Interspeech 2024
1-5 September 2024, Kos, Greece

Experimental evaluation of MOS, AB and BWS listening test designs

Dan Wells[1*], *Andrea Lorena Aldana Blanco*[1*], *Cassia Valentini-Botinhao*[1], *Erica Cooper*[2], *Aidan Pine*[3], *Junichi Yamagishi*[2], *Korin Richmond*[1]

[1]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[2]National Institute of Informatics, Japan [3]National Research Council Canada, Canada

{dan.wells, lorena.aldana, cvbotinh, korin.richmond}@ed.ac.uk,
{ecooper, jyamagis}@nii.ac.jp, aidan.pine@nrc-cnrc.gc.ca

Abstract

Mean Opinion Score (MOS) tests are the most widely used test type for subjective evaluation of speech samples. However, their use has been questioned, as results can vary significantly depending on the test material included. Forced-choice tests such as AB or Best Worst Scaling (BWS) can in principle mitigate some of these issues. Our aim here is to compare MOS, AB and BWS tests in 3 regards: 1) Which test type do listeners prefer in terms of ease, engagement and overall likeability? 2) How fast are listeners at each test type? 3) Does each test type provide the same pattern of results? To answer these questions we re-use a subset of stimuli from the Blizzard Challenge 2013 and conduct new MOS, AB and BWS tests. Overall, we conclude each test type is broadly equally valid, MOS may not in fact be the fastest or easiest test type for listeners, but the theoretical advantages of BWS are counterbalanced by it seeming less liked by our listeners here.

Index Terms: speech synthesis, preference evaluation, listening test, MOS, BWS, AB

1. Introduction

Text-to-speech (TTS) systems are most frequently evaluated through comparison of Mean Opinion Scores (MOS), i.e., average subjective ratings by listening test participants of the perceived 'quality' or 'naturalness' of synthesised stimuli from a given system. In this evaluation paradigm, audio samples from individual systems are typically presented one at a time with a Likert-type scale ranging from 1–5, with labels such as 'bad' and 'good' for audio quality or 'very unnatural' and 'somewhat natural'. Though MOS tests are widely used, researchers have pointed out several drawbacks and limitations [1, 2, 3, 4, 5]. While it is possible to aggregate such ratings for a particular system by averaging their numerical indices (hence *Mean* Opinion Score), it is important to note that they represent ordinal rather than interval data: while the ordering of responses on such a scale is meaningful, we cannot make any guarantees about what each point on the scale actually means in terms of speech 'naturalness' or 'quality', nor about the intervals between points being equal in any way. While it is therefore not statistically meaningful to compare average MOS scores, there are still appropriate statistical methods for analysing MOS ratings, such as the Wilcoxon signed-rank test [6]. Additional problems arise from listener-dependent biases with regard to where on the scale one might place a particular sample [2], how much of the scale one tends to use across samples, the granularity of the scale [5], the instructions given to participants [5] and how sample scores might vary depending on other stimuli involved in the test [7].

These problems may be avoided by using forced-choice comparison test designs instead, for example presenting pairs of audio samples from two different systems and asking participants to select which sounds more natural; we refer to these as AB or preference tests. Ratings from such tests tend to exhibit much less variance than MOS tests, making it easier to detect statistically significant differences between systems where they exist [1, 3]. However, for N systems to be compared against each other, there are $N(N-1)$ possible ordered pairs of systems (or half as many unordered pairs), which rapidly inflates the number of comparisons required in a listening test after factoring in multiple test utterances per system pair as well.

Of concern for all researchers is the resultant cost of carrying out a listening test. Recent recommendations for listening test design in [8] suggest recruiting at least 30 paid participants, and aiming for as much text coverage as possible with a test around 30 minutes in duration. Additionally, acquiring enough ratings to uncover any statistically significant differences between systems may be more difficult in contexts with few native speakers available for evaluation, as is the case with many endangered languages [9].

Another possibility for a forced-choice design is Best-Worst Scaling (BWS). This design has previously been shown to provide more consistent ratings for smaller numbers of annotations than Likert-type ratings in sentiment analysis [10] (comparing $2N$ ratings for N tuples vs. $5N$ ratings for N Likert questions) and evaluation of automatic summarisation models [11]. Applying this to TTS evaluation, we present sets of four audio samples to listeners at a time, representing the same text synthesised by different systems, and ask them to select the most and least natural. Listening to four samples and making two explicit choices of best and worst yields the equivalent of 5/6 possible pairwise comparisons between the systems involved, where the best selected system is known to beat the other three, and the two unselected systems to beat the worst; only the relationship between the two unselected systems remains unknown. This test design has previously been applied to TTS evaluation in [12], where BWS judgements were found to correlate with MOS scores in a small-scale listening test for Scottish Gaelic, where limited access to native speakers made participant recruitment difficult.

