A framework to incorporate aspects of social perception in synthetic voices

Sai Sirisha Rallabandi

Quality and Usability Lab, Technische Universität Berlin, Germany

s.rallabandi@tu-berlin.de

Abstract

In my dissertation, I investigate the social speaker characteristics: warmth and competence. I refer to them as the global aspects of social perception and analyse them in synthetic speech. Specifically, I hypothesize two things: (a) there are certain vocal cues responsible for evoking these characteristics and (b) it is feasible to incorporate these characteristics in the present day speech generation mechanisms. To validate my hypotheses, I propose a two stage framework: In the first stage, I perform extensive subjective evaluations in the form of listening tests to identify which synthetic voices demonstrate characteristics of warmth and competence. I also perform analysis on the voices to identify the vocal cues corresponding to these characteristics. In the second stage of my dissertation, I propose to incorporate the identified vocal cues into the speech generation mechanism. Employing two target application scenarios - *Customer Service* and *Health care*, I propose to show that incorporating social speaker characteristics can not only improve user satisfaction but also user trust.

Index Terms: Warmth, Competence, speech perception, subjective evaluation

1. Introduction

In the words of Aristotle, the Greek Philosopher, "*Man is a social animal*". There are studies which prove that social interactions reduce the levels of stress in humans [1, 2]. For instance, a recent research on children aged 7.5-12 shows that vocal interactions are comparable to that of the physical touch in stressful situations [1]. Psychologists state that such interactions can also have a long term impact on the individuals who experience them [2].

With the rapidly growing interest in human-machine interactions and spoken dialog systems, Chatbots have become increasingly predominant. Therefore, speech community is focusing on personalization of these chatbots for various applications such as customer service, screen-readers and personal assistants [3, 4, 5]. It is imperative that this goal can be achieved only through effective interaction of these bots with the humans. In human-human interactions, extra-linguistic information is commonly associated with the spoken content unlike in human-machine interactions. Therefore, these paralinguistic aspects, once considered only the "elegance" factors, have now become a pre-requisite in human-machine interactions.

Improved computing abilities have facilitated the analysis of paralinguistic information from speech. Accordingly, significant work on varied spoken content is available through Computational Parlinguistic Challenge (ComParE), annual Interspeech challenges [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. However, all these works were performed on human speech, and the perception of the paralinguistic phenomena, the social perceptions of synthetic voices is still obscure. To the best of our knowledge, this is the first attempt to analyse the synthetic speech for social speaker characteristics.

In my dissertation, I propose a two-stage framework,

- **Identify the acoustic correlates influencing the global aspects of social perception**

  I propose to conduct subjective evaluation based assessment. Here, I try to understand how people perceive the synthetic voices. Based on the evaluation results, I derive the acoustic correlates that influence the desired characteristics.

- **Modeling of acoustic features** Furthermore, I plan to modify the speech signals that displayed less warm (or competence) or unwarm (or incompetence). This could be either by 1) signal processing techniques or through 2) incorporating required modifications in the synthesis procedure.

2. Key issues identified

I target the application domains: health care, customer service and the desired characteristics are warmth and competence respectively. Firstly, I perform a study to verify if these characteristics can be perceived in synthetic voices. This study includes a research on what questions should we pose during the evaluation of synthetic voices.

2.1. Evaluation setup

For the evaluation, I choose to use multiple TTS voices. While doing so, I would also consider the gender balance in the synthetic voices being used. The subjective data would be collected from both native and non-native English speakers for a global response.

2.2. Analysis of results

The evaluation would be followed by a series of analysis steps:

- Does every sentence coming from a voice possess same characteristics? or Does the perception of different sentences from a single voice differ?
- Does the gender of the participant affect the speech perception? If yes, how do male participants perceive male speakers, female speakers and vice versa.
- What are the acoustic features that contribute to perception of warmth and competence in a speech utterance?

2.3. Modeling acoustic features

I believe that, the speaker characteristics or emotions or mood of the speaker are understood mostly based on the "way" the sentence is being spoken. Hence, I posit that, use of Global Style Tokens for speech generation might be a good start to investigate and generate the responsible vocal cues [17].

