Juvela, Lauri; Bollepalli, Bajibabu; Tsiaras, Vassilis; Alku, Paavo

GlotNet-A Raw Waveform Model for the Glottal Excitation in Statistical Parametric Speech Synthesis

Published in:
IEEE/ACM Transactions on Audio, Speech, and Language Processing

DOI:
10.1109/TASLP.2019.2906484

Published: 01/06/2019

Document Version
Peer reviewed version

Please cite the original version:
Juvela, L., Bollepalli, B., Tsiaras, V., & Alku, P. (2019). GlotNet-A Raw Waveform Model for the Glottal Excitation in Statistical Parametric Speech Synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(6), 1019-1030. [8675543]. https://doi.org/10.1109/TASLP.2019.2906484

GlotNet—A Raw Waveform Model for the Glottal Excitation in Statistical Parametric Speech Synthesis

Lauri Juvela, Bajibabu Bollepalli, Vassillis Tsiaras, and Paavo Alku, *Senior member, IEEE*

***Abstract*—Recently, generative neural network models which operate directly on raw audio, such as WaveNet, have improved the state of the art in text-to-speech synthesis (TTS). Moreover, there is increasing interest in using these models as statistical vocoders for generating speech waveforms from various acoustic features. However, there is also a need to reduce the model complexity, without compromising the synthesis quality. Previously, glottal pulseforms (i.e., time-domain waveforms corresponding to the source of human voice production mechanism) have been successfully synthesized in TTS by glottal vocoders using straightforward deep feedforward neural networks. Therefore, it is natural to extend the glottal waveform modeling domain to use the more powerful WaveNet-like architecture. Furthermore, due to their inherent simplicity, glottal excitation waveforms permit scaling down the waveform generator architecture. In this study, we present a raw waveform glottal excitation model, called GlotNet, and compare its performance with the corresponding direct speech waveform model, WaveNet, using equivalent architectures. The models are evaluated as part of a statistical parametric TTS system. Listening test results show that both approaches are rated highly in voice similarity to the target speaker, and obtain similar quality ratings with large models. Furthermore, when the model size is reduced, the quality degradation is less severe for GlotNet.**

***Index Terms*—Glottal source model, text-to-speech, WaveNet**

I. Introduction

STATISTICAL parametric speech synthesis (SPSS) has become a widely used text-to-speech technique due to its flexibility and good generalization to unseen utterances [1]. The adoption of deep neural networks for acoustic modeling has further improved the prosodic naturalness and clarity of the synthetic speech [2], and the further use of recurrent neural nets [3], [4], and most recently, sequence-to-sequence models with attention [5], [6] yield impressive results. Meanwhile, recent signal processing methods for vocoding have improved the synthetic speech quality. These techniques include source-filter models [7], [8], sinusoidal harmonic-plus-noise models [9], advanced aperiodicity models [10], [11], and direct modeling of the magnitude and phase spectra [12]. Furthermore, the ongoing emergence of neural network waveform generation models, i.e. "neural vocoders", has nearly closed the quality gap between natural and synthetic speech.

WaveNets are generative raw waveform audio models that use convolution neural nets to learn signal amplitude sequence distributions, enabling the model to generate new samples from the distribution [13]. WaveNets can be applied for text-to-speech (TTS) synthesis by using a time-varying local conditioning, and such a system has been reported to outperform state-of-the-art concatenative and statistical parametric TTS systems [13]. Although the TTS system described in [13] conditions the model using linguistic labels from an existing parametric TTS system, subsequent work has shifted toward using acoustic features for conditioning WaveNets. In this setting, a WaveNet replaces a vocoder in waveform generation from acoustic features. Various acoustic features have been utilized for WaveNet conditioning, including mel-filterbank energies [14] and mel-generalized cepstral coefficients (MGC) [15] in combination with fundamental frequency (F0) [16]. Recently, this kind of WaveNet vocoder (based on MGCs and F0) was reported to outperform various conventional waveform generation approaches in a large-scale TTS evaluation [17]. Additionally, acoustic conditioning enables the same models to be used in other (non-TTS) waveform generation tasks, such as voice conversion [18]. Despite their recent success, WaveNet models suffer from their need for substantial amounts of training data and large model sizes, making them expensive to train and use [19]. Furthermore, the autoregressive nature of WaveNets makes inference inherently slow. While WaveNet-like models with parallel inference have been proposed, these models require even larger amounts of computations and are difficult to train [20], [21]. Correspondingly, optimizing autoregressive inference for real-time requires considerable engineering effort and algorithmic compromises [14], [22].

Another important question with the WaveNet vocoders is how to condition the model so that the neural waveform generator faithfully reproduces the given acoustic conditions. Speech signal can often be predicted with high accuracy from the previous signal values, and if the control features (acoustics in a vocoder) are not informative enough, a WaveNet can learn to ignore the conditioning altogether and focus solely on the previous observed values. This can lead to mispronunciations, or in extreme cases complete babble [13], [23]. While the effect can be alleviated by, for example, contrastive training [20], more direct means of providing the model with suitable conditioning have been explored. Specifically, providing the waveform generator with information about future acoustics via a jointly trained non-causal conditioning model seems to be useful. For example, bidirectional recurrent neural networks (RNNs) have been proposed to encode the local conditioning in a non-causal manner [14], [17]. However, this considerably increases the computational complexity and training time of the model, as the RNNs require sequential processing of training utterances. As an alternative to RNNs, non-causal

Manuscript received October 01, 2018; revised January 14, 2019 and February 11, 2019; accepted March 17, 2019.

L. Juvela, B. Bollepalli, and P. Alku are with Department of Signal Processing and Acoustics, Aalto University, Finland; V. Tsiaras is with the University of Crete, Greece. (Corresponding author email: lauri.juvela@aalto.fi)

WaveNet-like convolutional networks have recently been applied to many tasks, including speech bandwidth extension [24] and denoising [25]. This type of architecture has also been applied in a machine translation encoder-decoder model for non-causal sequence encoding to condition a causal autoregressive decoder [26]. As a partial contribution of this paper, we adapt the idea to implement a non-causal encoding of the input acoustic sequences to condition the autoregressive causal waveform generator. This will act as a representative for highly expressive, non-causal conditioning models for the WaveNet vocoders.

In terms of more lightweight neural waveform generation methods, deep neural networks (DNNs) have previously been proposed for waveform generation in glottal vocoding, a family of TTS vocoders based on modeling the source of the human voice production mechanism, the glottal flow. The glottal excitation domain is lucrative for the DNN-based waveform generation, as the glottal source is decoupled from vocal tract filter resonances, and is thus more elementary and easier to model than the speech signal waveform itself. In the DNN-based glottal excitation generation, glottal inverse filtering (GIF) [27], a technique to estimate the glottal flow from speech, is needed in model training. The first TTS studies [28], [29] utilizing the DNN-based glottal excitation generation used a GIF technique proposed in [30]. More recently, improved TTS quality has been obtained by using a more accurate GIF method [31], jointly trained waveform and noise models [32], or alternative waveform representations [19]. All these previous studies are based on a fixed length pitch-synchronous waveform representation that enables the use of simple, fully connected networks for fast inference. However, this framework only enables modeling the voiced excitation and is somewhat sensitive to pitch-marking accuracy.

With the WaveNet-like models now available, it is possible to leverage these powerful neural models to generate glottal excitation waveforms. The motivation for shifting the WaveNet-like waveform model from the domain of speech signals to that of glottal flows is twofold. First, current statistical parametric speech synthesis (SPSS) systems can predict the parametric representation of the vocal tract relatively well [33], and explicitly using this envelope information for filtering adds controllability to the system. The remaining task, however, is the accurate and robust generation of the time-domain excitation waveform for the vocal tract filter. Second, due to the inherent simplicity of the glottal source signal, it should be possible to achieve similar TTS performance with smaller models and less training data compared to applying the WaveNet model directly to the speech signal waveform.

