Interspeech 2022
18-22 September 2022, Incheon, Korea

Production characteristics of obstruents in WaveNET and older TTS systems

Ayushi Pandey[1], *Sébastien Le Maguer*[1], *Julie Carson-Berndsen*[2], *Naomi Harte*[1]

[1]Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
[2]ADAPT Centre, School of Computer Science, University College Dublin, Ireland

pandeya@tcd.ie, lemagues@tcd.ie, julie.berndsen@ucd.ie, nharte@tcd.ie

Abstract

Segmental properties of Text-To-Speech (TTS) synthesizers have been studied for their influence on various perceived attributes of synthesized speech. However, they have received very limited attention for modern, neural vocoder-based TTS. In this paper, we compare segmental properties of WaveNET vocoder voices with a natural voice, and the best-performing non-neural synthesizers of the 2013 Blizzard Challenge. We extended the 2013 dataset with two new voices generated using a WaveNET vocoder. Acoustic-phonetic features of obstruent consonants and their neighbouring vowels were compared between the natural voice and each of these TTS systems. Statistical analysis was conducted using the Kruskal-Wallis test, and Dunn's test.

Compared to the reference natural voice, we find that the WaveNET vocoder performs very well in modelling vowels, but features like F0 at onset and spectral tilt show significant deviations from the natural voice. Among consonants, neural voices deviate most from natural in the context of voiceless fricatives. Compared to other TTS systems, several features (like vowel dispersions, and consonant duration) which had shown strong deviations from natural, were found to not differ from natural in the WaveNET vocoder systems.

Index Terms: WaveNET, obstruents, TTS evaluation

1. Introduction

Rapid advancements in TTS technology have enabled the emergence of high-quality, human-like synthetic speech. However, studies that discuss the contribution of segmental properties of speech in perceived naturalness of neural voices have been quite limited. Evidence from electroencephalography based studies [1, 2] have found that degradation in a signal can be detected by changes in neuronal activity, even when stimuli are as short as a vowel. These changes may not translate to conscious behaviour ratings, but can affect listener fatigue in long-term usage. Additionally, listeners' sensitivity to naturalness in synthetic speech [3], has been found even at a "microscopic" level (i.e, when stimuli were only a few glottal pulses from a vowel). Studies on diphone synthesis found a significant effect of segmental features on listener preferences [4], and their quality was reported to influence the naturalness of intonation [5]. In an investigation of unit-selection synthesisers, [6], segmental (or unit) appropriateness was reported to be an important dimension for listeners' judgment of naturalness.

These studies emphasise that the contribution of constituent segments cannot be ignored. However, most studies on segmental evaluation in speech synthesis either analyze phoneme-level stimuli, or evenly add a distortion throughout the utterance. Segments are usually not discussed as well-defined phonological classes. Such a definition is important, because it allows an array of features to be analyzed which are specific to those segments, and have been reported to be meaningful in listeners' perception of phonemic contrast. In our previous work [7], we had compared good-quality systems from the Blizzard Challenge 2013 (BC-2013) with poor-quality ones, on the basis of production characteristics of obstruent consonants. We reported that a feature-based analysis could provide insights into system naturalness, as a function of deviation from the natural voice. However, only gross characteristics of obstruent characteristics were presented, regardless of their position (pre-vocalic, post-vocalic) or voicing status. Secondly, transitional cues, contained in their surrounding vowels, were left out. Finally, no results from modern, neural TTS synthesizers were included.

In this paper, we use acoustic-phonetic properties of obstruent consonants, *and* their neighbouring vowels to compare WaveNET voices with the natural voice, and older TTS systems. We extend the BC-2013 corpus, to include two new voices generated using a WaveNET vocoder [8]. First, we compare WaveNET voices directly with the natural voice. We then compare WaveNET voices with older techniques (Hybrid, Unit-Selection and HMM) in TTS, on the basis of their deviation from natural voice. Finally we discuss the implications of these results on the perceptual attributes of synthesized speech. We find that analyzing production characteristics at a segmental level can provide insights into specific locations of distortions.

Section 2 discusses the importance of obstruents, and the reason for their choice in this study. Section 3 describes the experimental setup, and Section 4 presents our results. Section 5 discusses the relevance of this method within TTS, and Section 6 concludes the paper.

