Interspeech 2018
2-6 September 2018, Hyderabad

Speaker-independent raw waveform model for glottal excitation

Lauri Juvela[1], *Vassilis Tsiaras*[2], *Bajibabu Bollepalli*[1],
Manu Airaksinen[1], *Junichi Yamagishi*[3], *Paavo Alku*[1]

[1]Aalto University, Finland
[2]University of Crete, Greece
[3]National Institute of Informatics, Japan

lauri.juvela@aalto.fi, tsiaras@csd.uoc.gr

Abstract

Recent speech technology research has seen a growing interest in using WaveNets as statistical vocoders, i.e., generating speech waveforms from acoustic features. These models have been shown to improve the generated speech quality over classical vocoders in many tasks, such as text-to-speech synthesis and voice conversion. Furthermore, conditioning WaveNets with acoustic features allows sharing the waveform generator model across multiple speakers without additional speaker codes. However, multi-speaker WaveNet models require large amounts of training data and computation to cover the entire acoustic space. This paper proposes leveraging the source-filter model of speech production to more effectively train a speaker-independent waveform generator with limited resources. We present a multi-speaker 'GlotNet' vocoder, which utilizes a WaveNet to generate glottal excitation waveforms, which are then used to excite the corresponding vocal tract filter to produce speech. Listening tests show that the proposed model performs favourably to a direct WaveNet vocoder trained with the same model architecture and data.

Index Terms: Glottal source generation, WaveNet, mixture density network

1. Introduction

Recently, there has been a growing interest in WaveNet-based waveform generation in speech applications due to the high quality of generated speech. While the first WaveNet text-to-speech (TTS) model used linguistic features and fundamental frequency (F0) from an existing statistical parametric speech synthesis (SPSS) system [1], there seems to be a shift in focus towards using WaveNets as statistical vocoders. In the statistical vocoder approach, a WaveNet is conditioned with some acoustic features, such as mel filterbank energies [2, 3], or mel-generalized cepstrum (MGC) coefficients and F0 [4]. In context of TTS, high-quality systems have been built by separately training a WaveNet vocoder and a text-to-acoustic-features model, where the latter can be an end-to-end attention-based neural net [2, 3] or a more conventional frame-aligned SPSS system [5].

A clear benefit of acoustically conditioned WaveNets is that the same waveform generator model can be shared between multiple speakers, provided that the acoustic features contain sufficient information to capture the speaker identity. For example, multi-speaker WaveNets have been successfully conditioned on low-bitrate speech codec parameters [6], as well as on acoustic parameters typically used in parametric TTS (MGC, F0) [7]. Furthermore, previous research found no added benefit from using speaker codes to supplement the acoustic features [7], which suggests that the acoustic features themselves can be sufficient for high-quality speaker-independent waveform generation. However, training large-scale speaker-independent models that cover the acoustic space for various unseen speakers is expected to be costly in terms of data and computation. This problem can be mitigated by leveraging knowledge of the human speech production mechanism to reduce the data variability in speech.

Before WaveNets, waveform synthesis with neural networks has been applied, using simple fully connected networks [8, 9], to glottal excitations, i.e., time-domain signals corresponding to the volume velocity waveform generated by the vocal folds in the human speech production mechanism. In this approach, the target waveform is a glottal excitation signal estimated from speech using glottal inverse filtering (GIF), specifically quasi-closed phase (QCP) analysis [10]. GIF decomposes a speech signal into a vocal tract filter and a glottal source, effectively removing the vocal tract resonances from speech [11]. Due to the absence of vocal tract resonances, the glottal excitation signal is more elementary than the speech pressure signal, and thus easier to model and synthesize with simple neural nets. Similarly to the emerging WaveNet vocoders, previous glottal waveform synthesis models have mostly used acoustic features as the conditioning input. However, in contrast to the sample-by-sample generation of WaveNets, these glottal waveform models used a pitch synchronous frame-based waveform representation. While this representation facilitates learning (and is applicable to parallel inference), the approach is sensitive to pitch-tracking errors and is limited to producing voiced speech only. Furthermore, these models were trained using least-squares regression, which does not allow true stochastic sampling from the learned distribution. More recently, generative adversarial networks have been applied to the task to enable stochastic generation [12, 13], but these models are still constrained by the pitch synchronous windowing scheme.