With all these considerations in mind, in this study we explicitly compare these three listening test designs along three main axes. We asked participants to take part in a multi-paradigm listening test, with short MOS, AB and BWS sections presented sequentially in random order. Then, we asked participants of all three tests which one they preferred, which kept them most engaged, and which they found easiest to make judgements in. These preferences relate to considerations of data collection efficiency and stability of ratings insofar as we

[*]Equal contribution

10.21437/Interspeech.2024-1778

can assume that more engaged participants provide more consistent and considered responses, as they may be more focused on the task at hand. We evaluate this by checking the consistency of rankings in subsets of gathered responses, and by comparing discovered differences between systems to a much larger listening test previously conducted using the same test stimuli. We also measure how long participants spend in each test. Taken together, we hope these insights could help determine which test design offers the best balance between data collection efficiency and robustness of results.

2. Dataset

The 2013 edition of the Blizzard Challenge included two tasks for English TTS, with professionally-recorded audiobook corpora released for each task [13]. We follow previous work reusing data from task 2013-EH2 [4, 7], comprising 19 hours of WAV audio segmented into individual sentences with corresponding text transcripts; all recordings were made by a single professional female voice actor. The evaluation for this task included sentences from novels and newspaper extracts, but natural speech recordings were only released for the sentences from novels, so we use only these in our own listening tests. While [4] follows the original Blizzard Challenge protocol in only presenting a small subset of evaluation utterances (one per system under test), we use the full set of 100 synthesised test utterances released by the Blizzard Challenge organisers. Specifically, we use the files prepared by [4] with a consistent sampling rate of 16 kHz and sv56 amplitude normalisation to -26 dB [14].

Our test stimuli come from systems submitted to the EH2 task of the Blizzard Challenge 2013 [13] and the four modern neural systems trained in [4]. We replicate the system selection in [4, 7], taking unit selection (system N), HMM (C) and hybrid synthesis (K) systems representative of the state of the art in 2013. To these we add another hybrid synthesis system (M) which rated similarly to system K in the original Blizzard Challenge 2013-EH2 results, and another unit selection system (B) which was the Festival-based benchmark that year, and rated significantly worse than system N [7]. We also include neural systems trained in [4], namely Tacotron (TC) [15] and FastPitch (FP) [16] models, each paired with either a WaveNet (WN) [17] or Parallel WaveGAN (WG) [18] vocoder. This selection gives both a wide range of overall system quality and sub-groups of systems which are expected to be relatively close in quality, i.e., the four neural systems and some of the older systems. With this, we hope to test the ability of different listening test designs to discern small differences between similar systems.

3. Listening test design

The number of voices included here (9 TTS systems plus natural speech) would normally require a large listening test to ensure full coverage over system comparisons and test sentences to be synthesised. For MOS stimuli, following the Blizzard Challenge test protocol would need a 9×10 Latin square, with 9 blocks of 90 utterances to which individual listeners may be assigned. For an AB test, there are 45 unique pairs of systems to consider (or 90 if we account for presentation order), which must also be distributed over a set of test sentences. For BWS, there are $\binom{10}{4} = 210$ unique 4-tuples when selecting from 10 voices. However, given that our main goal in this study is to investigate participant preferences over listening test designs, rather than to determine the quality of the TTS systems involved, we instead came up with reduced listening test designs allowing each test to be completed in under 10 minutes, for a total of 30 minutes per participant. Nonetheless, we also attempt to balance our reduced test designs within these time constraints to compare the efficiency of different listening test paradigms in terms of how many ratings are required to determine statistically significant differences between systems. For each design, from the available 100 Blizzard test utterances, we assign a sample of 30 non-overlapping utterances. The remaining 10 utterances are used across all tests.