2. Obstruents and their importance

Obstruent consonants are characterized by complete or partial obstruction of air passing through the vocal tract. As a principal phonological class of consonants in English, they account for 6 unique phoneme categories for stops, [p, t, k, b, d, ɡ], 9 for fricatives, [f, v, θ, ð, s, z, ʃ, ʒ, h], and 2 for affricates [ʧ, ʤ].

Obstruent consonants have been documented [9] to account for between two-thirds and three-quarters of the consonantal population in cross-linguistic phoneme inventories. In the BC-2013 dataset, obstruents cover 64.01% of the consonantal population, on average, per utterance. This suggests that a large mass of acoustic cues in the utterance are comprised of obstruent consonants, which can be used by listeners while making perceptual judgments about the perceived attributes of synthetic speech.

Categorized along three manners and two voicing conditions, English obstruents also offer a set of distinctive phonetic attributes, which cannot be studied in other phonological classes. For example, effects of voicelessness may only be located within obstruents. Additionally, vowels in their neighbourhood are also influenced by their properties, and carry cues to their identification such as formant transitions [10], F0 per-

10.21437/Interspeech.2022-10606

turbations [11], and amplitude and duration changes [12, 13] as a function of manner and voicing.

Affricates and fricatives are known for their articulatory complexity, and their misarticulation is apparent in dysarthric speech [14] compromising intelligibility. From a speech perception perspective, obstruents are perceived less reliably in noise [15], and enhancing their target cues has resulted in improved recognition of speech [16]. These studies highlight the perceptual contribution of obstruents, and the critical role their precise production plays in the perception of speech. A targeted discussion on obstruents within TTS has been absent. Subjective evaluations attest that WaveNET voices sound much more natural than non-neural synthesizers [8]. Hence in this paper, we hypothesize that production characteristics of obstruents in WaveNET must come closer to natural speech. In other words, improvements in WaveNET may be connected to improvements in the segmental characteristics of obstruents.

3. Experimental setup

3.1. Description of dataset

In our study, we used the BC-2013 [17] corpus, extended with two neural voices for a comparative analysis of different techniques, based on their obstruent features. The Blizzard Challenge[1] is a well-known international challenge, annually organized to compare state-of-the-art TTS systems. We selected the natural voice, as well as 6 of the best-performing systems from the original challenge. Systems M and K were hybrid systems, I and C were HMM, while systems L and N were unit-selection systems. On a 5-point Mean Opinion Score (MOS) scale, the naturalness scores were as follows: M: 3.39, K: 3.31, I: 3.07, N: 2.93, L: 3.05 and N: 2.57. Two new systems, Fastpitch Wavenet (Y) and Tacotron Wavenet (Z) were included in the dataset, where waveform-generation in both systems was performed by the WaveNET vocoder. For acoustic modelling, Y used FastPitch [18], while Z used the Tacotron [19] model.

Thus, a dataset of 100 identical utterances synthesized by 8 systems, alongside the original natural voice, was created for comparative analysis. Acoustic-phonetic features were extracted from all obstruents and their neighbouring vowels, as described in the next subsection.

3.2. Feature extraction

We use acoustic-phonetic features which have been studied for contrastive properties of obstruent consonants, and their neighbouring vowels. Contrastive properties of obstruent consonants have been studied along the durational [20], amplitudinal and spectral [21, 22] attributes. The perceptual contribution of their surrounding vowels has also been discussed [23]. Our feature extraction closely follows the techniques presented in Redmon's [24] and Jongman's [25] studies on obstruents. To ensure scalability, we restricted our features to those that could be automatically extracted and did not include any that require detailed hand-correction.

