Confidence Intervals for ASR-based TTS Evaluation

Citation for published version:
Taylor, J & Richmond, K 2021, Confidence Intervals for ASR-based TTS Evaluation. in *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021.* Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, International Speech Communication Association, pp. 2791-2795, Interspeech 2021, Brno, Czech Republic, 30/08/21. https://doi.org/10.21437/Interspeech.2021-2203

Digital Object Identifier (DOI):
10.21437/Interspeech.2021-2203

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021

Confidence Intervals for ASR-based TTS Evaluation

Jason Taylor[1], *Korin Richmond*[2]

[1] The Centre for Speech Technology Research, The University of Edinburgh

jason.taylor@ed.ac.uk , korin@cstr.com

Abstract

Automatic speech recognition (ASR) is increasingly used to evaluate the intelligibility of text-to-speech synthesis (TTS). ASR is less costly than traditional listening tests, but questions remain about its reliability. We re-evaluate the Blizzard Challenge's intelligibility tasks in English since 2011 using ASR. Re-analysing transcriptions collected by paid in-lab participants, online volunteers and Amazon Mechanical Turkers (the latter used only in 2011), we compare their word error rates (WERs) and statistically-significant system-groupings with those generated by an open-source, Transformer-based ASR model. This ASR model consistently decodes test stimuli with more reliable WERs than the Blizzard Challenge's (mostly non-native) speech experts and online volunteers. The model also groups systems according to statistical significance similarly to the paid in-lab participants. Using surplus semantically unpredictable sentences (SUS) submitted every year to the challenge, we investigate how confidence intervals in ASR WERs change as the number of transcribed stimuli increases. We plot the Frobenius norm of pairwise significance matrices with increasing stimuli. We find that finer groupings of systems are detected as confidence intervals narrow. The number of stimuli where p-values start to converge ranges from 400-800 stimuli. We conclude that, with enough stimuli, ASR can be more reliable than humans.

Index Terms: Text-to-Speech, Objective Evaluation, Automatic Speech Recognition, Statistical Analysis

1. Introduction

The development of objective evaluation metrics is crucial to the field of text-to-speech synthesis (TTS). Traditional listening tests conducted under controlled conditions are expensive, and the data collected may require extensive quality control [1, 2]. The drive for simpler and less expensive means for evaluation have resulted in use of metrics such as PESQ [3], MCD [4] and ViSQOL [5]. Recent work has also focused on the prediction of Mean Opinion Scores (MOS) for TTS [6, 7, 8] and voice conversion [9, 10] systems using neural networks.

While MOS measures speech naturalness, another key factor in the perception of system quality is intelligibility [11]. Some of the previous work on objective intelligibility measurement has focused on speech in noise to evaluate speech enhancement algorithms [12]. This was the subject of the Hurricane Challenge [13]. Recent progress in Automatic Speech Recognition (ASR) has enabled the use of ASR transcription as a more interpretable metric for intelligibility. A phone-based ASR system outperformed other objective intelligibility measures for evaluating speech enhancement in [14].

The use of large, open vocabulary continuous speech recognition (LVCSR) to substitute human listening evaluations is a recent innovation. For instance, an open source LVCSR system available from [15] was also used to evaluate TTS intelligibility in [16]. Previously, only closed vocabulary ASR had been used for transcription tasks, as in [17]. Recently, ASR has also been used for other tasks in TTS such as the automatic selection of "clean" training utterances and speakers [18], and for transcription of training recordings in [19].

Little work has so far sought to establish the reliability of ASR for measuring TTS intelligibility. [20] found strong correlations between human word error rate (WER) collected from Amazon Mechanical Turk (MTurk) to the WER of 3 different ASR systems (IBM Watson, Google API and wit.ai). [21] also found correlations between MTurk, these ASR systems and MCD when building DNN-based TTS voices in Merlin. However, it remains unknown whether explicit ASR-derived rankings of multiple TTS systems correlate with those derived from paid, in-lab human transcribers.

The Blizzard Challenge [22] provides evaluation data for the development of objective metrics [23, 24, 25, 26, 27]. This annual challenge conducts human transcription evaluation with semantically unpredictable sentences (SUS). We use this resource here to compare WERs computed using in-lab and online human transcriptions with objective ASR transcriptions. Specifically, we compare rankings of systems submitted to the Blizzard Challenge in 2011, 2012, 2013, 2016, 2017 and 2018.

