DISCERNING DIMENSIONS OF QUALITY FOR STATE OF THE ART SYNTHETIC SPEECH

Fritz Seebauer[1], Michael Kuhlmann[2], Reinhold Haeb-Umbach[2], Petra Wagner[1]

Bielefeld University[1], Paderborn University[2]
{fritz.seebauer, petra.wagner}@uni-bielefeld.de, {kuhlmann, haeb}@nt.upb.de

ABSTRACT

This paper describes an approach for determining the dimensions of quality for state-of-the-art synthetic speech. We propose that current evaluation metrics do not fully capture the meaningful dimensions of text-to-speech (TTS) and voice conversion (VC) systems. In order to develop a revised paradigm for meaningful evaluation, we conducted two experiments. First, we determined descriptive terms by querying naïve listeners on their impressions of modern TTS and VC systems. In a second experiment, we refined these terms into dimensions of quality and similarity by showcasing a consolidation procedure of manual clusterings. The resulting dimensions contain the standard evaluation categories of “intelligibility” and “naturalness” for both conditions. We could additionally discern dimensions of “tempo” and “demographics” in both domains. The final two dimensions as well as the relationships between categories proved to be different between TTS and VC, suggesting the need for modified evaluation scales based on the target construct.

Keywords: voice quality, voice conversion, text-to-speech, dimensions of quality, evaluation

1. INTRODUCTION

Phonetic sciences often lean on speech technology to generate stimulus data or derive processes of speech production and perception [1, 2, 3]. It has been shown, however, that the quality of synthetic speech systems not only varies, but might also generate voice quality features that do not exist in natural voices. Consequently, there exists a need for evaluating the quality of different speech technology systems regarding their suitability in phonetic research. Traditionally, synthetic speech has been evaluated primarily on the dimensions of “intelligibility” and “naturalness” [4, 5] based on ITU-T P.85 [6]. It has been questioned whether these two dimensions are capable of exhaustively capturing the full extent of what constitutes a good synthesis [7, 8] and whether multidimensional approaches might yield more accurate representations of quality [9]. It has also been shown that the specific wording of a given evaluation task has a great influence on the resulting Mean Opinion Score (MOS) with or without context [10, 11]. Finally, it has been pointed out that the target construct might differ between different applications of speech synthesis [12]. To conciliate these findings with our research, we conducted experiments in two different domains of speech synthesis. The first one concerns pure text-to-speech, with the latent construct being “quality”, the second one represents voice conversion, with the latent construct being “similarity”.

2. METHODS

To elicit latent dimensions of a given construct, it is common practice to take preconceived items of potential attributes and reduce them into smaller subsets using factor analysis [8, 13]. To counteract an inherent researcher bias, we opted to follow an inductive approach of item-generation and complement it with a manual clustering of multiple participants to determine the underlying dimensions. The only step requiring expert knowledge in the newly proposed paradigm is the naming of resulting dimensions, rather than pre-selecting scales based on the generated items, as suggested in behavioral science and psychology [14, 15, 16]. The reasons for developing this method were two-fold: Firstly, the new procedure allowed us to use the whole set of terms elicited during the first experiment without introducing bias by pre-selection. Secondly, exploratory factor analysis is always dependent on the audio samples used during the weighting experiment [17]. Having participants cluster the semantic space directly instead of using an audio sample as an intermediate eliminates the chance of having an unrepresentative token.

2.1. Experiment 1: Terms of quality

80 participants (40f/40m, L1 English) were recruited over the crowdsourcing platform Prolific [18]. They

TTS-system	training corpus	gender
vits [21]	vctk	m
silero-tts [22]	private	m
tacotronDDC+hifiGAN [21]	sam	f
tacotron2+wavegrad [21]	ek1	f
glow-tts+multibandmel [21]	LJ	f
microsoft neural	private	f
google wavenet	private	m
amazon polly	private	m

Table 1: TTS system types used for generating the audio samples in the elicitation experiment.

were instructed to listen to a provided audio sample and note down terms that they felt best described its quality. There was a constraint to supply at least three different terms for each sample. The experiment was split into two subsets with 40 participants each. In the first subset the audios were constructed using 8 different neural state-of-the-art systems with varying vocoders, balanced by target gender. The different system architectures are listed in Table 1. They were chosen to generate a broad set of different architectures employed in current-day research and real-life application. Each system read the phonetically balanced "caterpillar story" [19], with four rotating sentences being presented to each participant. The experiment yielded 387 different descriptions of the quality of synthetic speech.

