ABSTRACT
This paper introduces a robust universal neural vocoder trained with 74 speakers (comprised of both genders) coming from 17 languages. This vocoder is shown to be capable of generating speech of consistently good quality (98% relative mean MUSHRA when compared to natural speech) regardless of whether the input spectrogram comes from a speaker, style or recording condition seen during training or from an out-of-domain scenario.

Together with the system, we present a full text-to-speech analysis of robustness of a number of implemented systems. The complexity of systems tested range from a convolutional neural networks-based system conditioned on linguistics to a recurrent neural networks-based system conditioned on mel-spectrograms. The analysis shows that convolutional neural networks-based systems are prone to occasional instabilities, while the recurrent approaches are significantly more stable and capable of providing universalizing robustness.

Index Terms—Neural Vocoder, Text-to-speech, Scalability, Statistical Waveform Speech Synthesis

1. INTRODUCTION
Statistical parametric speech synthesis (SPSS) has seen a paradigm change recently, mainly thanks to the introduction of a number of autoregressive models [1, 2, 3, 4, 5], turning into what could be called statistical speech waveform synthesis (SSWS). This change has closed the gap in naturalness between statistical text to speech (TTS) and natural recordings, maintaining the flexibility of statistical models.

In the case of traditional vocoding [6, 7, 8, 9], approaches commonly relied on simplified models (e.g. source-filter model [10]) that were defined by acoustic features such as the voicing decisions, the fundamental frequency (F0), mel-generalized cepstrum (MGC) or band aperiodicities. The quality of those traditional vocoders was limited by the assumptions made by the underlying model and the difficulty to accurately estimate the features from the speech signal.

Traditional waveform generation algorithms, while capable of generating speech from their spectral representation such as Griffin-Lim [11], are not capable of generating speech with acceptable naturalness. This is due to the lack of phase information in the short-time Fourier transform (STFT).

Neural vocoders are a data-driven method where neural networks learn to reconstruct an audio waveform from acoustic features [1, 2, 12]. They allow us to overcome the shortcomings of traditional methods [13] at a very significant cost in computation power and data requirements. However, due to sparsity (it is unlikely that we will ever be able to cover all possible human-generated sounds) the neural vocoder models are prone to over-fit to the training speaker characteristics and have poor generalization capabilities [14]. Several recent studies attempted to improve the adaptation capabilities of such models [15, 16], commonly using explicit speaker information (either as a onehot encoding or some other form of speaker embedding) [17]. There are however reports in literature of initial successes training neural vocoders without providing explicit speaker information [18], however the investigation did not cover the details of the robustness of this model to changes in domain and to unseen speakers.

This paper makes the following contributions: 1) it demonstrates that a speaker encoding is not required to train a top quality speaker-independent (SI) neural vocoder; 2) quality with a SI neural vocoder is as good, if not better than, that of a speaker dependent (SD); 3) with the SI neural vocoder we can synthesise with speakers that were unseen in the training data, which is not possible with vocoders trained with explicit speaker information. The SI vocoder in this paper appears to be very effective at reading mel spectrograms and producing the corresponding speech waveform, without requiring any supplementary information, behaving as a universal vocoder.

2. SYSTEMS DESCRIPTION
Even though CNN-based systems have been thoroughly researched and real-time implementations have been proposed [4, 19], it is known that they are prone to instabilities [20] which occasionally affect perceptual quality. RNN-based systems, on the other hand, can be expected to provide a more stable output due to the persistence of their hidden state, at least for the vocoding task in which context is not critical beyond the closest spectrogram frames (a known characteristic of RNNs). Table 1 overviews the systems. The linguistic RNN system was not up to par, so it was discarded.
Table 1. Description of the CNN and RNN SSWS systems.

| System | CNNLI | CNN_MS | SCNN_MS | RNN_MS |
|--------|-------|--------|---------|--------|
| Conditioning | Linguistics | Mel-spectrograms | Convolutional | Recurrent |
| Layer type | | | 4 blocks of 10 | 4 blocks of 5 |
| Layer structure | Convolutional | Recurrent | 1 |
| Parameters | 11.6M | 11.5M | 7.5M | 9.5M |
| Components | 198 | 130 | 70 | 5 |

Fig. 1. Block diagram of system RNN_MS

2.1. RNN models

The structure of system RNN_MS (heavily inspired by WaveRNN [2], only with minor changes in the conditioning network) is described in Figure 1. The autoregressive side consists of a single forward GRU (hidden size of 896) and a pair of affine layers followed by a softmax layer with 1024 outputs, predicting the 10-bit mu-law samples for a 24 kHz sampling rate. The conditioning network consists of a pair of bi-directional gated recurrent units (GRUs) with a hidden size 128. The mel-spectrograms used for conditioning the network were extracted using Librosa library [21], with 80 coefficients and frequencies ranging from 50 Hz to 12 kHz.