We find ASR gives similar transcription WERs and statistically significant pairings of systems as the paid human listeners. ASR is more reliable than the Online Volunteers and Speech Experts (the majority of whom are non-native English speakers) used in the Blizzard Challenge.

A key advantage of using ASR is that more stimuli may be transcribed without the considerable cost of recruiting human listeners. Obtaining human transcriptions for a large number of stimuli is expensive and [28] has shown participant retention drops after 20 stimuli, as participants struggle to stay engaged in speech quality evaluations. We analyse the confidence intervals relating to the WERs obtained as the number of stimuli increases using a bootstrap method [29]. With more stimuli, we find confidence intervals narrow and p-values decrease between systems.

2. Methods

2.1. Research Questions

Our following research questions aim to determine the extent to which ASR provides a reliable intelligibility metric for TTS:

1. How does ASR compare to paid listeners when transcribing synthetic speech in the Blizzard Challenge?
2. How do the confidence intervals over ASR WERs change as the number of TTS test stimuli is increased?
3. Do ASR transcriptions identify the same significant differences between system pairs as the paid listeners?
4. Are there any benefits to increasing the number of stimuli for ASR transcription?

2.2. ASR Model

We use a pretrained LibriSpeech Transformer model available from EspNet [15]. This has the advantages of being open-sourced, accessible and trained end-to-end (E2E) on a large (1,000 hours [30]) multi-speaker corpus. It performs with a WER of 4.9% on the LibriSpeech clean test set. As an E2E model, it does have the disadvantage though that extracting recognised phone strings to measure phone error rate (PER) is not possible, which may otherwise potentially offer insights into the reliability of ASR for TTS intelligibility. For example, [31] found ASR PER to be a superior means of TTS model selection than common loss functions.

2.3. Data

2.3.1. Blizzard Challenge

The Blizzard Challenge is an annual event where participants are provided with a speech dataset for voice building and are asked to submit a defined set of synthetic samples for evaluation. The focus of the challenge changes from year to year; for example, samples were evaluated at varying noise levels in 2010, while the challenge was focused on Mandarin TTS in 2019. Each year, a large-scale human listening evaluation is conducted. See the Blizzard Challenge summary papers [32, 33, 34, 35, 36, 37] for more detail. We exclude years 2014, 2015, 2019 and 2020 as these used languages other than English.

2.3.2. Listener Types

Each year a section of the evaluation focuses on measuring the intelligibility of submitted systems. Paid listeners are recruited who type-in transcriptions in purpose-built sound booths under controlled conditions. These listeners are known as EP or EE depending on the year of the challenge. Participating teams also recruit their own speech experts and online volunteers to conduct an evaluation. Known as ES and ER respectively, these are mainly composed of non-native speakers of English. In 2011, Amazon Mechanical Turk (AMT) was also used for evaluation.

2.3.3. TTS Test Stimuli and Systems

Each year a new test set of SUS stimuli is submitted as well as the test sets of the previous two years. For each challenge we analyse here there were 3 test sets used in the ASR *maximum stimuli* sets (henceforth *Extra ASR*): 2011 (700 stimuli), 2012 (800), 2013 (900), 2016 (600), 2017 (600), 2018 (600). We include data from two challenges in 2013 (EH1 and EH2). The test stimuli were created using a SUS generator and do not appear in the LibriSpeech training dataset for the ASR system we use for evaluation.

Systems are randomly allocated a different anonymized letter each year. Some systems did not submit the 3 SUS test sets in a given year (such as system N in 2017) and we exclude those systems from analysis. System A is always natural speech but since recordings of the SUS sentences do not exist we do not include them in our analysis. Years 2017 and 2018 included systems based on neural text encoders and WaveNet-based vocoders, with earlier years including previous Unit Selection and SPSS-based TTS. We computed statistics using the Scikit-learn Python package.