The second subset elicited similar terms of quality for the domain of voice conversion. Here, the data of the Voice Conversion Challenge 2020 (VCC2020) [20] served as a basis to elicit dimensions of synthetic voice similarity. The participants were asked to listen to a voice conversion sample and its corresponding natural target audio and instructed to note down terms that they felt best described *similarities* or *differences* between the samples. The dataset consisted of a subset from the VCC2020 data containing 33 different voice conversion systems. The subset was stratified by source and target gender, as well as the target content. Each participant was presented with one sample pair from each gender combination with each audio pair containing one of the 4 different contents. The voice conversion experiment yielded 359 unique terms of similarity.

2.2. Experiment 2: Clustering of terms

To aggregate the elicited terms of quality into meaningful dimensions, a second experiment was conducted. 10 Participants (5f/5m, L1 varied) were presented with a randomized list of all terms of quality for one domain. They were instructed to group those terms into 10 clusters of varying sizes on a graphical two-dimensional interface. The clusters were to be created according to the terms' semantic similarity, with closer terms being more similar. The participants were additionally asked to denote group similarity using 10 predefined colors. There were 5 participants each for both the text-to-speech and voice conversion terms. To ensure some level of English proficiency, a self-assessment pre-test was carried out based on the bilingual language profile of [23]. The experiment lasted about 2 hours, and the participants were instructed to take a mandatory break of 15min in between to avoid fatigue. The resulting clusters were evaluated both in the distance spacing and the color groupings denoting semantic similarity.

2.3. Finding latent Categories/Dimensions

The coordinate data were first normalized via min-max scaling, as some of the participants did not use the whole space. To find a representative view of their average groupings, a distance matrix was computed for each participant, denoting the relationship (distance) between two terms. Given those matrices, we obtain a mean Euclidean distance matrix by averaging across participants, from which we recalculate the plane coordinates by eigenvalue decomposition of $M_{ij} = \frac{D^2_{1j}+D^2_{i1}D^2_{ij}}{2}$, where D_{ij} denotes the average Euclidean distance between the i-th and j-th term [24]. This yields a single average representation of all terms in 2D space.

Regarding the color assignments, the participants each colored the terms in a self-defined order. This means that (1) they might not assign the same color to the same latent group (indicated by a large overlap of terms in the colored groups) and (2) terms might be assigned to different latent groups across participants. To solve (1), we compute the optimal cluster (color) match between participants. The problem was first turned into an assignment problem by counting how many terms were assigned the same 5-tuple of colors, and then transforming those counts into a cost matrix: Higher cost indicates better color matching. The resulting assignment problem was then solved by employing the modified Jonker-Volgenant algorithm [25]. However, the optimal assignment obtained this way is not universally valid as it may contain tuples where a specific color for a single participant is reused. To compensate for this, we used a greedy approach to prune tuples in the cost matrix by colors which already had been assigned in a previous step to a given participant. In each step, we compute the optimal assignment from the pruned cost matrix, take out the tuple with the highest cost, and perform the participant-color assignment followed by color pruning until the cost matrix contains no more valid color tuples. To solve (2), we performed a color-matched majority voting per term: If the majority of the participants assigned the same color to the term, it was also assigned this color. If the participants could not agree on a color, the term was dropped from the average representation. Given the newly clustered and reduced set of terms, we determined the centroids of all clusters and derived a measure of confidence for each term by calculating the Euclidean distance from its group centroid. Summing these across groups and inverting them yields a measure of internal certainty. An external certainty measure is derived by summing the distance of a given centroid to all other groups. These measures are variations of the subcomponents of the Calinski-Harabasz Index for measuring cluster dispersion [26].

3. RESULTS

Figure 1 shows the average distance results for the pure text-to-speech evaluation. The majority voting procedure only kept 33% of the original 382 terms. The most distinct clusters were those pertaining to demographic information of the inferred speaker as well as those regarding human likeness. The intelligibility group exclusively contains items describing the audio quality and clarity pointing to the fact that modern TTS systems do not suffer traditional understandability problems. This clarity dimension overlaps with the one regarding speech tempo, which might reflect the fact that some of the "tempo" terms actually describe difficulties with intelligibility due to speed. The voice conversion results are visualized in Figure 2. For this domain, the majority vote left 51% of data points. Their respective distance certainty measures are shown in Table 2. As is evident, the clearest group concerned terms relating to demographic associations with the perceived speakers, such as "accent" or "gender". The worst category regarding external and internal certainty contained terms regarding voice quality like "tone", "softness" or "vocal fry". Interestingly, the participants seemed to clearly distinguish between terms of "intelligibility" and "naturalness" regarding their color assignments, but associate and conflate them in their distance markings. The same seems to be true of the categories relating to "content" and "expression".

