10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

A Comparison of Recent Neural Vocoders for Speech Signal Reconstruction

Prachi Govalkar[1], *Johannes Fischer*[1], *Frank Zalkow*[2], *Christian Dittmar*[1]

[1]Fraunhofer IIS, Erlangen, Germany
[2]International Audio Laboratories Erlangen, Germany

{prachi.govalkar;johannes.fischer;christian.dittmar}@iis.fraunhofer.de,
frank.zalkow@audiolabs-erlangen.de

Abstract

In recent years, text-to-speech (TTS) synthesis has benefited from advanced machine learning approaches. Most prominently, since the introduction of the WaveNet architecture, neural vocoders have exhibited superior performance in terms of the naturalness of synthesized speech signals in comparison to traditional vocoders. In this paper, a fair comparison of recent neural vocoders is presented in a signal reconstruction scenario. That means we use such techniques to resynthesize speech waveforms from mel-scaled spectrograms, a compact and generally non-invertible representation of the underlying audio signal. In that context, we conduct listening tests according to the well established MUSHRA standard and compare the attained results to similar studies. Weighing off the perceptual quality to the computational requirements, our findings shall serve as a guideline to both practitioners and researchers in speech synthesis.

Index Terms: speech synthesis, neural vocoder, phase reconstruction, MUSHRA, listening test

1. Introduction

The aim of text-to-speech (TTS) synthesis is to convert a given text into a speech waveform. For many years, the state-of-the art technique for synthesizing natural sounding speech was to select and concatenate short speech segments from a large speech corpus, a technique commonly referred to as concatenative TTS or unit selection [1]. Alternatively, parametric TTS systems [2] try to predict acoustic speech features by employing machine learning techniques.

These features could be general time-frequency (TF) representations of speech or specialized *vocoder* control parameters. In this context, a vocoder is a signal processing system designed to synthesize the speech waveform from the feature representation. Classic vocoder parameters are motivated by an underlying speech production model and comprise suitable encodings of the fundamental frequency, spectral envelope and others. Usually, the feature sequences are defined on a much coarser temporal scale than the target audio signal. While parametric TTS approaches can synthesize intelligible and prosodically correct speech features, the attainable sound quality is often limited by the vocoder algorithm.

As in other fields of speech processing, TTS also has rapidly advanced with the advent of deep learning techniques. In their seminal paper [3], van den Oord et al. introduced the WaveNet architecture, a *neural vocoder* that predicts the speech waveform from past signal samples and can be controlled (conditioned) by above-mentioned feature representations. The key idea of such a neural vocoder is to implement an autoregressive (AR) probabilistic model that allows to predict a probability distribution of current waveform samples given previous samples. Although AR models can generate speech signals of high perceptual quality and naturalness, their generation speed is slow because they have to generate waveform samples in sequential manner. As a consequence, several other authors proposed alternative architectures, which we will briefly describe in Section 3.

Remarkably, there is relatively little work on using more general signal reconstruction techniques in case the acoustic features are TF representations, such as magnitude spectrograms or perceptually motivated representations (e.g., mel-scaled spectrograms). Often, the classic phase estimation method proposed by Griffin and Lim in [4] is used without further investigation. However, more elaborate methods have been proposed in recent years that might be more suitable for speech signal reconstruction tasks. We will discuss two such methods, which are not based on deep learning but classical signal processing in Section 4.

