SPEECH SYNTHESIS EVALUATION: REALIZING A SOCIAL TURN

Petra Wagner[1,2] *and Simon Betz*[1,2]

[1]*Fakultät für Linguistik und Literaturwissenschaft, Arbeitsgruppe Phonetik und Phonologie*
[2]*Center of Excellence Cognitive Interaction Technology (CITEC)*
Universität Bielefeld
petra.wagner@uni-bielefeld.de

Abstract: Based on a meta-analysis of the state-of-the-art in speech synthesis evaluations, we diagnose the following dilemma: Despite known drawbacks, evaluations predominantly rely on small-scale laboratory tests, typically capturing MOS-based global impressions based on isolated sentences, with the (resynthesized) human voice serving as a gold standard. The problem with such approaches is that synthesis quality can only reliably be estimated if presented in a contextualized manner, e.g. as part of an application and together with its embodiment as an artificial agent, robot or an disembodied voice. As most evaluations are carried out in parallel to system development, and as these tend to be concerned with small details in the developmental process, evaluations of fully fledged applications are often neither possible nor useful. We argue, that with a few modifications in standard evaluation protocols, i.e. by introducing simple interactive scenarios and by relying on both subjective impressionistic and behavioral or physiological measurements, the reliability of such evaluations could be significantly improved.

1 Introduction: The Dilemma in Estimating Speech Synthesis Quality

In his recent keynote presentation at the International Congress of the Phonetic Sciences, Simon King named the lack of suitable evaluation procedures as one of the key problems in current speech synthesis research [1]. In his textbook on estimating the quality of communicative technologies [2, p. 83], Sebastian Möller mentions a possible reason for the obviously lacking progress in this area: There exists a problematic trade-off between testing systems in realistic field applications and in the lab. While field tests are essential to gain the necessary ecological validity in order to find out whether a user is really content with the overall application or not, these tests are hardly fit to estimate the quality of the finer details of the system components, i.e. the adjustable parameters of a synthesis system. In order to accomplish such a fine-tuning of a speech synthesis application, it is necessary to accumulate large numbers of listeners' impressions on various system settings in order to determine an optimal solution for maximal system quality. This leads to an obvious dilemma: We are unable to make any a-priori claims on the quality of synthetic speech outside the lab, but we need to test quality (at least to some degree) inside the lab.

This dilemma actually mirrors a similar problem that has been recently identified in speech research, namely the necessity to carry out investigations in the lab in order to remain in control about the tested variables, while being unable to guarantee that ones findings generalize to situations outside the lab [3]. To some extent, similar problems have been noted in many research areas concerned with laboratory settings while attempting to predict and explain behavior in the field, and can be regarded as a typical instance of the general trade-off between external and

internal validity in all kinds of empirical studies. Interestingly, though, speech synthesis evaluations hardly seem to notice this problem, despite it is being mentioned in standard textbooks on the subject matter. When looking at the way speech synthesis evaluations are currently carried out, we find a strong reluctance to test whole systems rather than conducting small-scale laboratory tests, often comprising only a small number of participants and using coarse metrics of overall quality (cf. Section 2). We argue that the key to this problem is the dominance of the "conduit metaphor" in much speech technology-related thinking. That is, language's sole purpose is often regarded as being a means to transmit information between interlocutors [4]. Seen in this light, the quality of synthetic speech is largely determined by its comprehensiveness, plus a feature often called "naturalness". While we agree that comprehensiveness is a necessary key ingredient to well-sounding artificial speech, we argue that the notion of "naturalness" is ill-defined or even misguided, as it is not a suitable concept for describing speech production independently of the person or system who produces it and without making a reference to the situation in which it is uttered [3].

The predominant "conduit metaphor" of speech communication fails to address these dynamic aspects of what constitutes an appropriate (rather than natural) way of speaking. Similarly, it fails to take into account how language serves to establish a socio-pragmatic frame, e.g. by negotiating social roles, intentions, expressing attitudinal assessments, and how different communicative settings ask for specific speaking styles that may differ in phonetically subtle ways [3]. Much speech synthesis research appears to believe that synthetic utterances are immune to these social embeddings, but this assumption seems to be misguided: A well-know example for the effect of contextual embedding on perception of artificial systems is the so-called "uncanny valley"[5]. The "uncanny valley" refers to a small area alongside a continuum of robotic appearance patterns, where robots without any resemblance to humans are on one end (e.g. industrial robots), and systems indistinguishable from humans are on the other end. Human participants often perceive systems near to but not identical to humans as somewhat repelling, or even showing a resemblance to corpses or "zombies". Using a Bayesian modeling approach, [6] claims that the reason for this effect is an imbalance between human users' expectations (i.e. humans should behave and act exactly like humans) and appearance (a humanoid robot is very similar to but not identical to a human). On a similar vein, [7] argue that system designers in speech technology ought to meet these user's expectations rather than being guided by a human-like gold standard .