With WaveNets now available, it is natural to extend the generation of glottal excitation signals to utilize WaveNet-like models. This paper presents GlotNet, a speaker-independent neural waveform generator explicitly based on the source-filter model of speech production: a WaveNet conditioned on acoustic features generates a glottal source signal, which is then used to excite an all-pole vocal tract filter. The proposed system is compared with a direct speech pressure signal WaveNet vocoder trained using the same model architecture, acoustic conditioning and dataset. Additionally, we propose a simple but effective method for including a non-causal look-ahead into the acoustic conditioning. Although the paper scope is limited to copy-synthesis (i.e., natural acoustic features are used at test time), the proposed method should interface well with the ever-improving acoustic models in TTS systems.

The paper is structured as follows: Section 2 describes the

10.21437/Interspeech.2018-1635

waveform generator models, while the experiments and evaluation are described in Section 3. We discuss the results in Section 4 and conclude in Section 5.

2. Waveform generator models

An overview of the WaveNet and GlotNet vocoders is shown in Fig. 1. While a WaveNet vocoder learns a non-linear autoregressive (AR) model to predict next signal sample from previous samples signal and time-varying acoustic features, a GlotNet operates on a more simplistic glottal excitation signal. The excitation signal is then passed through an all-pole vocal tract (VT) filter to produce speech waveforms. The GlotNet model for the speech signal x_n can be viewed as a mixture of a low-order linear AR process (VT filter) and a non-linear residual excitation process e_n (glottal source)

$$x_n = \sum_{k=1}^{P} a_k x_{n-k} + e_n, \tag{1}$$

where the linear AR process of order P is described by the filter coefficients $a_1, \ldots, a_P$, while the excitation process e_n is modeled by a WaveNet with a receptive field of R samples. Specifically, we assume the excitation process to be a logistic mixture

$$e_n \sim \sum_{i=1}^{K} \pi_i \text{logistic}(\mu_i, s_i \mid e_{(n-R):(n-1)}, h_n) \tag{2}$$

with non-linear dependencies to past excitation samples, as parametrized by a WaveNet. Given previous excitation samples $e_{(n-R):(n-1)}$ and local (acoustic) conditioning h_n, the WaveNet predicts the current time-step logistic mixture parameters: mixture weight π_i, component mean μ_i and component scale s_i.

In this paper, the linear AR process parameters are estimated separately and kept fixed while training the excitation model. For this, we use QCP analysis, which utilizes time-weighted linear predictive analysis to attenuate the glottal contribution in the AR filter estimate [10]. The linear AR process order is relatively low (we use P=30), whereas the receptive field of a WaveNet can grow large due to its dilated convolution structure. Furthermore, the parameters of the two processes vary at different rates: the filter parameters are updated at a 200 Hz rate (or 5 ms frame shift), while the excitation process parameters are predicted for every sample at a 16 kHz rate.

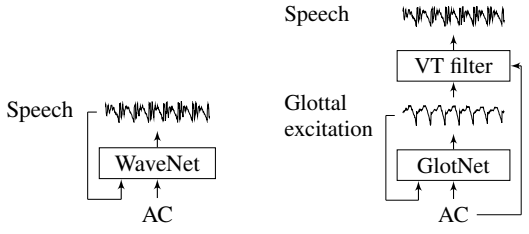

Figure 1: *WaveNet vocoder (left) uses acoustic features (AC) and past signal samples to generate the next speech sample. In contrast, GlotNet (right) operates on the more simplistic glottal excitation signal, which is filtered by a vocal tract (VT) filter already parametrized in the acoustic features.*

2.1. Network architecture

We use a WaveNet implementation based on [14]. The model architecture has two main parts: a stack of residual blocks, which acts as a multi-scale feature extractor, and a post-processing module, which combines the information from the residual blocks to predict the next signal sample distribution parameters. In each residual block, the key operation is a gated convolution given by

$$x_{\text{skip}} = \tanh(W_f * x_{\text{in}} + L_f) \odot \sigma(W_g * x_{\text{in}} + L_g), \tag{3}$$

where $*$ denotes dilated causal convolution and $\odot$ is element-wise multiplication. W_f and W_g are convolution weight tensors for filter and gate, respectively. Additionally, L_f and L_g are local conditioning vectors specific to the residual block. The skip path activations x_{skip} are connected to the post-processing module, while a residual block output $x_{\text{out}} = W x_{\text{skip}} + x_{\text{in}}$ is fed forward into the next layer of the residual stack.

