Are we using enough listeners? No!
An empirically-supported critique of Interspeech 2014 TTS evaluations

Mirjam Wester, Cassia Valentini-Botinhao, Gustav Eje Henter

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
{mwester,cvbotinh,ghenter}@inf.ed.ac.uk

Abstract

Tallying the numbers of listeners that took part in subjective evaluations of synthetic speech at Interspeech 2014 showed that in more than 60% of papers conclusions are based on listening tests with less than 20 listeners. Our analysis of Blizzard 2013 data shows that for a MOS test measuring naturalness a stable level of significance is only reached when more than 30 listeners are used. In this paper, we set out a list of guidelines, i.e., a checklist for carrying out meaningful subjective evaluations. We further illustrate the importance of sentence coverage and number of listeners by presenting changes to rank order and number of significant pairs by re-analysing data from the Blizzard Challenge 2013.

Index Terms: Subjective evaluation, text-to-speech, MOS test

1. Introduction

It is common to illustrate the performance of, for example, speech synthesis systems or voice conversion methods by presenting objective error measures such as mel cepstral distortion and the likelihood of the training set [1–3]. Although these give an indication of how well the synthesis model represents natural speech, automatically measuring the perceptual quality of synthetic speech is a challenge even when a reference natural speech signal is available, which often is not the case [4]. Although non-intrusive measures (measures that do not require a reference speech signal) have been proposed for synthetic speech [5, 6], subjective listening tests remain the gold standard for a true measure of quality.

The most commonly used listening tests are Mean Opinion Score tests (MOS) or Differential MOS (DMOS), preference tests, ABX-tests, transcription tasks, and MUSHRA tests. The synthesis attributes measured by these tests can range from quality to naturalness, intelligibility, similarity, expressiveness, pleasantness, and even emotions.

Through the years there have been many papers giving listening test guidelines [7–11]. However, contemporary evaluations of synthetic speech frequently do not take these guidelines to heart when designing and carrying out listening tests. This paper intends to present good practice in designing listening tests for subjective evaluation of synthetic speech systems, e.g., statistical parametric speech synthesis (SPSS), unit-selection, hybrid methods, and voice conversion. We detail some common shortcomings of current subjective evaluations and illustrate the importance of a sufficient amount of test material and participants in listening tests by an example using real data.

The paper is organised as follows: Section 2 begins by presenting a checklist of elements that must be considered when designing a good listening test. Following on from the checklist, we inspect the state of affairs pertaining to subjective evaluation at last year's Interspeech. Next, in Section 4, the importance of sentence coverage and number of listeners is illustrated by means of a re-analysis of a portion of the Blizzard 2013 data [12]. We conclude by discussing how the results may be interpreted and by presenting our final recommendations.

2. A checklist for successful testing

There are many factors to consider when designing a subjective evaluation. The first question to ask oneself is: "What do I want to measure?" This should be followed by: "How do I get the answer to my question using listeners?"

To help you answer the above two questions we present a checklist of points/questions you need to consider when designing a test for subjective evaluation. The checklist consists of a list of questions, with comments and references supporting the relevance of each item. There is no one correct answer to any of the questions, but if these points are addressed every time a listening test is designed it will result in more meaningful subjective testing of synthetic speech.

- What test to use? MOS, MUSHRA, preference, intelligibility, and same/different judgements all fit different situations.
- Which question(s) to ask? Be aware that the question you ask may influence the answer you get [13]. The terms you use may be interpreted differently by listeners, e.g., what does "quality" or "naturalness" actually mean?
- Which data to use for testing? Factor out aspects that affect the evaluation, but which are unrelated to the research question studied.
- What type of listeners? Native vs. non-native? Speech experts vs. naïve listeners? Age, gender, hearing impairments? Different listener groups can lead to different results [14–17]. See Section 4.1 for an analysis of the effect of listener type.
- Is a reference needed? Consider giving a reference or adding training material, particularly for intonation evaluation [18]. Also consider the case for including other anchors.
- How many listeners to use? See Section 4.1 for an analysis of the effect of listener numbers on Blizzard 2013 data.
- How many datapoints are needed? Section 4.2 investigates the effect of the number of datapoints.
- Is the task suitable for human listeners? Take into consideration listener boredom, fatigue, and memory constraints, as well as cognitive load [19].
- Can you use crowdsourcing? The biggest concern here is how to ensure the quality of the test-takers [20–22].
- How is the experiment going to be conducted? With headphones or speakers, over the web or in a listening booth?
- Is the evaluation material unbiased and free of training data?