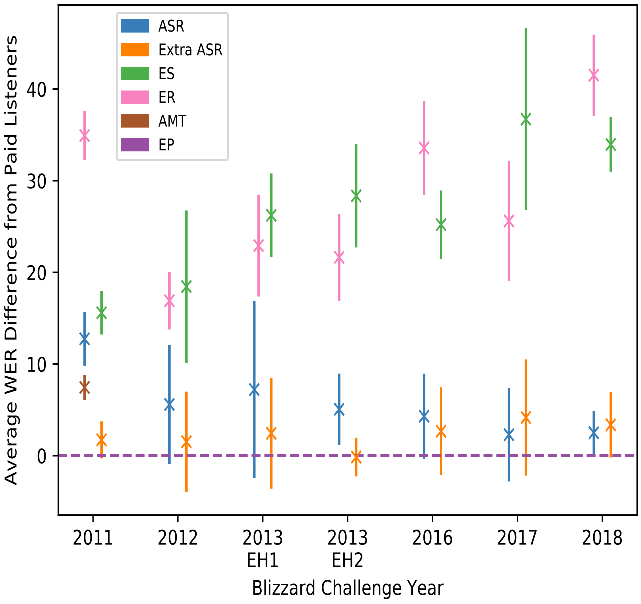


Figure 1: *Aggregate Difference in WER from Paid Participants. Each bar shows the mean and 2 standard deviations for each listener type: Speech Experts (ES), Online Volunteers (ER), Amazon Mechanical Turkers (MTurk), the same stimuli ranked by ASR (ASR) and the maximum number of test SUS synthesised each year (Extra ASR)*

2.4. Results

2.4.1. WER by listener type

Figure 1 shows the differences in WER from the paid listeners. For each year analysed, the cross represents the mean difference in WER and the bars 2 standard deviations. Each listener type is denoted by colour and the scores are offset around a year label to aid visualisation. The *ASR* bars in blue are the same stimuli as transcribed in the formal human evaluations, ranging between 25-40 stimuli depending on the year. As noted above, the *Extra ASR* bars in orange correspond to 3 SUS test sets (600-900 stimuli per year)

ASR performance is close to the paid listener WERs for every year except 2011, where the MTurk participants achieved lower WERs. ASR gives consistently lower WERs than the speech experts and online volunteers. The latter groups have high WERs as their evaluations are conducted more informally than for the paid listeners and non-natives are consistently above 60% of listeners each year. Similar WER averages and spreads are achieved by the ASR and *Extra ASR* sets. This was as expected since the genre of text was similar. The Extra ASR bar for 2013 EH2 outperformed the paid listeners.

2.4.2. Bootstrapping ASR Confidence Intervals

We wanted to know how statistically valid ASR transcriptions were to rank TTS systems. During our analysis, we noticed ASR WERs fluctuate considerably across the SUS stimuli. Measuring confidence in the WER metric is thus important. [29] aims to make the WER metric more reliable by bootstrapping [38] the WER of a system. A bootstrap of ASR WER involves sampling WERs from a bag of stimuli to remove any possible effects of ordering.

In our approach, the bootstrap was run successively in steps

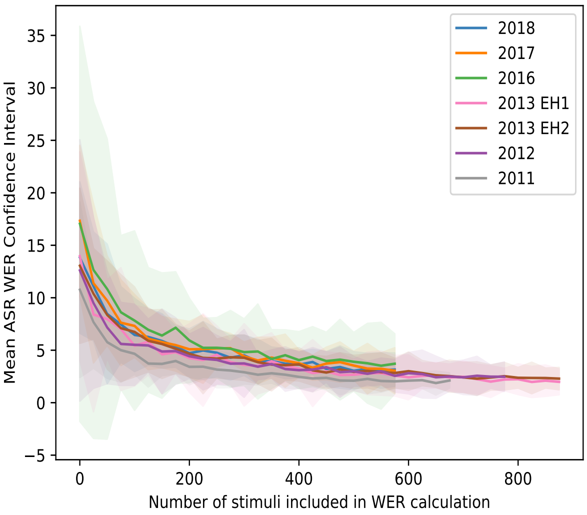


Figure 2: *Bootstrapped WER confidence interval averaged across ASR stimuli. The bootstraps were conducted in steps of 20 stimuli. At each step, the mean and variance confidence interval in WER of all systems in a year are computed. A solid line represents the mean confidence interval for a year as stimuli are increased. The shaded bands represent 2 standard deviations in the confidence intervals of all systems in a year.*

of 20 stimuli up to the size of the Extra ASR test sets for each year. For each step we re-sampled individual WER scores with replacement. We computed 1000 model simulations of WER in each step. To compute confidence intervals next we sorted the simulated WERs at 95% confidence by plotting the 25th and 975th scores at each step (2.5% either side of the distribution). These upper and lower bounds formed the bootstrapped confidence interval around the WER given a certain number of stimuli. The confidence intervals allowed visualisation of statistically valid differences between TTS systems and datasets as the number of stimuli under test increased.

