]: the verbal component and the prosody component. Verbal component focuses on the words alone and is defined using combinations of linguistic symbols. Most of the factors that make human speech more natural compared to monotone speech can be summarized under the prosody component. Emotions are expressed by variations in prosody component (like varying intonation, speech rate, stress) [10,11]. Empathetic behavior 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, and emotions govern the acoustic features pertaining to prosody [12]. Often empathy is incorporated into synthesized speech by the inclusion of words that convey an affective response (called dialog modelling). This research solely focuses on effectively modelling the prosody component of speech for the synthesis of empathetic speech. The objectives of this research are two-fold. Firstly, it is essential to understand what emotions should be included into speech to make it sound empathetic. Secondly, these identified emotions have to be modelled using prosodic features and good quality emotional speech has to be synthesised in New Zealand English to be used as the healthcare robot voice.

2. Issues Identified & Major Contributions

A major issue affecting the present speech synthesis research for social robots is the insufficient understanding about what type of voice is preferred by human users. Many experiments have varied the type of synthesised robotic voice to evaluate the users' responses [14-17], but they have not arrived at a final conclusion. Hence, before synthesizing an appropriate voice for social robots, it is essential to arrive at a decisive conclusion about the preferred voice by human users. For this, a large scale human perception experiment was conducted to identify what type of robotic voice is preferred by users as explained in [25]. This study tried to understand if people can perceive that a robotic companion is empathetic to them when empathetic behaviour is expressed only by the emotions in speech. The results show that humans are able to perceive empathy and emotions in robot speech, and prefer it over the standard robotic voice (monotone voice). It is important for the emotions in empathetic speech to be consistent with the language content of what is being said, and with the human users' emotional state. Analyzing emotions in empathetic speech using valence-arousal model [28] has revealed the importance of secondary emotions[1] in developing empathetically speaking social robots.

Another key constraint to be addressed here is the lack of sufficient resources to study empathetic speech. Currently available and widely used speech corpora [18-22] cover only a small set of Primary emotions, which are not sufficient to define empathetic speech. In order to address this need, an emotional corpus with 5 primary emotions and 5 secondary emotions was

[1] Primary Emotions are emotions that are innate to support fast & reactive response behavior like angry, happy, sad. Secondary emotions are assumed to arise from higher cognitive processes, based on an ability to evaluate preferences over outcomes & expectations like relief, hope

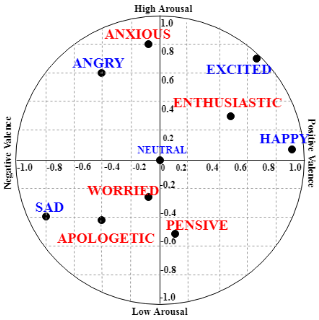

Figure 1: *V-A plane showing the emotions in the corpus*

developed [26]. The choice of the primary emotions has been made based on current research works focusing on these emotions [18-22,27], making it possible to compare the developed corpus with existing ones. The choice of the secondary emotions is based on the conclusions from [25], that these subtle emotions are needed for the empathetic speech synthesis of social robots in HRI. A visual representation of the emotions on a Valence-Arousal (V-A) 2D plot is shown in Figure 1. In the plot, the position of primary emotions has been marked in upper-case blue letters and secondary emotions in lower-case red letters. It can be seen that the corpus has been designed to span over a large part of V-A plane. In contrast to existing speech corpora, this was constructed by maintaining an equal distribution of 4 long vowels in New Zealand English. This balance is to facilitate emotion related formant and glottal source feature comparison studies. A large scale perception test with 120 participants showed that the corpus has approximately 70% and 40% accuracy in the correct classification of primary and secondary emotions respectively. The reasons behind the differences in perception accuracies of the two emotion types is further investigated. The corpus is made public at: github.com/tli725/JL-Corpus.

The next key issue identified was the lack of defined standards in literature to evaluate synthesised speech for social robots. The quality of synthesised speech is analyzed by its similarity to natural speech and also by human perception tests. But when it comes to synthesizing speech for robots that interact with users in social situations, such a quality analysis is not sufficient. HRI requires the study of other parameters likes the perception of empathy from the robotic speech by the human users and enhancement of acceptance of the robot due to its speech. The Motivational Interviewing Treatment Integrity (MITI) [23] has been currently used in some studies [24] to understand how well people are motivated by interviews. But there is a need for such standards to be translated to HRI speech as well, considering that a robot is the entity to which interaction is happening. Such an adaptation of standards was done as part of this work and reported in [25]. The Empathy-measuring scale from the MITI module which defines 5 scales to rate a clinician's empathy was modified to be suitable to evaluate peoples' reaction to the prosody component of synthesised speech used in HRI