Audio files from all the systems were force-aligned using the Montreal Forced Aligner (MFA) [26] to create phoneme-level boundaries. Sub-phonemic boundaries for the noise duration of stops and affricates were demarcated using a rule-based temporal boundary identification procedure described in [7]. Consonants were then separated into 3 positional contexts: pre-vocalic (CV), post-vocalic (VC) and consonant clusters. Consonant clusters were not analyzed for the present analysis, and will be followed up in future work. Vowels that appeared in the immediate neighbourhood of these consonants were also categorized into the CV and VC positional contexts. Then, the following feature set was extracted for the **analysis of vowels**:-

- *Vowel duration (V-Dur):-* The duration of the vocalic region, as returned by the MFA. In the CV position, the vowel onset is marked at the first 20% of this duration. Conversely, in the VC position, the offset is marked at the last 20% of this duration.
- *RMS amplitude (RMS-Amp):-* The root-mean-squared amplitude of the power spectrum of the vocalic region.
- *Formant values (F1-F5):-* The formant values of the first 5 formants at the onset/offset and midpoint of the vowel. These were extracted using the Burg formant-tracking algorithm in Praat [27]. The Escudero optimization procedure [28] was used to estimate the appropriate ceiling value.
- *Within-category dispersion:-* The absolute difference between formant values of individual instances of the vowel, and the mean of formants across all the instances of that vowel. Dispersion values were calculated for formants at both onset/offset *(On-Fn-disp)* and midpoint *(Mid-Fn-disp)*.
- *Relative amplitude (Fn-RA):-* The difference between amplitude of the vowel spectrum at F3, F4 and F5, and the consonant at the corresponding frequency.
- *Spectral tilt (Sp-Tilt):-* The slope of the least-squares regression line fitted after log-transforming the frequency domain of the spectrum.

For the **analysis of consonants**, the feature extraction procedure was identical to the 8 features described in [7]. The features were:- *consonant duration*; *noise duration*; *RMS amplitude*; *peak amplitude*; *peak frequency*; *dynamic amplitude*; and *spectral tilt*. In addition to these features, the present analysis also included *spectral shape* for consonantal analysis. Spectral shape has been described [29] as the difference between the spectral tilts below and above the mid-frequency region. (Tilt$<$ 2.5 kHz - Tilt$>$ 2.5 kHz). All of these features were analyzed separately in their positional contexts (CV, VC), as opposed to their global evaluation in [7].

3.3. Statistical analysis

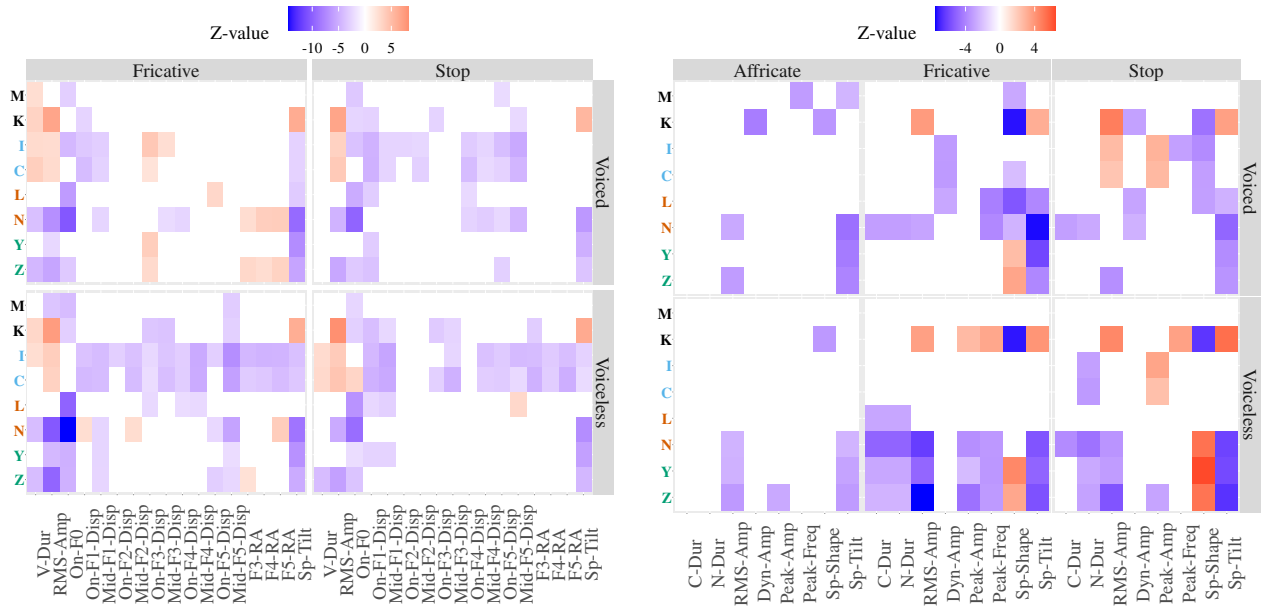
The feature set failed the Shapiro-Wilk test for normality, and also had unequal variances among the groups. Therefore, we decided to use non-parametric tests: the Kruskal-Wallis [30] test, and the Dunn Test [31].