Glottal excitation modeling with WaveNets has been previously proposed for building a speaker-independent neural glottal vocoder [34]. This kind of neural vocoder, dubbed "GlotNet", is trained to generate glottal source excitation waveforms conditioned on the acoustic features of an SPSS system. The generated excitation waveforms are then filtered using the vocal tract filter in order to produce synthetic speech. However, previous evaluations were done on ground truth acoustic features, i.e., copy-synthesis, whereas evaluation for speech synthesis requires both conditioning and filtering with generated acoustic features. Furthermore, we feel that the WaveNet vocoder was under-performing in [34], partially due to the high acoustic variability across multiple speakers and relatively limited training data. Additionally, difficulties in training a mixture density model led to further issues with the WaveNet. Compared to the original categorical softmax WaveNet, mixture density nets potentially give higher quality as they do not involve amplitude quantization (and have indeed given state-of-the art results [20], [21]). In the present paper, we limit our WaveNet and GlotNet models to the categorical softmax training, as this approach is more robust (in our experience) and still enables comparisons between similar WaveNet and GlotNet models.

The present paper has three main contributions. First, we train speaker-specific GlotNet and WaveNet vocoder models using the same datasets and acoustic conditioning, and compare their performances in an SPSS system. We further study the effect of reducing the model size for lighter computation. Second, we investigate the properties of non-causal conditioning models for WaveNet and GlotNet and discuss why GlotNet benefits more from acoustic look-ahead in conditioning. Finally, we perform extensive subjective evaluation on text-to-speech and copy-synthesis quality and on text-to-speech voice similarity, with comparisons including two more conventional vocoders STRAIGHT [35] and GlottDNN [33].

In this article, speech synthesis systems utilizing the raw waveform generation neural models WaveNet and GlotNet are first presented in section II. Then, two selected conventional vocoders, STRAIGHT and GlottDNN, are described in section III, after which experiments and evaluations are reported in sections IV and V, respectively. Results of the study show that all the tested WaveNet and GlotNet TTS systems achieve high quality and voice similarity to the target speaker, while systematically outperforming STRAIGHT. Moreover, while both WaveNet and GlotNet achieve high synthesis quality on a commonly used large model architecture, the proposed GlotNet vocoder outperforms WaveNet when using an architecture of reduced complexity. In a copy-synthesis experiment, a large GlotNet achieves the highest quality, while a smaller GlotNet is on par with the large WaveNet.

II. Speech synthesis system

Figure 1 depicts the flow diagram of the synthesis platforms studied in this article. Our text-to-speech synthesis system has two distinct neural network components. First, an acoustic model maps linguistic features to acoustic features, similarly to conventional SPSS systems. Second, a neural vocoder model (WaveNet or GlotNet) generates speech waveforms from the acoustic features. For comparison with conventional vocoders, the acoustic model is shared as far as possible (STRAIGHT uses a different acoustic feature set), and the neural vocoder block is replaced with GlottDNN or STRAIGHT. In this work, the acoustic model and neural vocoder components are trained separately, and the focus of experiments is on the neural vocoders and how they integrate to a SPSS system.

Our linguistic features are obtained via a conventional SPSS pipeline; for a detailed overview see [1]. The linguistic features

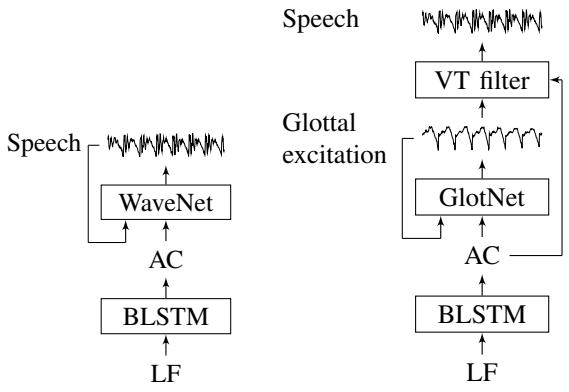

Fig. 1: The speech synthesis systems share a BLSTM acoustic model that maps linguistic features (LF) of input text to acoustic features (AC). A WaveNet vocoder (left) uses the acoustic features and past signal samples to generate the next speech sample. In contrast, a GlotNet (right) operates on a more simplistic glottal excitation signal, and produces speech by explicitly filtering the excitation with the vocal tract (VT) filter, whose parameters are included in the acoustic features.

include phoneme, syllable, word, phrase, and sentence level information and are created using the Flite speech synthesis front-end [36] and the Combilex lexicon [37]. The linguistic and acoustic features are aligned with the HMM-based speech synthesis system (HTS) [38] for training the neural net acoustic model. A bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) is used for the acoustic model, which maps frame-rate linguistic features to acoustic features.

A. Neural vocoders

Neural vocoders are neural networks that generate (speech) waveforms from acoustic feature inputs. While other neural waveform generators have been proposed [39], [22], this paper focuses on WaveNet-type models [13]. A WaveNet learns to predict the conditional distribution of a waveform sample, given previous samples and some external conditioning relevant to the signal. When trained, the model can be used to generate new samples from the conditional distribution. In this work, the target waveform is either the speech pressure signal (WaveNet) or the differential glottal excitation signal (GlotNet) estimated using the quasi-closed phase (QCP) glottal inverse filtering method [40]. The GlotNet output needs to be filtered with the vocal tract resonances to obtain speech, but vocal tract information is explicitly available in the acoustic features and the cost of this operation is negligible compared to the WaveNet run cost.

Wavenet-like model architectures have two main parts: a convolution stack and a post-processing module. The convolution stack consists of dilated convolution residual blocks and acts as a multi-scale feature extractor, while the post-processing module combines the information from the residual blocks to predict the next sample. The ith residual block applies a gated causal convolution between input $\boldsymbol{X}_i$ from previous time-steps and filter weights $\boldsymbol{W}_i^f$ and $\boldsymbol{W}_i^g$, where f and g denote filter and gate, respectively. Furthermore, each residual block receives a time-varying (local) conditioning vector $\boldsymbol{h}$, which is added to the convolution feature map after the projections with $\boldsymbol{V}_i^f$ and $\boldsymbol{V}_i^g$. The gated convolution output $\boldsymbol{Y}_i$ is given by the element-wise product

$$\boldsymbol{Y}_i = \tanh(\boldsymbol{W}_i^f * \boldsymbol{X}_i + \boldsymbol{V}_i^f \boldsymbol{h}) \odot \sigma(\boldsymbol{W}_i^g * \boldsymbol{X}_i + \boldsymbol{V}_i^g \boldsymbol{h}), \quad (1)$$

where σ is the logistic sigmoid function. Finally, $\boldsymbol{Y}_i$ is added to the residual block input $\boldsymbol{X}_i$ to produce the block output, which is then fed to the input of the next layer (similarly to ResNets [41]). Additionally, each residual block has a separate set of weights for a skip connection to the post-processing module.

The post-processing module combines the information from the residual blocks by summing the skip path outputs from each module. As the skip connections use 1×1 convolutions, they can also be seen as linear combination weights between different stages in the residual stack. The final output of the network is a 256-dimensional softmax layer, which is trained to predict the waveform amplitude class in 8-bit quantization (after μ-law companding).

B. Local conditioning mechanisms

A conditioning model for a Wavenet vocoder has two tasks: first, encode the acoustic feature sequence so that it is suitable for injection to the WaveNet convolution stack, and second, upsample the conditioning from an acoustic frame rate to sample rate (from 200 Hz to 16 kHz in our case). Learned upsampling via transposed convolution is known to cause checkerboard artifacts and linear interpolation upsampling has been suggested as an effective alternative [42]. Repetition upsampling also is commonly used [16], [17], but likewise exhibits discontinuities, while having virtually the same computational cost as linear interpolation. For these reasons, we use linear interpolation for all upsampling in our experiments.

At the simplest, there is no look-ahead mechanism for conditioning: the acoustic sequence is simply upsampled and the acoustic feature at current time instant is input as the conditioning to each residual block. This approach has been used, for example, in [16] and [43]. As a lightweight option for providing the waveform generator with a future look-ahead, [34] proposed stacking four past and future acoustic frames to the conditioning vector and then applying a global projection before injection into the WaveNet convolution stack. This frame stacking can be interpreted as a single non-causal convolution layer applied to the acoustic feature sequence. In this paper, we compare three conditioning methods that provide various degrees of complexity and acoustic look-ahead:

1) *Simple conditioning* directly inputs upsampled acoustic features to the waveform generator convolution stack. This approach does not allow any look-ahead in the waveform generator model and keeps the overall model fully causal.
2) *Frame stacking* allows a short acoustic look-ahead, e.g., four frames of future context gives a 20 ms look-ahead, similarly to [34]. This is equivalent to applying a single

convolution layer (with filter width 9) to the acoustic frame sequence before upsampling.