In Section 5, we present an experimental study that compares six neural vocoder methods and two phase reconstruction methods with respect to the perceptual quality of the synthesized speech signals. Recently, other researchers have published similar evaluations in which they compare different vocoder methods with respect to the attainable perceptual quality. For example, the authors of [5] conducted listening tests with classical signal-processing based vocoders. In [6], neural vocoders (WaveNet) were included in the benchmark. As is common in TTS research, mean opinion score (MOS) ratings were reported in those studies. Moreover, both papers also incorporated the prediciton of acoustic speech features into the test. In this paper, our experimental approach differs in two important aspects. First, we exclusively use speech features derived directly from natural speech recordings (this is sometimes referred to as copy synthesis). Second, we use the MUSHRA standard for conducting the listening tests and evaluating the results [7]. Similar attempts have been made by the authors of [8], who compare WaveNet against well-known speech codecs. Moreover, the authors of [9] evaluate the perceptual quality of two neural vocoder approaches against each other. In this contribution, we extend prior work by including six neural vocoder variants as well as two phase reconstruction methods into the test. This way, we try to provide a fair and neutral comparison of state-of-the-art approaches with respect to their synthesis quality as well as their computational requirements.

2. Task Definition

In this section, we formalize our task of speech signal reconstruction. This goes alongside the overview in Figure 1. We consider the real-valued, discrete time-domain signal $x : \mathbb{Z} \to \mathbb{R}$ to contain a natural speech recording as shown in Figure 1. For

10.21437/SSW.2019-2

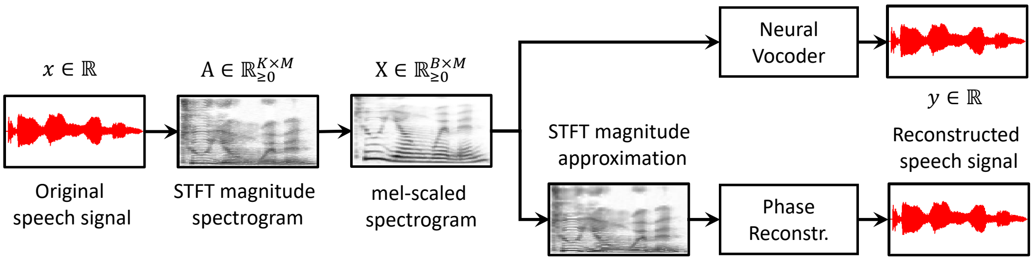
$x \in \mathbb{R}$
$\mathrm{A} \in \mathbb{R}_{\geq 0}^{K \times M}$
$\mathrm{X} \in \mathbb{R}_{\geq 0}^{B \times M}$
Neural Vocoder
$y \in \mathbb{R}$
Reconstructed speech signal
STFT magnitude approximation
Original speech signal
STFT magnitude spectrogram
mel-scaled spectrogram
Phase Reconstr.

Figure 1: *Overview of the signal processing flow employed in this paper for speech signal reconstruction.*

reasons explained later, x could optionally undergo a non-linear pre-processing such as μ-law compression and quantization. From x, we derive its mel-scaled spectrogram, a non-invertible, compressed TF representation that is widely used in audio and speech processing.

To this end, we transform x to the Short-Time Fourier Transform (STFT) domain. Let $\mathrm{A}\,(k, m)$ be the non-negative STFT magnitude at the k^{th} frequency bin and the m^{th} time frame, with $k \in [0 : K - 1]$ and $m \in [0 : M - 1]$. The number of frequency bins $K \in \mathbb{N}$ and frames $M \in \mathbb{N}$ determines the dimension of the magnitude spectrogram matrix $\mathrm{A} \in \mathbb{R}_{\geq 0}^{K \times M}$. Thereby, K depends on the STFT block size as $K = 1 + N/2$, with $N \in \mathbb{N}$. The number of frames M is determined by the number of signal samples in conjunction with the STFT hopsize $H \in \mathbb{N}$.

At this point, we would like to stress that we already introduce a first stage of information compression by disregarding the phase spectrogram, effectively throwing away all the information that is necessary for quasi-perfect signal reconstruction via the inverse STFT [4]. However, as will be detailed in Section 4, there are possibilities to recover from this situation via phase reconstruction methods.