First, systems of each type (hybrid, neural etc.) were grouped together with the natural voice. The **Kruskal-Wallis test** identified those features where significant differences were found within the group, on the basis of that feature. While this was important for identifying the efficacy of each feature, it did not tell us which pairs of groups show significant differences. Therefore, the **Dunn's test**, which is the recommended [32] complementary test was implemented for post-hoc pairwise comparison. This identified which individual system deviated from natural within the group, and by how much. Therefore, the *deviation* from natural is the comparative metric for our analysis.

4. Results

In this section, we describe the results of statistical analysis from the consonantal and vocalic feature-set. Their importance will be discussed in Section 5. We first describe those features and contexts where the greatest differences between the neu-

[1] https://www.synsig.org/index.php/Blizzard_Challenge



(a) *Feature set - vowels in CV context of obstruents*

(b) *Feature set - obstruent consonants in CV context*

Figure 1: *Deviation from natural compared across TTS techniques using Dunn's z-test statistic. White cells indicate no significant difference (p-val> 0.05). TTS Techniques:* ***M K - Hybrid****,* ***I C - HMM****,* ***L N - Unit-Selection****,* ***Y Z - WaveNET***

ral and natural voice can be observed. Then, we compare the WaveNET voices with other systems using deviation from natural voice as the comparative metric. Results in both cases are presented from the CV contexts, as similar trends were observed in VC.

4.1. WaveNET voices vs natural voice - vowels

Figure 1a displays significant deviation from natural across vocalic features, categorized by the voicing and manner of articulation of the obstruent they follow, i.e. fricative and stops. We can see that WaveNET voices performed overall quite well in modelling vowels. System Y in particular shows minimal divergence from the natural voice in each of the voiced and voiceless contexts, across all manners of articulation. **Spectral tilt** is reduced in all contexts of manner and voicing, indicating a broad tendency of neural voices with respect to high-frequency amplitude dampening. This effect is particularly strong for vowels following voiceless fricatives. Here, group differences are shown by the Kruskal-Wallis test [χ^2 = 56.94, p-val $<$ 0.001], while Dunn-test showed that both systems Y and Z exhibit significant differences from the natural voice (p-val$<$0.001). Spectral tilt in system Y drops by a median difference of -3.38 dB/log(Hz) and in system Z, it drops by -2.51 dB/log(Hz).

We also found a significant lowering of the **F0 onset**, for all contexts in system Z, and for voiceless contexts in system Y. This effect is particularly strong for vowels following voiceless fricatives. Here, group differences are shown by the Kruskal-Wallis test [χ^2 = 29.94, p-val $<$ 0.001], while Dunn-test showed that both systems Y and Z exhibit significant differences from the natural voice (p-val$<$0.001). F0 onset in system Y drops by -15.07 Hz and in system Z, it drops by -16.52 Hz.

Finally, we found that **RMS amplitude** impacted the vowels in each context for system Z, and in vowels following fricatives. Once again, vowels in the context of voiceless fricatives show the strongest effect. Here, group differences are shown by the Kruskal-Wallis test [χ^2 = 58.4, p-val $<$ 0.001], while Dunn-test showed that both systems Y and Z exhibit significant differences from the natural voice (p-val$<$0.001). The median drop in amplitude for system Z by -3.74 dB.

4.2. WaveNET voices vs other techniques - vowels

Compared to HMM systems, WaveNET voices show a variety of important improvements. In HMM systems I and C, a blanket lowering of **formant dispersion values** is visible, particularly for vowels following voiceless stops and fricatives. A decrease in dispersion values signifies reduced within-vowel variation, or a shrinkage in the vowel space. This is a strong trend in HMM systems, which has been rectified in WaveNET voices. The second feature of note is **relative amplitude**, i.e, amplitude difference between consonant and vowel at the same frequency. This difference shows a strongly significant decrease in HMM systems (p-val $<$ 0.001), especially for voiceless stops and fricatives. But in WaveNET systems, it resembles the natural voice.