Empathetic speech in some HRI experiments have been synthesised by modelling of verbal component of speech alone. There have been no modelling of the prosody component and this need is being addressed here. To address the modelling of prosody component for empathetic speech, the acoustic analysis of the emotional speech corpus developed has been conducted. The F_0 contour of the emotional speech have been analyzed in terms of parameters like mean, maximum, range and minimum. Also, the contour was labelled using a Tones and Break Indices (ToBI) model. Along with this, other acoustic features like intensity, speech rate, glottal source features and vocal tract features have also been analyzed.

A major difficulty in synthesizing empathetic speech is the limitation imposed by the current open-source speech synthesizers. When it comes to synthesizing expressive speech in synthesizers, there are only a few acoustic parameters that can be modelled and varied. These parameters are sufficient to model primary emotions (like angry, happy, sad) that are well apart in the V-A domain. While, in synthesizing subtle emotions there is a need to model voice quality features and also change the properties of the vocal source as well. The flexibility offered in current synthesizers do not cater to these features. The synthesizer being used for this research is MaryTTS and experimental changes have been made to the structure of the TTS for incorporating subtle emotions. For this, currently changes are being made to the ToBI modelling implemented in the TTS and also to the prosody component to synthesize secondary emotions.

3. Results & Discussion

The major result obtained from the human perception test is the significance of secondary emotions for synthesizing empathetic speech. These emotions are not easy to model as they are not well separated on the V-A plane (seen in Fig. 1). However, these emotions are the ones that people use to empathetically interact with each other in social situations and these emotions have to be modelled for HRI as well. This knowledge about the emotions that portray empathy in speech is a pre-requisite for synthesis of empathetic speech in social robots. A thorough acoustic analysis of the secondary emotions has revealed the need for modelling more features that can differentiate emotions on the valence level. The arousal level is often well represented by features like intensity and F_0. While for the valence level only features like Long term average spectrum, harmonic richness factor and some voice quality features were obtained as significant. It is easier to model intensity, F_0 and implement it in synthesizers compared to the voice quality features. A ToBI-based tone marking of the secondary and primary emotions has revealed that the same F_0 variation rules cannot be used for various emotion types, as implemented in current synthesizers. Even though tones like $L + H^*$ are present in all emotions, there are variations in their shape(like the point where the contour turns with respect to time) and F_0 range as well. These findings are essential in synthesizing subtle emotions.

4. Future Plan

MaryTTS is the synthesizer being used and the Healthcare Robots are the application of this research. As a thorough modelling of the acoustic features of the secondary emotions needed for empathetic speech have been done. Now, the requirement is to effectively implement them in the synthesizer. The limitations of the synthesizer model restricts a lot of changes like variation of the voice quality features. Currently, the aim is to model the F_0 contour effectively via the ToBI analysis completed. Further, the synthesised voice need to be implemented in the Healthcare robots and human perception tests need to be conducted to evaluate whether human users are perceive the subtle emotions and whether the robot is perceived to be empathetic.

5. Acknowledgements

This work is funded by Centre for Automation & Robotic Engineering Science, University of Auckland & supervised by Dr. Catherine Watson & Dr. Bruce MacDonald. We thank the perception test participants & the professional actors (emotional speech corpus) for their time and effort.

6. References

[1] Broadbent, E., Stafford, R., MacDonald, B., ”Acceptance of Healthcare Robots for the Older Population: Review and Future Directions”, *Int J of Soc Robotics 1: 319*,2009

[2] S. R. Fussell, S. Kiesler, L. D Setlock, V. Yew, ”How people anthropomorphize robots”, *in Proc. ACM/IEEE International Conference on Human-Robot Interaction*, 2008

[3] P. Fung, D. Bertero, Y. Wan, et al., ”Towards Empathetic Human-Robot Interactions”, *arXiv preprint arXiv: 1605.04072,*2016.

[4] A. Stephan , ”Empathy for Artificial Agents”, *in Int J of Soc Robotics 7: 111*,2015.