From the magnitude spectrogram matrix A, it is straightforward to compute the mel-scaled [10] spectrogram matrix $\mathrm{X} \in \mathbb{R}_{\geq 0}^{B \times M}$ as

$$\mathrm{X} := \mathrm{M} \cdot \mathrm{A}, \tag{1}$$

with $\mathrm{M} \in \mathbb{R}_{\geq 0}^{B \times K}$ being a suitable transformation matrix that contains $B \in \mathbb{N}$ rows with the typical, triangular shaped mel-filter weights. Obviously, mapping the spectral content in A to the mel-scaled version in X is a non-invertible process in which we compress information for the second time due to weighted summation of several frequency bins into one mel-filter band. This is especially pronounced for the high frequency content. The great achievement of neural vocoder methods (see Section 3) is that they can consume such a strongly compressed (both in time and frequency) signal representation as conditioning information and still reconstruct a plausible speech waveform y.

If we do not use neural vocoders, we can still try to revert the loss of data by approximating the inverse mapping from mel-scale to linear frequency spacing as

$$\mathrm{A} \approx \mathrm{M}^{\dagger} \cdot \mathrm{X}, \tag{2}$$

with $\mathrm{M}^{\dagger} \in \mathbb{R}^{K \times B}$ being the pseudo-inverse of M. Doing so typically introduces negative values in the approximated magnitude STFT matrix. As a simple remedy to this problem we apply half-wave rectification to each TF bin (similar to a ReLU nonlinearity). As shown in the lower right signal flow of Figure 1, the resulting spectrogram normally looks smoothed in the upper frequency region but can potentially still be used for phase reconstruction followed by an inverse STFT to reconstruct the time-domain speech waveform y.

In the preceding paragraphs we have described the different signal representations involved in our speech signal reconstruction task. The following sections will provide a brief overview of the different methods we employ for our experiments, both from the neural vocoder perspective and the phase reconstruction perspective.

3. Neural Vocoder Methods

A common feature of the different neural vocoder methods used in this study is that they can be conditioned with acoustic features (e.g., mel-scaled spectrograms) described earlier. As the feature matrix runs on a much lower temporal resolution than the target waveform, it is necessary to first upsample the conditioning information. It has been proposed to either replicate the vectors across time or to learn an upsampling kernel in the sense of transposed convolutional layers. Moreover, the representation of the target waveform may differ across systems, since some algorithms rely on μ-law compression and quantization, while others operate on the raw waveform.

3.1. WaveNet

The autoregressive WaveNet (henceforth abbreviated as WNET) [3] learns to predict realistic speech waveforms based on past waveform samples. The architecture is based on stacks of causal, dilated convolutional layers to achieve a wide receptive field. In the original publication, it was proposed to train WaveNet to predict one-hot encoded μ-law quantized waveforms. In follow-up work [13], it was proposed to use a mixture of logistic distributions from which to draw the current output sample.

3.2. nv-WaveNet

The nv-WaveNet (NVWN) is a reference implementation of a CUDA-enabled autoregressive WaveNet inference engine [14]. In particular, it implements the WaveNet variant described in [15], which learns to predict 8 bit μ-law quantized waveforms. The big advantage over the vanilla WaveNet is the ability to generate signals faster than real-time. Depending on the choice of hyper parameters, this advantage might come at the cost of inferior audio quality.

Year	References	Abbr.	Open-source implementation (implementation by original authors in bold font)	Real-time factor	Signal representation	Pre-trained
2015	Beauregard et al. [11]	SPSI	**https://anclab.org/software/phaserecon**	1.08	mel-Spectrogram	-
2016	Pruša et al. [12]	PGHI	**https://github.com/ltfat/phaseret**	81.25	mel-Spectrogram	-
2016	van den Oord et al. [3, 13]	WNET	https://github.com/r9y9/wavenet_vocoder	0.0032	mel-Spectrogram	Yes
2018	Pharris [14], Arik et al. [15]	NVWN	**https://github.com/NVIDIA/nv-wavenet**	1.064	mel-Spectrogram	No
2018	Jin et al. [16]	FFTN	https://github.com/azraelkuan/FFTNet	0.0238	mel-Spectrogram	No
2018	Kalchbrenner et al. [17]	WRNN	https://github.com/fatchord/WaveRNN	0.072	mel-Spectrogram	No
2018	Prenger et al. [18]	WGLO	**https://github.com/NVIDIA/waveglow**	6.46	mel-Spectrogram	Yes
2019	Valin & Skoglund [9]	LPCN	**https://github.com/mozilla/LPCNet**	8.19	Bark-Cepstrum & f_0	No