3) *WaveNet-on-WaveNet* uses a non-causal WaveNet-like model to encode the acoustic feature sequence, which is upsampled before inputting it to the waveform generator convolution stack. This conditioning setup resembles the byte-net architecture [26] with an added upsampling operation between the encoder and decoder models.

The experiments are described in section IV-C.

III. Conventional vocoders

A. STRAIGHT

STRAIGHT [35], [44] is a classical source-filter vocoder widely used in parametric TTS. The vocoder decomposes the speech signal into a smooth spectral envelope and a spectrally white excitation. The excitation is further characterized by spectral harmonics arising from the fundamental frequency and an aperiodicity spectrum that measures deviation from an ideal harmonic series structure.

Spectral envelope analysis in STRAIGHT is designed to minimize harmonic interference in the envelope estimate by using two complementary pitch-adaptive analysis windows: a primary analysis window consists of a Gaussian window convolved with a triangular b-spline window, while a complementary asymmetric window is created by multiplying the primary window with a sine window. These windows are then used to obtain two complementary spectral estimates that are finally combined via a weighted quadratic mean, whose weights have been optimized to minimize the harmonic interference [35].

Aperiodicity spectrum is estimated by comparing upper and lower envelope values and smoothing them using a look-up table of known aperiodicity values [44]. The resulting aperiodicity spectrum is then parametrized for SPSS by log-averaging over equivalent rectangular bands (ERB), where we used 25 bands. Similarly, the spectral envelope is parametrized as mel cepstral coefficients [15]. At synthesis, a mixed excitation of an F0 controlled impulse train modified to match the aperiodicity statistics is generated for voiced speech, while white noise excitation is used for unvoiced speech. Finally the spectral envelope is applied as a minimum phase version of the envelope magnitude.

B. Glottal vocoders

GlottDNN [33] is a glottal vocoder that uses quasi closed-phase analysis [40] for spectral envelope estimation and utilizes a relatively simple neural network for a pitch-synchronous glottal excitation pulse generator [31]. QCP analysis uses weighted linear prediction (WLP) with the attenuated main excitation (AME) window [40] to reduce the effect of harmonic bias in vocal tract estimates. Another major component in GlottDNN is its glottal excitation model. Glottal pulses are extracted in segments of two pitch periods from the glottal excitation signal, cosine windowed and zero-padded to a constant length vector, such that a glottal closure instant (GCI) is located at the middle of the frame [31]. Then a DNN is trained to predict these glottal pulse vectors from input acoustic features.

There are two key high-level differences between GlotNet and GlottDNN. First, GlotNet generates excitation signals sample-by-sample, whereas GlottDNN operates on frames of pitch-synchronous pulses. Therefore GlotNet needs no explicit F0 information for synthesis pitch-marks, but its inference speed is limited by the autoregressive generation. Second, GlotNet learns a signal distribution and allows generating stochastic components via random sampling at synthesis time, while GlottDNN outputs average waveforms that require separate addition of shaped noise, constructed using the harmonic-to-noise ratio (HNR) parameter. Recently, an alternative non-parametric method of generating stochastic components has been proposed, utilizing generative adversarial networks (GANs) [45], [46]. These GAN-based approaches still rely on pitch-synchronous windowing, but manage to integrate random sampling into the neural waveform model. However, a comparison between GlotNet and "GlotGAN" is not in the scope of this paper and is left for future work.

In the current experiments, GlottDNN uses the glottal excitation model configuration described in [47]. To simplify comparisons, the acoustic features used for the TTS acoustic model targets and neural vocoder inputs is the GlottDNN acoustic feature set (for both the GlotNet and WaveNet neural vocoders). Five kinds of glottal vocoder parameters are used as the acoustic features: 1) frame energy, 2) fundamental frequency with voicing information, 3) vocal tract filter line spectral frequencies (LSFs) (order 30), 4) glottal source spectral envelope LSFs (order 10), and 5) equivalent rectangular bandwidth (ERB) harmonic-to-noise ratios (HNRs) (five bands).

IV. Experiments

A. Speech material

In the experiments, we use speech data from two speakers (one male, one female), who both are professional British English voice talents. The dataset for the male speaker "Nick" comprises 2 542 utterances, totaling 1.8 hours, and the dataset for the female speaker "Jenny" comprises 4 314 utterances, totaling 4.8 hours. For both speakers, 100-utterance test and validation sets were randomly selected from the data, while the remaining utterances were used for training. The material was down-sampled to a 16 kHz sample rate from the original 48 kHz rate. The pitch range for "Nick" is from 78 Hz to 136 Hz with an average of 106 Hz, while pitch of "Jenny" ranges from 101 Hz to 226 Hz, averaging at 161 Hz. These pitch ranges were estimated as two standard deviations from the average on the mel scale.

B. Model specifications

We built speaker-specific systems for the two voices, such that the TTS acoustic models are shared between GlotNet and WaveNet. For the classical vocoders, GlottDNN uses the same acoustic model and shares the acoustic features, while STRAIGHT uses a distinct feature set and a separate acoustic model with a similar architecture. Both acoustic models consist

TABLE I: Model configurations for WaveNet-like models. Wave/Glot models are the sample-rate waveform generator models for the neural vocoders. Cond. net is the non-causal frame-rate model used to condition the "WaveNet-on-WaveNet" systems in section IV-C.

	Wave/Glot 9	Wave/Glot 30	Cond. net
Residual channels	64	64	64
Post-net channels	256	256	64
Filter width	2	2	3
Causal	yes	yes	no
Dilated layers	9	30	8
Dilation cycle length	-	10	4
Number of parameters	602K	1.56M	282K
FLOPs per second	19.3G	50.0G	113K

TABLE II: Model configurations for acoustic models. Input linguistic features (LF) are the same on all systems, while the output acoustic features (AC) are vocoder-dependent. "Glot" acoustic model produces glottal vocoder acoustic features and these features are used by WaveNet, GlotNet and GlottDNN vocoders.

Layer	Glot	STRAIGHT
Input (LF)	396	396
Dense	512	512
Dense	512	512
BLSTM	256	256
BLSTM	256	256
Output (with $\Delta, \Delta\Delta$)	142	259
Output (AC)	48	87

of two fully connected layers (size 512) followed by two BLSTM layers (size 256). The acoustic models are trained to minimize the mean squared error between their output and acoustic features and their deltas and delta-deltas. Dynamic features are omitted for the binary VUV flag. Table II summarizes the acoustic model configurations. Maximum likelihood parameter generation (MLPG) [48] is used to produce the final outputs. Models were trained with the Adam optimizer [49] with an initial learning rate of 1e-4 for a maximum of 100 epochs using a ten epoch early stopping criterion. Both the input and output features were normalized by mean and standard deviation in a feature-wise manner, and Gaussian white noise (with $\sigma = 0.01$) was added to the inputs as regularization.

The waveform generator models use 64 channels with a filter width of 2 in the residual blocks, while the skip connections to the post-processing module have 256 channels, similarly to [50]. The models were trained using the Adam optimizer with exponential learning rate decay. The maximum number of epochs was set to 150 with an early stopping criterion of 10 epochs of no improvement. The input acoustic features to the neural vocoders did not include deltas or delta-deltas.

C. Conditioning models for waveform generators

To study the effects of the conditioning mechanisms described in section II-B, we trained 30-layer WaveNet and GlotNet models conditioned with *simple conditioning*, *frame stacking*, and a non-causal WaveNet-like model (*WaveNet-on-WaveNet*) for our two test speakers "Jenny" and "Nick". Average categorical cross entropy losses on the validation set during training are shown in Figure 2 and the related network configurations are listed in Table I. Based on the validation losses, WaveNet models perform similarly regardless of the conditioning model. This is likely due to the high prediction power the previous speech samples have for the next sample. In contrast, for the GlotNet, the glottal inverse filtering has removed most of the short-time linear dependency, and the waveform generator has to rely more on the conditioning. As a result, GlotNet models show increased performance when rich non-causal conditioning is available. A similar pattern emerges for both speakers. This is further reflected in relatively lower MFCC distances in copy-synthesis, as detailed in section V-B.