Compared to unit-selection systems, WaveNET systems show clear improvement in terms of F0 onset. Both unit-selection systems L and N show a lowered pitch at the onset, of all the manner and voicing combinations. Although this lowering of F0 can be seen in neural system Z too, the magnitude of lowering is less. The maximum reduction in L and N is -24.6 Hz and -36.1 Hz respectively, while that in Z is only -16.52 Hz.

4.3. WaveNET voices vs natural voice - consonants

Figure 1b displays statistically significant deviation from the natural voice across consonantal features, displayed across manner (affricates, fricatives and stops) and voicing. Here, we see that **voiceless** fricatives and stops show divergence from the natural voice across several features. This trend is visible in both CV and VC contexts, although only CV is displayed here due to space constraints. On the other hand, there is less divergence of features in terms of voiced obstruents. This indicates a broad, overall tendency of neural systems to model characteristics of voiced obstruents better than voiceless ones.

Secondly, we observe significant differences between the

natural voice and neural systems in **spectral tilt**, in all voicing and manner combinations. This means that in neural voices, high frequency regions of the consonants are more damped than they are in the natural voice. Although explicit in every context, this effect is the strongest in voiceless fricatives. Group differences were shown by the Kruskal-Wallis test [χ^2 = 42.06, p-val $<$ 0.001], while Dunn-test showed that both systems Y and Z exhibit significant differences from natural voice (p-val$<$0.001). System Y lowers their spectral tilt by -7.19 dB/log(Hz), while system Z by -8.36 dB/log(Hz).

Next, significant differences were found from the natural voice on the basis of **spectral shape**, particularly in voiced and voiceless fricatives, and in voiceless stops. This means that there is a greater difference between the spectral tilts above and below the mid-frequency range. This phenomenon can be seen most clearly in voiceless stops. Here, group differences were shown by the Kruskal-Wallis test [χ^2 = 47.84, p-val $<$ 0.001], while Dunn-test showed that both systems Y and Z exhibit significant differences from natural voice (p-val$<$0.001). System Y increases their spectral shape by 1.73 dB/Hz, while system Z by 1.35 dB/Hz.

Finally, we observe differences on the basis of **RMS amplitude**, especially for system Z. Voiceless fricatives were the most important site for this difference, where both systems significantly differed from the natural voice. Kruskal-Wallis shows the feature to be significant in this context, [χ^2 = 58.4, p-val $<$ 0.001], while Dunn test shows both systems Y and Z to be significantly different from natural (p-val$<$0.001). RMS amplitude drops in system Y by -2.84 dB and in system Z by -4.43 dB.

4.4. WaveNET voices vs other techniques - consonants

With respect to hybrid system K, both system Y and Z differ less from natural on the basis of **spectral shape**. Systems Y and Z do not differ from natural in affricates, nor in voiced stops. In voiceless fricatives, they show relatively lesser deviation from natural, compared to system K. System K shows a strong lowering of spectral shape, with a median difference of -3.11 dB/Hz, but systems Y and Z show an increase only by 1.59 and 0.98 dB/Hz respectively.

For HMM systems I and C, **voiced stops** show the maximum deviation from natural. RMS amplitude increases (p-val $<$ 0.05), and so does peak amplitude (p-val $<$ 0.05), while spectral shape shows a strong reduction (p-val $<$ 0.001). It appears that the characteristics of voicing during short, transient regions is not reproduced well in HMM systems. In WaveNET voices, we see that these features do not differ from natural.

Compared to unit-selection voices, WaveNET voices show less deviation from natural in noise and consonant durations. Unit-selection systems significantly shorten the consonant duration, across fricatives and stops. In WaveNET voices, we only see this effect for voiceless fricatives. Here, systems Y and Z show a reduction by -8.1 ms and 7.2 ms respectively, while system N reduces the duration by -12.1 ms.

5. Discussion

In the previous section, we analyzed production characteristics of obstruents and their neighbouring vowels across different TTS systems. We found that WaveNET voices deviate from the natural the most in the context of voiceless fricatives. This observation suggests that consonants with a periodic source excitation are modeled more closely to the natural voice than those with an aperiodic excitation. This problem has been identified [33] where WaveNET, when enhanced with separate periodic/aperiodic decomposition, scores higher in naturalness. Since this effect was found in both Tacotron and FastPitch voices, it indicates a vocoder-specific tendency. However, resynthesizing human voice with WaveNET will present a clearer picture in future work.

