Synthetic speech evaluation: Description of a more fine grained MOS framework and proposal for a more context specific paradigm

Fritz Seebauer

Phonetics Work Group, Bielefeld University, DE
fritz.seebauer@uni-bielefeld.de

Abstract

This abstract describes the efforts of finding new avenues of synthetic speech evaluation. It proposes a theoretic framework with which to describe different evaluation paradigms. One key insight lies in the fact that improving upon one part of evaluation procedure seems to often degrade its validity in some other dimension. After describing the efforts of generating a more granular evaluation procedure based on mean opinion score (MOS), an outline is given for improving upon contextual accuracy.

Index Terms: Synthetic speech evaluation, factor analysis, synthesis quality

1. Introduction

The thesis will be concerned with developing new and improved techniques for evaluating the quality of synthetic speech systems. When defining what it means for a system to be of high quality, multiple suggestions can be found in common literature [1, 2, 3]. This paper will assume the definition of: "The anticipated and explicitly stated needs of an individual in a given context.", based off of the former ISO definition of quality [4]. The resulting frameworks which try to capture this quality with respect to a system can be roughly described in six non-orthogonal axes: *cost*, *accuracy*, *objectivity*, *comparability*, *granularity* and *reliability*. The *cost* dimension refers to the feasibility of conducting a chosen test, given that there are usually limited resources regarding participants and time. *Accuracy* describes how accurately a proposed test measures the target of synthetic speech quality. Many works on the reliability of synthetic evaluation procedure have highlighted the need for awareness of the listening context and as such, higher accuracy. While some only propose to ensure a silent room without distractions [5] or to make sure there is no undue influence of the immediate textual context [6], others acknowledge that these conditions should actually represent the intended listening context [7, 8, 9]. While the taxonomy of potential context dimensions suggested in the latter is still experimental and subject of discussion, the general notion that lab evaluations can not be extrapolated to all listening contexts seems to be persistent. The dimension of *objectivity* describes a similar level of abstraction on which a given test framework operates regarding the generalizability of a given test result. One general assumption is that the subjective answer we elicit from participants points towards a common quality that is shared in the whole population. A second assumption is made insofar that our sample size is large enough to be representative of that whole. A gain in objectivity usually involves abstracting the process of evaluation so that there are no subjects involved, like in PESQ [10], POLQA [11], ANIQE [12], or more recent attempts such as AutoMOS predictors [13, 14, 15] or the auditory nerve modeling in [16]. These are also very cost efficient because they are able to circumvent the sample size problem by applying a sufficiently large training set to all subsequent evaluations, achieved through modeling an abstracted listener or directly comparing acoustic or other properties against a pre-specified benchmark. This comes at the cost of accuracy, however, as they are not perfectly able to model and abstract the human responses (yet). *Comparability* describes the ability of a measure to differentiate between different systems of the same kind. Again a certain overlap with the other dimensions can be denoted, as a very subjective measurement is unlikely to yield comparable results across systems and test instances. It is one of the major shortcomings of comparison based approaches like AB-Tests [17] and MUSHRA [18], which are not necessarily transitive across testing instances. They do, however, carry the advantage of circumventing the generalization problem regarding subject's internal expectations as is described for voice quality in [19]. This is achieved via grounding of the participants' inner framework by presenting all stimuli at the same time, so a direct ranking can be obtained. *Granularity* refers to the level of precision at which quality is being extracted. The traditional Mean Opinion Score approach used in the blizzard challenge [20] does not yield much insight into what the underlying problems of a given system are. A more granular testing framework would be the older ITU rec. P. 85 [21] regarding signal degradation which computed overall scores of naturalness by eliciting answers on multiple subscales. While both testing paradigms have been criticized and reworked over the years [22, 23, 7, 24], many of these critiques aimed at a different level of granularity with [25], for example, suggesting to include the standard deviation in reports. The last axis *reliability* is a well known measure in the fields of test construction from social sciences and psychology, referring to the consistency in test results across different instances and homogeneity within its sub-parts. As is obvious, the objective approaches described above e.g. also carry the advantage of being very reliable and consistent.

Note that these axes are purely theoretical in nature and an expert panels opinion on their merits and shortcomings would be very useful to ground their applicability.

2. Results so far

Considering the theoretical framework described above, the goal of this thesis is to improve upon the quality paradigms used in the employ of synthetic speech evaluation. The axes used to describe the validity of a given evaluation procedure seem to be at least partially inversely correlated. The goal of defining a better paradigm can as such be framed as one that increases one or multiple of these qualities while commanding no or little negative effect on the others. The efforts so far have been constrained to improving on the granularity of evaluation, as this promises to yield more insights as to the underlying problems of a given system.