3. Case study: Discussion of preliminary experiments and results

3.1. Dataset preparation

I performed a case study in the last one year for more insights in this regard. Various TTS voices that evoke diverse subjective speaker attributions were employed in the study [18, 19, 20, 21, 20, 22]. The objective is to gather the synthesized voices spanning the social speaker characteristics warmth and competence. Thus, the prompts were generated for each of the application domains: health care and customer service reflecting the desired social characteristics. The text used to synthesize such prompts is provided below.

- *Is there anything I can do to help?*
- *I am sure we can reach a solution.*

A detailed description of number of male and female voices synthesized from each of the TTS systems is discussed in the Table 1.

Table 1: *Details of synthetic voices used in the study*

TTS System	Female		Male	
	#voices	MOS	#voices	MOS
Neural TTS	13	3.98	14	3.23
Clustergen	2	2.98	3	2.51
USS	2	1.63	4	2.37
HTS	1	2.8	2	3.4

The Mean Opinion Scores are calculated over the speech quality and naturalness of the synthetic voices. The scores provided are averaged across all the voices per each system.

3.2. Subjective evaluation

A systematic strategy for the perception of the desired characteristics from synthetic voices is unknown. Therefore, I perform a 3-phase evaluation of TTS systems in order to reach the perceptual aspects contributing to the targeted characteristics, i) pre-tests to see if the characteristics can be perceived with the sentences chosen, ii) questionnaire preparation based on the observations and iii) Semantic scaling test and Exploratory Factor Analysis (EFA) to determine the underlying perceptual dimensions.

3.3. Observations

3.3.1. ***Comparison with experiments on natural speech***

I compared my experimental results with that of the speech perception studies performed on natural speech [23, 24]. Ignoring the context of the speech utterance, I found, a) slightly different set of attributes for warmth and b) new set of attributes for competence.

3.3.2. ***Predicted perceptual dimensions***

Through this case study, I have determined the perceptual dimensions underlying social speaker characteristics, warmth and competence. I found that male TTS voices require slightly more number of questions during the evaluation than that of the female voices. The finalised questions/perceptual dimensions common for male and female voices are further presented. **Warmth:** Hearty, Pleasant, Emotional, Trusting, Agreeable, Non-likable, Sympathetic. **Competence:** Calm, Relaxed, Anxious, Tense.

3.3.3. ***Gender-bias***

We observed that the female participants found female voices to be more pleasant than the male voices. We understood this based on the two sample t-test, performed on the participant ratings collected separately for male and female voices. The p-value calculated over the attribute "pleasant" was 0.38 (< 0.5) for female participants. All the other attributes were almost correlated in both the genders' participant ratings ($p > 0.5$).

3.3.4. ***Acoustic features***

I have derived 88 acoustic features for each of the synthetic voices using OpenSMILE toolkit [25]. Further, I employed backward selection in linear regression, for feature selection. The finalised acoustic features are presented in the Table 2.

Table 2: *Finalised acoustic features in male and female speaker for each of warmth and competence. f0= fundamental frequency, f1, f2, f3= formants, HI= Hammerberg Index, HNR= Harmonics to Noise ratio, mfcc= mel frequency cepstral coefficients*

Female		Male	
Warmth	Competence	Warmth	Competence
f0	f0	loudness	loudness
f2	mfcc4	mfcc3	mfcc4
HI	f1	HNR	f1
loudness	f3	f3	HI

From the above Table, it is evident that the acoustic features responsible for the desired characteristics are not consistent across the gender. Based on the observations, I posit that the characteristics warmth and competence are perhaps not dependent on one feature, but are distributed across multiple features.

4. Future work

4.1. Real-time data

From my case study, I observed that the provided sentences could not properly display the social speaker characteristics. Hence, I choose to modify 1) the sentence length and 2) the context of the sentence, to enable the perception of warmth and competence from speech alone. For instance, this could be achieved by the use of twitter posts and comments.