Figure 2 shows the average WER confidence interval after bootstrapping. The confidence interval for each year is an average of the confidence interval of all systems at each step.

The lines are the mean interval at each step of 25 stimuli, the shaded area shows 2 standard deviations around the mean. We see the means begin to stabilise around 500 stimuli to around 4%. In 2016, the range of confidence intervals was more diverse than other years, but its mean score was similar to other years.

Narrower confidence intervals show systems may be more reliably scored with Extra ASR stimuli. Below we analyse the effect extra stimuli have on identifying significant differences.

2.4.3. Rankings by pairwise wilcoxon p-values

The Blizzard Challenge tests significance using pairwise Wilcoxon signed-rank tests. Initially, we compared rankings from each listener type using the Kendall-Tau rank correlation statistic [39]. However, since many systems exhibited no significant differences between one another, the statistic was misleading. It would have been indicative if every system had a significant difference between itself and its neighbouring ranked systems.

We therefore computed aggregate statistics using the pairwise Wilcoxon signed-rank matrices for each listener type. In a matrix, each cell is a p-value between a pair of systems. Figure 3 simplifies two such matrices using a p-value threshold of 0.005, where blue indicates a significant pair, while red represents pairs above the threshold. The top heatmap shows the rankings and p-values for paid listeners in 2017. The bottom one shows the ranking and p-values according to Extra ASR.

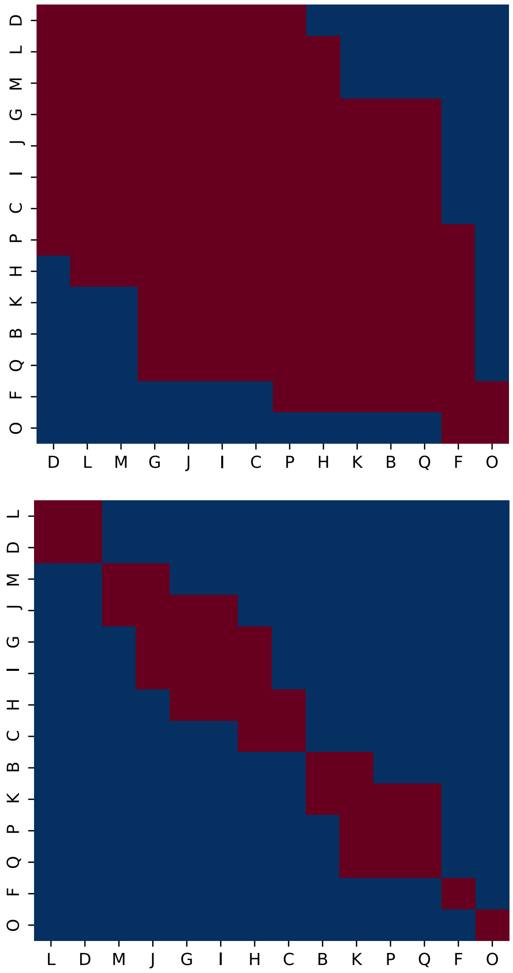
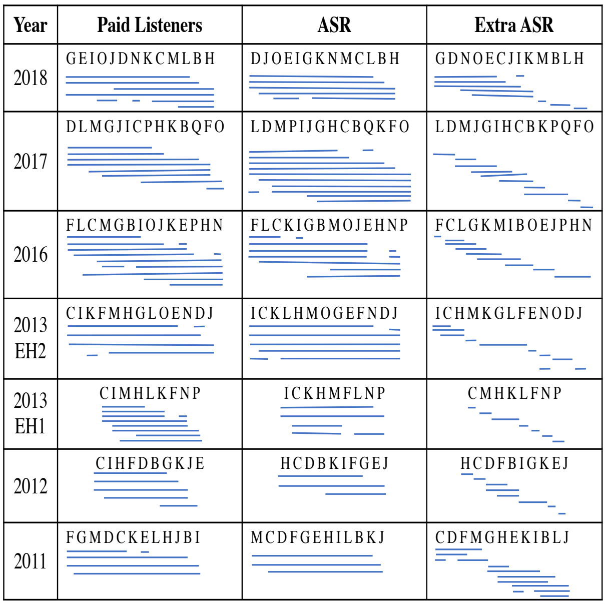