4. DISCUSSION

The amount of variability in the participants' group assignments highlights the need for an objective procedure. They clearly differed in their category assignments as was evident by low overall Jaccard agreement scores of 32% in the TTS condition and 31% in the VC condition. This underlines our premise that having the reduction of original items into scales done by a single individual is questionable. The findings regarding the dimensions of quality are very much in line with the current status quo. The main categories we could elicit in the domain of speech technology did include the often-tested intelligibility and naturalness, while additionally uncovering that many laypeople seem to take perceived demographic information into account when judging quality. The space measures also allow us to relate the revealed dimensions to each other. For the TTS domain, we can discern that the human-likeness does seem to correspond to evoked emotional qualities and audio distortions, while intelligibility conflates with the perceived tempo and overall pleasantness. In the VC condition, we see semantic associations of naturalness, intelligibility, and voice quality features and a clear separation of tempo terms. This first step in the search for the underlying dimensions to modern day synthetic speech opens up many directions for future research. In particular, a direct comparison of our procedure and a classical Exploratory Factor Analysis is still pending. Post-experiment feedback from the participants in the second experiment revealed that they felt constrained by the two-dimensional space to accurately model the group relationships. While it would be feasible to extend the experimental setup and analysis procedure by a third dimension, care should be taken to ensure naïve participants are actually able to navigate a three-dimensional space on a digital interface. The dimensions proposed here could also be compared to quality dimensions of natural voices such as those described in [27, 28].

Domain	Category	Ext. certainty	Int. certainty
VC	expression	0.35	0.84
	intelligibility	0.05	0.22
	naturalness	0.08	0.82
	voice quality	0.0	0.0
	tempo	0.52	0.65
	content	0.54	0.85
	demographics	1.0	1.0
TTS	pleasantness	0.24	0.35
	emotion	0.0	0.07
	audio quality	0.14	0.41
	tempo	0.08	0.64
	clarity	0.11	0.0
	human-likeness	0.93	1.0
	demographics	1.0	0.54

Table 2: External and internal certainty measures for each category as calculated by the distance to other categories and averaged distance to the centroid, respectively.The values are scaled to range between 0 and 1 for easier comparison