3.1. Online testing and response validation

Recruiting participants and administering listening tests online is a common practice in TTS evaluation, for example in cases where a large number of listeners is required [19], or where researchers are not in the same location as speakers of the target language. In our case, we recruited 30 native US English speakers with self-reported normal hearing through the crowdsourcing platform Prolific Academic. Although this removes much of the control over listening environment and listener attention possible in a lab environment, online results have been found consistent with in-person tests for evaluating the intelligibility of TTS systems [20]. Moreover, we aim here to provide recommendations that fit common practice, including addressing some of the challenges of online listening tests; we refer the reader to [21] for further recommendations. Validating participant responses and excluding any listeners who do not appear to be taking the test properly is a large part of that, and we take several measures to achieve this.

First, we include attention-check questions with speech samples that contain an explicit instruction synthesised by the FastPitch-WaveNet voice. There are two such questions in the MOS test with the instruction "please give this sample a score of one", and one in the AB and BWS tests, instructing participants that the relevant sample should be selected as the best system. We also present the same two stimuli at the beginning and end of the MOS and AB tests, and one in the BWS test, to check the consistency of participant responses – we expect that they should provide identical scores or system rankings given identical stimuli. Finally, we expect that any natural speech samples should always be rated as highly as possible across test designs, so we also check for any deviation on those stimuli.

Participants are required to listen to the full duration of every audio sample at least once and provide a rating before they are able to progress through the listening test. Only one audio sample can play at a time, preventing participants from starting playback on all samples in an AB or BWS question at once in an attempt to speed up the test. We also record the time spent on each question, the number of times each audio sample is played through fully and the number of clicks on the question page. These measurements are potentially useful to detect disengaged participants who sit idle on a particular question page, or more deliberate listeners who review audio samples multiple times, perhaps adjusting their ratings as they do so.

3.2. MOS

We assign all 40 MOS test utterances to a single block to be taken by all participants. Systems and utterances are presented in a fixed random order, with 4 utterances randomly assigned to each system. Additionally, we include attention check questions $1/4$ and $3/4$ of the way through the test, and repeat two stimuli (i.e. one particular system synthesising one particular utterance) at the beginning and end of the test to check for consistency of ratings. This gives 44 MOS questions in total.

3.3. AB

We create two blocks of 20 test utterances to assign to AB test stimuli. Each participant gets assigned one block. Since participants must listen to two audio samples at a time compared to one in a MOS test, we expect half as many AB questions to take roughly the same amount of time to complete. Because each test includes fewer than the total possible number of pairs of systems under test, we used a feature-based submodular optimisation approach [22, 23] to select two subsets which are evenly balanced in terms of how many times each individual system appears across selected pairs. In this way, $40/45$ possible pairs of systems are represented across the two test blocks. Again, we included two repeated pairs at the beginning and end of the test, plus one explicit attention check halfway through, for a total of 23 AB questions per block.

3.4. BWS

For the BWS test, there is a much larger set of possible combinations of systems to cover, with 210 possible 4-tuples from 10 voices. Presenting 10 BWS questions to each participant to match the time taken for the equivalent 40 MOS questions would require 21 distinct question blocks to be presented across participants, making the test setup quite cumbersome. Instead, we use the minimal set of 9 4-tuples which between them include comparisons of every pair of systems at least once [24]. In this set, three pairs of systems appear together in 4/9 tuples, while the remaining four systems are presented in three tuples each, appearing with each other system only once; the repeated pairs also appear with each other system only once. The repeated pairs provide an opportunity to gather more comparisons for some pairs than others, which we exploit by assigning systems we expect to be relatively close in quality, namely the two pairs of neural systems using either WaveNet or Parallel WaveGAN vocoders, and the two hybrid synthesis systems K and M. As in the other test designs, we repeat one tuple at the beginning and the end, and also include one explicit attention check question in the middle of the test, for a total of 11 BWS questions.

3.5. Participant questionnaire

After participants had completed all three listening tests, we asked them to rank the different test designs with respect to the following questions:

- If you had to take part in one of these listening tests again, which would you prefer?
- Which test design kept you most engaged until the end?
- In which test design did you find it easiest to make decisions about the quality of the speech samples?

We also asked participants which, if any, of the listening test designs they had prior experience with, in case their preferences might be influenced by familiarity with one test over another, and provided an open text box for any additional comments.

4. Results

We excluded 2 participants who failed attention check questions and continued data collection until we had 30 participants in total. After initial results analysis, we additionally excluded 3 participants who were not self-consistent across different parts of the test (Section 4.2), leaving 27 for further analysis.