Table 1: *Overview of the speech signal reconstruction methods that we compare in our listening test. The real-time factor indicates how much faster than real-time each method could synthesize. It is based on the average synthesis time of the first 10 test items in the LJ Speech corpus [19] for each algorithm on a single Nvidia GTX 1080 Ti GPU. Note that both SPSI and PGHI do not make use of GPU computations. Furthermore, SPSI was used for phase initialization before running 60 additional GL iterations which counted into the synthesis time.*

3.3. FFTNet

The FFTNet architecture (FFTN) is also inspired by WaveNet but uses a shuffling of the audio data as in an FFT butterfly graph [16]. It is reported by the authors to yield faster generation time than WaveNet and can achieve comparable quality, if a noise-reduction post-processing is applied. Other tricks such as changing the temperature of the distribution sampling in a signal-dependent way has also been reported to be beneficial.

3.4. WaveRNN

The architecture of WaveRNN (WRNN) consists of a single layer recurrent neural network with a dual softmax layer [17]. A generation scheme based on subscaling folds a long sequence into a batch of shorter sequences and thus can generate multiple samples at once. Furthermore, sparsification of the weight matrices has been proposed as a means for increased efficiency.

3.5. LPCNet

The LPCNet [9] (LPCN) is a WaveRNN variant that combines linear prediction with recurrent neural networks to synthesize audio with almost three times lower complexity than WaveRNN for comparable network sizes. The conditioning parameters are obtained using 20 features (Bark-cepstrum and pitch) from input of the synthesis, which pass through a series of convolutional and fully-connected layers. The target waveform is also 8 bit μ-law quantized, but an additional linear pre-emphasis filter leads to better masking of high-frequency quantization noise.

3.6. WaveGlow

One approach to accelerate the generation speed is to change an AR model into an inverse-AR one. For example, inverse-autoregressive-flow (IAF) can be used to transform a noise sequence into a speech waveform without the sequential generation process. However, in order to learn the parameters of such an IAF model, the waveform must be sequentially transformed into a whitened, noise-like signal. WaveGlow [18] (WGLOW) is a combination of flow-based neural networks and autoregressive models. It is inspired from the flow-based approach of Glow [20] and the simplicity of the autoregressive WaveNet [3] architecture. The flow-based generative model provides tractability of exact log-likelihood and efficiently parallelizes both training and inference.

4. Phase Reconstruction Methods

In their well-known paper [4], Griffin and Lim proposed the so-called LSEE-MSTFTM (often called GL) algorithm for iterative, blind signal reconstruction from magnitude spectrograms. According to [21], LSEE-MSTFTM belongs to a class of algorithms called Projection onto Convex Sets (POCS), also known as *Alternating Projections*. Sturmel and Daudet [22] provided an in-depth review of iterative phase reconstruction methods and pointed out that convergence problems often result from the random initialization of the phase spectrogram. Since the iterative phase estimates can only converge to local optima, several publications were concerned with finding a good initial estimate for the phase information [23, 24]. The authors of [25] showed that constraining the intermediate signal reconstructions by means of a step-like target envelope could lead to improved reconstruction for transient signals.

4.1. Single Pass Spectrogram Inversion

The phase-locked vocoder is a signal analysis and synthesis approach that is already several decades old [26]. Two key ingredients of this method are the estimation of instantaneous frequencies of sinusoidals and phase locking around those components [27]. The authors of [11] picked up the idea of the phase-locked vocoder for obtaining initial phase spectrogram estimates. Their corresponding Single Pass Spectrogram Inversion (SPSI) algorithm is relatively straightforward to implement. Although the underlying signal model is too simplistic for speech signals, it can still serve as a very efficient baseline method in conjunction with GL iterations. For this paper, we used 60 GL iterations as this number has been reported by other authors as well [18].