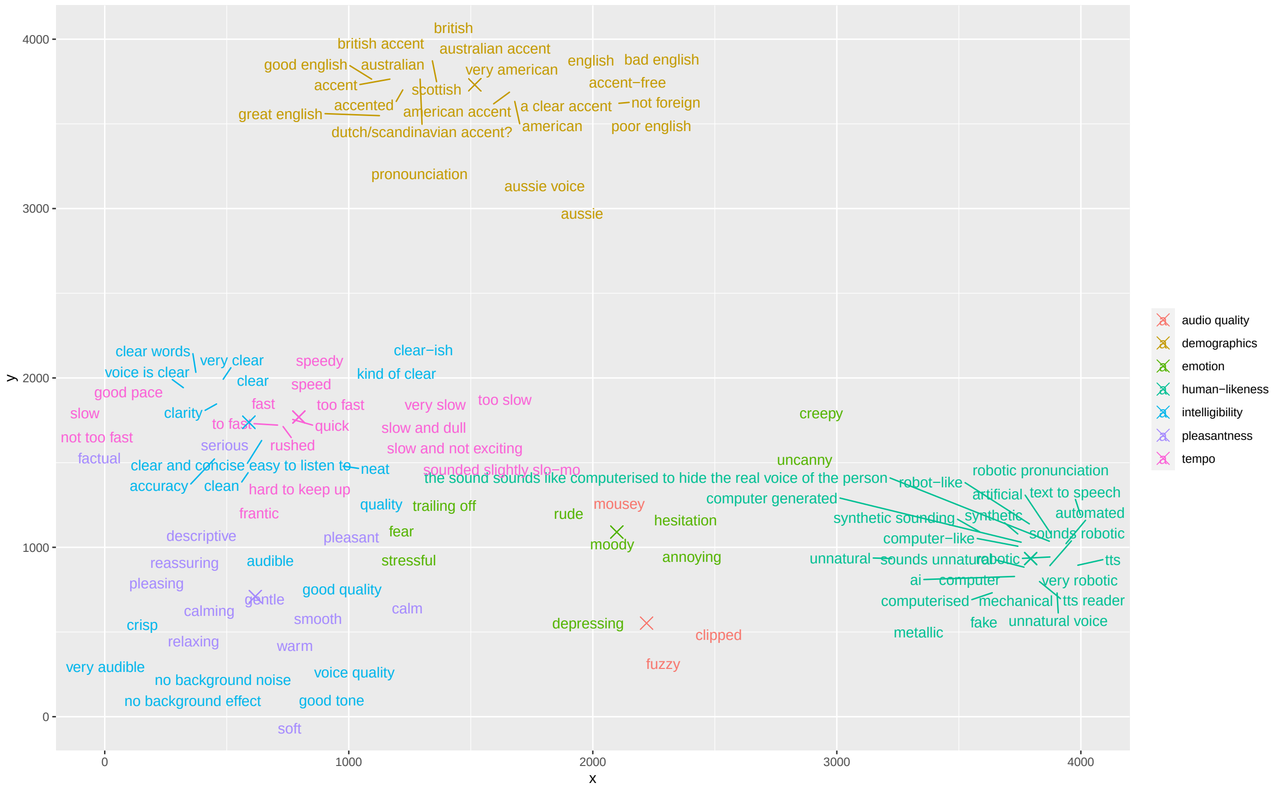

Figure 1: Relative distance assignment of text-to-speech terms averaged over all participants. Color denotes group assignment after consolidating cluster information across participants. The medians are marked in their respective colors

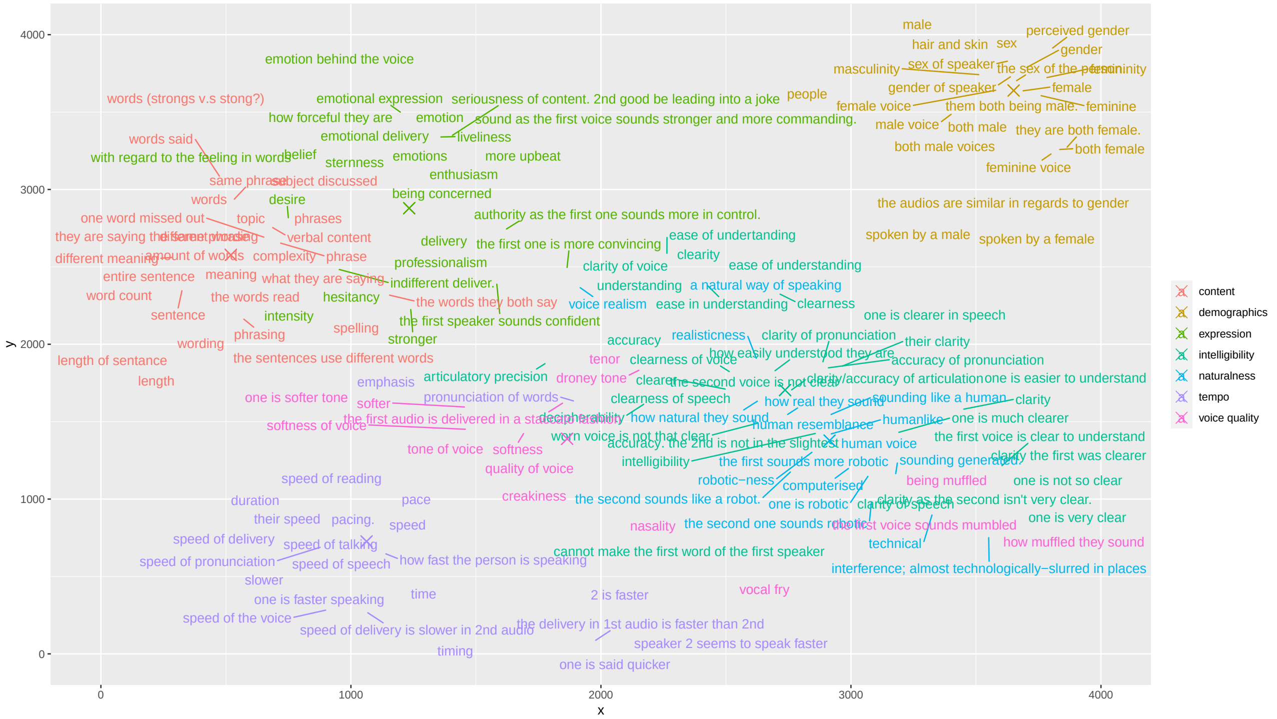

Figure 2: Relative distance assignment of voice conversion terms averaged over all participants. Color denotes group assignment after consolidating cluster information across participants. The medians are marked in their respective colors

5. REFERENCES

[1] Z. Malisz, G. E. Henter, C. Valentini-Botinhao, O. Watts, J. Beskow, and J. Gustafson, "Modern speech synthesis for phonetic sciences: A discussion and an evaluation," in *International Congress of Phonetic Sciences ICPhS*, 2019, pp. 487–491.