We trained system RNN_MS in 4 different configurations, whose details are shown in Table 2.1. First three SD systems were trained on American English speakers, two female (F1 and F2) and 1 male (M1) from our internal corpora.

We also trained 3 multi-speaker vocoders, one with all the training data from the 3 SD voices (3Spk), another one with 7 American English speakers (7Spk) comprising 4 females, 2 males and 1 child but with restricted amounts of training data per speaker (5000 utterances). This 7Spk neural vocoder aims to check whether variability or data (i.e. 3Spk) are more important for robustness in general. Finally we trained what is introduced as our universal neural vocoder with 74 different voices, in a 22 males and 52 females, extracted from 17 languages, with approx. 2000 utterances per speaker. This neural vocoder was designed with the expectation of being generalizable to any incoming speaker regardless of whether it was seen during training or not.

Table 2. Summary of the training data of the different RNN-based vocoders.

| Vocoder | Speakers | Utterances | Language |
|---------|----------|------------|----------|
| F1(SD)  | 1        | 22000      | US English |
| F2(SD)  | 1        | 15000      | US English |
| M1(SD)  | 1        | 15000      | US English |
| 3Spk    | 3        | 52000      | US English |
| 7Spk    | 7        | 35000      | US English |
| Univ    | 74       | 149134     | Multiple (17) |

Early implementations of the universal neural vocoder explicitly included speaker information via a variational encoding of the input spectrograms to handle the inherent variability in the training data better, it proved to be unnecessary. Such encodings did not provide any naturalness improvements, which can be explained by the mel-spectrogram already containing sufficient information about the speaker, making it redundant in our scenario.

2.2. CNN models

The trained CNN-based models follow the topology introduced in [5] and were trained only on a single speaker, F1. The only variations from those in [5] are the input to the model (linguistic features as used in [5] or mel-spectrograms as an addition in the present paper) and the structures of the CNN layers. The mel-spectrograms and audio specifications were analogous to those of the RNN models.

The CNN-based approaches implemented were: 1) an end-to-end linguistics-conditioned system (CNNLI, the system evaluated in [5]), 2) a vocoder (CNN_MS) and 3) a smaller vocoder (SCNN_MS). The purpose of training a smaller CNN vocoder was to increase inference time by reducing model complexity.

3. EXPERIMENTAL PROTOCOL

3.1. Naturalness analysis

To properly characterize the generalization capabilities of the different vocoders in terms of naturalness we considered a number of scenarios. First of all a topline scenario in which we generated speech from speakers present in the training data of all the vocoders, but with utterances not seen during training (section 5.2). Then, we also generated speech for speakers not present in the training data of any of the vocoders, with a mixture of male and female speakers neutral extracted from VCTK [22], with a speaking style similar to the training data but out of domain in terms of speaker and recording conditions. Finally we also evaluated highly expressive children audiobook speech extracted from Blizzard2013 development set [23], which was out of domain in speaker, speaking style and recording conditions. Experiments were carried out with oracle spectrograms unless stated otherwise.
The naturalness perceptual evaluation was designed as a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test [24], where the participants were presented with the systems being evaluated side-by-side, asked to rate them in terms of naturalness from 0 (very poor) to 100 (completely natural). The test consisted of 200 randomly-selected utterances for the different speakers, these were not included in the training data. The evaluation was balanced so that every utterance was rated by 5 listeners, and each listener rates 20 screens. The evaluations were conducted with self-reported native American English speakers using Amazon Mechanical Turk. In total 50 listeners participated in each evaluation.

Paired Student T-tests with Holm-Bonferroni correction were used to validate the statistical significance of the differences between systems, considering it validated when $p-value < 0.01$. We use the ratio between the mean MUSHRA score of a system and natural speech, from now on ‘relative MUSHRA’, to illustrate the quality gap with the reference.

### 3.2. Glitch analysis

This evaluation is aimed at gauging the nature of speech errors present in the evaluated SSWS systems. 4 internal listeners were asked to rate 50 different samples for each of the 4 considered systems, for a total of 200 rated utterances per system. For each sample the evaluators decided if a speech error was present, the perceived nature of the error and its severity (i.e. minor, medium, critical).