The post-processing module takes in the skip-outputs from each residual block and concatenates them along their channel dimension. This is followed by two 1×1 convolution layers with contenated rectifier activations [15], whose output is finally projected to the mixture density network output of size $3K$ (where K is the number of mixture components).

2.2. Local conditioning

For local conditioning, both models use the same acoustic feature set of glottal vocoder parameters [16]: the vocal tract filter, estimated by QCP analysis [10], and the corresponding glottal source spectral envelope are parametrized by line spectrum frequencies (LSFs), using orders 30 and 10, respectively. Fundamental frequency in log-scale (LF0) and a binary voicing flag (VUV) describe the pitch contour, whereas the average harmonic-to-noise ratio (HNR) in 5 ERB frequency-bands characterizes the signal aperiodicity. Finally, the frame energy (in dB) is used to indicate the signal level.

In initial experiments, we found that the waveform generator reliability is improved when the model is allowed to use a small look-ahead into future conditioning. Previous work has proposed using various bi-directional recurrent structures for encoding the future of the conditioning [2, 5]. However, training these kind of structures jointly with a WaveNet notably increases the computational cost. Instead, we first stack adjacent past and future frames to the current frame to provide context, after which we use linear interpolation to upsample the conditioning from 200 Hz to 16 kHz. Finally, we apply a global projection to embed the conditioning into smaller dimensionality before injecting the embedded conditioning into the residual blocks, as shown in Fig. 2. In the experiments, we use 4 frames of context to both directions, corresponding to 20 ms look-ahead.

2.3. Discretized logistic mixture density loss

WaveNets have commonly used 8-bit quantization, which requires 256-dimensional softmax output if trained as a classifier. However, this often results in quantization-noise like artefacts, whereas using the full 16 bits of amplitude levels would require prohibitively large softmax layers. To overcome this limitation, a discretized logistic mixture density loss was proposed to improve PixelCNN [17]. The approach was quickly adopted to improving WaveNet fidelity [18, 3]. Furthermore, mixture density networks extend more easily to multivariate modeling: for example, a WaveNet-like architecture with Gaussian mixtures

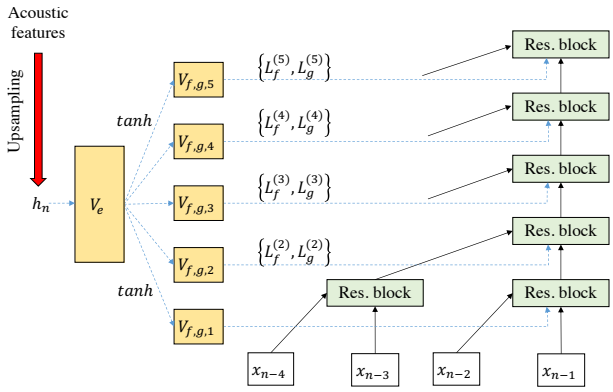

Figure 2: *A five-level residual stack of a WaveNet vocoder. The residual stack shares a global embedding for the acoustic features, which is transformed to block-specific local conditioning vectors.*

has been proposed for generating vocoder parameters in singing synthesis [19].

To train a mixture density network, one has to be able to evaluate likelihoods for observations. For the logistic distribution, the cumulative distribution function (CDF) is the logistic sigmoid, and the probability of a quantized observation x is a Δ-wide slice of the CDF

$$p(x) = \sum_{i=1}^{K} \pi_i [\sigma((x + \frac{\Delta}{2} - \mu_i)/s_i) - \sigma((x - \frac{\Delta}{2} - \mu_i)/s_i)], \tag{4}$$

where Δ is the quantization bin width and σ is the logistic CDF. This formulation is then used to minimize the negative log-likelihood for the observations [17]. In practice, the network outputs are treated as mixture weight logits, component means and log-scale parameters. Notably, the log-scales should be floored to avoid variance collapse during training, but the floor level simultaneously acts as a noise floor in generation. If the floor is set too high, this property may lead to exaggerated background noise or roughness in the synthetic voiced speech.