Notably, the categorical cross entropy of the GlotNet models converges at a higher level compared to the WaveNets. This was expected, as glottal inverse filtering whitens (decorrelates) the speech signal and results in the glottal excitation signal with higher entropy (c.f. maximum entropy corresponds to having white noise as the signal). While this does not have a large effect on training a WaveNet on the cross-entropy of quantized amplitudes (as we do in this paper), the situation is different when using mixture density networks (MDN) of location-scale family distributions (such as logistic or Gaussian). On a properly trained WaveNet, the speech signal amplitude distributions become very peaky (i.e., of low entropy), which easily causes numerical problems. This is due to the $(x - \mu)/s$ term in the likelihood function, where a peaky distribution corresponds to small values of the scale parameter s. A common failure mode in MDN training is related to exploding gradients when the predicted s reaches a low level, characteristic to the appropriate low entropy. In contrast, the higher entropy target signal provided by the glottal excitation effectively regularizes the MDN training, which partially explains a GlotNet mixture density model significantly outperforming an otherwise equivalent WaveNet model in [34]. The use of continuous distributions in MDN training remains theoretically appealing, as it avoids the amplitude quantization and large output layer inherent to the softmax approach. However, since in the present work we train multiple models with relatively limited data, we opted to use softmax due to its relative robustness and ease of training. Unlike the MDN likelihood function, the softmax cross entropy is bounded at all entropy levels, which avoids the exploding gradient problem.

In addition to comparing likelihoods, we conducted listening tests comparing the subjective quality of the different conditioning methods (similarly to section V-C). However, no statistically significant differences were found, and we therefore omit plots and further analyses from this paper. We chose to use the *frame stacking* conditioning option in all further experiments, as it improves the GlotNet losses over simple conditioning and does not add much computational cost.

D. Effect of reducing model size

The effect of reducing a WaveNet vocoder model size has been previously studied in [51], which reports successful

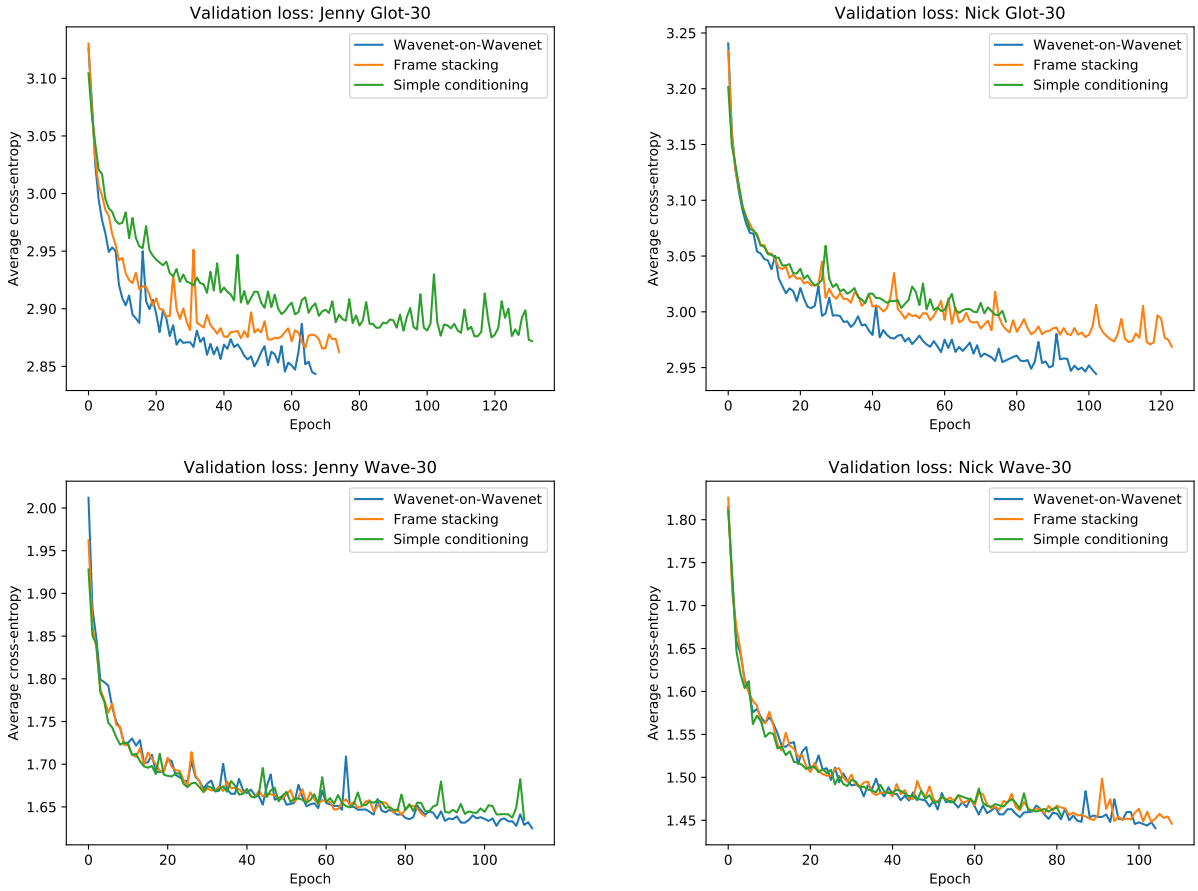

Fig. 2: Validation set losses during training of the GlotNet and WaveNet models with different conditioning methods. The WaveNet models (bottom) behave similarly regardless of conditioning, while the GlotNet models (top) benefit from richer acoustic context provided by "WaveNet-on-WaveNet" conditioning or simple context frame stacking. The loss values are relatively higher for GlotNet, as glottal inverse filtering has removed the effect of vocal tract resonances, thereby reducing the predictability from previous values. As a result, the GlotNet models have to rely more on the acoustic conditioning.

use of 12-layer and 24-layer models, with only a slight quality degradation compared to a 30-layer model. In this paper, we examine two different residual stack depths: first, a commonly used 30-layer configuration that consists of the dilation pattern 1, 2, 4, ..., 512 repeated three times [16], [17], [51], resulting in a receptive field size of 3 071 samples. We call these models "Wave-30" and "Glot-30" for WaveNet and GlotNet, respectively. Second, we study smaller models of 9 layers with dilations 1, 2, 4, ..., 256. These models have a receptive field of 513 samples, which is similar to the window size in previous glottal excitation models [31]. The models are referred to as "Wave-9" and "Glot-9" for WaveNet and GlotNet, respectively.

V. Evaluation

A major focus in this paper is placed on the subjective evaluation of neural and conventional vocoders used in a TTS system, measured in terms of pair-wise comparative quality (section V-C), voice similarity to natural reference (section V-D), and mean opinion scores on quality (section V-E). All systems under evaluation are speaker-dependent TTS systems, where their respective neural vocoder models are trained only on the available speaker-specific data. This is in contrast to previous research on GlotNet [34], which evaluated speaker-independent neural vocoders in a copy-synthesis setting. Additionally, the present paper includes comparisons with two conventional vocoders, STRAIGHT and GlottDNN.

The listening tests were performed in two groups. The first group contains tests where all TTS systems use either WaveNet or GlotNet neural vocoders, and the purpose is to obtain high-resolution snapshots of the fairly subtle differences between the neural vocoders. Meanwhile, the second group contains all pairings of the studied neural and conventional vocoders, and the aim is to provide an overall picture of the performance across systems.

Listening tests were conducted on the Figure Eight (formerly CrowdFlower) crowd-sourcing platform [52]. The tests were made available in English-speaking countries and the top four countries in the EFI English proficiency rating [53]. Each test case was evaluated by 50 listeners, and the test subjects were screened with null pairs and artificially corrupted low-quality anchor samples. The anchor samples were created with the GlottDNN vocoder using a simple impulse train excitation, and by over-smoothing all acoustic features using

a 20-point equal weight moving average filter, and by scaling up the aperiodic energy in voiced segments by a factor of ten. Additionally, listeners with zero-variance responses were excluded from analysis as post-screening. We invite the reader to listen demonstration samples at http://tts.org.aalto.fi/glotnet-demo-tasl/.