4.2. Phase Gradient Heap Integration

The Phase Gradient Heap Integration (PGHI) method was proposed in [12]. As a key difference to SPSI, it uses estimates of the instantaneous frequency as well as instantaneous time (i.e., the full phase gradient) for phase reconstruction. Furthermore, PGHI exploits the principle that the phase gradient can be approximated by the log-magnitude gradient under certain conditions. A clever gradient integration scheme based on self-sorting lists is used to implement this algorithm in a very efficient way. Although PGHI lends itself to be used as phase initialization for refinement with GL iterations, we use its output directly for signal reconstruction as the improvement is usually only minor.

5. Quality Evaluation

As has been reported in [8], perceptually-motivated quality metrics such as PESQ and POLQA are problematic to adequately evaluate neural vocoders. Thus, we conducted a subjective listening test following the MUSHRA methodology for this paper. In that test, five utterances by the same female speaker were rated by 20 participants. A MUSHRA compliant web audio API based open source software called webMUSHRA [28] was used to carry out the listening tests. In the following we provide more details about the experimental settings and results.

5.1. Corpus and Features

In this work, we used the well-known LJ Speech corpus [19]. This public domain dataset comprises almost 24 h of speech recordings by a single female speaker. In total, there are $13,100$ utterances with an average length of 6.6 s. The content of these recordings includes passages from seven non-fiction books. Although the recordings exhibit a certain degree of room reverberation, it is still used by many TTS researchers.

For training of the different neural vocoder methods, we used almost the entirety of the LJ Speech corpus. We only held out the first ten items from the corpus for measuring the time needed for synthesis (see Table 1). The original recordings are encoded as 16 bit PCM audio, with a sampling rate of 22.05 kHz. For our MUSHRA listening test, we resampled all utterances to 16 kHz since this was the upper limit that some of the algorithms under test allowed for. As described in Section 2, the acoustic features we extracted are mel-scaled spectrograms. The dimension of these feature vectors was $B := 80$ mel-filter bands, extracted at a feature rate of 80 Hz (12.5 ms hopsize, $H := 200$), with an underlying STFT frequency resolution of 20 Hz (50 ms blocksize, $N := 800$).

Only for training the LPCNet, we used a different set of features. As specified in the paper [9] and the reference implementation of the original authors, the conditioning features consisted of 18 Bark-scale cepstral coefficient and two features representing the period duration and the correlation of fundamental frequency estimate per frame. The feature rate in this case was 100 Hz (10 ms hopsize, $H := 160$).

5.2. Listening Test Method

We carried out the listening tests using the MUSHRA methodology (MUltiple Stimuli with Hidden Reference and Anchor) following the ITU-R BS.1534-3 recommendation [7]. During the MUSHRA test, participants are asked to rate the basic audio quality for a set of processed signals with respect to the reference signal. The grading scale is continuous from 0 to 100 with 5 categories: Bad (0–20), Poor (20–40), Fair (40–60), Good (60–80) and Excellent (80–100). It is a double-blind multi-stimulus test method with a *hidden reference* and an additional *anchor* signal. This standard anchor is a low-pass filtered version of the original signal with a cut-off frequency of 3.5 kHz. According to the ITU-R BS.1534-3 recommendation, there should be two phases in the test, a training phase and a testing phase. The training phase allows the participant to familiarize with all the different audio quality ranges as well as to learn to use the test equipment. Only the grades given during the test phase are taken into consideration for the final evaluation.