3. Experiments

3.1. Speech material

We use a multi-speaker database originally released for speech enhancement research [20], and only take the clean speech subset for these experiments. The voice talents in the dataset are non-professional native British English speakers. The full training dataset consists of 56 speakers, but to scale the task for our available computational resources, we use a 28-speaker subset provided in the data. We treat these data as our seen speakers dataset, which contains 11571 utterances in total, amounting to 9.4 hours of speech, i.e., about 20 minutes per seen speaker. The ten first utterances from each seen speaker were reserved for testing, and 500 of the remaining utterances were randomly chosen for validation. Additionally, two speakers (one female, one male) from the database testset were held out as unseen.

3.2. Training the models

For both WaveNet and GlotNet, we used 64 channels within the residual blocks (residual and skip channels) and 128 channels in the post-processing module. The convolution filter width is two everywhere in the residual stack, in which the dilation pattern $1, 2, 4, \ldots, 512$ is repeated three times, resulting in a total of 30 residual blocks and a receptive field length of 3071 samples. The training criterion for the models was to minimize the discretized logistic mixture negative log-likelihood for their respective observed signals, where we used 5 mixture components. The models were trained for 70 epochs (with a 10 epoch early stopping criterion) using the Adam optimizer [21] and exponential moving average weight smoothing [22].

The prediction of signal sample probability distributions allows manual adjustment of the sampling strategy at test time. Maximum posterior sampling in voiced regions has been reported to improve perceived synthetic speech quality [5, 23], and we observed a similar effect in our informal experiments. Nevertheless, we chose to sample directly from the predicted distributions as we feel this reflects the learned model quality more accurately.

3.3. Listening tests

For subjective evaluation of the system performances, we conducted listening tests on speaker similarity and speech quality.[1] The tests were run on the CrowdFlower crowd-sourcing platform [24], where the tests were made available in English-speaking countries and the top four countries in the EFI English proficiency rating [25]. Each test case was evaluated by 50 listeners, while the listeners were screened using natural reference null pairs and artificially corrupted anchor samples.

To evaluate the subjective quality of the synthetic speech, we conducted pairwise category comparison rating (CCR) tests [26], where the listeners were presented with a pair of samples and asked to rate the comparative quality on a 7-level scale, ranging from -3 (much worse) to 3 (much better). Combined scores are shown in Fig. 3. The scores were calculated by reordering the ratings for each system and pooling together all ratings the system received. Natural speech target utterance was included in the tests as a reference system. The plots show mean ratings with 95% confidence, corrected for multiple comparisons.

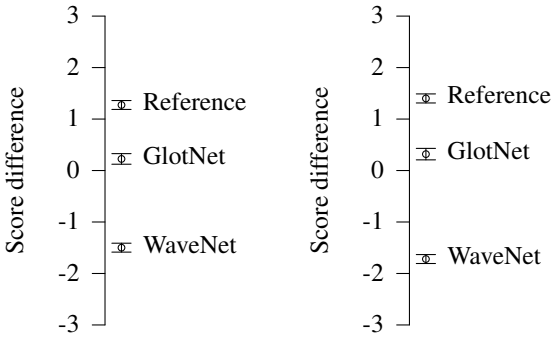

Figure 3: *Combined score differences obtained from the quality comparison CCR test for seen speakers (left) and unseen speakers (right). Error bars are t-statistic based 95% confidence intervals for the mean.*

Synthetic speech voice similarity to a natural reference was measured in a DMOS-like test [26]. The listeners were presented with a test sample and asked to rate the voice similarity

[1] Samples available at http://tts.org.aalto.fi/interspeech18_glotnet

to the target natural speech utterance on a 5-level absolute category rating scale, ranging from 1 (bad) to 5 (excellent). Results are shown in Fig. 4. The plot shows mean ratings with 95% confidence intervals, as well as stacked score distribution histograms in the background.