A. Computational complexity

The main hyperparameters of a WaveNet that determine the computational complexity are the number of dilated convolution blocks, and the amount of residual and skip channels. Additional fixed costs stem from the input and conditioning embedding layers, and the post-processing module. Table I lists the number of trainable parameters and the estimated floating point operations per second (FLOPS). To estimate the computational complexity of the studied models, we used TensorFlow's inbuilt profiler at 16kHz audio sample rate by performing a parallel forward pass of the model. The parallel FLOPS count serves as a lower bound for the sequential inference computation, and we omit here any additional cost required to maintain the dilation queues in the fast sequential inference algorithms [54]. Furthermore, we excluded the sampling operation from the FLOP profiling. We note that the practical real time factor further involves GPU kernel invocation overheads and other implementation issues [22], but these are independent of the algorithmic focus in the current work.

Our estimates are 19.3G FLOPS for the nine-layer model and 50.0G FLOPS for the 30-layer model. These numbers include a four-frame past and future context in the acoustic conditioning. For the separate frame-rate conditioning model, the parameter count is similar to the WaveNets, but the FLOPS count is significantly lower due to the operation at 200Hz rate (as opposed to 16kHz). For comparison, [19] reported their WaveNet running at 209G FLOPS, although the paper lacks information on the exact model configuration and how the FLOPS count was estimated. For further comparison, our estimation method gives 1008G FLOPS for the model configuration proposed in [16] (30 dilation layers, 256 residual channels, 2048 skip channels).

Furthermore, to estimate the added computational cost from the vocal tract synthesis filter in GlotNet, we implemented the AR synthesis filter in STFT domain using TensorFlow, and ran similar profiling. We assume that the filter coefficients are provided in the direct polynomial form. Using a 5-ms hop size and 512 point FFTs, the filtering uses 511K FLOPS. Increasing the hop size to 10 ms reduces the complexity estimate to 281K FLOPS. Both are orders of magnitude smaller than the WaveNet computational cost, and can be considered close to negligible. Filtering directly in time domain would further reduce the cost, but STFT domain filtering has the added benefit of being fully parallelizable for future work.

B. Objective metrics

To evaluate how well the neural vocoders follow their acoustic conditioning, we computed acoustic error measures between copy-synthesis waveforms and natural reference signals from the test set. We used MFCC distance to measure the spectral distortion and three F0 related features: voicing accuracy, gross pitch error and fine pitch error. For computing the MFCCs, we use a HTK-style [55] triangular mel filterbank with 24 filters and 20 DCT coefficients. The pitch related measures are estimated in the following hierarchy: first we check the voicing estimates for the reference and the generated signal and compute the voicing accuracy as the proportion of frames where the estimates agree. After this, we calculate the gross pitch error (GPE) by counting the number of frames where both estimates are voiced, but the F0 differs more than 20%. Finally, from the remaining voiced frames, we compute the mean absolute fine pitch error (FPE) in cents (100 cents is one semitone, 12 semitones is one octave), i.e.

$$\mathrm{FPE} = 1200 \frac{1}{N} \sum_{n=1}^{N} \left| \log_2 F_0^{\mathrm{ref}}(n) - \log_2 F_0^{\mathrm{gen}}(n) \right|, \quad (2)$$

where N is the number of voiced frames with no gross pitch error. A single data point in Fig. 3 then represents an utterance average.

Fig. 3 shows box plots of the acoustic measures. As expected, the MFCC distances are smallest for the non-generative vocoders STRAIGHT and GlottDNN. Between WaveNets and GlotNets, the latter consistently obtain lower MFCC distances, which indicates a more faithful reproduction of the (spectral) acoustic conditioning.

There are a few notable restrictions on evaluating WaveNet-like models with acoustic measures computed at frame level. A neural waveform generator may drift slightly out of sync with the conditioning and still produce results that are perceptually very close to the reference signal. Furthermore, since we are re-estimating the acoustic features from the generated waveforms, the estimation process adds noise to the error measures.

C. Quality comparison test

A category comparison rating (CCR) [56] test was performed to evaluate the synthetic speech quality. The listeners were presented with pairs of test samples and asked to rate the comparative quality from -3 ("Much worse") to 3 ("Much better"). The CCR scores are obtained by re-ordering and pooling together all ratings a system received. The plots show mean CCR score with 95% confidence intervals, while statistical significance was tested pairwise with the non-parametric Mann-Whitney U-test with Bonferroni correction for the number of pairings. For each test, 15 utterances were selected randomly from the test set, such that a different subset from the 100 utterance test set is used for each test.

To obtain a high-resolution picture of the differences between GlotNet and WaveNet, a first set of evaluations was conducted only with the natural reference and neural vocoders. These tests were further split between the large 30-layer and small nine-layer models to focus on the difference between the GlotNet and WaveNet models of the same size. Figures 4 and 5 show the mean scores with 95% confidence intervals for speakers "Jenny" and "Nick", respectively. U-tests (with

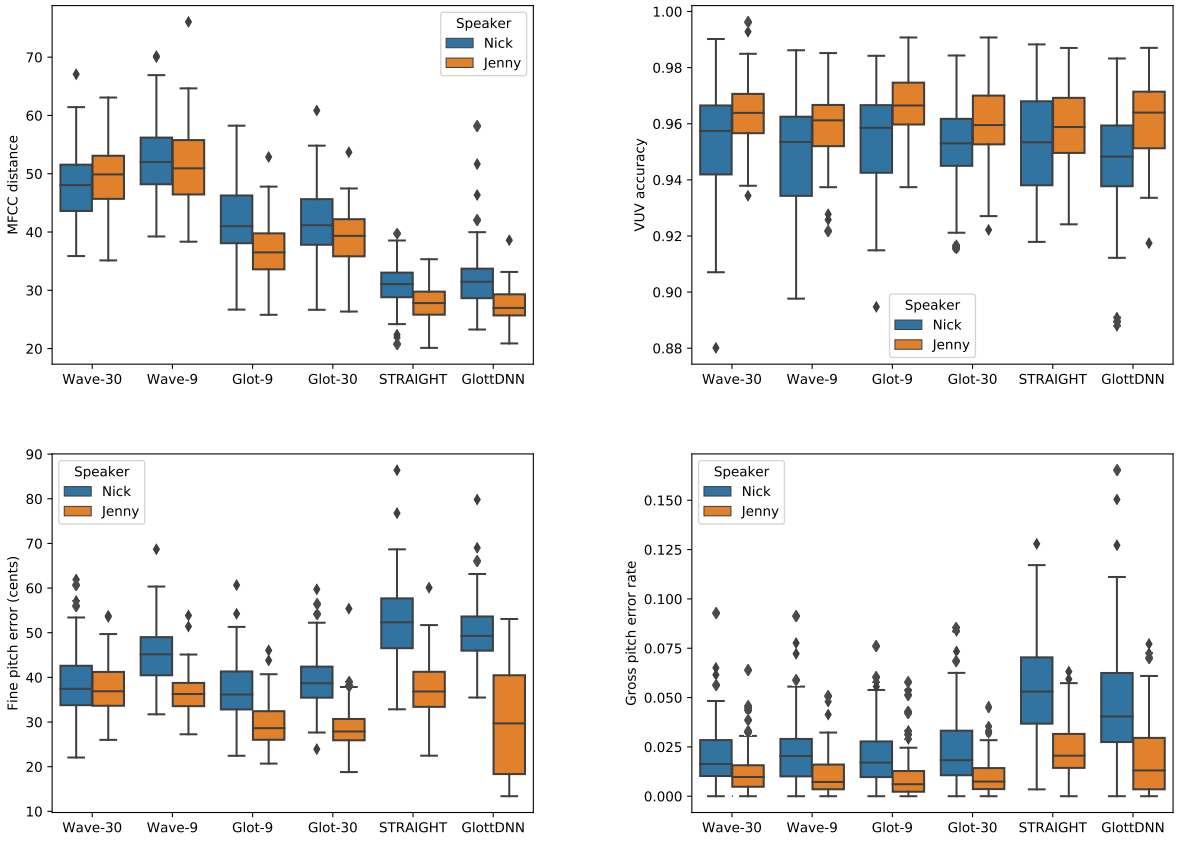

Fig. 3: Objective acoustic measures computed between natural reference speech and copy-synthesis waveforms. Conventional vocoders (STRAIGHT, GlottDNN) achieve the lowest *MFCC distance* (top left), while GlotNet models outperform WaveNets in this metric. All models result in similar *voicing accuracy* (top right), and the neural vocoders achieve similar scores in *gross pitch error* (GPE) (bottom left) and *fine pitch error* (FPE) (bottom right).