4.2. Gender independent framework

From Table 2, we observe that the acoustic features derived were different for both male and female speakers. However, it is desirable to have a gender-independent system that can generate these characteristics with high fidelity. Therefore, through my future experiments, I hypothesize to bridge this gap and propose a synthesis mechanism that is common to both the genders.

5. Acknowledgements

I sincerely thank Sebastian Möller and Benjamin Weiss for their invaluable guidance. This work is being supported by the German Research Foundation (DFG), under funding MO 1038/29-1, TU PSP-Element: 1-50001062-01-EF.

6. References

[1] L. Seltzer, A. Prososki, T. Ziegler, and S. Pollak, "Instant messages vs. speech: hormones and why we still need to hear each other," *Evolution and human behavior : official journal of the Human Behavior and Evolution Society*, vol. 33, pp. 42–45, 01 2012.

[2] B. L. Fredrickson, "The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions." 2001.

[3] Potluri, Venkatesh and Rallabandi, SaiKrishna and Srivastava, Priyanka and Prahallad, Kishore, "Significance of Paralinguistic Cues in the Synthesis of Mathematical Equations," in *Proc. of the 11th International Conference on Natural Language Processing*, 2014.

[4] A. Wilkinson, A. Parlikar, S. Sitaram, T. White, A. W. Black, and S. Bazaj, "Open-Source Consumer-Grade Indic Text To Speech," in *Proc. SSW*, 2016.

[5] Yaniv, Leviathan and Yossi, Matias, "Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone," in *Google AI Blog*, 2018.

[6] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. Interspeech*, 2009.

[7] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proc. Interspeech*, 2010.

[8] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proc. Interspeech*, 2011.

[9] B. W. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. J. J. H. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge," in *Proc. Interspeech*, 2012.

[10] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," *in Proc. Interspeech*, 2013.

[11] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive Physical Load," *in Proc. Interspeech*, 2014.

[12] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hoenig, J. Orozco-Arroyave, E. Noeth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's Eating Condition," *in Proc. Interspeech*, 2015.

[13] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity and Native Language," in *Proc. Interspeech*, 2016.

[14] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. Warlaumont, G. Hidalgo Gadea, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, and S. Zafeiriou, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold Snoring," in *Proc. Interspeech*, 2017.

[15] B. Schuller, S. Steidl, A. Batliner, P. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E. Messner, K. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical and Self-Assessed Affect, Crying and Heart Beats," in *Proc. Interspeech*, 2018.

[16] B. Schuller, A. Batliner, C. Bergler, F. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl, A. Warlaumont, L. Yankowitz, E. Noeth, S. Amiriparian, S. Hantke, and M. Schmitt, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds Orca Activity," in *Proc. Interspeech*, 2019.

[17] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *arXiv preprint:1803.09017*, 2018.

[18] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, and et al., "Tacotron: Towards End-to-End Speech Synthesis," *in Proc. Interspeech*, 2017.

[19] H. Zen, K. Tokuda, and A. W. Black, "Statistical Parametric Speech Synthesis," *Speech communication*, 2009.

[20] M. Schroder, M. Charfuelan, S. Pammi, and O. Turk, "The MARY TTS entry in the Blizzard Challenge 2008," *in Proc. Blizzard Challenge Workshop*, 2008.

[21] P. Taylor, A. W. Black, and R. Caley, "The Architecture of the Festival Speech Synthesis System," in *Proc. Speech Synthesis Workshop*, 1998.

[22] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 1996.

[23] L. Fernández Gallardo and B. Weiss, "The Nautilus Speaker Characterization Corpus: Speech Recordings and Labels of Speaker Characteristics and Voice Descriptions," in *Proc. Eleventh International Conference on Language Resources and Evaluation (LREC)*, 2018.

[24] L. Fernández, Gallardo, and B. Weiss, "Towards Speaker Characterization: Identifying and Predicting Dimensions of Person Attribution," *in Proc. Interspeech*, 2017.

[25] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, 2013.