INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Listener sensitivity to deviating obstruents in WaveNet

Ayushi Pandey[1], *Jens Edlund*[2], *Sébastien Le Maguer*[1], *Naomi Harte*[1]

[1]Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
[2]Department of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden

pandeya@tcd.ie, edlund@speech.kth.se, lemagues@tcd.ie, nharte@tcd.ie

Abstract

This paper investigates the perceptual significance of the deviation in obstruents previously observed in WaveNet vocoders. The study involved presenting stimuli of varying lengths to 128 participants, who were asked to identify whether each stimulus was produced by a human or a machine. The participants' responses were captured using a 2-alternative forced choice task. The study found that while the length of the stimuli did not reliably affect participants' accuracy in the task, the concentration of obstruents did have a significant effect. Participants were consistently more accurate in identifying WaveNet stimuli as machine when the phrases were obstruent-rich. These findings show that the deviation in obstruents reported in WaveNet voices is perceivable by human listeners. The test protocol may be of wider utility in TTS.

Index Terms: WaveNet, obstruents, TTS evaluation, perception, distortion

1. Introduction

Human-likeness is one of the primary and longstanding goals for Text-To-Speech (TTS) synthesizers. Voice is an indispensable medium of communication and social exchange in human communities, and contains rich information in addition to the intended message. Our reliance on spoken language explains the many findings where human, or human-like, voices are consistently rated as more socially acceptable [1], pleasant [2] and trustworthy [3]. Recent work also shows that while the phenomenon of the uncanny valley holds true for visual stimuli, the likeability of speech stimuli increases with human-likeness [4]. Thus it is important to assess the human-likeness of synthetic voices.

In targeted applications of TTS, the concept of human-likeness is closely linked with naturalness, which is a widely tested attribute in TTS evaluation. For example, in the Blizzard Challenge series[1], where the purpose is to compare TTS techniques through a common evaluation platform, the word "natural" is implicitly used to refer to the original human voice. However, the question of naturalness, presented usually as "how natural does this utterance sound?", has been considered "nebulous" [5], or "poorly defined" [6], as it relies on listeners' own interpretation of naturalness. As a consequence, naturalness ratings can differ with context [7, 8], application specific expectations [9], instructions to listeners [10] and interaction conditions [11]. The question of human-likeness, on the other hand, is more precise, and can thus offer more diagnostic information about system-weakness, especially in applications where naturalness and human-likeness are linked.

[1]http://festvox.org/blizzard/

While these studies establish the importance of studying human-likeness of speech synthesizers, an evaluation framework also plays a major role. In ASVSpoof[12] settings, although the question of human-likeness is straightforward, the utterance is presented as a whole. This limits its diagnostic abilities, because the exact source of perceivable distortion is unclear. Perceptual judgments are commonly assumed to increase in accuracy, as sensory input accumulates[13]. On the other hand, judgements for speech stimuli are found to be more variable with limited, or shorter input[14, 15]. This suggests that the longer the exposure to a stimulus, the more accurate or consistent a participant becomes in their response. Therefore, in this paper, we design a test of human-likeness as a function of length of the stimulus. This gives us the **first research question** explored in this paper: does the accuracy of judgment of human-likeness increase with the length of the stimulus?

Recent work on the segmental evaluation of WaveNet voices [16] showed that their voiceless obstruents deviate strongly from the human voice, in most of their contrastive features. Features of voiced obstruents and vowels were shown to be more similar to the human voice. Since the autoregressive nature of WaveNet vocoders was a potential cause for this deviation, it may be extrapolated that characteristics of sonorant consonants may be well reproduced by WaveNet vocoders. The perceptual significance of this deviation was not established. Contrastive features encode information that is phonemically meaningful, and human listeners (at least, native speakers) may be attuned to expect them in human-like speech. If this deviation is perceivable, then obstruent-rich utterances should provide more clues as to whether a stimulus is human or from a machine. On the other hand, if the phrase contains more non-obstruents, or is sonorant-rich, then the clues should be weaker. Hence, we pose our **second research question**: does accuracy of judgment of human-likeness increase more with the length of the stimulus when the utterance is obstruent-rich?

