Dimensions of quality for state of the art synthetic speech

Fritz Seebauer, Petra Wagner

Faculty of Linguistics and Literary Studies & CITEC, Bielefeld University

Synthetic speech has a long standing tradition of being employed for experiments in phonetics and laboratory phonology. The choice of synthesis method and system is commonly made by the researcher(s) to fit the specific quality criteria and study design. The overall quality of a given system, however, remains as a confound that is difficult to control for [1]. In speech technology newly proposed systems are usually compared across specific dimensions e.g., 'Intelligibility' and 'Naturalness'. These dimensions have already been extensively studied and evaluated within the context of old diphone and formant synthesis networks [2]. We contend, however, that these traditional dimensions need to be re-examined in the context of state of the Art Text-to-Speech (TTS) systems, as those newer models exhibit different quality deteriorations. Our work aims to bridge the conflicting demands for quality criteria that are easily computed and applied during TTS development, while at the same time remaining descriptive and meaningful for phonetic research. As a first step in this endeavor, we carried out an experiment to find suitable dimensions of TTS quality with a bottom-up approach based on descriptions provided by 11 participants (phonetic experts). The participants were instructed to label speech samples generated by 8 different state of the art Text-to-speech systems (varieties of English). Each system produced a stimulus consisting of two sentences of the phonetically balanced 'caterpillar story' [3]. In order to ensure that all systems were evaluated across different phonetic contexts in a balanced way, the sentences were rotated between participants so that each participant heard the complete story but with different parts read by different systems. The experimental setup is loosely based on the work in [4]. The participants were instructed to write down nouns, adjectives or sentences describing the quality of a given stimulus. Using embeddings generated by a pretrained BERT model [5] for semantic distances, we determined which of the participants terms were semantically similar. A subsequent affinity propagation clustering revealed there to be 39 meaningfully different clusters, each representing a dimension of quality for synthetic voices. Keeping in mind that these dimensions are later to be used for ratings in actual evaluation experiments, it was decided to reduce the number of clusters to a more practical number of 10 and re-calculate the spectral clustering with a precomputed cosine affinity matrix. The resulting clusters and their respective quality descriptions are depicted in fig. 1. A manual analysis of the resulting dimensions led to the following descriptive labels: 'artificiality/voice quality', 'intonation/noise/prosody', 'voice/audio quality', 'audio cuts', 'style/recording quality', 'emotion/voice quality/attitude', 'engagedness', 'human likeness', 'hyperarticulation'. From the assigned cluster descriptions it is evident, that the semantic embeddings sometimes conflated several seemingly unrelated quality features into single dimensions (e.g. prosody and background noise), while occasionally splitting almost synonymous terms into multiple clusters (e.g. 'artificiality', 'roboticness' and 'metallicness'). To evaluate these shortcomings of the semantic model, two independent manual clusterings were carried out. They were both limited to 10 clusters and reported a modified jaccard agreement index of 63,44, while agreeing with the automatic computed clusters with 54.48 and 57.93, respectively. The low interrater agreement between the manual clusters suggests that a panel decision process might be needed to determine the final quality dimensions. Subsequent research will evaluate clusters created by naïve listeners and quality dimensions of different sub-tasks in synthetic speech, such as voice conversion.

doi:10.11576/pundp2022-1016

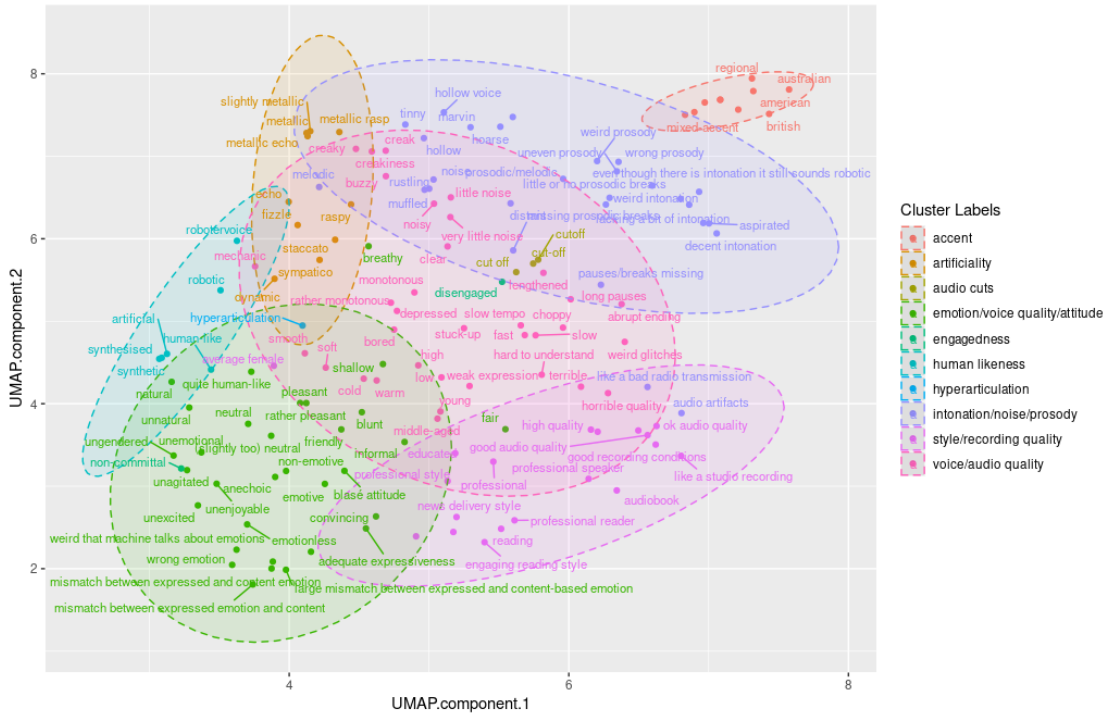

Figure 1: *Automatic clustering of synthetic quality dimensions along semantic dimensions produced by a pretrained BERT model. The multidimensional semantic embeddings were reduced to 2D using UMAP, which are denoted on the x and y-axis.*

References

[1] Zofia Malisz, Gustav Eje Henter, Cassia Valentini-Botinhao, Oliver Watts, Jonas Beskow, and Joakim Gustafson. Modern speech synthesis for phonetic sciences: A discussion and an evaluation. In *ICPhS 2019, Melbourne*, 2019.

[2] Mahesh Viswanathan and Madhubalan Viswanathan. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale. *Computer speech & language*, 19(1):55–83, 2005.

[3] Rupal Patel, Kathryn Connaghan, Diana Franco, Erika Edsall, Dory Forgit, Laura Olsen, Lianna Ramage, Emily Tyler, and Scott Russell. 'the caterpillar': A novel reading passage for assessment of motor speech disorders. *American Journal of Speech-Language Pathology*, 22(1):1–9, 2013.

[4] Florian Hinterleitner, Sebastian Möller, Christoph Norrenbrock, and Ulrich Heute. Perceptual quality dimensions of text-to-speech systems. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[5] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

doi:10.11576/pundp2022-1016