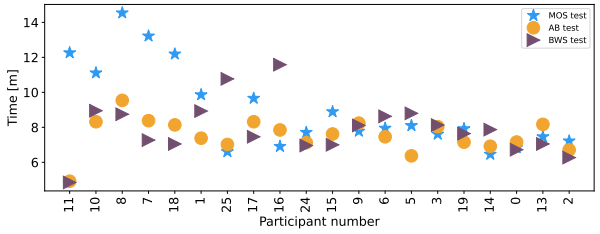

Figure 1: *Average duration per participant in each test. Results are sorted from longest (left) to shortest (right) test duration.*

4.1. Test duration

On average, participants took $31.9\,\text{min} \pm 6.1\,\text{to}$ complete the listening test. For the test duraton analysis only, we excluded data from another 7 participants who spent more than one standard deviation above the mean time in each test type and the overall test. Participants spent $9.0\,\text{min} \pm 2.4$ going through the MOS test, $7.5\,\text{min} \pm 1.0$ for AB and $7.9\,\text{min} \pm 1.5$ for BWS. A paired t-test showed that the MOS test took significantly longer than the AB test $t\,(23) = 2.75, P = 0.02, P < 0.05$. No significant differences were found among the other test comparisons. Figure 1 shows the total duration per participant in each test. Results are sorted from the longest to shortest duration in the whole listening test. Interestingly, we observe a trend whereby listeners who are slower in general tend to be particularly slower on the MOS test versus the other two types (i.e. 8 participants on the left half of the plot).

4.2. Analysis of repeated stimuli

To determine how consistent responses were from participants, we first analyse results of the repeated stimuli explained in Section 3.1. We noticed that the responses from a number of participants were not consistent across those questions. We found inconsistent responses to repeated stimuli in all tests. Specifically, for 50% of participants in the BWS and for 27% of participants in the AB test. However, there was only one repeated stimuli question in BWS and two questions in the AB.

Intra-annotator inconsistencies have been shown to increase as a function of the time difference between two annotations [10]. To gain further insight about the reliability of responses we computed a ranking of systems per participant in each test. We then computed a Spearman's rank correlation between the listener-specific rankings and the final ranking derived from combining all responses. This can be viewed as similar to, though much simpler than, previously reported approaches to estimating participant reliability, such as [25]. Based on intra-annotator correlation values we excluded responses from 3 participants whose correlation value was below the mean minus one standard deviation in two or more tests.

4.3. MOS, AB and BWS results

MOS test scores are presented in Figure 2 and significant differences are shown in Figure 3. AB and BWS rankings are computed using a Plackett-Luce model, which is a probabilistic model commonly used in fields such as psychology or marketing to derive the underlying ranking of a set of items based on observed preference or choice data [26]. A higher worth value indicates preference for a system, that is, that a system with higher worth is ranked higher when compared to the other systems. Rankings computed from (a) AB and (b) BWS scores are

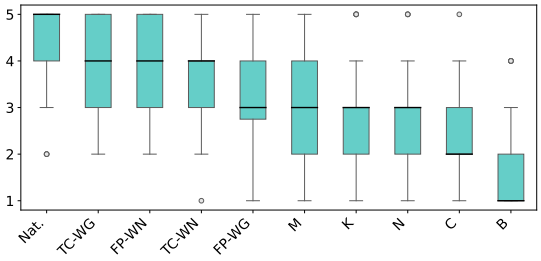

Figure 2: *MOS test results ordered by mean value. The median score is represented by a solid line across each box.*

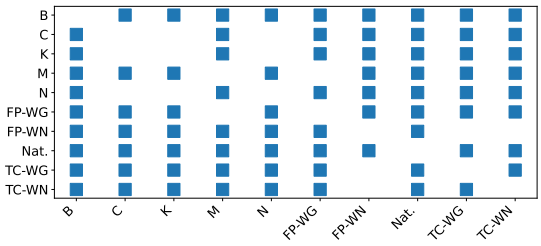

Figure 3: *MOS test significance results of a pairwise Wilcoxon signed-rank test. Solid boxes indicate a significant difference.*

shown in Figure 4. Significant differences are represented by non-overlapping error bars. Error bars indicate 95 % CI.