In Section 2, we describe the experimental details, detailing stimuli creation and their presentation to participants. Section 3 describes the results of our experiments, demonstrating the effect of increasing stimulus length in a) randomly selected, or b) specially selected obstruent or sonorant-rich utterances. Section 4 presents a discussion on the important observations, and the relevance of these findings to other fields of speech perception and technology. Section 5 concludes the paper.

2. Experimental Design

2.1. Description of dataset

The source material for our study comes from the recently extended [17] Blizzard Challenge 2013 (BC-2013) [18] corpus. The human voice in the original challenge came from audio-

10.21437/Interspeech.2023-1843

book renditions by an American, female voice artist. All participating teams had to develop their own TTS systems based on this human voice as a common training data. However, no team had used neural TTS in 2013. The extended version [17] contributes 4 neural voices, which are trained on the same human speaker as in the original challenge. Tacotron [19] and FastPitch [20] were used as acoustic models for mel-spectrogram generation, and WaveNet [21] and WaveGAN [22] as vocoders for waveform-generation. Since the previous findings of voiceless obstruent deviation [16] were limited to WaveNet, we only selected two of the neural voices for experiments presented in this paper: FastPitch WaveNet (System Y), and Tacotron WaveNet (System Z), along with the human voice.

All our stimuli were derived from the 100 utterances that originally formed the test corpus in both the original and extended versions of BC-2013. The next subsection explains the design and creation of the stimuli and presentation strategy.

2.2. Phrase extraction: text and audio

We developed a refined set of audio stimuli from the 100 utterances by taking the following aspects into consideration:-

Grammatical well-formedness:- First, we divided each utterance into its constituent phrases using the Stanford NLP parser. The grammatical well-formedness of the resultant phrases (e.g. a noun phrase "*big, solemn oaks*", instead of a roughly cut up *"before but she"*) ensured that our participants could focus only on the audio. All duplicates were removed.

Phrase length:- This was determined in terms of the number of syllables per phrase. We only preserved unique phrases of *2, 4, 8, 16*, and *32* syllables to maintain a sufficiently perceivable "doubling" of their lengths. The number of phrases selected at each length is described in Table 1, as is the distribution of the stimuli between human and the two WaveNet systems. A total of 124 phrases was heard by each participant.

Phrase length (in #syllables)	#phrases	Human	FastPitch (Y)	Tacotron (Z)	Total phrases
2	64	32	16	16	
4	32	16	8	8	
8	16	8	4	4	**124**
16	8	4	2	2	
32	4	2	1	1	
Total	124	62	31	31	

Table 1: *Number of phrases at each phrase-length heard by each participant across the human voice, and systems Y and Z.*

Audio extraction:- The corresponding audio for the selected phrases was extracted from System Y, System Z and the human voice, and hand-corrected for phrases boundaries. Additionally, a fade of 50 ms was also added before and after each utterance, to minimise any audible clicks. The sampling rate was 44.1 kHz, bitrate 320 kbps, and the format was `.mp3`. A high bitrate of 320 kbps is maintained [23] to ensure that the recording format does not influence our results.

2.3. Experimental conditions and groups

As shown in Table 1, each participant evaluated 124 phrases. First, 62 human stimuli were extracted, in accordance with the phrase distribution in Table 1. These were maintained identically throughout the experiments. Synthetic stimuli were extracted based on one of the two conditions now described.

The baseline condition:- 62 synthetically produced stimuli of the required phrase length were selected randomly, with no particular constraints on their lexical content. This condition was designed for the first research question, in Section 1, relating to length.