[5] A. Lim, H. G. Okuno, ”A Recipe for Empathy: Integrating the Mirror System”, *in Int J of Soc Robotics 7:3549*,2015

[6] M. Asada, ”Towards Artificial Empathy: How Can Artificial Empathy Follow the Developmental Pathway of Natural Empathy?”, *in Int J of Soc Robotics 7: 19*,2015

[7] L. Damiano,P. Dumouchel,H. Lehmann, ”Towards HumanRobot Affective Co-evolution Overcoming Oppositions in Constructing Emotions and Empathy”, *in Int J of Soc Robotics 7: 7*,2015

[8] H. Boukricha, I. Wachsmuth, M. N. Carminati, P. Knoeferle, ”A computational model of empathy: Empirical evaluation”, *Proc. ACII, pp. 1-6,* 2013.

[9] Li X, Watson CI, Igic A, Macdonald BA, Expressive speech for a virtual talking head, *in Proc. Australasian Conference on Robotics and Automation,* 2009

[10] P.Taylor, ”Text-to-Speech Synthesis”, *Chapter 13, Cambridge*, 2009

[11] Crumpton, J. Bethel, ”Validation of vocal prosody modifications to communicate emotion in robot speech”, in Proc CTS), 2015

[12] Alam, F.; Danieli, M., Riccardi, G., ”Annotating and modeling empathy in spoken conversations”,*Computer Speech Language 50* , 2018

[13] S. R. Fussell, S. Kiesler, L. D Setlock, V. Yew, ”How people anthropomorphize robots”, *in Proc. ACM/IEEE International Conference on Human-Robot Interaction*, 2008

[14] Duffy, Brian R, ”Anthropomorphism and the Social robot”, *in Robotics and autonomous systems 42:3*,177-190,2003

[15] Eyssel, F, Kuchentrandt, D, Bobinger, S, de Ruiter, L, Hegel, F, ”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 Proc. International Conference on Human-Robot Interaction*, 2012.

[16] Scheutz M, Schermerhorn P, KramermJ, Middendorff C, ”The utility of affect expression in natural language interactions in joint human-robot tasks”, *in Proc. ACM International Conference on Human-Robot Interaction*,2006

[17] Goetz, J. Kiesler, S, Powers, A,” Matching robot appearance and behavior to tasks to improve human-robot cooperation”, *in Proc. 12th IEEE RO-MAN*, Vol. IXX,2003

[18] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B, ”A database of German emotional speech,” *Interspeech, Lisbon, Portugal,*, pp. 1517–1520, 2005.

[19] Steiner I, Schrder M, Klepp A. ”The PAVOQUE corpus as a resource for analysis and synthesis of expressive speech,” *In: Proc. Phonetik and Phonologie*, 2013.

[20] S. Haq and P.J.B. Jackson, ”Multimodal Emotion Recognition,” *In W. Wang (ed), Machine Audition: Principles, Algorithms and Systems, IGI Global Press, ISBN 978-1615209194*, pp. 398–423, 2010.

[21] Koolagudi, S. G., Maity, S., Kumar, V. A., Chakrabarti, S., and Rao, K. S, ”IITKGP-SESC: speech database for emotion analysis,” *Communications in computer and information science, LNCS, Berlin, Springer*, 2009

[22] Engberg, Inger S and Hansen, Anya Varnich and Andersen, Ove and Dalsgaard, Paul, ”Design, recording and verification of a Danish emotional speech database,” *Fifth European Conference on Speech Communication and Technology*, 1997.

[23] T. Moyers, T. Martin, J. Manuel, and W. Miller, ”The motivational interviewing treatment integrity (MITI) code: Version 2.0”, Unpublished.Albuquerque, NM: University of New Mexico, Center on Alcoholism, Substance Abuse and Addictions, 2008.

[24] B. Xiao, P. G. Georgiou, S. Narayanan, ”Analyzing the language of therapist empathy in motivational interview based psychotherapy”, *in Proc. Asia Pacific Signal Inf. Process. Assoc.*, pp. 1-4, 2012.

[25] Jesin James, Catherine Watson, Bruce MacDonald, ”Artificial Empathy in Social Robots: An analysis of Emotions in Speech,” *In Proc. IEEE International Conference on Robot and Human Interactive Communication* , 2018

[26] J. James, L. Tian and C. Watson, An Open Source Emotional Speech Corpus for Human Robot Interaction Applications, *in Proc. Interspeech*, 2018.

[27] Ververidis, D., Kotropoulos, ”A review of emotional speech databases,” *In: PCI 2003, 9th Panhellenic Conf. on Informatics,Greece*, pp. 560-574, 2003.

[28] G. Paltoglou and M. Thelwall, ”Seeing Stars of Valence and Arousal”, *IEEE Transactions of Affective Computing* , 2013.