In all test types, natural speech obtained the highest ranking, followed by neural systems, then hybrid and unit selection systems and lastly system B. The rankings resemble the same trends found in [4]. Nonetheless, the ranking of neural systems varies across the three test types. This variation in rankings can also be observed between hybrid (K, M) and unit selection (N) systems. Specifically, K was found to be significantly more natural than N in the original Blizzard Challenge 2013 evaluation and by our BWS test, but not in either MOS or AB test. Differences between natural speech and neural systems, and systems B and C were also significant in the MOS and BWS tests, but not in the AB. These differences were also significant in [4].

4.4. Consistency of annotations

To gain further insight about the consistency of responses we calculated average split-half reliability (SHR) across 50 trials. We computed a Spearman's rank correlation among the rankings obtained after splitting the set of total responses in halves, as in [10]. Average split-half reliability ρ across 50 trials was 0.98 for MOS, 0.95 for AB and 0.96 for BWS.

4.5. Listener preference for test type

Results to the preference questions presented in Section 3.5 are shown in Figure 5. We computed rankings from a Plackett-Luce model to determine significant differences. Participants found the AB test significantly easier than the BWS test. They also preferred the AB test over MOS or BWS, however the difference in preference was not significant. Some of the comments added in the open comment box about the BWS test indicated that memory recall played a role, which made the test longer. Specifically, one participant said *"the BWS test takes much longer to do properly, as the earliest examples get partially forgotten."* Another participant mentioned that they view

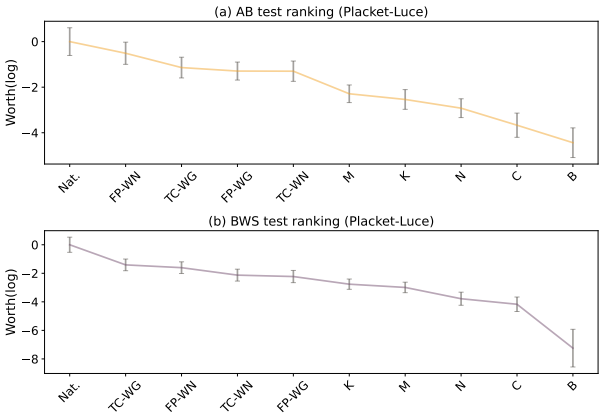

Figure 4: *(a) AB and (b) BWS rankings from a Plackett-Luce model. Error bars are the 95% confidence intervals.*

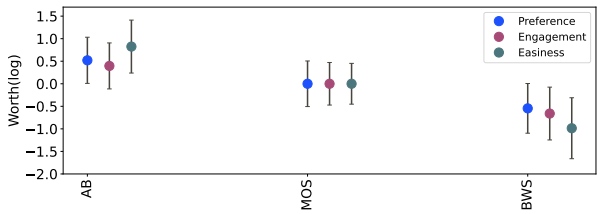

Figure 5: *Test design preference questionnaire rankings from a Plackett-Luce model.*

BWS as *"basically the same as A vs B, but with additional unnecessary complexity"*. These comments highlight one difference between the BWS task applied to speech vs. text modalities (as in the sentiment analysis task of [10]), namely that text options can be viewed in parallel at all times during the test, whereas participants must rely on their own working memory when comparing multiple speech samples in a single question.

4.6. Conclusions

This paper aimed to critically evaluate and compare the MOS, AB and BWS test designs in three main respects: 1) listeners' subjective experience when performing the tests; 2) listeners' relative speed at completing each test type; 3) test reliability, in terms of outcomes and conclusions drawn from the data gathered in each test. Overall, we found all tests give broadly similar outcomes, suggesting they can be viewed as equally valid. Though MOS is most commonly used, we found it is not necessarily seen as easiest or most engaging, and over one third of our participants were greatly slower at MOS than the other two test types. Though BWS has benefits in principle (i.e. finding the same significant differences between systems as a much larger previous equivalent test), our results indicate listeners may find it more taxing. There are caveats to the work here: each test was of neccessity shorter (≈8 min per listener) than we would normally aim for when conducting a single test type (≈30 min), limiting conclusions about the potential for each test type to find significant differences between systems and so provide insights; people might also become more accustomed to BWS with more questions, thus changing their mind from an initial reaction based on a relatively small number of questions. Addressing such caveats requires longer tests to be done, which is the subject of our ongoing work.

5. Acknowledgements

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1), the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences, and the National Research Council of Canada's Ideation Fund: 'Small Teams – Big Ideas'.

6. References

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

[2] S. Zieliński, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests-a review," *Journal of the Audio Engineering Society*, vol. 56, no. 6, pp. 427–451, 2008.