Bonferroni corrected $p < 0.05/3$) found all differences to be statistically significant, except for the comparison between Wave-30 and Glot-30 for "Jenny".

A separate large-scale test was conducted to get an overall picture of the different systems, now including conventional vocoders. This test includes all pairwise comparisons between natural reference, Glot-9, Glot-30, Wave-9, Wave-30, GlottDNN, and STRAIGHT. The CCR score plots are shown in Fig. 6. Introducing conventional vocoders to the test seems to re-calibrate the listener judgments, resulting in the already subtle differences between neural vocoders becoming even smaller. For plot clarity, Fig. 6 groups the nine-layer and 30-layer systems to the left and right columns, respectively. However, the scores were computed from the same data including all pairings, and the plots are directly comparable. U-tests indicate significant differences for all pair comparisons, excluding Glot-30 vs. Wave-30, Glot-30 vs. Wave-9, Wave-30 vs. Wave-9 for "Nick", and Glot-30 vs. Wave-30, Glot-30 vs. Wave-9 for "Jenny". Generally, the neural vocoders tend to outperform the conventional vocoders, but somewhat surprisingly GlottDNN tops the quality comparison for "Nick".

Fig. 7 shows copy-synthesis CCR test results. Pairwise differences are statistically significant between all systems,

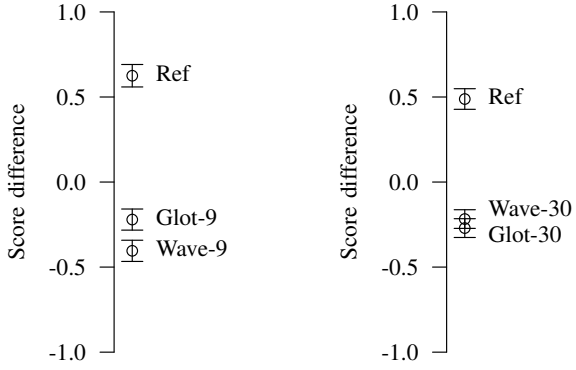

Fig. 4: Combined text-to-speech score differences obtained from the quality comparison CCR test for the female speaker "Jenny". Error bars are t-statistic based 95% confidence intervals for the mean.

except Glot-9 vs. Glot-30, Glot-9 vs. Wave-30 for "Nick" and GlottDNN vs. STRAIGHT, GlottDNN vs. Wave-9, Glot-9 vs. Wave-30, STRAIGHT vs. Wave-9 for "Jenny". In a copy-

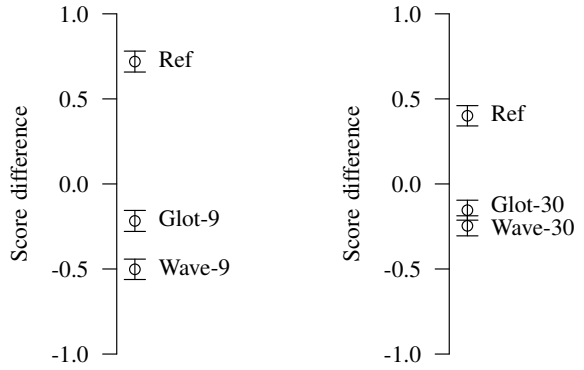

Fig. 5: Combined text-to-speech score differences obtained from the quality comparison CCR test for the male speaker "Nick". Error bars are t-statistic based 95% confidence intervals for the mean.

synthesis setting, Glot-30 is the best performing vocoder for both speakers, followed by Wave-30 and Glot-9 at roughly equal comparative quality. The Wave-9 model falls behind the other neural vocoders and receives scores similar to GlottDNN and STRAIGHT.

D. Voice similarity test

Voice similarity between synthetic speech and a natural reference was evaluated in a DMOS-like test [56]. The listeners were asked to rate the voice similarity of a test sample to a natural reference (with the same linguistic contents) using a 5-level absolute category rating scale, ranging from "Bad"(1) to "Excellent"(5). For the tests, 20 utterances were selected randomly from the test set. Figure 8 shows system mean ratings with 95% confidence intervals and stacked histograms for the rating levels. The Mann-Whitney U-test found no significant differences between systems for "Jenny", whereas all differences for "Nick" were statistically significant, except between Wave-30 and Glot-30. All U-tests and the plotted confidence intervals use Bonferroni correction ($p < 0.05/6$) for the 6 pair-comparisons between the four systems. Figure 9 shows voice similarity ratings for the Wave-30 and Glot-30 systems when the test also included STRAIGHT and GlottDNN. All pairwise differences were statistically significant, except Glot-30 vs. Wave-30 and Glot-30 vs. GlottDNN for "Jenny", and Glot-30 vs. Wave-30 for "Nick". The neural vocoders achieve higher ratings than the conventional vocoders, but are also rated worse than in the previous similarity DMOS test. The effect may be due to different random sampling of utterances or listener judgment recalibration in the presence of conventional vocoders.

E. Mean opinion score quality test

To get a general view of the synthetic speech quality on the absolute category rating scale, we conducted additional mean opinion score (MOS) tests [56]. The listeners were asked

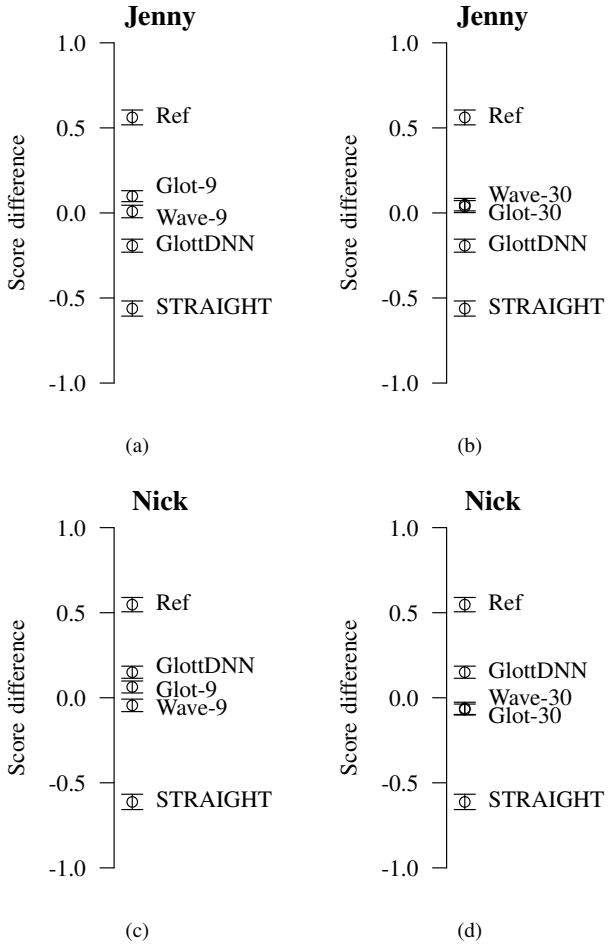

Fig. 6: Text-to-speech CCR quality scores comparing different neural and conventional vocoders for "Jenny" (upper panes) and "Nick" (lower panes). For clarity, the nine-layer WaveNet and GlotNet models are grouped to the left and the 30-layer WaveNet and GlotNet models are grouped to the right, but the scores are directly comparable across columns.

to rate the quality of the presented stimulus on a five-point scale ranging from "Bad"(1) to "Excellent"(5). For the tests, 16 utterances were selected randomly from the test set and 50 listener ratings were collected for each system. Figure 10 shows the MOS test results. Overall, all neural vocoders and GlottDNN received high ratings, while the natural reference and STRAIGHT stand out above and below this group, respectively. For "Jenny", U-tests with Bonferroni correction found no statistically significant differences between the neural vocoders and GlottDNN. A similar result holds for "Nick", with the exception that Wave-9 received slightly (but statistically significant) lower rating.

