Can Auditory Nerve models tell us what's different about WaveNet vocoded speech?

Sébastien Le Maguer, Naomi Harte

Sigmedia Lab, ADAPT Centre, EE Engineering, Trinity College Dublin, Ireland

{lemagues, nharte}@tcd.ie

Abstract

Nowadays, synthetic speech is almost indistinguishable from human speech. The remarkable quality is mainly due to the displacing of signal processing based vocoders in favour of neural vocoders and, in particular, the WaveNet architecture. At the same time, speech synthesis evaluation is still facing difficulties in adjusting to these improvements. These difficulties are even more prevalent in the case of objective evaluation methodologies which do not correlate well with human perception. Yet, an often forgotten use of objective evaluation is to uncover prominent differences between speech signals. Such differences are crucial to decipher the improvement introduced by the use of WaveNet. Therefore, abandoning objective evaluation could be a serious mistake. In this paper, we analyze vocoded synthetic speech re-rendered using WaveNet, comparing it to standard vocoded speech. To do so, we objectively compare spectrograms and neurograms, the latter being the output of AN models. The spectrograms allow us to look at the speech production side, and the neurograms relate to the speech perception path. While we were not yet able to pinpoint how WaveNet and WORLD differ, our results suggest that the Mean-Rate (MR) neurograms in particular warrant further investigation.

Index Terms: Speech synthesis analysis, Wavenet, AN model

1. Introduction

In recent years, the advances in speech synthesis have led to a highly realistic synthesized speech signal which is almost indistinguishable from human speech. A crucial step in this evolution was the introduction of WaveNet [1] and its application as a neural vocoder. As several studies [2, 3] showed, using WaveNet significantly improved the MOS scores of the resulting speech compared to signal-processing based vocoders. The impressive results achieved by WaveNet resulted in the displacement of signal-processing based vocoders in favor of neural vocoders. This shift is apparent in the latest edition of the Blizzard Challenge [4] in which a resounding majority of the participating systems include a neural vocoder. In addition, researchers are exploring the replacement of the signal-processing based vocoder in every possible architecture. One detailed study is presented by Wang et al. [5] which proposes a source-filter neural vocoder.

Despite the predominance of neural vocoders in speech synthesis, the number of studies that focus on the analysis of the speech produced by WaveNet remains limited. Vít et al. present an useful analysis which investigates the influence of the training data on the results of WaveNet in [6]. The purpose of this study was to evaluate the robustness of WaveNet on noisy data. In [7]; the authors propose to evaluate multiple combination of vocoders and back-end available in speech synthesis using a MOS test. In [8], the authors compares multiple neural vocoders using a MUSHRA test. Finally, in [2] and in [3], the authors validated their own implementation of a neural vocoder based on WaveNet. To do so, the authors conducted experiments to compare, both objectively and subjectively, the quality of WaveNet to the standard vocoder STRAIGHT [9]. In both studies, the objective evaluation was conducted on spectrograms. However, the results were used for a validation purpose and there is no insight into what WaveNet learns. Thus, it is apparent that there is a lack of investigation into what WaveNet actually learns about speech.

As discussed by Wagner et al. in [10], the evaluation and, hence, the analysis of synthetic speech is no easy task. While subjective evaluation is critical to get human feedback on synthetic speech, the results are difficult to interpret in detail. Furthermore, objective evaluation results do not correlate well with listener tests, and are mainly used for validation purposes. Thus, the current state of the art does not provide protocols to pinpoint the improvement brought about by the introduction of neural vocoders. Yet an alternative use-case of objective evaluation methodologies is currently neglected by the community: identifying key differences to further analyze. While some studies [11] have acknowledged this asset of objective evaluation, they restricted its application to only improving the significance of the subjective evaluation results.

In this paper, we analyse the impact of WaveNet on synthetic speech compared to a standard signal-processing based vocoder: WORLD. To do so, and despite the flaws, we argue that objective evaluation is a good starting point to decipher the improvement from the use of WaveNet in place of signal-processing based vocoders. We conduct the analysis using complementary time-frequency representations of the speech signal. Our hypothesis is that differences prominent in both representations can help focus a subjective analysis in the future. The focus of this paper is the objective comparison.