In short, *think before you test!* Don't treat subjective evalu-

Number of listeners	Number of studies	
	Preference test	MOS
1–10	10	8
11–20	5	5
21–30	0	1
31–50	4	5
> 50	3	3
Not stated	2	0
Total studies	24	22

Table 1: *Number of speech synthesis studies at Interspeech 2014 using a particular amount of listeners.*

ation as a fishing expedition: By all means, carry out pilot tests to learn how long experiments take, if they are feasible and doable for subjects, but conform to good scientific practice by deciding your hypothesis and how many subjects you need before running your experiment, and correct for multiple comparisons in your statistical analysis. Finally, report on the design of your experiment and motivate the choices made – just showing an MOS plot is not enough information.

3. Listening tests at Interspeech 2014

With the checklist from Section 2 in mind, we carried out a survey of synthetic speech evaluations at Interspeech 2014. Searching the Interspeech 2014 proceedings for the term "synthesis" returns 188 matching papers. Of these, 64 actually deal with speech synthesis and include some form of subjective evaluation. A further ten also perform speech synthesis, but only include objective evaluations. The remaining 114 papers include the search term "synthesis" either as part of a reference or in an example, but do not actually deal with synthesised speech.

The papers that performed subjective evaluation were divided into groups depending on the number of listeners included in their subjective evaluation. Table 1 summarises the number of studies that used a particular amount of listeners in either preference tests or MOS tests. Note that DMOS studies are included in the "MOS" column, while "Preference test" also includes ABX tests and pairwise comparisons. Studies that carried out MUSHRA tests, transcription tasks, or reaction time experiments have not been included here, as there were too few of these to carry out a meaningful analysis.

Studies with both MOS and preference tests have been included in the table twice; once for each column. A study with more than one MOS counts as a single MOS test and a study with multiple preference tests counts as one preference test.

The first striking bit of information in Table 1 is that around 40% of studies include only between one and ten listeners. Studies with more than 50 subjects are all studies that use crowdsourcing platforms (Amazon Mechanical Turk or Crowdflower) for their experiments. A few studies fail to mention the number of listeners altogether. Actually, the publications frequently omit other relevant information as well, for example:

- The demographics of listeners (native or non-native, age, accent, possible hearing impairments).
- The language of the synthesised speech.
- The domain of the sentence material (training and test).
- The number of test samples (sentences, words, paragraphs).
- The specific question participants were asked to answer.
- The listening conditions (headphones or speakers, listening booth or on the web).

The lack of relevant detail in papers could suggest that little thought has gone into designing the experiments. Following the checklist in Section 2 would largely remedy such issues.

4. Re-analysing the Blizzard Challenge

To illustrate the importance of sentence coverage and the number and type of listeners, we re-analysed listener response data from the Blizzard Challenge 2013 evaluation [12]. This particular year was chosen since it is the most recent challenge involving English synthetic speech. We focus on results of MOS tests for naturalness and similarity on the main task (EH1).

In 2013, the Blizzard Challenge evaluated eleven systems, including natural speech. Each listener scored each system five times for naturalness and once for similarity, except natural speech, which was only scored three times for naturalness. The final scores published in [12] were obtained with 50 paid participants (EE), 92 volunteers (ER), and 52 speech experts (ES). Paid participants were native English speakers performing the task using headphones seated in sound isolated booths. Volunteers and speech experts were recruited online and took the test over the Internet, with no control over their listening conditions or nativeness status.

To assess the robustness of MOS test conclusions to the number of participants and sentences used, we re-analysed progressively larger subsets of the Blizzard data. For each analysis, we computed two things: the number of significantly different system pairs, and the rank correlation between the ranking given by the current data subset and the ranking obtained when considering all participants for the test in question. To compute the number of significantly different pairs we used Bonferroni-corrected pairwise Wilcoxon signed-rank tests at a 1% level. This is the same procedure used to analyse MOS test data in Blizzard [23]. To calculate the correlation between two rankings we used the Kendall τ rank correlation coefficient [24].