VI. Conclusion

In this study, we apply GlotNet, a WaveNet-like glottal excitation model, for statistical parametric speech synthesis. The method is compared with a WaveNet vocoder of an equivalent architecture by using both approaches as neural vocoders in speaker-specific TTS systems (one female, one male voice).

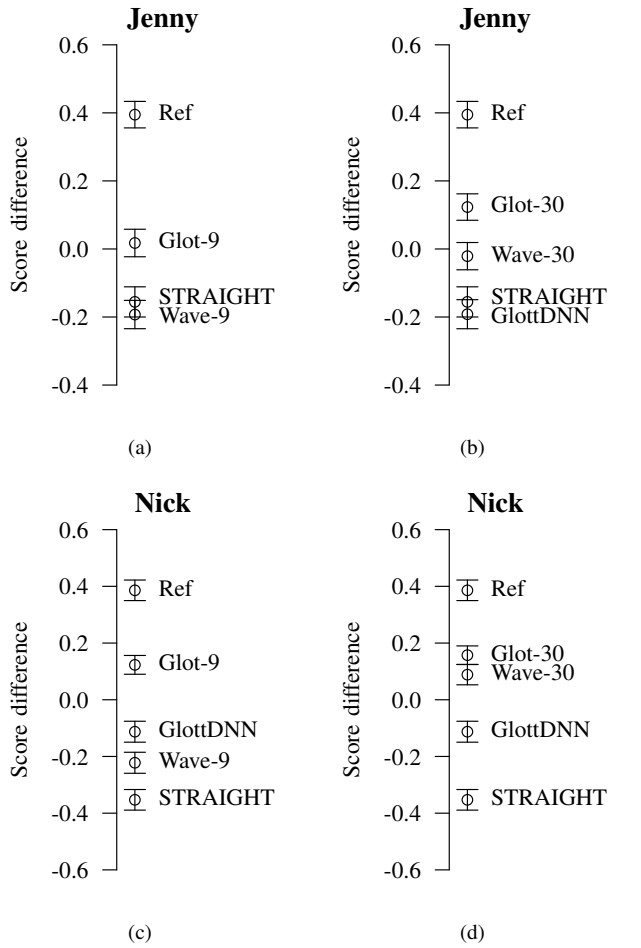

Fig. 7: Copy-synthesis CCR quality scores comparing different neural and conventional vocoders for "Jenny" (upper panes) and "Nick" (lower panes). For clarity, the nine-layer WaveNet and GlotNet models are grouped to the left and the 30-layer WaveNet and GlotNet models are grouped to the right, but the scores are directly comparable across columns.

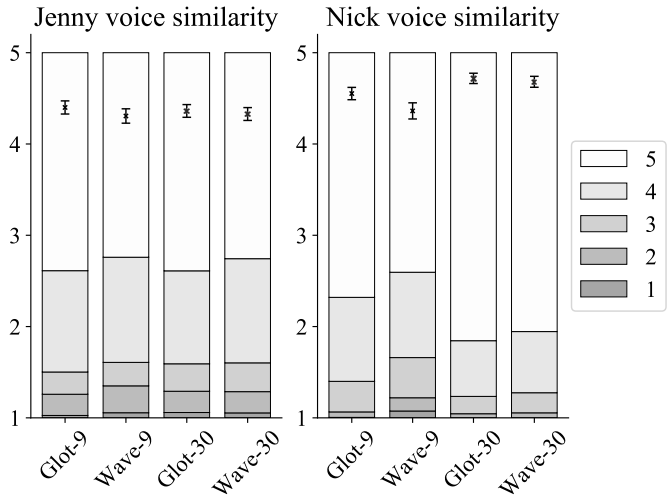

Fig. 8: Voice similarity ratings in text-to-speech on different size GlotNet and WaveNet models for "Jenny" (left) and "Nick" (right). The plot shows mean ratings with 95% confidence intervals, along with stacked rating distribution histograms.

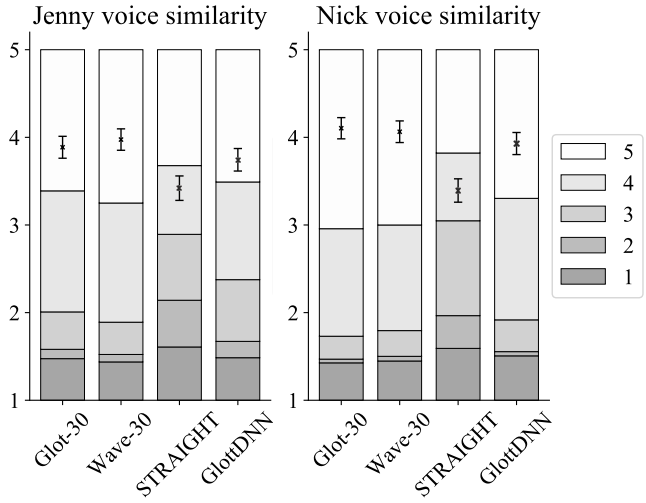

Fig. 9: Voice similarity ratings in text-to-speech on different vocoders for "Jenny" (left) and "Nick" (right). The plot shows mean ratings with 95% confidence intervals, along with stacked rating distribution histograms.

Furthermore, both our GlotNet and WaveNet neural vocoders perform favorably to conventional vocoders STRAIGHT and GlottDNN in subjective evaluations (except that GlottDNN gave the best quality score for "Nick"). Additionally, we study the effect of different conditioning mechanisms of increasing complexity for the neural vocoders, and find that GlotNet benefits from a future look-ahead in the acoustic conditioning.

To reduce the amount of computation needed, we intentionally limit ourselves to relatively small datasets of 1.8 and 4.7 hours of speech. This effectively rules out elaborate local conditioning strategies with, e.g., recurrent neural nets. Nevertheless, both the proposed GlotNet model and WaveNet achieve high speech quality, which (based on our informal opinion) is mostly limited by the over-smooth prosody provided by the acoustic model of the TTS system. The listening test evaluation for TTS shows that all our waveform generator variants are rated highly in terms of voice similarity to the original speaker. In pairwise quality comparisons, GlotNet and WaveNet obtain similar ratings with the widely used 30-layer residual stack architecture. However, when the residual stack depth is reduced to nine layers, GlotNet retains a high quality. Furthermore, in copy-synthesis listening experiments using ground truth acoustic features, the 30-layer GlotNet is consistently rated highest and a nine-layer GlotNet shows similar performance to a 30-layer WaveNet. Not surprisingly, the quality gap between the natural reference and the generated speech is narrower when using natural acoustic features in copy-synthesis, compared to the synthetic acoustic features given by the TTS acoustic model. Furthermore, the relative performance of the studied neural vocoders is more consistent (also across speakers) when doing copy-synthesis. In contrast, using generated acoustic features leads to varying degrees