[3] S. Shirali-Shahreza and G. Penn, "MOS Naturalness and the Quest for Human-Like Speech," in *SLT Workshop*, 2018, pp. 346–352.

[4] S. Le Maguer, S. King, and N. Harte, "Back to the Future: Extending the Blizzard Challenge 2013," in *Interspeech 2022*. ISCA, 2022, pp. 2378–2382.

[5] A. Kirkland, S. Mehta, H. Lameris, G. E. Henter, E. Szekely, and J. Gustafson, "Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation," in *SSW*, 2023.

[6] R. A. J. Clark, M. Podsiadło, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *The Blizzard Challenge 2007*, 2007, pp. 1–6.

[7] S. Le Maguer, S. King, and N. Harte, "The limits of the Mean Opinion Score for speech synthesis evaluation," *Computer Speech & Language*, vol. 84, p. 101577, 2024.

[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 *Interspeech 2015*. ISCA, 2015, pp. 3476–3480.

[9] A. Pine, D. Wells, N. Brinklow, P. Littell, and K. Richmond, "Requirements and Motivations of Low-Resource Speech Synthesis for Language Revitalization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2022, pp. 7346–7359.

[10] S. Kiritchenko and S. Mohammad, "Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2017, pp. 465–470.

[11] J. Steen and K. Markert, "How to Evaluate a Summarizer: Study Design and Statistical Analysis for Manual Linguistic Quality Evaluation," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Association for Computational Linguistics, 2021, pp. 1861–1875.

[12] D. Wells, K. Richmond, and W. Lamb, "A Low-Resource Pipeline for Text-to-Speech from Found Data With Application to Scottish Gaelic," in *Interspeech*, 2023, pp. 4324–4328.

[13] S. King and V. Karaiskos, "The Blizzard Challenge 2013," in *The Blizzard Challenge Workshop*, 2013. [Online]. Available: http://www.festvox.org/blizzard/bc2013/summary_Blizzard2013.pdf

[14] International Telecommunication Union, "Recommendation ITU-T G.191: Software tools for speech and audio coding standardization," International Telecommunication Union, Tech. Rep., 2005.

[15] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," in *Interspeech 2017*. ISCA, 2017, pp. 4006–4010.

[16] A. Łańcucki, "Fastpitch: Parallel Text-to-Speech with Pitch Prediction," in *ICASSP*, 2021, pp. 6588–6592.

[17] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *CoRR*, vol. abs/1609.03499, 2016, arXiv: 1609.03499. [Online]. Available: http://arxiv.org/abs/1609.03499

[18] R. Yamamoto, E. Song, and J. Kim, "Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," in *ICASSP*, 2020, pp. 6199–6203.

[19] A. W. Black and K. Tokuda, "The blizzard challenge - 2005: evaluating corpus-based speech synthesis on common datasets," in *Interspeech 2005*. ISCA, Sep. 2005, pp. 77–80.

[20] M. K. Wolters, K. B. Isaac, and S. Renals, "Evaluating Speech Synthesis Intelligibility using Amazon Mechanical Turk," in *Proc. 7th ISCA Workshop on Speech Synthesis (SSW 7)*, 2010, pp. 136–141.

[21] International Telecommunication Union, "Recommendation ITU-T P.808: Subjective evaluation of speech quality with a crowdsourcing approach," International Telecommunication Union, Tech. Rep. ITU-T P.808, 2021.

[22] K. Wei, Y. Liu, K. Kirchhoff, C. Bartels, and J. Bilmes, "Submodular subset selection for large-scale speech training data," in *ICASSP*, 2014, pp. 3311–3315.

[23] J. Schreiber, J. Bilmes, and W. S. Noble, "apricot: Submodular selection for data summarization in Python," *Journal of Machine Learning Research*, vol. 21, no. 161, pp. 1–6, 2020.

[24] W. H. Mills, "On the covering of pairs by quadruples. II," *Journal of Combinatorial Theory, Series A*, vol. 15, no. 2, pp. 138–166, Sep. 1973.

[25] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz, "Pairwise ranking aggregation in a crowdsourced setting," in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 193–202.

[26] H. L. Turner, J. van Etten, D. Firth, and I. Kosmidis, "Modelling rankings in R: the PlackettLuce package," *Computational Statistics*, vol. 35, no. 3, pp. 1027–1057, 2020.