2.1. More granular TTS dimensions

To gain a more fine grained picture of naïve participants' perception of what constitutes synthetic speech quality other than the vague term of "naturalness", a line of tests has been carried out. Other people have taken the ceiling effect observed for naturalness since 2011 [26, 27] as a reason to switch the focus of investigation on more difficult tasks and data sets. This thesis follows the reasoning in [27, 28] that these effects are actually a shortcoming of the framework being used and that there might very well be finer differences which cannot be captured by the paradigm. Seeing that the original measures and subsequent revisions of proposed quality scales date back to a time of diphone-synthesis, the re-examination of underlying quality dimensions seemed warranted on modern text-to-speech (TTS) systems. In order to find original terms of quality, a bottom-up pre-experiment was conducted in which participants were presented two-sentence samples of the caterpillar story [29] and asked to provide free text input on what they thought described the quality of what they had just perceived in an online interface. The resulting terms were then transformed into unilateral interval scales. One of the main differences to the work in e.g. [30] is that there was no assumption regarding the prominence of a scale being dependent on its frequency, keeping all original scales for further analysis. The scales were subsequently presented to 88 participants in a web-framework. A different corpus of 12 different state-of-the-art TTS systems was constructed with 3 harvard sentences per sample separated each by 500ms of silence. The analysis revealed 8 significant factors of synthetic speech quality of which the dimensions of "naturalness" and "audio degradation" overlapped with previously found factors in literature [31]. A subsequent retest in which 18 participants each re-evaluated a subset of the same 4 samples for all scales was conducted to gather a sense of reliability of these rating scales. Computing the intra class correlation coefficient revealed very poor consistency between participants' ratings on all sub-scales except for gender. This runs counter to the previous findings on the reliability of MOS for signal degradation in [32], the consistency of listeners ratings in the blizzard challenge [7] or the retest reliability for naturalness MOS reported in [26]. Using the framework described above, the divergence could point to the fact that the level of abstraction is inversely correlated to the reliability of the results with [26] also reporting utterance level correlations which were much lower than system level correlation to previous evaluations of the same data. The simulation experiment in [8] also shows higher stability in MOS ratings with the inclusion of more participants, suggesting that our sample size for the reliability experiment might have been too small. A correlation analysis of the obtained perceptual ratings with computed acoustic measures revealed very low correlations except for the gender dimension. While this can be for one attributed to the spread distribution within the perceptual ratings noted above, it can also be explained through the necessary averaging of acoustic features. Since the MOS for each dimension is computed on the whole utterance, the acoustic correlates were also averaged over all three sentences in a sample. It has been explicitly stated that such procedures muddle the discriminative capabilities of many acoustic properties such as cepstral peak prominence [33].

2.2. Even more fine grained

Following these insights, a new evaluation paradigm was devised to offer even more granular responses in the time domain. After asking the participants' opinion on a standard ACR scale over the whole sample they were instructed to mark the parts of a signal which they felt were most detrimental. To evaluate the validity of such an approach, the inter annotator agreement described in [34] was computed on the overlapping marked regions. The experiment did yield promising results regarding the agreement between participants of which portions of the signal were detrimental to its quality, with κ=0.60 and κ=0.55. The second value yields from a separate group which was given the same task querying a different dimension of quality. The first group was asked to denote unnatural segments and the second to mark emotionally negative parts. A correlation analysis using a General Additive Mixed Model (GAMM) confirmed a significant effect of the question on participants' markings. Subsequent investigations revealed, however, that this effect was primarily due to magnitude and participants did indeed overlap quite heavily in their markings between conditions. Referring back to the meta-framework, we take this as further evidence that untrained listeners do not have the awareness necessary to attribute signal properties to sub-dimensions of their listening impressions, which is a generalization of the statement regarding prosody made in [6].

3. New concepts for evaluation procedure

Having explored the possibilities of increasing examination granularity, the following steps are dedicated to investigating different axes of synthetic speech evaluation.