4.1. Participants

To begin with, we consider the effect of the number of test participants, as well as how results differ depending on the type of listener used. To quantify how the number of listeners affects the ability of the test to discriminate between different systems, we computed the number of system pairs that were found to be significantly different when gradually increasing the number of listeners included in the analysis. To eliminate potential effects of sentence material, listeners were subsampled such that all system-sentence combinations were always covered. The results are presented in Fig. 1, where each point is an average across independent 1 000 resamplings (hence the minor amount of sampling noise). In this, as in all our graphs, cubic interpolation has been used between datapoints to better visualise the shapes of the curves. Solid curves correspond to naturalness results, while dashed graphs refer to the similarity task.

From Fig. 1, it is clear that the Blizzard similarity tests overall resulted in fewer significant differences than the naturalness evaluation. We will investigate the cause of this difference in the next section. There also appear to be some differences between the various types of listeners, particularly for naturalness.

Apart from the discriminative power of the test, it is also important that appropriate distinctions are made. To assess this, we calculated the rank correlation between system rankings based on the subsampled data and the final ranking obtained when averaging the full dataset including all listener types. (Since data is shared between the two rankings, the results may be biased to be overly optimistic, especially for large

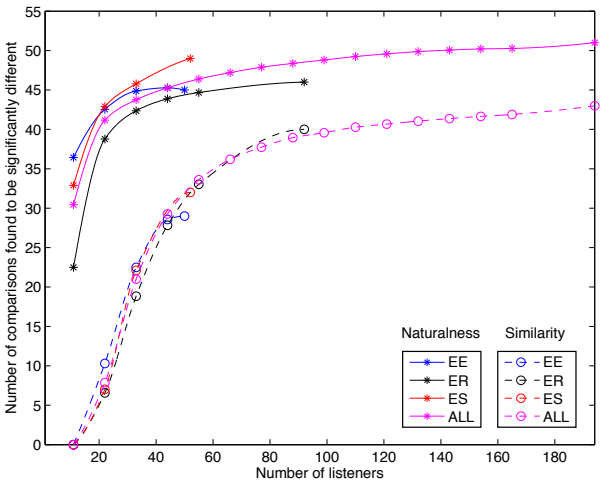

Figure 1: *Number of significantly different pairs in MOS tests, as a function of the task and the number of listeners. The theoretical maximum is 55.*

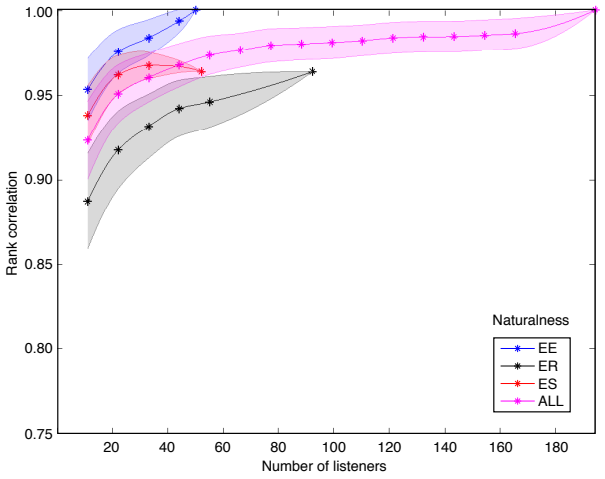

Figure 2: *Rank correlation of naturalness ratings with final rank obtained from all participants.*

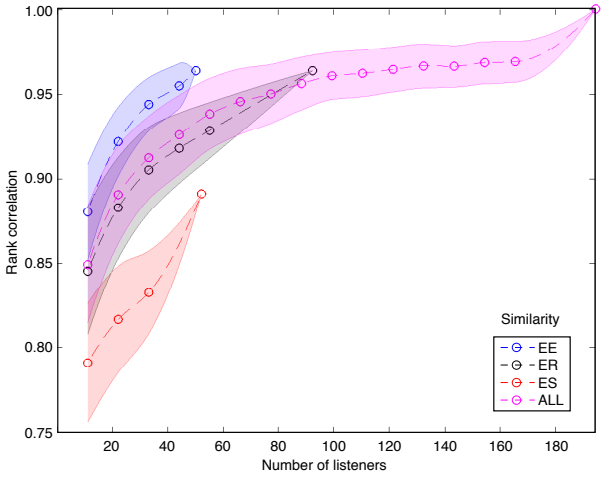

Figure 3: *Same as Fig. 2, but for similarity ratings.*

numbers of listeners.) The results are displayed in Figs. 2 and 3 for naturalness and similarity, respectively. The shaded bands have widths of one standard deviation, estimated from 1 000 resamplings as before.