The first representation used is the log-spectrogram, as it corresponds to the standard input of WaveNet. The second representation is the neurogram, which is the output of an AN model. The research in AN modelling is a crucial field in audiology, with medical applications which require a critical understanding of the hearing process [12]. Thereby, considering the capabilities of AN models, we believe they provide a useful tool to simulate part of the speech perception pipeline. AN models have been applied successfully as an enhancer for the concatenation cost in Unit Selection (US) systems in [13]. In [14], we applied a state-of-the-art AN model to analyze vocoded synthetic speech. For all these reasons, we propose an objective analysis protocol which compares aligned synthetic speech using spectrograms and neurograms on different scales.

2. Experimental protocol

The experimental protocol developed to conduct the analysis is presented in Figure 1. There are three main blocks. The first

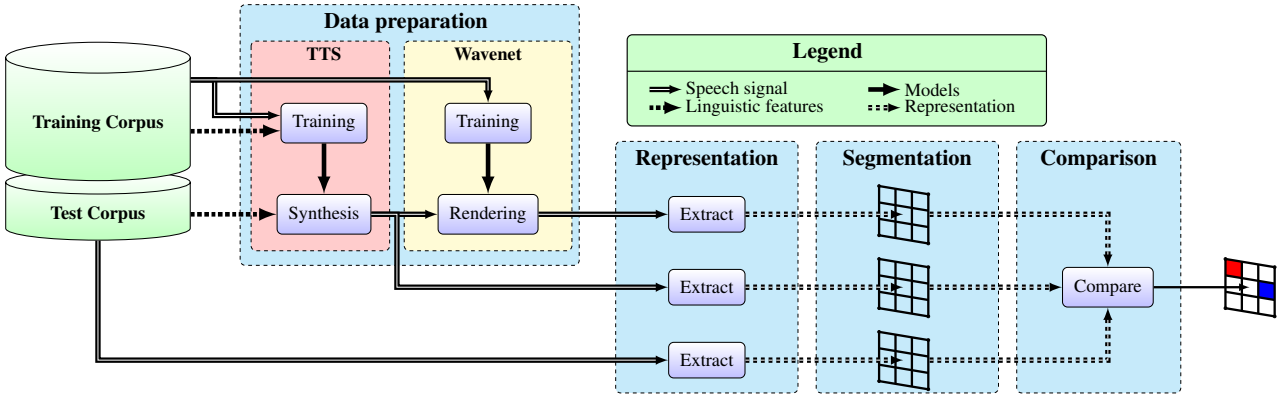


Figure 1: *The evaluation protocol architecture. The "data preparation" block has two stages. First, it trains a TTS model from the training corpus. In parallel, a WaveNet model is trained using only the speech signal from the training corpus. The second stage is the generation of the samples for analysis. Using the linguistic labels of the test corpus, the TTS system synthesizes the vocoded samples. These samples are then re-rendered using wavenet to obtain the second set of samples. The ground truth is the natural samples from the test corpus. From these 3 samples, the "representation" block extracts the time-frequency representations (spectograms and neurograms) used for the analysis. The third block ("segmentation") segments the representations into phonetic segments and frequency bands to focus the analysis. The final "comparison" block computes a set of metrics (RMSE, NSIM) to complete the comparison.*

block *(data preparation)* generates the speech samples for the evaluation. The second block (*representation*), extracts the representations from the speech samples for further analysis. To focus the analysis, the third block (*segmentation*), segments the representation into dedicated areas. Finally, the last block (*comparison*), applies metrics to compare the different representations of the speech. The following sections describe each block in detail, according to their position in the pipeline.

2.1. Data preparation: TTS toolkits and corpora

The first step is to obtain the speech samples. A critical aspect of our analysis is a fully reproducible protocol. Consequently, to synthesize the samples need for the analysis, our study relies on a dataset, as well as toolkits, which are freely available.

The corpus used is the standard CMU ARCTIC dataset [15], specifically the SLT voice. The SLT voice is US female speaker containing 1132 utterances which represents a total of about 50 min of speech sampled at 16 kHz. This corpus is widely used in the community and has the advantage that all the tools used in this study provide dedicated recipes. The only modification we apply to these recipes is to randomly extract 100 files which become the test corpus.