In human speech, voiceless obstruents raise the F0 of vowels that follow them [11, 34]. This cues the voicing contrast between obstruents (intelligibility), and also contributes to the overall intonation of the utterance (naturalness). In both systems Y and Z, F0 in voiceless obstruents is indeed higher than voiced ones. But, compared to natural, while system Z uniformly lowers the F0, system Y does so only in voiceless contexts. This reduces the necessary difference for cuing the voicing contrast in voiceless fricatives, and may compromise its intelligibility. Additionally, since this deviation in pitch is localized to the voiceless context, it may also negatively influence the overall intonation contour. This can have consequences on the perceived naturalness.

Next, we found that the lowering of spectral tilt is a consistent trend in WaveNET voices across all contexts, both in consonants and vowels. Previous studies have highlighted the importance of flatter spectral tilt on intelligibility [35]. Enhancing strongly negative tilts for voiced frames has resulted in improved naturalness and speaker similarity for synthetic speech [36]. Recent studies on masked speech also suggest a lowering of spectral tilt [37] results in a muffled speech output. Additionally, attributes such as pleasantness have been associated with energy in the high-frequency regions [38]. Therefore, dampening high frequencies may result in degraded perception of voices.

Carefully designed listener tests are needed to firmly establish a relationship between these features of obstruents and perceived attributes of synthetic speech. In the long-term, we envisage that the segmental analysis presented in this paper can be used to detect those system-specific weaknesses that may not be diagnosed by subjective evaluations alone. Additionally, using controllable neural TTS architectures like Wavebender [39], specific locations of distortion can be improved.

6. Conclusions

In this paper, we use acoustic-phonetic properties of obstruent consonants and their neighbouring vowels to compare the WaveNET vocoder with the natural voice and other non-neural systems of BC-2013. We find that WaveNET voices reproduce the characteristics of vowels quite well, showing several improvements from older TTS techniques, in terms of deviation from natural. However, analysis of the consonantal portion reveals that voiceless fricatives show maximal deviation from the natural voice. This indicates that neural voices can model segments with a periodic structure better than noise regions. We also find systematic lowering of spectral tilt across consonants and vowels, indicating that WaveNET vocoders exhibit high-frequency dampening across phonological segments. For future work, we will extend the feature-set to sonorants, and conduct dedicated evaluation tests for various perceived attributes of synthesized speech.

7. 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.

8. References

[1] A. K. Porbadnigk, J.-N. Antons, B. Blankertz, M. S. Treder, R. Schleicher, S. Möller, and G. Curio, "Using erps for assessing the (sub) conscious perception of noise," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 2690–2693.

[2] J.-N. Antons, R. Schleicher, S. Arndt, S. Moller, A. K. Porbadnigk, and G. Curio, "Analyzing speech quality perception using electroencephalography," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 721–731, 2012.

[3] H. C. Nusbaum, A. L. Francis, and A. S. Henly, "Measuring the naturalness of synthetic speech," *International journal of speech technology*, vol. 2, no. 1, pp. 7–19, 1997.

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

[5] M. Vainio, J. Järvikivi, S. Werner, N. Volk, and J. Välikangas, "Effect of prosodic naturalness on segmental acceptability in synthetic speech," in *IEEE Workshop on Speech Synthesis*. Citeseer, 2002, pp. 143–146.

[6] C. Mayo, R. A. Clark, and S. King, "Multidimensional scaling of listener responses to synthetic speech," 2005.

[7] A. Pandey, S. Le Maguer, J. Carson-Berndsen, and N. Harte, "Mind your p's and k's–comparing obstruents across tts voices of the blizzard challenge 2013," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 166–171.

[8] 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

[9] B. Lindblom and I. Maddieson, "Phonetic universals in consonant systems," *Language, speech and mind*, vol. 6278, 1988.

[10] P. Delattre, F. S. Cooper, A. M. Liberman, and L. Gerstman, "Acoustic loci and transitional cues for consonants," *The Journal of the Acoustical Society of America*, vol. 26, no. 1, pp. 137–137, 1954.