We see that rank correlations at first improve rapidly with the number of listeners, but that the rate of growth generally decreases as higher listener numbers are reached. For the naturalness task in Fig. 2, using 30 paid participants was sufficient to achieve strong correlation (more than 0.98) with the final ranking. The minor correlation gap when using 30 rather than 50 paid participants was due to frequent rank changes between two pairs of Blizzard systems: (I, L) and (H, F). For similarity (Fig. 3), the correlations are generally a bit lower, it takes a larger number of listeners to make the rank correlation rise to similar levels, and the results never quite reach stability.

It is also interesting to contrast different types of listeners. For this purpose, the graphs in Figs. 1 through 3 have all been broken down across the three different Blizzard listener types. In terms of rankings, paid participants (EE), despite being the smallest group, correlated the best with the full-data rankings for both naturalness and similarity. For naturalness ratings, volunteers (ER) consistently gave low rank correlations and the least number of number of significant pairs for a given number of listeners, suggesting a greater inherent noise level and lower discriminative power in their ratings.

Interestingly, expert listeners (ES) identified a large number of significant differences in naturalness as the number of participants grew; however, their rank correlation with the overall, full-data picture was either close to average (for naturalness) or the lowest observed (for similarity). This could be interpreted as expert listeners having strong and clear opinions, though these opinions may diverge significantly from the general population.

4.2. Data coverage

Of course, having a large number of listeners is not sufficient for achieving generalisable conclusions. When evaluating synthesis systems it is essential to also use a large variety of sentences, as the output quality might vary dramatically from one sentence to the next. In a MOS test, where it is not considered desirable to present the same sentence to the same listener more than once, this typically calls for balanced designs such as a latin square, to ensure that every system-sentence combination is evaluated. The number of listeners required to cover all combinations is then equal to the number of systems being tested, and the set of listeners that listen to exactly the same system-sentence combinations (the same stimuli) will here be referred to as a *listener group*.

To illustrate the importance of covering all system-sentence combinations we computed the average score of each system for each listener group (wherein everyone scored the same stimuli). These scores are presented in Fig. 4. It can be seen that the judgments change substantially between listener groups, particularly for the similarity scores.

The need to adequately sample both listener and sentence variation puts a lower bound on the number of datapoints required. With too few samples, stochastic variation perturbs rankings and makes it impossible to confidently tell systems apart. To investigate the effect of the size of the statistical material, and to put the naturalness and similarity results on a more equal footing, Fig. 5 graphs the number of significant pairs for the two tasks as a function of the total number of ratings used per system. Like before, the plots are averages over a large number of data subsets, but for convenience and to get better granularity, this figure was created by successively adding entire listener groups (in all possible combinations), rather than selecting a certain number of listeners within each group as in

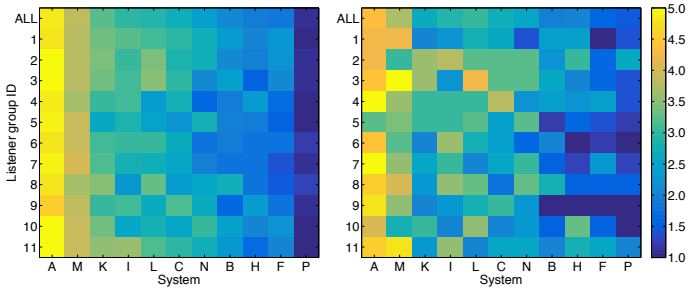

Figure 4: *Naturalness (left) and similarity (right) results for paid participants. Colours indicate the mean score of each system for each listener group.*

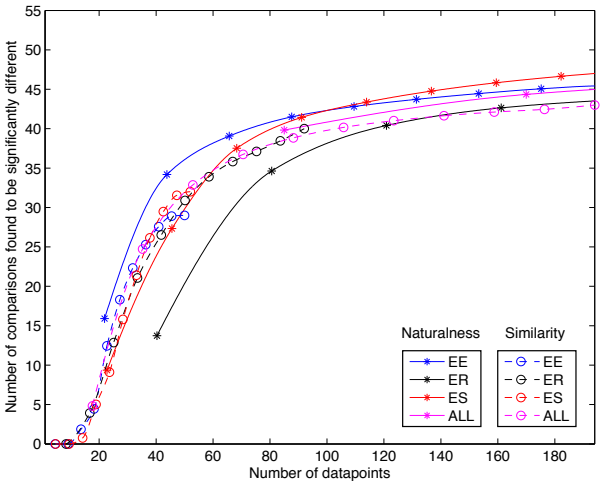

Figure 5: *Number of significantly different pairs in MOS tests, as a function of the task and the number of datapoints.*

previous graphs.