Taking this argument one step further, we can expect a human system user's impression of a particular voice to be shaped by her expectations of this voice, and these expectations will be influenced by the overall system's appearance and behavior. Hönemann and Wagner [8] indeed found a metallic sounding artificial voice to be more acceptable when combined with a typical "robot head" as compared to a more humanoid artificial head.

It is thus highly likely that humans expect a close correspondence between an artificial agent's embodiment or verbal behavior and the sound of its voice. A synthetic voice's gold standard may therefore be different from a human voice. Given this, the evaluation process needs to trigger the user's expectations and therefore makes little sense in a context-free setting.

2 Analysis: Assessing the State-of-the-Art in Speech Synthesis Evaluation

In order to get a realistic picture of the established standard methods in speech synthesis evaluation, we carried out a meta-analysis of the evaluation methodologies reported in the proceedings of the last three ISCA Speech Synthesis Workshops (SSW7 in Kyoto, Japan, SSW8 in Barcelona, Spain, SSW9 in Sunnyvale, USA). For each research paper in the proceedings, we assessed whether there was an evaluation. In line with the general high quality of this work-

shop series, this was the case for the vast majority of papers, and only missing in such cases, where due to a particular research focus, an evaluation of speech synthesis quality made no immediate sense (e.g. in papers on grapheme-to-phoneme conversion or a new synthesis-related resources). For all of the remaining papers, we then assessed the methods that have been applied in order to evaluate the synthesis quality. As many papers employed more than one evaluation method (typically a combination of objective and subjective evaluation or a combination of several subjective methods), the total raw counts are not identical to the number of workshop papers reporting an evaluation.

This meta-analysis revealed that even though there exist a few quasi-standards (e.g. MOS or DMOS scales for subjective impressionistic ratings, pairwise comparisons or preference tests (e.g. ABX, AB-tests) for system comparisons, WER or Levenshtein edit distance for detailed intelligibility assessment, RMSE for objective evaluations), there is much heterogeneity (cf. Table 1). In some cases, the different approaches are explicable by the specific research goal: When aiming at a convincing synthesis of various basic emotions, an intelligibility test is obviously not the best choice. MOS-type scales are employed as a multi-purpose tool, most often used to rate naturalness and intelligibility, but also other dimensions of quality and occasionally also as a tool for system comparison (DMOS). In recent years, more fine-grained alternatives to MOS are found (e.g. MUSHRA). Few publications use more complex forms of evaluation, e.g. MDS (Multidimensional Scaling) in order to isolate quality dimensions of synthetic speech, but those are in the vast minority. Recently, there seems to be a trend to rely more on objective evaluations only, while "informal listening tests" have mostly been replaced by more reliable methods. Intelligibility appears to be less of a problem in more recent years, as it tends to be assessed only for selective problems (e.g. synthesis of minority languages, synthesis under noisy conditions)

Workshop	WER/edit dist.	MOS-type	preference tests	objective tests	other
SSW7	5	22	17	22	6
SSW8	5	28	16	10	6
SSW9	-	13	8	18	2

Table 1 – Raw count of speech synthesis evaluation methods employed in recent ISCA Speech Synthesis Workshops (only papers considered where evaluation was indicated, several counts per paper possible)

Epecially for fine-grained system modifications as part of a longer system development process, objective evaluations are the preferred form of evaluation, as they are less time consuming, less costly and can be repeated as a consistent diagnostic at various stages of system development. If additional subjective tests are performed in parallel to system development, they tend to rely on comparatively small participant samples, often $n < 10$, in some cases even $n < 5$! In recent years, web-based or crowdsourcing-based evaluations have somewhat gained popularity, but are still far from representing the standard evaluation approach. Despite the availability of web-based platforms that should be able to target large numbers of participants, only very few evaluations rely on participant numbers $n > 30$, the typical sample size for listening tests ranging from $n = 8$ to $n = 20$.

In almost all publications we examined, the test material used were isolated sentences played to listeners under "out of the blue" conditions. The sentences were sometimes custom-designed in the form of semantically unpredictable sentences (SUS) if intelligibility was at stake. In most cases, the sentences to be rated by listeners were randomly selected out of a larger database or training set. In many cases, the way the sentence material was selected was not revealed by the researchers and hardly any paper actually gives precise information about the exact test material used.