[11] J. P. Kirby and D. R. Ladd, "Effects of obstruent voicing on vowel f 0: Evidence from "true voicing" languages," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2400–2411, 2016.

[12] I. Lehiste and G. E. Peterson, "Vowel amplitude and phonemic stress in american english," *The Journal of the Acoustical Society of America*, vol. 31, no. 4, pp. 428–435, 1959.

[13] V. L. Gracco, "Some organizational characteristics of speech movement control," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 1, pp. 4–27, 1994.

[14] H. Kim, K. Martin, M. Hasegawa-Johnson, and A. Perlman, "Frequency of consonant articulation errors in dysarthric speech," *Clinical linguistics & phonetics*, vol. 24, no. 10, pp. 759–770, 2010.

[15] N. Li and P. C. Loizou, "Masking release and the contribution of obstruent consonants on speech recognition in noise by cochlear implant users," *The Journal of the Acoustical Society of America*, vol. 128, no. 3, pp. 1262–1271, 2010.

[16] F. Li and J. B. Allen, "Manipulation of consonants in natural speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 496–504, 2011.

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

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

[19] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *international Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[20] A. Jongman, "Duration of frication noise required for identification of english fricatives," *The Journal of the Acoustical Society of America*, vol. 85, no. 4, pp. 1718–1725, 1989.

[21] E. Chodroff and C. Wilson, "Burst spectrum as a cue for the stop voicing contrast in american english," *The Journal of the Acoustical Society of America*, vol. 136, no. 5, pp. 2762–2772, 2014.

[22] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *The Journal of the Acoustical Society of America*, vol. 64, no. 5, pp. 1358–1368, 1978.

[23] H. M. Sussman, D. Fruchter, and A. Cable, "Locus equations derived from compensatory articulation," *The Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3112–3124, 1995.

[24] C. Redmon, "Lexical acoustics: Linking phonetic systems to the higher-order units they encode," *PhD dissertation, University of Kansas, Lawrence*, 2020.

[25] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of english fricatives," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1252–1263, 2000.

[26] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *International Conference on Speech Communication and Technology (Interspeech)*, 2017, pp. 498–502.

[27] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [computer program]. version 6.0. 37," *RetrievedFebruary*, vol. 3, p. 2018, 2018.

[28] P. Escudero, P. Boersma, A. S. Rauber, and R. A. Bion, "A cross-dialect acoustic description of vowels: Brazilian and european portuguese," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1379–1393, 2009.

[29] V. Evers, H. Reetz, and A. Lahiri, "Crosslinguistic acoustic categorization of sibilants independent of phonological status," *Journal of phonetics*, vol. 26, no. 4, pp. 345–370, 1998.

[30] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[31] O. J. Dunn, "Multiple comparisons using rank sums," *Technometrics*, vol. 6, no. 3, pp. 241–252, 1964.

[32] A. Dinno, "Nonparametric pairwise multiple comparisons in independent groups using dunn's test," *The Stata Journal*, vol. 15, no. 1, pp. 292–300, 2015.

[33] 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.

[34] H. M. Hanson, "Effects of obstruent consonants on fundamental frequency at vowel onset in english," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 425–441, 2009.

[35] Y. Lu and M. Cooke, "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, no. 12, pp. 1253–1262, 2009.

[36] B. Sharma and S. M. Prasanna, "Enhancement of spectral tilt in synthesized speech," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 382–386, 2017.

[37] M. Magee, C. Lewis, G. Noffs, H. Reece, J. C. Chan, C. J. Zaga, C. Paynter, O. Birchall, S. Rojas Azocar, A. Ediriweera *et al.*, "Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols," *The Journal of the Acoustical Society of America*, vol. 148, no. 6, pp. 3562–3568, 2020.

[38] G. VaroSanec-SkariC, "Relation between voice pleasantness and distribution of the spectral energy," in *Proceedings of the XIVth International Congress of Phonetic Sciences (ICPhS 99), San Francisco*, 1999.

[39] G. T. D. Beck, U. Wennberg, Z. Malisz, and G. E. Henter, "Wavebender gan: An architecture for phonetically meaningful speech manipulation," *arXiv preprint arXiv:2202.10973*, 2022.