Fig. 5 is interesting, because it shows that the the big gap between naturalness and similarity tasks in previous figures can largely be explained by the difference in the number of scores collected per listener. The Blizzard Challenge only collects a single similarity judgement per system and participant, which really is too few. Perhaps the biggest outlier in the graph is the low number of significant naturalness differences identified by the volunteers. This can possibly be explained by unnaturalness cues being quite subtle artefacts, which may be difficult to perceive on poor equipment or with impaired hearing.

5. Discussion

In general, the number of participants required to accurately identify significant differences, as well as to assess the magnitudes of these differences (effect sizes), will of course depend on the task and the stimuli considered. Some caution is thus advised when generalising the results of our re-analysis of the Blizzard data to other situations. If differences between stimuli are minor, a substantial amount of data might be required to tease them apart. Often, this can be mitigated by using other testing paradigms that emphasise pairwise or parallel comparisons, such as MUSHRA, ABX, or preference tests. Among these, the MUSHRA methodology [25] has the advantage of also allowing straightforward estimation of effect sizes, and has been found to yield more significant differences than a MOS test would, other things being equal [26].

On the other hand, a large amount of acoustic variation is not necessarily helpful, either. It is likely of importance how well the stimulus variation correlates with the listener's internal perceptual model, to which synthetic stimuli are compared when a listener scores them. As an example, the artefacts in unit selection synthesis (e.g., bad joins) are typically quite distinct from artefacts in SPSS (e.g., vocoder buzz), and arguably mostly orthogonal to each other. Since preferences and internal perceptual models may vary from listener to listener, having many listeners is important in order to accurately sample the space of internal listener models and converge on the population average. The Blizzard Challenge analysed here is an example of a test with a highly heterogeneous pool of systems, and the associated stimuli may be acoustically quite distinct.

All things considered, the results of the Blizzard Challenge re-analysis strongly suggest that synthetic speech naturalness evaluations, particularly MOS tests, should include more listeners compared to the numbers commonly used today (cf. Section 3). For reliability, we would recommend using at least 30 listeners. Moreover, each listener should listen to several examples of each system evaluated. 150 total judgements per MOS value computed should probably be considered a minimum.

The above numbers are for paid participants in carefully controlled conditions. In less controlled scenarios, such as crowdsourcing, behaviour closer to the online volunteers (ER) in Blizzard may be expected. In these situations, our advice would be to collect significantly more data and listeners. Even so, the power to draw conclusions may be limited, for instance because participants may not be using proper listening equipment and therefore not be able to discern minor differences between systems. For expert listeners, as may be recruited in a lab of speech researchers, one should keep in mind that their preferences may differ from those of the general public.

The numerical analysis in this paper has mostly focussed on requirements for identifying significant and stable differences. Of course, statistical significance is not the same as a practical significance, and the end goal is not to always tell all systems apart. Somewhere in the long tails of our figures, it makes sense to stop testing, and instead direct resources towards improving the systems involved. However, this is not an excuse for using an unsound testing methodology. Moreover, being able to identify many significant differences is generally a pre-requisite for accurately estimating effect sizes. Effect sizes quantify the subjective advantages of one system over another, and so are a step towards a more meaningful difference measure.

For truly meaningful results, a system should be evaluated for the task and context where it is ultimately used. This ideal is however at odds with the plight of the researcher or engineer making a general-purpose speech synthesiser, who wants to achieve results that are as broadly applicable as possible. Typical speech synthesis systems are in a sense designed for any task, and thus, paradoxically, for no task at all. While more meaningful tests are conceivable, they generally require a prohibitive amount of time and resources.

Until improved benchmarks and better objective measures arrive, differences in generic measures such as naturalness ratings remain our best indicators of synthesis adequacy. However: to get the answers we seek, and to convince fellow researchers and practitioners of their validity, we have to get better at asking the right questions, in the right way, to a good set of listeners. In other words, we need to pay attention to the points in Section 2. Only when correct thoughts and design go into our experiments will the correct answers be sure to emerge.

Acknowledgements This work was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

6. References

[1] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Computer Speech and Language*, vol. 21, no. 1, pp. 153–173, 2007.