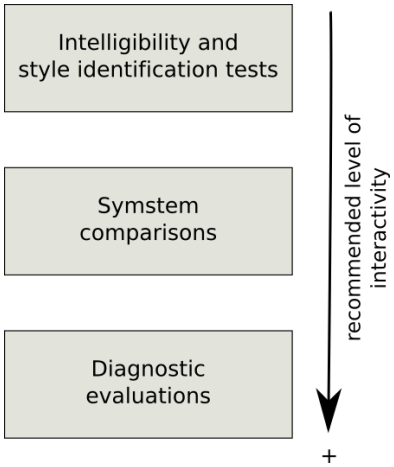

Figure 1 – Recommendations for level of interactivity depending on evaluation focus.

Maybe most strikingly, and despite the known disadvantages of the approaches described above (cf. Section 1), *not a single paper in our meta analysis* employed a method where the subjective assessment was performed based on impressions gathered in a more realistic, interactive application. This was not even the case when the synthesis quality of fully-fledged dialogue systems was at stake. A plausibe reason for this finding is the system developer's concern that such an approach would result in listener's assessment of the entire system rather than the speech synthesis quality only.

With very few notable exceptions (e.g. Edlund et al. [9]), the dilemma of speech synthesis evaluations, namely the difficulty to predict a system user's reaction to a realistic application based on lab-based testing is obviously not familiar to researchers or simply ignored by them. The reasons for this may relate to the difficulty and cost involved in performing field tests, or the concern to evaluate the overall system rather than the synthesis voice. We argued above that these concerns are misguided as the assessment of speech synthesis quality is likely to not work without contextualization. In the next section, we make suggestions for laboratory-based testing that can still be considered an improvement over the predominant approaches currently followed.

3 Discussion: Sketching an Alternative Research Programme

Even if performed under laboratory conditions, we believe that the subjective evaluation results may become more informative and robust, if more care is taken to choose a suitable approach instead of simply going for a "safe bet" such as having isolated sentences judged by an MOS or in a pairwise comparison task. More specifically, we suggest to generally embed the synthetic sentences in interactive settings, e.g. web-based games, or interactive tasks, thereby modeling the design, layout and embodiment of the finalized system (e.g. the appearance of a robot or artificial agent, an disembodied voice) and shaping user expectations as part of the evaluation process (cf. Section 1).

Based on our meta-analyis in Section 2, we define three main types of (subjective) evaluation foci. We argue that these are in need of different amounts of interactivity, contextualization and diagnostic control in increasing order (cf. Figure 1):

- Intelligibility tests/identification tests (e.g. of target emotions)
- System comparisons

- Diagnostic evaluations, to derive strategies for further system development

3.1 Intelligibility and style identification tests

If an evaluation targets the intelligibility of synthetic speech, or the performance in mimicking a certain speaking style, voice quality or target emotion, traditional laboratory tests based on isolated sentences may probably be still appropriate, e.g. to estimate the overall performance. However, we still believe that impressions of voice qualities, emotional expression, speaking styles or gender may be strongly influenced by their contextual embeddings. Therefore, interactive settings may still be preferable for their overall assessment, especially if it can already be estimated in which applications the synthesis voices are likely to be employed.

3.2 System comparisons

If the target of an evaluation lies in the comparison of two systems, the potential improvement of the novel approach to synthesis is decided on best in interactive scenarios rather than a pairwise comparison of isolated sentences. Here, suitable scenarios can be found in HCI and HRI evaluations, where two systems interact with one user in a task, e.g. by acting as museum or tourist guides, making suggestions for interesting things to look at. In such scenarios, user preferences have shown to be indicative of preferring one system over another. Furthermore, these preferences may depend on the users themselves, e.g. their gender [10]. An additional advantage of such task-based interaction is the fact that besides user preferences for a particular voice, other behavioral measures can be assessed as well, e.g. task completion time, response times [11] or gaze behaviour [12]. These additional measures can be seen as suitable operationalizations of processing ease, intelligibility or listener attention, all of which can be expected to be indicative of overall synthesis quality. It can even be expected that some of these behavioral metrics may be able to measure synthesis quality in a more fine-grained manner than subjective quality assessments. As for subjective judgments, it seems a promising idea to not entirely rely on MOS-like ratings, which are time consuming and distract the listeners from their task. Rather, it seems to be profitable to search for methods of assessing subjective impressionistic responses in a more online fashion, e.g. by giving the listener the option to intervene if the synthesis quality falls below an acceptable standard [9].