The ObSon condition:- Each unique phrase among the well-formed phrases was assigned a score, based on the obstruent or sonorant concentration in its lexical content. Based on this score, phrases were categorized as obstruent-rich, or sonorant-rich[2]. Of the required 62 synthetic phrases, we selected 31 obstruent-rich phrases (OBS-P), and 31 sonorant-rich phrases (SON-P). This condition was designed for the second research question in Section 1. Based on previous work [16], we expected the obstruent-rich stimuli to increase the accuracy of the human-or-machine responses.

The ObSon condition required us to also confirm whether the effect we hypothesized was truly an effect of obstruent-richness, and not that of a specific type of TTS system. In other words, we wanted to examine if this effect was consistent across both the acoustic models. Therefore, we designed our stimuli such that, for one group of participants, we retained those OBS-P which were produced by FastPitch (Y) and SON-P produced by Tacotron (Z). Then for another group, these pairings were reversed. To maintain consistency between the two conditions, we also split the baseline stimuli equally, and paired them alternately with each of the acoustic models. But this split was completely arbitrary. The details of the individual participant groups are described in Table 2. Participants were assigned to one of the 4 groups listed. No participant was repeated in any group. Their details are described in the next section.

Participant group	FastPitch (Y)	Tacotron (Z)	Human
Baseline_Group1	R1-P	R2-P	HM-P
Baseline_Group2	R2-P	R1-P	HM-P
ObSon_Group1	SON-P	OBS-P	HM-P
ObSon_Group2	OBS-P	SON-P	HM-P

Table 2: *Phrases paired with acoustic model for each group. R1-P, R2-P = Random Phrases 1, Random Phrases 2 . OBS-P = Obstruent-rich phrases; SON-P = Sonorant-rich phrases.*

2.4. Participant details

We recruited 128 participants ($32_{participants}$ x $2_{condition}$ x 2_{group}) through Prolific. Gender balance was maintained for each group. All participants were native speakers of English (UK or US English speakers, only), and reported no history of hearing impairment. Their informed consent was obtained prior to the experiment, and the following demographic information was collected: a) age, b) sex at birth, c) speaker of UK or US English, d) experience with Alexa or other TTS devices, and e) professional experience in speech/audio processing. The median time for completion was 25 minutes, and their remuneration rate was 7 GBP/hour.

2.5. Presentation of the stimuli

The stimuli were presented in a random order, to remove any effect of sequencing on length. In every trial, we presented only one stimulus to the participant, and requested their response to the question: "*Did this sound like a human, or a machine?*". Stimuli were only played once. Their responses were captured in a 2-alternative forced choice task: "Human" or "Machine". The experiment was designed entirely in Psychopy [24], and

[2]obstruent-rich: "mo**st** se**l**f **p**o**s**se**s**se**d**"; sonorant-rich:"**m**ea**n**i**n**g i**n** it."

hosted online on the Pavlovia server [3].

3. Results

Listener responses are coded as a binary variable where `0=wrong` i.e, the participant was wrong, and `1=correct` i.e, they were correct in their judgment of human or machine. Figure 1 shows the relationship between stimulus-length (x-axis) and the predicted probability of the correct response (P_{ACC}), as a result of a Generalized Linear Model (GLM) model fit (y-axis). The underlying GLM model is described by the following model equation:-

`CorrectResponse` $\sim$ `Number of Syllables, (family = binomial).`

In Figure 1(a), we show the **baseline** condition, exploring whether the P_{ACC} increases with increasing stimulus-length. In Figure 1 (b), we show the pattern of response in the **ObSon** condition. Here we explore how responses change on the basis of the concentration of obstruents or sonorants in the lexical content of the stimuli. If the reported deviation is perceivable in obstruents, listeners should be more accurate, i.e. show a further increase in P_{ACC} for obstruent-rich phrases.