The central tool of this block is Merlin [16]. Merlin is a widely used deep neural network (DNN) backend which provides a full synthesis pipeline from the descriptive features to the signal rendered using WORLD [17]. To get the descriptive features, we used MaryTTS [18] as a frontend. All the features predicted are those proposed in [19] excluding the intonation information at the sentence level (i.e. ToBI tag). Lastly, the implementation of WaveNet is an open-source implementation available on github [20]. WaveNet is trained based on the provided recipe to obtain a Mixture Of logistics (MOL) model [20].

In order to measure the sensitivity of WaveNet and WORLD to back-end behaviour, we condition Merlin by different linguistic descriptive feature subsets. To do so, we followed an incremental approach already proposed in [14]. We distinguish six subsets which can be summarized in two categories. The first category focuses on the influence of the direct phonetic context on the synthetic signals. It comprises the monophone subset (`min`) which is the minimal information; the triphone subset `p3` which incorporates the direct phonetic context (previous and following phoneme); and the quinphone subset (`p5`) which expands the context window of an additional phone in both directions. The second category is based on `p5` and focuses on higher-level features. More specifically it targets the influence of position as well as prosodic features at the syllable level (`p5-sy_full`), at the word level (`p5-wrd_full`) and finally at a phrase and utterance level `full`. We also consider four additional conditions focused on the analysis/synthesis: the natural signal (`nat`); the analysis/synthesis using respectively WaveNet (`wavenet`) and WORLD (`world`); the re-rendering of the condition `world` using wavenet. This last one, identified by `w-wavenet`, is used to investigate what WaveNet can do with the optimal WORLD rendering. All of these conditions are summarized in Table 1.

Table 1: *Descriptive feature sets. "prev." = previous; "re-render." = re-rendering and "AS" for analysis/synthesis. For a detailed description of position and prosodic features, see [19].*

	Identifier	**Description**
	`nat`	Natural signal
	`world`	AS based on WORLD
	`wavenet`	AS based on WaveNet
	`w-wavenet`	WaveNet re-render. of `world`
1	`min`	Current phoneme
2	`p3`	1 + previous/next phonemes
3	`p5`	2 + prev.-prev./next-next phonemes
4	`p5-sy_full`	3 + syllable informations
5	`p5-wrd_full`	4 + word informations
6	`full`	5 + phrase/utterance informations

2.2. Speech representation: spectrograms and neurograms

To investigate which differences are the most salient between speech synthesized using WORLD and its WaveNet re-

rendering, we rely on two time-frequency representations: spectrograms and neurograms.

A spectrogram is the most commonly used time-frequency representation of the speech signal. It is used in speech science for analysis purposes, and in speech technology as an input feature or as a feature to predict. Its widespread usefulness is from the ability to correlate characteristics of the articulatory production of speech units to visual patterns in their acoustic realisation [21]. Furthermore, these patterns are generally localised in some specific frequency sub-bands. For example, the analysis of vowels shows that the first formant position is generally below 1 kHz, while the majority of the energy of some fricatives is condensed in the higher frequencies. Duration properties are also exhibited by the spectrograms, e.g. the patterns related to vowels are longer than the ones related to unvoiced plosives. Consequently, by comparing spectrograms, we are not only able to visualize where two speech signals differ but also interpret this difference to a certain extent. Spectrograms are also commonly used as the input of WaveNet when it is used as a neural vocoder (e.g. Tacotron [22]).

Narrowband spectrograms are computed using a window size of above 30 ms and are commonly used to visualize formants. In contrast, wideband spectrograms are computed with a window size of only a few milliseconds and reveal the harmonic structure of the speech signal [21]. We use both narrowband and wideband log-spectrograms in our analysis. The narrowband spectrogram is computed using a window size of 40 ms, while the wideband spectrogram is computed using a window size of 5 ms.

Neurograms are the time-frequency representation of the output of an AN model. An AN model can be described (simplistically) as a sequence of filters that model the behaviour of the human auditory system from the external ear to the auditory nerve fibres. An AN model generates a Post Stimulus Time Histogram (PSTH) for a set of Characteristic Frequencies (CF) in response to an input audio signal. A neurogram is the 3D representation of the PSTH across time for each CF. Therefore, we can visualize the speech neurogram the same way as we do for a spectrogram.