Figure 3: *Heatmap of pairwise p-values for systems ranked by Paid Listeners with 25 stimuli (Top) and* Extra ASR *(600 total - bottom) for the Blizzard Challenge 2017. Blue indicates a p-value below 0.005 between the pair. Red indicates no-significance. With extra stimuli p-values are lower.*

The heatmaps display overlap of no-significance between systems with partial rows of red cells. We extracted each of the unique partial rows of no-significance for every listener type in each challenge in Figure 4. The first partial row in the heatmaps above spans systems D to P (the top line in Figure 3), the second partial row spans systems D to H. Hence these are the first two lines in the 2017 Paid Listeners cell in Figure 4. For each challenge we show 3 rankings (EP, ASR and *Extra ASR*). Consistently, the rows of no-significance are similar for EP and ASR but longer than in *Extra ASR*. Figure 4 consistently shows that ASR indeed finds similar significance groups to EP and also more significant differences when extra stimuli are used.

Figure 4 also shows that some systems which are further than 1 step away in a rank may be in a similar group of no-



Year	Paid Listeners	ASR	Extra ASR
2018	GEIOJDNKCMLBH	DJOEIGKNMCLBH	GDNOECJIKMBLH
2017	DLMGJICPHKBQFO	LDMPIJGHCBQKFO	LDMJGIHCBKPQFO
2016	FLCMGBIOJKEPHN	FLCKIGBMOJEHNP	FCLGKMIBOEJPHN
2013 EH2	CIKFMHGLOENDJ	ICKLHMOGEFNDJ	ICHMKGLFENODJ
2013 EH1	CIMHLKFNP	ICKHMFLNP	CMHKLFNP
2012	CIHFDBGKJE	HCDBKIFGEJ	HCDFBIGKEJ
2011	FGMDCKELHJBI	MCDFGEHILBKJ	CDFMGHEKIBLJ

Figure 4: *Groups of no-significance for Paid Listeners, ASR and* Extra ASR. *Each line represents a unique grouping of systems as found in the p-value heatmaps. For the pairwise Wilcoxon signed-rank test, the significance level was set at a p-value of 0.005. When stimuli only used in the formal evaluation were included (Paid Listeners and ASR), the groups of no-significance encompass more systems than with extra stimuli (*Extra ASR*). We omit the the Online Volunteers, Speech Experts and Mechanical Turk groupings in this Figure due to space considerations and ease of visualisation - these also had long groups of no-significance such as the Paid Listeners and ASR. Note that systems which did not have 3 SUS test sets available were excluded from our analysis.*

significance. For example, system Q in the top line of ASR 2017. Although the mean performance of a system gives a particular ranking, the spread in its performance might result in no statistical significance when tested. The mean score of system Q was skewed by 2 low quality outlier stimuli. Such stimuli may be very important to examine for systems in deployment, and wide variance is observed even when using more stimuli with higher confidence such as in system I in Extra ASR 2018 and systems C and E in Extra ASR 2011. The problems resulting from smoothing out the effects of certain individual stimuli was the focus of [40] where the authors proposed evaluating systems based on test sets that target differences in the output of systems.

2.4.4. Frobenius norm of p-value matrices

We next sought to find out how significance levels improved as we increased the number of stimuli under consideration, potentially to find an optimum where significance was maximised with as few stimuli as possible. To visualise how significance varied we calculated the Frobenius norm of each pairwise Wilcoxon p-value matrix as we increased the number of ASR stimuli. The Frobenius norm is the square root of the sum of all the squared values of a matrix. In Figure 5 we plot the Frobenius norm as we increased the stimuli for each challenge.

The falling curves reflect falling p-values overall. There is a fall for all challenges, but to a differing degree for each. The absolute value of the norm is dependent upon the total number of systems (e.g. 2013 EH1 contained the fewest systems and has the lowest Frobenius norm curve). More noteworthy is the relative gradient change for each and finding where they converge - this level indicates where we find optimal significant differences between systems in a challenge. We see the curves fall the most in the first 200 stimuli. 2013 EH1 falls further after 400 stimuli when reaching a subset of the Extra ASR stimuli. Each curve has its own relative convergence level arising from the performance on the Extra ASR stimuli, the number of systems, and the relative quality of each in a challenge. Levelling can be observed from between 400-800 stimuli, although this is less clear for 2013 EH2 where the curve increases after 400 stimuli until it drops further around 700 stimuli. Thus, in addition, the test stimuli included can have an effect on whether a convergence level is found.