[2] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[3] A. Cohen and S. G. Nooteboom, *Structure and process in speech perception*. Springer, 1975.

[4] N. Campbell, "Evaluation of speech synthesis," in *Evaluation of text and speech systems*. Springer, 2007, pp. 29–64.

[5] C. Mayo, R. A. Clark, and S. King, "Listeners weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis," *Speech Communication*, vol. 53, no. 3, pp. 311–326, 2011.

[6] I. T. Union, "ITU-T Rec. P.85, A method for subjective performance assessment of the quality of speech voice output devices," 1994.

[7] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale," *Computer speech & Language*, vol. 19, no. 1, pp. 55–83, 2005.

[8] F. Hinterleitner, C. Norrenbrock, and S. Möller, "Is intelligibility still the main problem? a review of perceptual quality dimensions of synthetic speech," in *Eighth ISCA Workshop on Speech Synthesis*, 2013, pp. 148–151.

[9] J. Benesty, M. M. Sondhi, Y. Huang *et al.*, "Speech quality assessment," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 84–96.

[10] R. Clark, H. Silen, T. Kenter, and R. Leith, "Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs," *arXiv preprint arXiv:1909.03965*, 2019.

[11] J. O'Mahony, P. Oplustil-Gallegos, C. Lai, and S. King, "Factors Affecting the Evaluation of Synthetic Speech in Context," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 148–153.

[12] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, Č. Székely, C. Tånnander *et al.*, "Speech synthesis evaluation state-of-the-art assessment and suggestion for a novel research program," in *Proc. of the 10th Speech Synthesis Workshop*, 2019, pp. 105–110.

[13] S. S. Rallabandi, B. Naderi, and S. Möller, "Identifying the vocal cues of likeability, friendliness and skilfulness in synthetic speech," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 1–6.

[14] G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quiñonez, and S. L. Young, "Best practices for developing and validating scales for health, social, and behavioral research: a primer," *Frontiers in public health*, vol. 6, p. 149, 2018.

[15] T. R. Hinkin, "A review of scale development practices in the study of organizations," *Journal of Management*, vol. 21, no. 5, pp. 967–988, 1995.

[16] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan, "Evaluating the use of exploratory factor analysis in psychological research." *Psychological methods*, vol. 4, no. 3, pp. 272–299, 1999.

[17] F. Hinterleitner, S. Möller, C. Norrenbrock, and U. Heute, "Perceptual quality dimensions of text-to-speech systems," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[18] S. Palan and C. Schitter, "Prolific. ac—a subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2018.

[19] R. Patel, K. Connaghan, D. Franco, E. Edsall, D. Forgit, L. Olsen, L. Ramage, E. Tyler, and S. Russell, "'the caterpillar': A novel reading passage for assessment of motor speech disorders," *American Journal of Speech-Language Pathology*, vol. 22, no. 1, pp. 1–9, 2013.

[20] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020 database v1. 0," 2020.

[21] Coqui AI, "Coqui tts," https://github.com/coqui-ai/TTS, 2022.

[22] Silero Team, "Silero models: pre-trained enterprise-grade stt / tts models and benchmarks," https://github.com/snakers4/silero-models, 2021.

[23] D. Birdsong, L. M. Gertken, and M. Amengual, "Bilingual language profile: An easy-to-use instrument to assess bilingualism," *University of Texas at Austin*, 2012.

[24] G. Young and A. S. Householder, "Discussion of a set of points in terms of their mutual distances," *Psychometrika*, vol. 3, no. 1, pp. 19–22, 1938.

[25] D. F. Crouse, "On implementing 2d rectangular assignment algorithms," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 4, pp. 1679–1696, 2016.

[26] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[27] B. Weiss, D. Estival, and U. Stiefelhagen, "Non-experts' perceptual dimensions of voice assessed by using direct comparisons," *Acta Acustica united with Acustica*, vol. 104, no. 1, pp. 174–184, 2018.

[28] J. Kreiman and B. R. Gerratt, "Validity of rating scale measures of voice quality," *The Journal of the Acoustical Society of America*, vol. 104, no. 3, pp. 1598–1608, 1998.