3.1. The baseline: randomly selected phrases

Figure 1(a) shows that the P_{ACC} of the **human** stimuli have a consistent positive relationship with the increase in stimulus length. The GLM model across both groups predicts an overall *increase* of +23% in P_{ACC} between stimulus length 2 and 32 [Slope(SE) +1.7(0.32) ,p-val$<$0.001]. This means that for human stimuli, across both groups in the baseline, we see an increase in P_{ACC} as the length of the stimulus increases.

In **WaveNet** stimuli (labelled "Machine"), Figure 1(a) shows that P_{ACC} either remains constant with stimuli-length, or shows a slight *decrease*. When individually evaluated for every length, we find that although P_{ACC} does show some increasing trends upto length 16, the difference is not significant. However, at stimulus-length 32, we see a statistically significant lowering in [Slope(SE)–0.084(0.26), p-val$<$0.01]. At a preliminary level, this may suggests that P_{ACC} decreases at length 32. However, since this was contrary to expectations, we further explored other factors that may have influenced this behaviour.

3.1.1. Interaction of stimulus-length with other variables

We introduced the following variables as interaction effects to the underlying model: a) AGE, b) EXPERIENCE with TTS devices, c) SEX, and d) PHRASE SETS, and e) ACOUSTIC MODEL. While (a-c) were demographic variables, (d) refers to R1-P and R2-P (see Section 2.3). The model equation was: `CorrectResponse` $\sim$ `Number of Syllables * InVAR, (family = binomial)`, where `InVAR` was one of (a-e). Between each of these variables, we found clear effects of (d) PHRASE SETS to be the most consistent effects among both groups of the baseline. This means that even though the phrases were randomly selected, we can see consistent drops in accuracy in one set of phrases over the other in both groups. The interaction between stimulus-length and R2-P shows a statistically significant lowering of accuracy as length increases [Slope(SE)-1.27(0.38), p-val$<$0.001].

This effect, although consistent between both groups, is stronger for BASELINE_GROUP1, where R2-P is produced by Tacotron. This means that when linguistic content is maintained identically, Tacotron produced voices are harder to tell apart from the human voice. Further analysis on the linguistic content of the phrases in R2-P, reveals that even though the selection was completely random, this set has a relatively lower concentration of voiceless obstruents, particularly in stressed syllables. Taken together, these observations may indicate listeners' preference towards the voice produced by Tacotron WaveNet.

[3] https://pavlovia.org/

3.2. The ObSon: obstruent-rich vs sonorant-rich phrases

As in the baseline condition, in Figure 1(b) we see that the P_{ACC} of the **human** stimuli **increases** with increase in stimulus length overall. The GLM model across both groups predicts an overall *increase* of +19% in P_{ACC} between stimulus length 2 and 32 [Slope(SE) +1.19(0.28), p-val$<$0.001]. Therefore, in both baseline and the ObSon condition, participant responses become more accurate for human stimuli as the length of the stimulus increases.

For WaveNet stimuli, it can be clearly seen in Figure 1(b) that the obstruent-rich stimuli ***rise faster*** and ***reach higher values*** of P_{ACC} in both groups. This indicates that the deviation in WaveNet obstruents, first reported in [16], is perceptible and contributes to the perceived machine-likeness of WaveNet stimuli. Combined over both groups, the P_{ACC} shows a strongly significant rise of +32% between stimulus length 2 and 32 [Slope(SE) +1.5(0.34) ,p-val$<$0.001], when the stimuli are obstruent-rich. This effect is stronger in Group II, where sonorant-rich stimuli are produced by Tacotron.