Multiple AN models have been developed, but, for this study, we are using the model proposed by Bruce et al. [23]. This model is the result of research started over 15 years ago and is still in active development. Furthermore, it is freely available for research purposes[1]. We use the default configuration of this model adapted for the human ear. The signal is converted to an instantaneous pressure waveform, resampled at 100 kHz and the CFs are logarithmically spaced from 250 Hz to 16 kHz. Analogous to spectrograms, the detail in a neurogram is influence by the bin size used to compute the PSTH. As in [14], we use two kind of neurograms: the Fine-Timing (FT) neurograms which include spike timing of the neural responses by using a window size of 320 µs with an overlap of 160 µs; the MR neurograms which give only a mean discharge rate over time with a window size of 12.80 ms and an overlap of 6.40 ms.

Our intention is that the spectrograms will allow us to look at the speech production aspects, and the neurograms relate to the speech perception path.

2.3. Segmentation

With natural speech as our reference, our goal is to identify which part of the speech signal is improved, or simply changed, by WaveNet in comparison to WORLD. Based on the properties of the spectrograms presented in the previous subsection, we divide the image along the time and the frequency axes. For the time axis, we use the phone as a unit as it is the easiest unit to interpret. For the frequency axis, we use fixed subbands from 250 Hz to 8 kHz with steps at 1 kHz, 2 kHz, 4 kHz and 8 kHz. These bands are anchored to exact values in the CF vector and only diverge by couple of hertz from their closest values in the frequency vector associated with the spectrograms. Future work could examine aligning these directly with critical bands.

2.4. Comparison

The last block computes the actual metrics used in the analysis. We use two metrics: the Root Mean Square Error (RMSE) and the Neurogram Similarity Index Measure (NSIM) [24]. The RMSE is used as standard in previous studies [2, 3] to compare spectrograms extracted from speech signals generated by both WaveNet and signal-processing based vocoder. The NSIM approach is slightly different as it evolved from the Structural Similarity Index Measure (SSIM) [25] designed to compare two aligned images. The structural hypothesis which is the foundation of the SSIM is that spatially close pixels have strong interdependencies. Consequently, a Gaussian window is introduced to "look" at each pixel and its context. By employing such a design, this metric is able to exhibit better discrimination than the RMSE [25, 24]. NSIM adapted SSIM to the specificities of a time-frequency speech representation. NSIM has been applied to compare both neurograms [24] and spectrograms [26], and the metric is used in audio quality [27].

3. Results of the analysis

To conduct the analysis, 1032 utterances from the SLT voice (about 45 min) of data were used to train the WaveNet and WORLD systems. The test corpus is composed of 100 test sentences (about 3 min). These synthetic speech samples were each compared to their natural counterpart by applying the NSIM and RMSE metrics to the aligned and segmented neurogram and spectrogram representations.

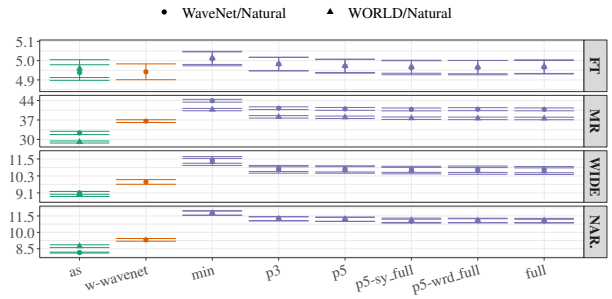
3.1. Global analysis results

The global results are presented in Figure 2. The results using the RMSE[2] to perform the comparison are presented in Figure 2a. The analogous results using the NSIM are presented in Figure 2b. For the RMSE metric, lower is better, whereas for the NSIM metric, closer to 1 is better.

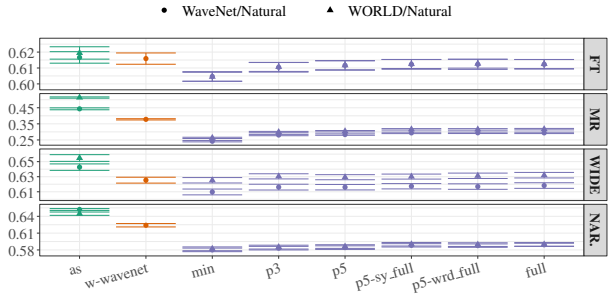
Foremost, if we compare the evolution of the distances along the condition axis (x-axis), we can see that our results are consistent with those presented in [14]. For both metrics, the analysis/synthesis signal (`as`) is more similar to the natural signal than the synthesis conditions. We also can observe a statistically significant improvement from `min` to `p5-sy_full` for the MR neurograms. When applying NSIM to the MR neurograms and the RMSE metric to the wideband spectrograms, we find that WORLD is more similar to the natural voice than WaveNet is. It is notable that the NSIM values associated with the MR neurograms are much lower overall. These values between 0.2 and 0.4 are indicating major differences are found between the natural and synthetic speech. Despite the metric's advantages over RMSE, NSIM may be less sensitive in

[1] The implementation used is available at `http://www.ece.mcmaster.ca/~ibruce/zbcANmodel/zbcANmodel.htm`

[2] For the spectrogram, the y-axis is in dB while for the neurogram it is a count so there is no unit.