3.3 Diagnostic evaluations

In diagnostic evaluations, it is considerably harder to place a finger on the exact causes for a degrading synthesis quality (it could be anything, or anywhere!). In such cases, it is certainly tempting to go back to simplistic listening tests enabling complex analyses such as Multidimensional Scaling as to get a clear picture of the problem areas. However, we argue that diagnostic evaluations may profit from interactive designs, as these enable the evaluator to assess the behavioral metrics as suitable indicators pointing to problems in voice design. In addition to potential impressionistic responses, interactive systems may be combined with methods for assessing physiological indicators of speech quality. While these could in principle be used as well in non-interactive tasks, they have a huge potential in interactive settings, as they could be used for monitoring the interaction quality in an online fashion without the user having to actively make a subjective assessment. Potentially useful physiological measures may be EEG, skin conductance response or pupillary response [13]. These may be even more sensitive than behavioral responses and have the further advantage that they do not distract the user while interacting with the system. Such physiological metrics may guide the system designer directly to problematic aspects in the interaction quality that the users themselves are not even aware of.

4 Conclusion

Summing up, we suggest that synthesis evaluations should preferably take place in socially embedded, interactive settings and ought to monitor user responses to a system throughout these interactions, e.g. by assessing behavioral (e.g. reaction times, eye tracking) and physiological (e.g. EEG, skin conductance response, pupillary response) responses which are indicative of changes in the overall interaction quality, possibly together with a simultaneous online assessment of subjective listeners' assessments.

References

[1] KING, S.: *What speech synthesis can do for you (and what you can do for speech synthesis)*. In *Proceedings of the 18th International Congress of the Phonetic Sciences (ICPhS 2015)*. Glasgow, Scotland, 2015.

[2] MÖLLER, S.: *Quality Engineering: Qualität kommunikationstechnischer Systeme*. Berlin, Heidelberg: Springer-Verlag, 2010.

[3] WAGNER, P., J. TROUVAIN, and F. ZIMMERER: *In defense of stylistic diversity in speech research*. *Journal of Phonetics*, 48, pp. 1–12, 2015.

[4] SHANNON, C.: *A mathematical theory of communication 27:379–423, 623–656, 1948*. *Bell System Technical Journal*, 27, pp. 379–423, 623–656, 1948.

[5] MORI, M.: *Bukimi no tani: the uncanny valley*. *Energy*, 7, pp. 33–35, 1970.

[6] MOORE, R.: *A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena*. *Scientific Reports*, 56(2), p. 864, 2012.

[7] MOORE, R.: *From talking and listening robots to intelligent communicative machines*. In J. MARKOWITZ (ed.), *Robots That Talk and Listen*. Boston, MA: De Gruyter, 2015.

[8] HÖNEMANN, A. and P. WAGNER: *Adaptive Speech Synthesis in a Cognitive Robotic Service Apartment: An Overview and First Steps Towards Voice Selection*. In *Tagungsband Elektronische Sprachsignalverarbeitung ESSV 2015*, pp. 135–142. 2015.

[9] EDLUND, J., C. TÅNNANDER, and J. GUSTAFSON: *Audience response system-based assessment for analysis-by-synthesis*. In *Proceedings of the 18th International Congress of the Phonetic Sciences (ICPhS 2015)*. Glasgow, UK, 2015.

[10] STRUPKE, E., O. NIEBUHR, and K. FISCHER: *Influence on robot gender and speaker gender on prosodic entrainment in HRI*. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communiction (RO-MAN 2016)*. New York, 2016.

[11] BETZ, S., S. ZARRIESS, and P. WAGNER: *Synthesized lengthening of function words – the fuzzy boundary between fluency and disfluency*. In *Proceedings of the International Conference Fluency & Disfluency Across Languages and Language Varieties ((DIS)FLUENCY 2017)*. Louvain-la-Neuve, Belgium, 2017.

[12] RAJAKRISHNAN, R., M. WHITE, S. R. SPEER, and K. ITO: *Evaluating prosody in synthetic speech with online (eye tracking) and offline (rating) methods*. In *Proceedings of the 7th Speech Synthesis Workshop (SSW7)*, pp. 276–281. Kyoto, Japan, 2010.

[13] ANTONS, J.-N., R. SCHLEICHER, S. ARNDT, S. MÖLLER, A. K. PORBADNIGK, and G. CURIO: *Analyzing speech quality perception using electro-encephalography*. *IEEE Journal of Selected Topics in Signal Processing*, 6(6), pp. 721–731, 2012.