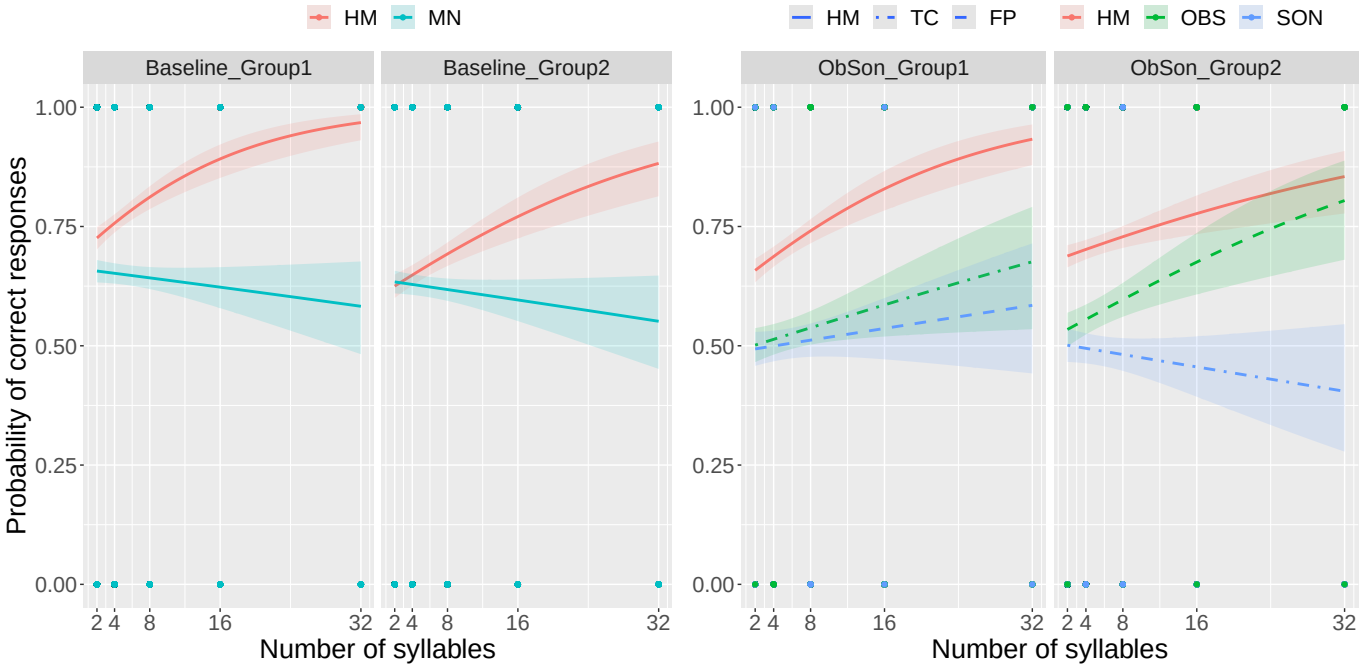
3.2.1. Interaction of stimulus-length with other variables

The underlying model was modified as described in Section 3.1.1. The other variables were the same as above, except that (d) PHRASE SET was composed of OBS-P and SON-P. An interaction between stimulus-length and the PHRASE SET showed consistent trends among both groups. While SON-P phrases lower the P_{ACC} in all lengths, the effect is most significant at stimulus-length 16. The P_{ACC} for OBS-P is 0.54, while that of SON-P is 0.35. This 20% lowering of the P_{ACC}, indicates that the largest perceivable difference in human-likeness between OBS-P and SON-P occurs at stimulus-length 16. However, a deeper analysis, especially which accounts for demographic variables, will be required before confirming and reproducing this result on stimulus-length.

4. Discussion

We have analyzed the effect of increasing stimulus length on listeners' judgement in a 2-alternative forced choice task of detecting human vs machine. Listener accuracy was expected to increase with increasing length, because longer exposure to stimuli has been shown to increase participants' consistency and accuracy in tasks [15, 14] In judging the human-likeness of speech samples, our participants grew steadily more accurate with increasing length, when the stimulus was human speech or obstruent-rich synthetic speech from the WaveNet vocoder. When the stimuli were sonorant-rich synthetic speech, their length had a smaller (ObSon-Group1) or a negative effect (ObSon-Group2), compared to obstruent-rich utterances.