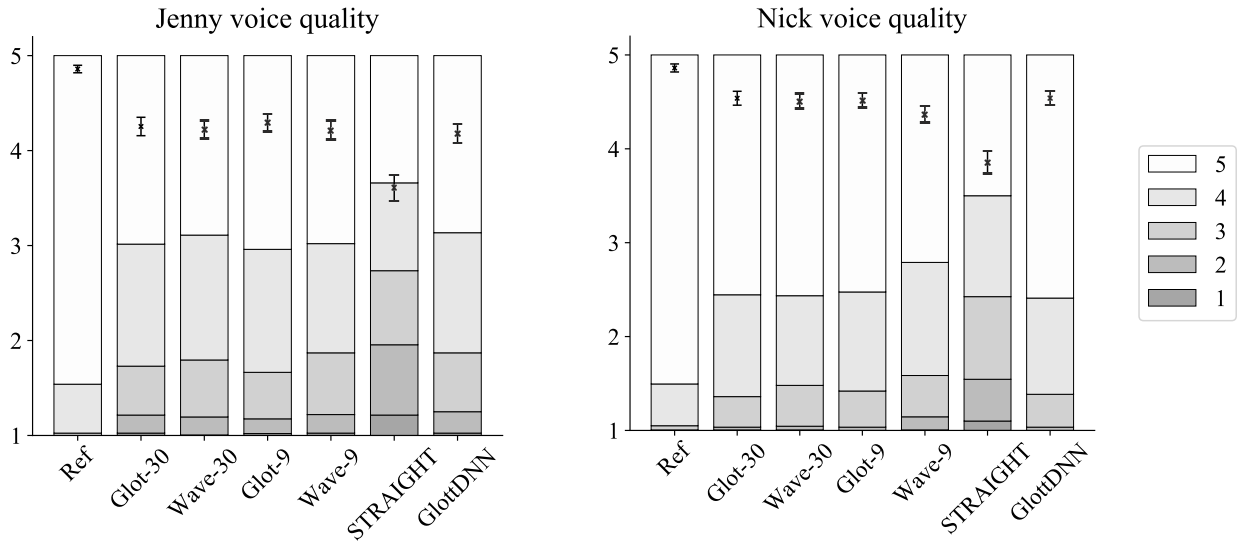

Fig. 10: Mean opinion score (MOS) quality ratings for text-to-speech systems. The plot shows mean ratings with the 95% confidence intervals (corrected for multiple comparisons), along with stacked score distribution histograms.

of degradation depending on the speaker and neural vocoder model. The current results correspond to using neural vocoders in an existing SPSS system, and the results may change when using a combination of more advanced acoustic models, joint training, and larger datasets.

We further found that the GlottDNN vocoder, which uses a simple feed-forward neural net to generate excitation waveforms in a frame-wise manner, is occasionally rated on par with the WaveNet-type of neural vocoders (i.e., ones that operate on raw signal sample-by-sample). This is not completely unexpected, as GlottDNN has previously been found to perform well, especially on the "Nick" dataset [33]. In addition, due to using a trainable neural net in excitation generation, GlottDNN can be regarded as a kind of hybrid neural vocoder, which combines properties of conventional and neural vocoders. Meanwhile, a different glottal neural vocoder has recently been reported to achieve comparative performance to a WaveNet [19], which makes frame-based glottal neural vocoders a viable option in limited resource systems.

While recent work has found ways to speed up WaveNet inference by parallelization [20], at the current large model sizes WaveNets are still arguably expensive to use in production systems. Exploring smaller, faster models (without trading off system performance too much) is useful not only for developing new ideas, but also in the large-scale deployment of WaveNet-like models. To this effect, the results in this paper suggest that switching the speech waveform domain to the domain of the glottal excitation facilitates the use of smaller models and is worth investigating further, as the generative neural network methods keep evolving. The present results encourage future research in training compact parallel WaveNet-like [20] excitation models. Another potential research avenue is building sequential glottal excitation models with efficient WaveRNN-based models [22] instead of WaveNet.

ACKNOWLEDGMENT

This study was supported by the Academy of Finland (project 312490). We acknowledge the computational resources provided by the Aalto Science-IT project.

REFERENCES

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, May 2013, pp. 7962–7966.

[3] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks." in *Interspeech*, 2014, pp. 1964–1968.

[4] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. ICASSP*. IEEE, 2015, pp. 4470–4474.

[5] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Waw: End-to-end speech synthesis," in *ICLR, workshop track*, 2017, https://openreview.net/pdf?id=B1VWyySKx.

[6] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1452

[7] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions*, vol. 99-D, pp. 1877–1884, 2016.

[8] 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.

[9] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, April 2014.

[10] G. Degottex, P. Lanchantin, and M. Gales, "A log domain pulse model for parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 57–70, Jan 2018.

[11] Y. Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in *Proc. ICASSP*, April 2015, pp. 4230–4234.

[12] F. Espic, C. V. Botinhao, and S. King, "Direct modelling of magnitude and phase spectra for statistical parametric speech synthesis," in *Proc. Interspeech*, 2017, pp. 1383–1387. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1647

[13] 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: http://arxiv.org/abs/1609.03499

[14] 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.

[15] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – a unified approach to speech spectral estimation." in *Proc. ICSLP*, 1994, pp. 18–22.

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

[17] 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.

[18] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," in *Proc. Interspeech*, 2017, pp. 1138–1142.

[19] Y. Cui, X. Wang, L. He, and F. K. Soong, "A new glottal neural vocoder for speech synthesis," in *Proc. Interspeech*, 2018, pp. 2017–2021.

[20] 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

[21] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," *arXiv pre-print*, 2018. [Online]. Available: https://arxiv.org/abs/1807.07281

[22] 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," *arXiv pre-print*, 2018. [Online]. Available: http://arxiv.org/abs/1802.08435

[23] J. Vít, Z. Hanzlíček, and J. Matoušek, "On the analysis of training data for WaveNet-based speech synthesis," in *Proc. ICASSP*, 2018, pp. 5684–5688.

[24] Y. Gu and Z.-H. Ling, "Waveform modeling using stacked dilated convolutional neural networks for speech bandwidth extension," in *Proc. Interspeech*, 2017, pp. 1123–1127.

[25] D. Rethage, J. Pons, and X. Serra, "A WaveNet for speech denoising," in *Proc. ICASSP*, 2018, pp. 5069–5073.

[26] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. van den Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," *arXiv pre-print*, 2016. [Online]. Available: http://arxiv.org/abs/1610.10099

[27] 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.

[28] T. Raitio, H. Lu, J. Kane, A. Suni, M. Vainio, S. King, and P. Alku, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *Proc. EUSIPCO*, 2014.

[29] T. Raitio, A. Suni, L. Juvela, M. Vainio, and P. Alku, "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort," in *Proc. Interspeech*, 2014, pp. 1969–1973.

[30] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2–3, pp. 109–118, 1992, Eurospeech '91.

[31] 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*, 2016, pp. 5120–5124.

[32] M.-J. Hwang, E. Song, J.-S. Kim, and H.-G. Kang, "A unified framework for the generation of glottal signals in deep learning-based parametric speech synthesis systems," in *Proc. Interspeech*, 2018, pp. 912–916.

[33] 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, Sept 2018.

[34] L. Juvela, V. Tsiaras, B. Bollepalli, M. Airaksinen, J. Yamagishi, and P. Alku, "Speaker-independent raw waveform model for glottal excitation," in *Proc. Interspeech*, 2018, pp. 2012–2016.

[35] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, "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.

[36] A. W. Black and K. A. Lenzo, "Flite: a small fast run-time synthesis engine," in *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

[37] K. Richmond, R. A. Clark, and S. Fitt, "Robust LTS rules with the Combilex speech technology lexicon," in *Proc. Interspeech*, Brighton, September 2009, pp. 1295–1298.

[38] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proc. ISCA SSW6*, 2007, pp. 294–299.

[39] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. C. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," in *Proc. ICLR*, 2017.

[40] 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, 2014.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, June 2016, pp. 770–778.

[42] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard

[43] T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Mel-cepstrum-based quantization noise shaping applied to neural-network-based speech waveform synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1177–1184, 2018.

[44] 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," in *Proc. MAVEBA*, 2001.

[45] 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.

[46] 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.

[47] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "Reducing mismatch in training of DNN-based glottal excitation models in a statistical parametric text-to-speech system," in *Proc. Interspeech*, 2017, pp. 1368–1372.

[48] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, vol. 3, 2000, pp. 1315–1318.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

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

[51] 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.

[52] Figure Eight Inc., "Crowd-sourcing platform," https://www.figure-eight.com/, accessed: 2018-09-13.

[53] "EF English proficiency index," http://www.ef.com/epi/, accessed: 2017-10-24.

[54] P. Ramachandran, T. L. Paine, P. Khorrami, M. Babaeizadeh, S. Chang, Y. Zhang, M. A. Hasegawa-Johnson, R. H. Campbell, and T. S. Huang, "Fast generation for convolutional autoregressive models," in *Proc. ICLR (Workshop track)*, 2017.

[55] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, *The HTK book*. Cambridge University, 2002, vol. 3.

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