### 4. RESULTS

#### 4.1. CNN vs. RNN evaluation

To narrow down the number of systems to be analyzed in the robustness investigations we compared first of all the complete set of introduced systems, aiming only the best amongst them. For this evaluation we considered predicted spectrograms in the case of vocoders, to provide a fair comparison ground for CNN_LI. The spectrograms were predicted with our internal sequence-to-sequence model, based on Tacotron2 and explained in detail in [25]. The naturalness evaluation of the systems (figure 2 showed no significant difference between the vocoders, with CNN_LI falling behind, which we hypothesize is likely due to its less dynamic prosody.

From the glitch analysis (table 3) the following conclusions can be drawn: First of all, system CNN_LI showed to be overall more prone to glitches (51 out of 200 utterances) due to the much more complex nature of the task, as there is no support module capable of generating an intermediate acoustic representation such as the mel-spectrograms. Secondly, we can see how reducing the size of the CNN-based vocoder (system SCNN_MS) significantly increased the medium + critical glitches compared to the full-size implementation (system CNN_MS) (25 vs. 14). Finally, this analysis confirms the hypothesis that the RNN-based system (RNN_MS) is the most stable approach in both absolute terms (20) and when only looking at medium + critical (8).

From this point onwards we will focus only on the RNN-based system (RNN_MS).

#### 4.2. In-domain speakers

This evaluation considered 2 female speakers and 1 male speaker (the same speakers used to train the 3Spk vocoder). The results in Figure 4 show a clear story. There is no significant difference in terms of evaluated naturalness when using any of the trained vocoders as long as the speakers were part of the training data. This is a strong result for the proposed universal vocoder, as it showed no degradation when compared to the highly specific SD neural vocoder. Moreover, while there was a statistically significant difference, between vocoded and natural naturalness, it was minimal (98.5% relative MUSHRA scores). It must be noted that while there were inter-speaker differences, those did not affect the rank-order of the systems’ ratings, so results are presented only as averages.

#### 4.3. Out-of-domain speakers

This evaluation considered 2 female speakers and 1 male speaker (the same speakers used to train the 3Spk vocoder). The results in Figure 4 show a clear story. There is no significant difference in terms of evaluated naturalness when using any of the trained vocoders as long as the speakers were part of the training data. This is a strong result for the proposed universal vocoder, as it showed no degradation when compared to the highly specific SD neural vocoder. Moreover, while there was a statistically significant difference, between vocoded and natural naturalness, it was minimal (98.5% relative MUSHRA scores). It must be noted that while there were inter-speaker differences, those did not affect the rank-order of the systems’ ratings, so results are presented only as averages.
4.3.1. Selecting speaker dependent vocoders

In order to limit the perceptual evaluation complexity we tried to predict the best SD vocoder to run inference with. For that we built multi-variate Gaussian Mixture Models (GMMs) of the training data of the different vocoders and of the target speaker, then obtained the Kullback-Leibler divergence (KLD) between the GMMs. This could be improved by applying conventional speaker similarity measures, but the measure proves to be reliable with a correlation of 0.81 between the KLD and the naturalness degradation obtained through the perceptual evaluations, with degradation being defined as the difference between natural speech and system mean MUSHRA. The analysis concluded that F1 was closest to the female out-of-domain speakers, and M1 to the males.

4.3.2. Neutral style speakers

Results (Figure 4) show an interesting picture. The more variety in number of training speakers the better the achieved quality, to the point where the proposed universal system is capable of providing practically the same relative MUSHRA score as for in-domain speakers (98% vs. 98.5%). This speaks very strongly about the generalization capabilities of such a system. Moreover, we can see how the vocoder trained with more speakers but with less training data (7Spk) is capable of providing better quality than the other two systems (SD and 3Spk), suggesting that variability is more important than quantity for generalization. Similarly to what happened in the in-domain evaluation, the inter-speaker variability did not affect the rank-ordering, so we only show averaged results.

4.3.3. Audiobook style speakers

In the case of highly expressive data, including disfluencies and onomatopoeias, (see Figure 5) the universal vocoder is still capable of proving steady quality, once again maintaining the relative MUSHRA scores at 98%. Both SD and 7Spk show comparatively poor performance, while 3Spk breaks the trend. This is confirmed by the KLD between the audiobook speaker and those of the vocoders (2.64 against Univ, 5.42 against 3Spk, 14.45 against 7Spk and 14.62 against SD).

4.3.4. Other evaluations

We carried out evaluations analogous to those presented but with Japanese, which was an in-domain language. Interestingly, results were extremely similar to those in US English (98% relative MUSHRA), hence we do not report them in detail. This comforts us in the observation that the universal neural vocoder can be expected to be robust not only to speaker, style and recording conditions but also language.