5.3. Test setup

In our tests, we only used the first six items instead of all ten test items to avoid fatiguing of listeners and to reduce the total duration of the listening tests. The six items were further divided into: one for listener training phase, five for testing phase. The names of our test phase items with their respective conditions are shown in Figure 2. For each item 11 conditions had to be evaluated. These included the hidden reference (REF) and the low-anchor (Anchor35) signal. Similar to [9], we also included a 8 bit μ-law quantized version of reference signal as additional condition (8bit μ-law REF). This is to assess how much the resulting quantization noise matters to the listeners, as it is likely to be present in the reconstruction of any method that uses this kind of waveform representation (such as LPCN and NVWN). The remaining 8 conditions consisted of a mixture of phase reconstruction and neural vocoder methods as laid out in Table 1. Neural vocoder methods like WNET, FFTN and WGLO have been trained on the LJ Speech Corpus with a samping rate of 22.05 kHz as compared to NVWN, WRNN and LPCN at 16 kHz. We provide a table on our accompanying webpage[1], enlisting the different training hyperparameters for these methods. In order to have a fair comparison of real-time factors for neural vocoder methods, we multiply the WNET, FFTN and WGLO factors by 1.38 (22.05 kHz/16 kHz) to overcome their disadvantage of having to synthesize more samples per second.

As required in the MUSHRA test, the participants needed to switch between the different signals near instantaneouly to allow for a fine-grained comparison between the systems. The synthesized signals obtained from various implementations for our experiments often exhibited different number of samples. Hence it was important to align the signals in order to allow seamless switching between them in the MUSHRA test. This temporal alignment was achieved by converting both the synthesized and the reference signal to the magnitude STFT domain with an extremely small hopsize of $H := 1$ sample and subsequently shifting the synthesized signal along the temporal axis of reference signal. Since phase differences are ignored in this procedure, the position of the minimal Euclidean distance could then be used to shift the synthesized time-domain signal to the optimal alignment in time. In order to achieve normalization of volume, all signals under test were normalized with respect to the root mean square level of the reference signal.

5.4. Selection of Participants

The tests were performed by 20 participants (15 male and 5 female), with an average age of 28.8 (standard deviation of

[1] https://www.audiolabs-erlangen.de/resources/NLUI/2019-SSW-NeuralVocoders/

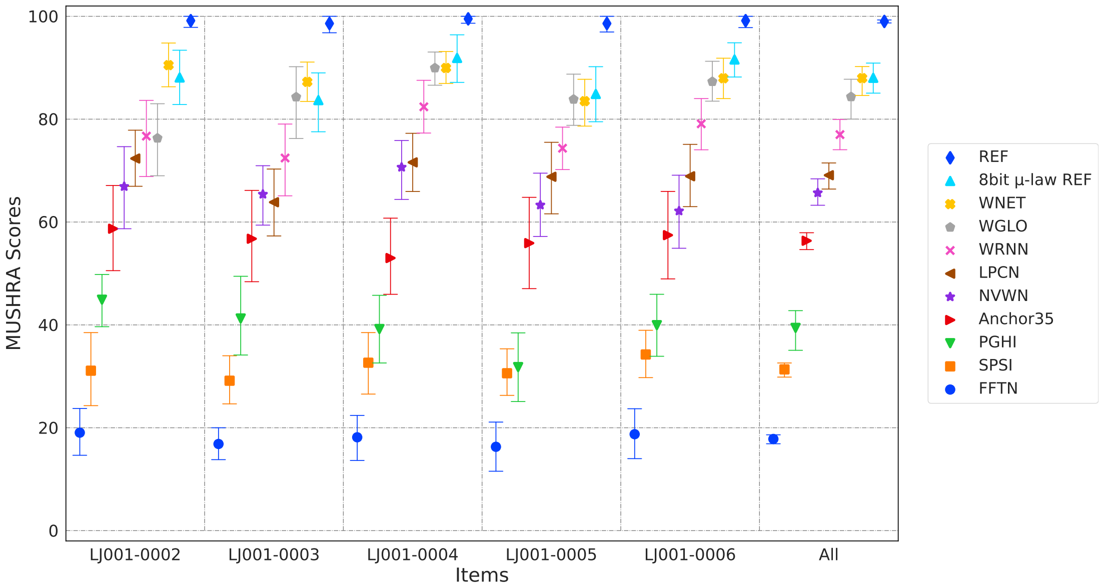

Figure 2: *Overview of the experimental results from our MUSHRA listening test. The colored markers represent the mean ratings, enclosed by the* 95% *confidence intervals.*