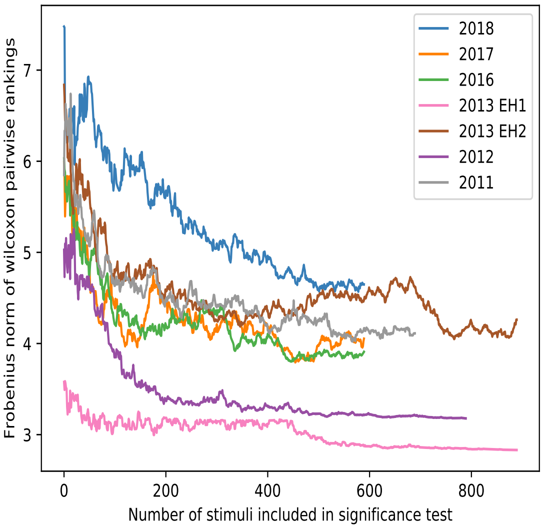


Figure 5: *Frobenius norm of pairwise Wilcoxon rankings varying according to number of stimuli included in significance test. We wish to find the convergence level for each curve. This level demonstrates the number of stimuli where the amount of significance discoverable is optimised with as few stimuli as possible.*

3. Conclusion

We used ASR to re-evaluate the Blizzard Challenge SUS tasks in 2011, 2012, 2013, 2016, 2017 and 2018. We found ASR performed reliably for evaluating intelligibility of TTS systems in the Blizzard Challenge, indeed on a comparable level to the challenge's paid listeners. Using extra stimuli, ASR also detected more statistically significant differences between pairs of systems, which would have been expensive to find in the human evaluations. We conclude that our analysis of the large and varied Blizzard Challenge data sets confirms that ASR can be a reliable and convenient metric to measure intelligibility, as long as a sufficiently large number of stimuli are used.

4. Acknowledgements

The organisers of the Blizzard Challenge, and Gustav Eje Henter for useful feedback and suggestions. This work was supported by an ESRC doctoral training grant provided via the SGSSS.

5. References

[1] R. Jiménez *et al.*, "Outliers detection vs. control questions to ensure reliable results in crowdsourcing.: A speech quality assessment case study," in *Proc. of WWW*, 2018, p. 1127–1130.

[2] R. Jimenez *et al.*, "Intra- and inter-rater agreement in a subjective speech quality assessment task in crowdsourcing," in *Proc. of WWW*, 2019, p. 1138–1143.

[3] A. Rix *et al.*, "Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs," vol. 2, 2001, pp. 749–752.

[4] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," *Proc. of PACRIM*, vol. 1, pp. 125–128, 1993.

[5] M. Chinen *et al.*, "ViSQOL v3: An open source production ready objective speech and audio metric," in *Proc. of QoMEX*, 2020.

[6] R. Gupta, A. Avila, and T. Falk, "Towards a neuro-inspired no-reference instrumental quality measure for text-to-speech systems," in *Proc. of QoMEX*, 2018.

[7] A. Avila *et al.*, "Non-intrusive speech quality assessment using neural networks," in *Proc. of ICASSP*, 2019, pp. 631–635.

[8] J. Williams and others., "Comparison of speech representations for automatic quality estimation in multi-speaker text-to-speech synthesis," in *Proc. Interspeech 2020*, 2020.

[9] C. Lo *et al.*, "MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion," in *Proc. of Interspeech 2019*, 2019, pp. 1541–1545.

[10] Y. Choi, Y. Jung, and H. Kim, "Deep MOS predictor for synthetic speech using cluster-based modeling," in *Proc. of Interspeech*, 2020.

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

[12] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?" in *Proc. of Interspeech*, 2011, pp. 1837–1840.

[13] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proc. of Interspeech*, 2013.

[14] K. Arai *et al.*, "Predicting Speech Intelligibility of Enhanced Speech Using Phone Accuracy of DNN-Based ASR System," in *Proc. of Interspeech*, 2019, pp. 4275–4279.

[15] T. Hayashi *et al.*, "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proc. of ICASSP*, 2020, pp. 7654–7658.