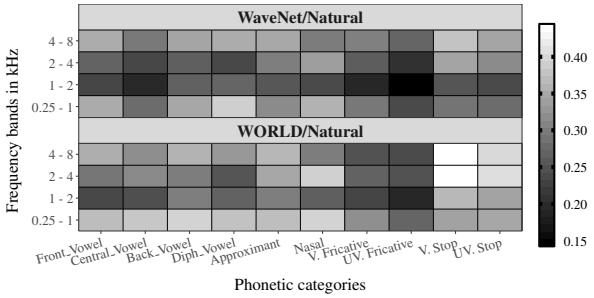


(a) *Root Mean Square Error (RMSE)*

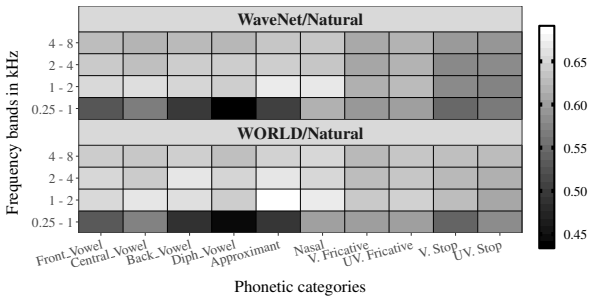


(b) *Neurogram Similarity Index Measure (NSIM)*

Figure 2: *Comparison across test data to the Natural signal*[2]*. FT and MR = FT and MR neurogram; WIDE and NAR. = wideband and narrowband spectrogram; Green (as) denotes analysis/synthesis; red (w-wavenet) denotes resynthesis of WORLD speech using WaveNet; blue denotes the synthesis configurations. Error bars represent the confidence intervals at 95 %.*



(a) *MR neurogram*



(b) *wideband spectrogram*

Figure 3: *Decomposition of NSIM values for MR neurogram representation of the "*full*" condition. Lower NSIM values, i.e. less similar, have darker cells.*

this range [26]. In these MR neurograms, the RMSE is able to flag these large differences from the natural voice, and significant differences between WORLD and WaveNet, with WORLD more similar to the natural speech. The other representation that uncovers significant differences between WORLD and WaveNet is the wideband spectrogram, when comparing signals to natural speech with NSIM. Here WORLD is more similar to the natural speech across all the synthetic conditions.

Overall, these results suggest that key differences between WORLD and WaveNet may occur in the temporal envelope captured in the MR neurograms and the formant structure captured in the wideband spectrogram. The narrowband spectrograms and FT neurograms don't appear to uncover any differences.

3.2. Local analysis

We conducted a focused analysis of the MR neurograms and wideband spectrogram using NSIM. This analysis is broken down into individual phones and frequency subbands. We only consider the `full` synthesis condition for this analysis. Results are presented in Figure 3a for the MR neurograms and in Figure 3b for the wideband spectrogram. Each cell in these matrices represents how different NSIM indicates the synthetic voice is from the natural voice, for that phonetic group, in that specific frequency range.

From Figure 3b, we can see that the primary source of dissimilarities is in the first frequency band for vowels and vocalic consonants. This supports the idea that the wideband spectrogram uncovers differences in formant structure, specifically below 1 kHz.

The MR neurograms in Figure 3a uncover differences across multiple frequency bands. The fricatives exhibit more differences than other phone categories across all frequency bands. The second and third bands (1 kHz to 4 kHz) show the greatest dissimilarities, which are even more pronounced for WaveNet. This suggests than the MR neurograms are more sensitive to the variations of the second and third formants. Thus we have identified MR neurograms and wideband spectrograms as useful representations to signpost interesting properties of synthetic speech.