In both test types, GlotNet performs favourably to WaveNet. Furthermore, GlotNet ratings remain largely unaffected by testing on unseen speakers, whereas WaveNet scores slightly decrease. It should be noted that both tests involve paired comparisons to a natural speech reference, which makes the tests quite sensitive to small degradations.

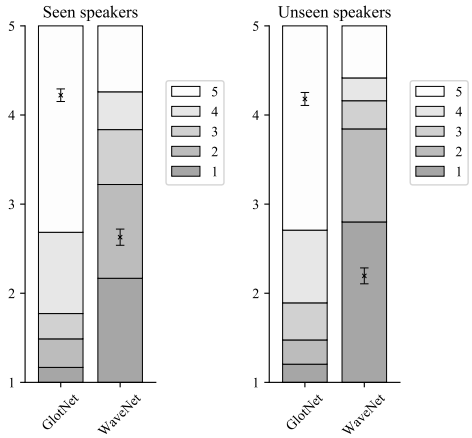

Figure 4: *Voice similarity ratings in a DMOS test for seen speakers (left) and unseen speakers (right). Mean scores are shown with 95% confidence intervals, while relative score distribution histograms are shown in the background.*

3.4. Objective measures

To quantify how reliably the different waveform generation methods follow their acoustic conditioning, we computed various objective metrics. Fig. 5 shows objective measures for different systems, computed with respect to the original signal. The box-and-whiskers plots show the medians, along with the 25% and 75% quantiles. A deterministic glottal vocoder which uses the same acoustic feature set is included as a reference method. Mel spectral distortion (MCD, in dB) was calculated by applying a 24-band mel filterbank matrix to FFT magnitude spectrum, and taking the root-mean-squared error of the log-differences over frames and mel-bands. F0 was estimated from the synthetic signals using the RAPT algorithm [27], and log-domain F0 difference (in cents: 100 cents is one semitone, 12 semitones is one octave) is reported over frames where the voicing estimates agree. Finally, we report the voicing error percentages between the local conditioning and the one estimated from the synthetic signals.

4. Discussion

In the present experiments, the direct waveform WaveNet vocoder performance appears lacking both in terms of subjective quality and objective reliability. This can be largely attributed to the multi-speaker task combined with the relatively small dataset and computation budget. Furthermore, we feel that the logistic mixture density network training is more demanding than the softmax-based approach. Previously, high-quality logistic mixture WaveNets have been trained using more

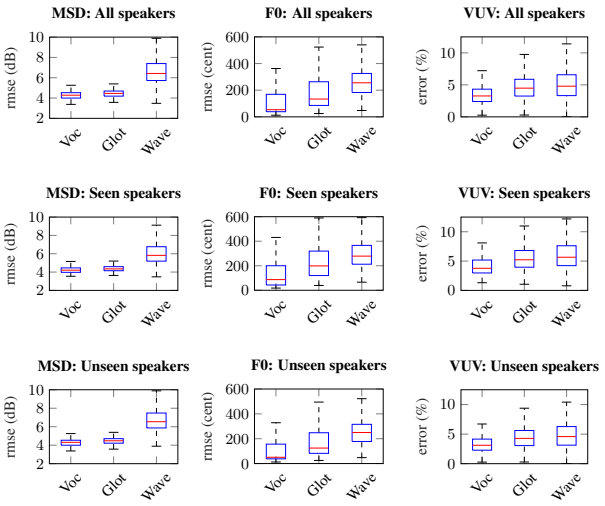

Figure 5: *Objective measures for mel spectral distortion (MSD), log-F0 RMSE (in cents) and voicing decision error (%). 'Voc' denotes a deterministic glottal vocoder, while 'Glot' and 'Wave' are GlotNet and WaveNet vocoders, respectively.*

data and speaker-specific models [18, 3], whereas previous speaker-independent models have used the softmax training approach [7, 6]. We also note that our models use relatively few parameters compared to previous research. As such, the WaveNet vocoder performance would likely improve by using more training data and larger models.