5. CONCLUSIONS

We have introduced a robust universal neural vocoder conditioned on mel-spectrograms, without any form of speaker encoding. The performance of this vocoder has been shown to be as good, if not better, than a SD neural vocoder. The universal vocoder has proven to be capable of generating speech for speakers not present in the training corpora, regardless of speaking style or language, with a relative MUSHRA score of 98%. A universal neural vocoder allows for future work to focus on spectrogram estimation from text to any new speaker, language or style without being constrained by training a specific neural vocoder.
6. REFERENCES

[1] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, et al., “WaveNet: A generative model for raw audio,” CoRR abs/1609.03499, 2016.

[2] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, et al., “Efficient neural audio synthesis,” arXiv preprint arXiv:1802.08435, 2018.

[3] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, et al., “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” arXiv preprint arXiv:1712.05884, 2017.

[4] Wei Ping, Kainan Peng, and Jitong Chen, “Clarinet: Parallel wave generation in end-to-end text-to-speech,” arXiv preprint arXiv:1807.07281, 2018.

[5] Thomas Merritt, Bartosz Putrycz, Adam Nadolski, Tianjun Ye, Wei Ping, Kainan Peng, and Jitong Chen, “Clarinet: Parallel wave generation in end-to-end text-to-speech,” arXiv preprint arXiv:1807.07281, 2018.

[6] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight,” Proc. MAVEBA, pp. 13–15, 2001.

[7] 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. E99-D, no. 7, pp. 1877–1884, 2016.

[8] Thomas Drugman and Thierry Dutoit, “The deterministic plus stochastic model of the residual signal and its applications,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 3, pp. 968–981, 2012.

[9] M. W. Macon and M. A. Clements, “Speech concatenation and synthesis using an overlap-add sinusoidal model,” in Proceedings of the Acoustics, Speech, and Signal Processing, 1996. On Conference Proceedings., 1996 IEEE International Conference - Volume 01, Washington, DC, USA, 1996. ICASSP '96, pp. 361–364, IEEE Computer Society.

[10] Gunnar Fant, Johan Liljencrants, and Qi-guang Lin, “A four-parameter model of glottal flow,” STL-QPSR, vol. 4, no. 1985, pp. 1–13, 1985.

[11] Daniel Griffin and Jae 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.

[12] Zeyu Jin, Adam Finkelstein, Gautham J Mysore, and Jingwan Lu, “Fftnet: A real-time speaker-dependent neural vocoder,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 2251–2255.

[13] Xin Wang, Jaime Lorenzo-Trueba, Shinji Takaki, Lauri Juvela, and Junichi Yamagishi, “A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis,” arXiv preprint arXiv:1804.02549, 2018.

[14] Sercan O Arik, Heewoo Jun, and Gregory Diamos, “Fast spectrogram inversion using multi-head convolutional neural networks,” arXiv preprint arXiv:1808.06719, 2018.

[15] Xixin Wu, Yuwen Cao, Mu Wang, Songxiang Liu, Shiyin Kang, et al., “Rapid style adaptation using residual error embedding for expressive speech synthesis,” Proc. Interspeech 2018, pp. 3072–3076, 2018.

[16] Berrak Sisman, Mingyang Zhang, and Haizhou Li, “A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder,” Proc. Interspeech 2018, pp. 1978–1982, 2018.

[17] Li-Juan Liu, Zhen-Hua Ling, Yuan Jiang, Ming Zhou, and Li-Rong Dai, “Wavenet vocoder with limited training data for voice conversion,” Proc. Interspeech 2018, pp. 1983–1987, 2018.

[18] Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al., “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” arXiv preprint arXiv:1806.04558, 2018.

[19] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, et al., “Parallel wavenet: Fast high-fidelity speech synthesis,” arXiv preprint arXiv:1711.10433, 2017.

[20] Yi-Chiao Wu, Kazuhiro Kobayashi, Tomoki Hayashi, Patrick Lumban Tobing, and Tomoki Toda, “Collapsed speech segment detection and suppression for wavenet vocoder,” arXiv preprint arXiv:1804.11055, 2018.

[21] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, “librosa: Audio and music signal analysis in python,” in Proceedings of the 14th python in science conference, 2015, pp. 18–25.

[22] Junichi Yamagishi and Keith Edwards, “Voice cloning toolkit for festival and hts,” 2010.

[23] Simon King and V Karaiskos, “Blizzard Challenge 2013,” 2013, http://www.festvox.org/blizzard/.

[24] ITUR Recommendation, “Bs. 1534-1. method for the subjective assessment of intermediate sound quality (mushra),” International Telecommunications Union, Geneva, 2001.

[25] Javier Latorre, Jakub Lachowicz, Jaime Lorenzo-Trueba, and Thomas Drugman, “Effect of data reduction on sequence-to-sequence acoustic models for speech synthesis,” in Submitted to ICASSP 2019, 2019.