4. Conclusion

We have designed an objective evaluation protocol and applied it to compare WaveNet rendered speech, and its WORLD counterpart, to natural speech. Frustratingly, initial comparisons show that WORLD yields speech that is more similar as a signal to natural speech than WaveNet. This may contradict MOS tests showing the superiority of WaveNet [7], but reminds us that the overall perceptual effect of a signal is much more complex. Closely reproducing a signal at a local level does not guarantee naturalness. While we were not yet able to pinpoint how these two systems differ, we believe that MR neurograms in particular warrant further investigation as they relate to the perception of speech.

5. Acknowledgements

The authors thank Dr. Michael Wirtzfeld and Prof. Ian Bruce for valuable input regarding the AN model. This research was conducted with the financial support of Irish Research Council (IRC) under Grant Agreement No. 208222/15425 at the ADAPT SFI Research Centre at Trinity College Dublin. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant #13/RC/2106.

6. References

[1] 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

[2] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder." in *Interspeech*, 2017, pp. 1118–1122.

[3] T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "An investigation of subband wavenet vocoder covering entire audible frequency range with limited acoustic features," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5654–5658.

[4] Zhizheng Wu, Zhihang Xie and S. King, "The blizzard challenge 2019," 2019.

[5] X. Wang, S. Takaki, and J. Yamagishi, "Neural Source-filter-based Waveform Model for Statistical Parametric Speech Synthesis," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5916–5920, May 2019.

[6] J. Vít, Z. Hanzlíček, and J. Matoušek, "On the Analysis of Training Data for Wavenet-Based Speech Synthesis," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5684–5688, Apr. 2018.

[7] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4804–4808.

[8] P. Govalkar, J. Fischer, F. Zalkow, and C. Dittmar, "A Comparison of Recent Neural Vocoders for Speech Signal Reconstruction," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 7–12. [Online]. Available: http://dx.doi.org/10.21437/SSW.2019-2

[9] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous frequency-based F0 extraction: Possible role of a repetitive structure in sounds." *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[10] 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.

[11] J. Chevelu, D. Lolive, S. Le Maguer, and D. Guennec, "How to compare TTS systems: A new subjective evaluation methodology focused on differences," in *Interspeech*, 2015.

[12] M. Takanen, I. C. Bruce, and B. U. Seeber, "Phenomenological modelling of electrically stimulated auditory nerve fibers: A review," *Network: Computation in Neural Systems*, vol. 27, no. 2-3, pp. 157–185, 2016, pMID: 27573993. [Online]. Available: https://doi.org/10.1080/0954898X.2016.1219412

[13] J. H. Hansen and D. T. Chappell, "An auditory-based distortion measure with application to concatenative speech synthesis," *IEEE transactions on speech and audio processing*, vol. 6, no. 5, pp. 489–495, 1998.

[14] S. Le Maguer and N. Harte, "Investigation of auditory nerve model based analysis for vocoded speech synthesis," in *Quality of Multimedia EXperience (QoMEX)*, 2020, (in press).

[15] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *Workshop on Speech Synthesis (SSW)*, 2004.

[16] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system." in *Workshop on Speech Synthesis (SSW)*, 2016, pp. 202–207, https://github.com/CSTR-Edinburgh/merlin/tree/master/egs/slt_arctic/s1.

[17] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[18] I. Steiner and S. Le Maguer, "Creating New Language and Voice Components for the Updated MaryTTS Text-to-Speech Synthesis Platform," in *International Conference on Language Resources and Evaluation (LREC)*, 2018.

[19] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to english," in *IEEE Speech Synthesis Workshop*, 2002, pp. 227–230.

[20] R. Yamamoto. Wavenet vocoder. [Online]. Available: https://github.com/r9y9/wavenet_vocoder/tree/master/egs/mol

[21] P. Lieberman and S. E. Blumstein, *Speech physiology, speech perception, and acoustic phonetics*. Cambridge University Press, 1988.

[22] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15–20, 2018.

[23] I. C. Bruce, Y. Erfani, and M. S. Zilany, "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites," *Hearing research*, vol. 360, pp. 40–54, 2018.

[24] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Communication*, vol. 54, no. 2, pp. 306–320, 2012.

[25] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[26] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOL: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–18, 2015.

[27] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An open source production ready objective speech and audio metric," in *Quality of Multimedia EXperience (QoMEX)*. IEEE, 2020, (in press).