Nevertheless, adding the low-order linear AR component to the signal model in GlotNet considerably improves the model performance with the same data and equivalent model architecture and training procedure. This is well motivated by the prevalent use of linear predictive models in speech applications. Furthermore, GlotNet-like excitation models should be well applicable to existing parametric TTS systems, as their acoustic features often include spectral envelope information interpretable as a filter. Among these spectral features, glottal inverse filtering based models are physiologically motivated and aim to consistently separate the excitation signal from the linear AR envelope filter.

5. Conclusions

This paper proposed a speaker-independent neural waveform generator which combines a linear autoregressive (vocal tract filter) process with a non-linear (glottal source) excitation process parametrized by a WaveNet. Listening tests and objective measures show that the proposed method outperforms directly modeling speech with a WaveNet vocoder, when both models use identical architectures and training data. While the current work focuses on copy-synthesis experiments, future work includes integrating the waveform generator models into parametric text-to-speech systems.

6. Acknowledgements

This work was supported by the Academy of Finland (proj. no. 284671 and 312490) and MEXT KAKENHI Grant Numbers (15H01686, 16H06302, 17H04687). We acknowledge the computational resources provided by the Aalto Science-IT project.

7. 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," *arXiv pre-print*, 2016. [Online]. Available: https://arxiv.org/pdf/1609.03499

[2] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep Voice: Real-time neural text-to-speech," in *Proc. ICML*, 2017.

[3] J. Shen, R. Pang, R. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

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

[5] 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 *Proc. ICASSP*, 2018, pp. 4804–4808.

[6] 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 *Proc. ICASSP*, 2018, pp. 676–680.

[7] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *Proc. ASRU*, Dec 2017, pp. 712–718.

[8] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *Proc. ICASSP*, March 2016, pp. 5120–5124.

[9] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, "GlottDNN—a full-band glottal vocoder for statistical parametric speech synthesis," in *Proc. Interspeech*, 2016, pp. 2473–2477.

[10] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 596–607, March 2014.

[11] P. Alku, "Glottal inverse filtering analysis of human voice production – a review of estimation and parameterization methods of the glottal excitation and their applications. (invited article)," *Sadhana – Academy Proceedings in Engineering Sciences*, vol. 36, no. 5, pp. 623–650, 2011.

[12] B. Bollepalli, L. Juvela, and P. Alku, "Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis," in *Proc. Interspeech*, 2017, pp. 3394–3398.

[13] L. Juvela, B. Bollepalli, X. Wang, H. Kameoka, M. Airaksinen, J. Yamagishi, and P. Alku, "Speech waveform synthesis from MFCC sequences with generative adversarial networks," in *Proc. ICASSP*, 2018, pp. 5679–5683.

[14] N. Adiga, V. Tsiaras, and Y. Stylianou, "On the use of WaveNet as a statistical vocoder," in *Proc. of ICASSP*, 2018, pp. 5674–5678.

[15] W. Shang, K. Sohn, D. Almeida, and H. Lee, "Understanding and improving convolutional neural networks via concatenated rectified linear units," *arXiv pre-print*, 2016. [Online]. Available: http://arxiv.org/abs/1603.05201

[16] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 153–165, January 2011.

[17] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications," *arXiv pre-print*, 2017. [Online]. Available: http://arxiv.org/abs/1701.05517

[18] 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," *arXiv pre-print*, 2017. [Online]. Available: http://arxiv.org/abs/1711.10433

[19] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer," in *Proc. Interspeech*, 2017, pp. 4001–4005.

[20] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models," 2017. [Online]. Available: http://dx.doi.org/10.7488/ds/2117

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[22] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Control Optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[23] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "FFTNet: a real-time speaker-dependent neural vocoder," in *Proc. ICASSP*, 2018, pp. 2251–2255.

[24] CrowdFlower Inc., "Crowd-sourcing platform," https://www.crowdflower.com/, accessed: 2018-03-22.

[25] "EF English proficiency index," http://www.ef.com/epi/, accessed: 2018-03-22.

[26] "Methods for Subjective Determination of Transmission Quality," ITU-T SG12, Geneva, Switzerland, Recommendation P.800, Aug. 1996.

[27] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.