Previous work on spoofing detection has demonstrated the realism of synthetic voices, and has been rated as indistinguishable from human speech [12] in many listening conditions [25]. In our study, we find that stimuli accumulation, i.e, longer utterances, nudge the participant responses more towards its human-likeness. However, in the presence of segmental distortion, par-



(a) ***Baseline****: randomly selected utterances*

(b) ***ObSon****:utterances selected for obstruent/sonorant-richness*

Figure 1: *A GLM-model fit between predicted probability of correct responses (0=incorrect, 1=correct) and increasing stimulus-length for human and machine-generated sentences (a). In (b), machine sentences further divided into obstruent- and sonorant-rich. Responses obtained from a 2-AFC task where "human" or "machine" judgements are made on audio stimuli. HM:Human, MN:Machine, TC:Tacotron, FP:FastPitch, OBS:Obstruent-rich, SON:Sonorant-rich*

ticipants determine the machine-likeness more accurately. This is also in line with previous research, where segmental distortion has pointed to higher-level attributes like naturalness and system-preferences [26]. It must be noted, however, that more variance can be seen in participant responses for synthetic speech. A potential reason is the imbalance between short and long utterances. This is a limitation of the dataset, as naturally occurring corpora do not contain utterances that are neatly balanced for obstruent/sonorant-richness, unless specially designed. In future work, it will be useful to redesign these experiments, with equal numbers of long and short stimuli, which are not "cut-outs" from running speech.

A second observation from our study is that Tacotron voices showed reduced accuracy in both conditions. Participants were better at detecting machine-likeness in FastPitch utterances. A potential explanation is that the Tacotron-WaveNet combination is auto-regressive both in the acoustic model and the vocoder [19, 21]. Therefore, it is possible that sonorants, which are characterised by vowel-like resonances, are faithfully reproduced. This perceivable difference between the synthesizers points to a limitation of a subjective listening test: both systems in [17] had obtained an identical MOS score of 4 in the source paper.

The perceptual significance of deviating obstruents in WaveNet systems has implications for multiple fields. First, it may motivate TTS engineers to focus on segmental attributes of a system, or even perform a post-processing of their audio. For example, [27] demonstrate that the use of WaveNet vocoders with distinct periodic/aperiodic decomposition, scores higher naturalness. From a TTS evaluation perspective, the test methodology presented may offer a more fine-grained insight into localizing the source and perceptual significance of distortion, compared to traditional, MOS-based listening tests. Finally, if segmental characteristics of sonorants are indeed indistinguishable from human speech, then analysis of synthetically produced sonorants may generalize well to human speech. This could accelerate research in phonetics, because of the reduced reliance on speech data collection.

5. Conclusion

In this paper, we explore whether accuracy of detecting human-likeness increases with increase in length of stimuli. We also investigate whether the segmental distortion, previously reported for obstruents in WaveNet, is perceivable by listeners, and increases their accuracy of human vs machine detection.

The central finding here is that accuracy in detecting obstruent-rich phrases consistently improves with longer stimuli. This shows that human listeners can perceive segmental distortion in high-quality neural TTS synthesizers. We also found that Tacotron voices were judged to be human more frequently than FastPitch voices, contrary to their previously reported equivalence in MOS based evaluations. These findings show that MOS-based evaluations are not sufficiently diagnostic, and assert a greater need in better methodologies for neural TTS evaluation.

6. Acknowledgements

This work has the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224, the ADAPT Centre (Grant 13/RC/2106), and a Google Faculty Award.

7. References

[1] S. Schreibelmayr and M. Mara, "Robot voices in our daily lives: Vocal human-likeness and application context as determinants of user acceptance," *Frontiers in Psychology*, p. 1843, 2022.

[2] K. Kühne, M. H. Fischer, and Y. Zhou, "The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study," *Frontiers in neurorobotics*, vol. 14, p. 105, 2020.

[3] L. Weidmüller, "Human, hybrid, or machine?: Exploring the trustworthiness of voice-based assistants," *Human-Machine Communication*, vol. 4, pp. 85–110, 2022.