3.1. Context aware

As was pointed out in the introduction, one major dimension of contention for synthetic speech quality evaluation seems to be the *accuracy* of our measures. Including more specific context into our paradigms promises better applicability of the results obtained. To counteract a major influx in cost that would be tied to this gain in accuracy, I propose to carry out these context aware evaluations in a simulation environment. A clear distinction has to be made here between the full interactive environments as they are used in Human machine interaction and a more shallow computer based version which might not offer as much immersion, but still offers the ability to assimilate the use case at lower cost. A training phase might be needed for participants to get used to the environment and diminish the effects of the simulation. While passive measurements like interaction time to gage the systems pleasantness have their inherent own problems (it could also denote bad intelligibility), they could easily be recorded and serve as complimentary indicators. To test the effect of simulation immersion, two different tests are proposed to be carried out: One in the simulation environment and a second in a mimicking real life environment, with an optional intermediate level being artificial reality. If the simulation proves to have no significant distortion effect, different application contexts should be evaluated to see if this is generalisable (i.e. car navigation vs. sheduling task). One big challenge lies also in the naïve participants ability to navigate a two dimensional simulation space, without a break in immersion, which might in turn invalidate the results for a real life setting. As it is the goal to have the environment be web-distributable so that experiments can be carried out in an efficient fashion, this also constrains the set of use cases for which it can be applied.

4. References

[1] U. Jekosch, *Voice and speech quality perception: assessment and evaluation*. Springer Science & Business Media, 2006.

[2] W. D. Voiers, A. D. Sharpley, and I. L. Panzer, "Evaluating the effects of noise on voice communication systems," in *Noise reduction in speech applications*. CRC Press, 2002, pp. 125–152.

[3] K. Kondo, *Subjective quality measurement of speech: its evaluation, estimation and applications*. Springer Science & Business Media, 2012.

[4] ISO8402:1986, "Quality," International Organization for Standardization, Geneva, CH, Standard, 1986.

[5] I. T. Union, "ITU-T Rec. P.808, Subjective evaluation of speech quality with a crowdsourcing approach," 2018.

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

[7] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, 2014.

[8] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? no!—an empirically-supported critique of interspeech 2014 tts evaluations," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] 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. SSW*, 2019, pp. 105–110.

[10] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE international conference on acoustics, speech, and signal processing.*, vol. 2. IEEE, 2001, pp. 749–752.

[11] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment," *J. Audio Eng. Soc*, vol. 61, no. 6, pp. 366–384, 2013. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16829

[12] D.-S. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 821–831, 2005.

[13] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," *arXiv preprint arXiv:1611.09207*, 2016.

[14] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proc. Interspeech*, 2021, pp. 2127–2131.

[15] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8442–8446.

[16] S. Le Maguer and N. Harte, "Investigation of auditory nerve model based analysis for vocoded speech synthesis," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2020, pp. 1–6.

[17] V. Kraft and T. Portele, "Quality evaluation of five german speech synthesis systems," *Acta acustica (Les Ulis)*, vol. 3, no. 4, pp. 351–365, 1995.

[18] I. T. Union, "ITU-R BS.1534-1, method for the subjective assessment of intermediate quality level of coding systems," 2015.

[19] J. Kreiman, B. R. Gerratt, and M. Ito, "When and why listeners disagree in voice quality assessment tasks," *The Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. 2354–2364, 2007.

[20] A. Black and K. Tokuda, "The blizzard challenge 2005: Evaluating corpus-based speech synthesis on common databases," in *Proceedings of interspeech*, 2005, pp. 77–80.

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

[22] Y. V. Alvarez and M. Huckvale, "The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems," in *Proc. 7th International Conference on Spoken Language Processing*, 2002, pp. 329–332.

[23] 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.

[24] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (mos) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[25] T. Hoßfeld, R. Schatz, and S. Egger, "Sos: The mos is not enough!" in *2011 third international workshop on quality of multimedia experience*. IEEE, 2011, pp. 131–136.

[26] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8442–8446.

[27] S. Shirali-Shahreza and G. Penn, "Mos naturalness and the quest for human-like speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 346–352.

[28] P. Wagner and S. Betz, "Speech synthesis evaluation: Realizing a social turn," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017*, pp. 167–173, 2017.

[29] 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.

[30] F. Hinterleitner, S. Möller, C. Norrenbrock, and U. Heute, "Perceptual quality dimensions of text-to-speech systems," in *Proc. Interspeech*, 2011, pp. 2177–2180.

[31] F. Hinterleitner, C. Norrenbrock, and S. Möller, "Is intelligibility still the main problem? a review of perceptual quality dimensions of synthetic speech," in *Proc. SSW*, 2013, pp. 147–151.

[32] R. Zequeira Jiménez, A. Llagostera, B. Naderi, S. Möller, and J. Berger, "Intra-and inter-rater agreement in a subjective speech quality assessment task in crowdsourcing," in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 1138–1143.

[33] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomedical Signal Processing and Control*, vol. 14, pp. 42–54, 2014.

[34] C. Fournier and D. Inkpen, "Segmentation similarity and agreement," in *Proc. of the conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2012, pp. 152–161.