4.7). The participants included 9 experienced listeners and 11 inexperienced listeners. The average time taken by the participants to perform the listening test (including training phase) was 25 min. The MUSHRA test recommends post-screening of participants where a participant should be excluded if he or she rates the hidden reference condition for greater than 15% of the test items, a score lower than 90. No partipant was excluded in the post-screening process concerning our tests.

5.5. Results

The results of our MUSHRA test are presented in Figure 2. The first five columns represent the individual ratings per test item, the final column gives the averaged results. We always show mean value and the 95% confidence intervals. As can be seen, the ranking of different systems under test is quite consistent accross utterances. As also reported in [8], WNET is rated slightly below the reference and is on par with the 8 bit μ-law quantization. Remarkably, for some items WGLO achieves comparable ratings as WNET at a much higher synthesis speed. The slightly lower ratings for WRNN and the slower than real-time synthesis speed are probably due to the open source implementation which differed in various aspects from the original paper [17].

Our results for LPCN are approximately 10 MUSHRA points lower from what has been reported in [9], possibly due to the exposure of our listeners to a wider range of conditions. Still, LPCN is a worthwile approach since it allows for synthesis eight times faster than real-time and uses only one quarter of the conditioning information provided to the other methods. We find that NVWN delivers almost the same quality as LPCN, at the cost of slower synthesis speed.

On the lower end of the quality range, we have PGHI, SPSI and FFTN. The methods based on phase reconstruction have a clear disadvantage in that they do not have the capability to learn signal specific features. Most of the artifacts they introduce are caused by the pseudo-inverse mapping from mel-scaled to linearly-spaced spectrogram as described in Section 2. For the FFTN, we assume that the open source implementation under test performed so poorly because it was not by the original authors and did not include the additional improvements suggested in the paper [16].

To make these numbers more comprehensible, we provide one item from our test set in all different conditions on our accompanying webpage[1]. With the integrated player, it is possible to switch seamlessly between the different conditions and get a visualization of the corresponding mel-scaled spectrogram at the same time. To justify our above statement that the phase reconstruction methods were mostly affected by the pseudo-inverse mapping, we added a second player with audio examples synthesized using PGHI and SPSI. For those items, both methods were applied on the magnitude spectrogram A, i.e., omitting the round trip to the mel-scale and back. As can be heard, the quality is better, with only minor degradations. However, we did not include those items in the listening test since the feature representation would be drastically different to the other methods, rendering the comparison unfair.

6. Conclusions

In this paper, we presented the results of a listening test experiment with neural vocoder and phase reconstruction methods. Our aim was to provide a neutral overview of the subjective audio quality that can be obtained with different approaches. Overall, only two methods (WNET, WGLO) achieved excellent ratings, whereas three neural vocoder methods (WRNN, LPCN, NVWN) achieved good MUSHRA ratings. Three among these leading approaches have the potential to synthesize in real-time or faster (LPCN, WGLO, NVWN).

The phase reconstruction methods (PGHI, SPSI) achieved poor to fair MUSHRA ratings, mainly due to artifacts introduced by the pseudo-inverse mapping from the mel-scale. However, they are still worth considering when synthesis much faster than real-time has to be achieved.

7. Acknowledgements

This work is part of the SPEAKER project (FKZ 01MK19011), funded by the German Federal Ministry for Economic Affairs and Energy. The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

8. References

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing Conference (ICASSP)*, Atlanta, Georgia, USA, May 1996, pp. 373–376.