[4] A. Baird, E. Parada-Cabaleiro, S. Hantke, F. Burkhardt, N. Cummins, and B. Schuller, "The perception and analysis of the likeability and human likeness of synthesized speech," in *Interspeech*, 2018.

[5] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. E. Henter, S. L. Maguer, Z. Malisz, É. Székely, C. Tånnander *et al.*, "Speech Synthesis Evaluation—State-of-the-Art Assessment and Suggestion for a Novel Research Program," in *Speech Synthesis Workshop (SSW)*, 2019.

[6] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, pp. e006–e006, 2014.

[7] R. Clark, H. Silen, T. Kenter, and R. Leith, "Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs," *arXiv preprint arXiv:1909.03965*, 2019.

[8] J. O'Mahony, P. O. Gallegos, C. Lai, and S. King, "Factors affecting the evaluation of synthetic speech in context," in *The 11th ISCA Speech Synthesis Workshop (SSW11)*. International Speech Communication Association, 2021, pp. 148–153.

[9] E. Roesler, L. Naendrup-Poell, D. Manzey, and L. Onnasch, "Why context matters: the influence of application domain on preferred degree of anthropomorphism and gender attribution in human–robot interaction," *International Journal of Social Robotics*, vol. 14, no. 5, pp. 1155–1166, 2022.

[10] R. Dall, J. Yamagishi, and S. King, "Rating naturalness in speech synthesis: The effect of style and expectation," in *Proceedings of Speech Prosody*. Citeseer, 2014.

[11] S. Betz, B. Carlmeyer, P. Wagner, and B. Wrede, "Interactive hesitation synthesis: modelling and evaluation," *Multimodal Technologies and Interaction*, vol. 2, no. 1, p. 9, 2018.

[12] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[13] M. Inglis and C. Gilmore, "Sampling from the mental number line: How are approximate number system representations formed?" *Cognition*, vol. 129, no. 1, pp. 63–69, 2013.

[14] W. R. Thurlow and J. R. Mergener, "Effect of stimulus duration on localization of direction of noise stimuli," *Journal of speech and hearing research*, vol. 13, no. 4, pp. 826–838, 1970.

[15] K. Apeksha, B. H. Mahadevaswamy, S. Mahadev, and M. T. Shivananda, "Pattern perception in quiet and at different signal to noise ratio in children with learning disability," *The Journal of International Advanced Otology*, vol. 15, no. 2, p. 263, 2019.

[16] A. Pandey, S. L. Maguer, J. Carson-Berndsen, and N. Harte, "Production characteristics of obstruents in wavenet and older tts systems," in *INTERSPEECH*, 2022.

[17] S. L. Maguer, S. King, and N. Harte, "Back to the future: Extending the blizzard challenge 2013," in *Interspeech*, 2022.

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

[19] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. A. J. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *ArXiv*, vol. abs/1703.10135, 2017.

[20] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," *arXiv preprint arXiv:2006.06873*, 2020.

[21] A. Van den 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. [Online]. Available: http://arxiv.org/abs/1609.03499

[22] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[23] B. Bollepalli and T. Raito, "Effect of mpeg audio compression on vocoders used in statistical parametric speech synthesis," in *2014 22nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2014, pp. 1237–1241.

[24] J. Peirce, R. Hirst, and M. MacAskill, *Building experiments in PsychoPy*. Sage, 2022.

[25] C. Terblanche, P. Harrison, and A. J. Gully, "Human spoofing detection performance on degraded speech." in *Interspeech*, 2021, pp. 1738–1742.

[26] H. T. Bunnell, S. R. Hoskins, and D. Yarrington, "Prosodic vs. segmental contributions to naturalness in a diphone synthesizer." in *ICSLP*, 1998.

[27] T. Fujimoto, T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Speech synthesis using wavenet vocoder based on periodic/aperiodic decomposition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 644–648.