[2] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[3] Z.-H. Ling and L.-R. Dai, "Minimum Kullback-Leibler divergence parameter generation for HMM-based speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 5, pp. 1492–1502, 2012.

[4] S. Möller and T. H. Falk, "Quality prediction for synthesized speech: Comparison of approaches," in *International Conference on Acoustics*, 2009, pp. 1168–1171.

[5] T. H. Falk, S. Möller, V. Karaiskos, and S. King, "Improving instrumental quality prediction performance for the Blizzard Challenge," in *Proceedings of the Blizzard Challenge Workshop*, 2008.

[6] C. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, "Towards perceptual quality modeling of synthesized audiobooks," in *Proceedings of the Blizzard Challenge Workshop*, 2012.

[7] K. Morton, "Expectations for assessment techniques applied to speech synthesis," *Proceedings of the Institute of Acoustics*, vol. 13, no. 2, 1991.

[8] D. B. Pisoni, H. C. Nusbaum, and B. G. Greene, "Perception of synthetic speech generated by rule," *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1665–1676, 1985.

[9] *A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices*, ITU Recommendation ITU-T P.85, International Telecommunication Union Telecommunication Standardization Sector, Geneva, Switzerland, June 1994.

[10] M. D. Polkosky and J. R. Lewis, "Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X," *International Journal of Speech Technology*, vol. 6, no. 2, pp. 161–182, 2003.

[11] N. Campbell, *Evaluation of Text and Speech Systems*. Springer, 2007, no. 2, ch. Evaluation of Speech Synthesis: From Reading Machines to Talking Machines.

[12] S. King and V. Karaiskos, "The Blizzard Challenge 2013," in *Proceedings of the Blizzard Challenge Workshop*, 2013.

[13] R. Dall, J. Yamagishi, and S. King, "Rating naturalness in speech synthesis: The effect of style and expectation," in *Proceedings of the 7th International Conference on Speech Prosody*, 2014, pp. 1012–1016.

[14] M. L. García Lecumberri, M. Cooke, and A. Cutler, "Non-native speech perception in adverse conditions: A review," *Speech Communication*, vol. 52, no. 11, pp. 864–886, 2010.

[15] M. Reynolds, Z. S. Bond, and D. Fucci, "Synthetic speech intelligibility: Comparison of native and non-native speakers of English," *Augmentative and Alternative Communication*, vol. 12, no. 1, pp. 32–36, 1996.

[16] C. Watson, W. Liu, and B. MacDonald, "The effect of age and native speaker status on intelligibility," *Proceedings of the 8th ISCA Speech Synthesis Workshop (SSW8)*, 2013.

[17] S. Gordon-Salant and P. J. Fitzgibbons, "Selected cognitive factors and speech recognition performance among young and elderly listeners," *Journal of Speech, Language and Hearing Research*, vol. 40, no. 2, pp. 423–431, 1997.

[18] J. Latorre, K. Yanagisawa, V. Wan, B. Kolluru, and M. J. Gales, "Speech intonation for TTS: Study on evaluation methodology," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech)*, 2014, pp. 2957–2961.

[19] D. B. Pisoni, "Perception of synthetic speech," in *Progress in Speech Synthesis*. Springer, 1997, pp. 541–560.

[20] M. K. Wolters, K. B. Isaac, and S. Renals, "Evaluating speech synthesis intelligibility using Amazon Mechanical Turk," in *Proceedings of the 7th ISCA Speech Synthesis Workshop (SSW7)*, 2010, pp. 136–141.

[21] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "Crowdmos: An approach for crowdsourcing mean opinion score studies," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 2416–2419.

[22] S. Buchholz, J. Latorre, and K. Yanagisawa, "Crowdsourced assessment of speech synthesis," in *Crowdsourcing for Speech Processing*. John Wiley & Sons, 2013, pp. 173–216.

[23] R. A. J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proceedings of the Blizzard Challenge Workshop*, 2007.

[24] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1–2, pp. 81–93, 1938.

[25] *Method for the subjective assessment of intermediate quality level of coding systems*, ITU Recommendation ITU-R BS.1534-1, International Telecommunication Union Radiocommunication Assembly, Geneva, Switzerland, March 2003.

[26] M. S. Ribeiro, J. Yamagishi, and R. A. J. Clark, "A perceptual investigation of wavelet-based decomposition of $f0$ for text-to-speech synthesis," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech)*, 2015.