[16] M. Ribeiro *et al.*, "Tal: A synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos," in *Proc. of SLT*, 2021, pp. 1109–1116.

[17] R. Vích, J. Nouza, and M. Vondra, "Automatic speech recognition used for intelligibility assessment of text-to-speech systems," in *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Springer Berlin Heidelberg, 2008, pp. 136–148.

[18] T. Godambe *et al.*, "Developing a unit selection voice given audio without corresponding text," *EURASIP*, vol. 2016, no. 1, Dec. 2016.

[19] J. Fong *et al.*, "Investigating the Robustness of Sequence-to-Sequence Text-to-Speech Models to Imperfectly-Transcribed Training Data," in *Proc. Interspeech 2019*, 2019, pp. 1546–1550.

[20] E. Cooper *et al.*, "Utterance selection for optimizing intelligibility of tts voices trained on asr data," in *Proc. of Interspeech*, 2017, pp. 3971–3975.

[21] K. Lee, E. Cooper, and J. Hirschberg, "A comparison of speaker-based and utterance-based data selection for text-to-speech synthesis," in *Proc. of Interspeech*, 2018, pp. 2873–2877.

[22] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, 2014.

[23] F. Hinterleitner *et al.*, "Comparison of approaches for instrumentally predicting the quality of text-to-speech systems: Data from blizzard challenges 2008 and 2009," pp. 1325–1328, 2010.

[24] C. Norrenbrock *et al.*, "Towards perceptual quality modeling of synthesized audiobooks – Blizzard Challenge 2012," in *Proc. of Blizzard Challenge Workshop*, 2012.

[25] R. Ullmann *et al.*, "Objective intelligibility assessment of text-to-speech systems through utterance verification," in *Proc. of Interspeech*, 2015, pp. 3501–3505.

[26] L. Latacz and W. Verhelst, "Double-ended prediction of the naturalness ratings of the blizzard challenge 2008-2013," in *Proc. of Interspeech*, 2015, pp. 3486–3490.

[27] T. Yoshimura *et al.*, "A hierarchical predictor of synthetic speech naturalness using neural networks," in *Proc of Interspeech*, 2016, pp. 342–346.

[28] R. Jimenez, L. Gallardo, and S. Moller, "Influence of number of stimuli for subjective speech quality assessment in crowdsourcing," in *Proc. of QoMEX*, 2018.

[29] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. of ICASSP*, 2004, pp. 409–411.

[30] V. Panayotov *et al.*, "Librispeech: An ASR corpus based on public domain audio books," *Proc. of ICASSP*, pp. 5206–5210, 2015.

[31] A. Baby *et al.*, "An ASR guided speech intelligibility measure for TTS model selection," *arXiv e-print: 2006.01463*, 2020.

[32] S. King and V. Karaiskos, "The Blizzard Challenge 2011," in *Proc. of Blizzard Challenge Workshop*, 2011. [Online]. Available: http://www.festvox.org/blizzard/bc2011/summary_Blizzard2011.pdf

[33] S. King and V. Karaiskos, "The Blizzard Challenge 2012," in *Proc. of Blizzard Challenge Workshop*, 2012. [Online]. Available: http://www.festvox.org/blizzard/bc2012/summary_Blizzard2012.pdf

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

[35] S. King and V. Karaiskos, "The Blizzard Challenge 2016," in *Proc. of Blizzard Challenge Workshop*, 2016. [Online]. Available: http://www.festvox.org/blizzard/bc2016/blizzard2016_overview_paper.pdf

[36] S. King, L. Wihlborg, and W. Guo, "The Blizzard Challenge 2017," in *Proc. of Blizzard Challenge Workshop*, 2017. [Online]. Available: http://www.festvox.org/blizzard/bc2017/blizzard2017_overview_paper.pdf

[37] S. King *et al.*, "The Blizzard Challenge 2018," in *Proc. of Blizzard Challenge Workshop*, 2018. [Online]. Available: http://www.festvox.org/blizzard/bc2018/blizzard2018_overview_paper.pdf

[38] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statist. Sci.*, vol. 1, no. 1, pp. 54–75, 1986.

[39] M. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, 11 1945.

[40] J. Chevelu *et al.*, "How to compare TTS systems: A new subjective evaluation methodology focused on differences," in *Proc. of Interspeech*, 2015, pp. 3481–3485.