[2] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System," in *Proceedings of the ISCA Speech Synthesis Workshop*, Sunnyvale, CA, USA, September 2016, pp. 202–207.

[3] 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," in *Proceedings of the ISCA Speech Synthesis Workshop*, Sunnyvale, CA, USA, September 2016, pp. 125–125.

[4] D. W. Griffin and J. S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[5] M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "A Comparison Between STRAIGHT, Glottal, and Sinusoidal Vocoding in Statistical Parametric Speech Synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1658–1670, 2018.

[6] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, April 2018, pp. 4804–4808.

[7] International Telecommunications Union, "ITU-R Rec. BS.1534-3: Method for the subjective assessment of intermediate quality levels of coding systems," 2015.

[8] W. B. Kleijn, F. S. C. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, "WaveNet Based Low Rate Speech Coding," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, April 2018, pp. 676–680.

[9] J. Valin and J. Skoglund, "LPCNet: Improving Neural Speech Synthesis Through Linear Prediction," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 5891–5895.

[10] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Springer, 2007.

[11] G. T. Beauregard, M. Harish, and L. Wyse, "Single Pass Spectrogram Inversion," in *Proceedings of the IEEE International Conference on Digital Signal Processing (DSP)*, Singapore, July 2015, pp. 427–431.

[12] Z. Průša, P. Balázs, and P. L. Søndergaard, "A Noniterative Method for Reconstruction of Phase From STFT Magnitude," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1154–1164, 2017.

[13] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel WaveNet: Fast High-Fidelity Speech Synthesis," in *Proceedings of the International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018, pp. 3915–3923.

[14] B. Pharris, "Nv-WaveNet: Better Speech Synthesis Using GPU-Enabled WaveNet Inference," https://devblogs.nvidia.com/nv-wavenet-gpu-speech-synthesis/, 2018.

[15] S. Ö. Arik, M. Chrzanowski, A. Coates, G. F. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Y. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep Voice: Real-time Neural Text-to-Speech," in *Proceedings of the International Conference on Machine Learning (ICML)*, Sydney, NSW, Australia, 2017, pp. 195–204.

[16] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "FFTNet: A Real-Time Speaker-Dependent Neural Vocoder," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, April 2018, pp. 2251–2255.

[17] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient Neural Audio Synthesis," in *Proceedings of the International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018, pp. 2415–2424.

[18] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 3617–3621.

[19] K. Ito, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[20] D. P. Kingma and P. Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, Montréal, Canada, December 2018, pp. 10 215–10 224.

[21] K. Jaganathan, Y. C. Eldar, and B. Hassibi, "STFT Phase Retrieval: Uniqueness Guarantees and Recovery Algorithms," *Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 770–781, 2016.

[22] N. Sturmel and L. Daudet, "Signal Reconstruction from STFT magnitude : A State of the Art," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Paris, France, September 2011, pp. 375–386.

[23] X. Zhu, G. T. Beauregard, and L. L. Wyse, "Real-Time Signal Estimation From Modified Short-Time Fourier Transform Magnitude Spectra," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1645–1653, July 2007.

[24] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Phase Initialization Schemes for Faster Spectrogram-Consistency-Based Signal Reconstruction," in *Proceedings of the Acoustical Society of Japan Autumn Meeting*, September 2010, pp. 601–602.

[25] C. Dittmar and M. Müller, "Towards Transient Restoration in Score-informed Audio Decomposition," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Trondheim, Norway, December 2015, pp. 145–152.

[26] M. R. Portnoff, "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 243–248, 1976.

[27] M. Puckette, "Phase-locked Vocoder," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, October 1995, pp. 222–225.

[28] M. Schoeffler, F.-R. Stöter, B. Edler, and J. Herre, "Towards the Next Generation of Web-based Experiments: A Case Study Assessing Basic Audio Quality Following the ITU-R Recommendation BS.1534 (MUSHRA)," in *Proceedings of the Web Audio Conference*, Paris, France, January 2015.