A Literature Review of WaveNet: Theory, Application and Optimization

Jonathan Boilard
Ubisoft
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Philippe Gournay
Université de Sherbrooke
79 PUBLICATIONS **318** CITATIONS
SEE PROFILE

R. Lefebvre
Université de Sherbrooke
65 PUBLICATIONS **782** CITATIONS
SEE PROFILE

A Literature Review of WaveNet: Theory, Application and Optimization

Jonathan Boilard, Philippe Gournay, and Roch Lefebvre

Speech and Audio Research Group,
Université de Sherbrooke
Sherbrooke (Québec) J1K 2R1
Canada

Correspondence should be addressed to Jonathan Boilard (jboilard1994@gmail.com)

ABSTRACT

WaveNet is a deep convolutional artificial neural network. It is also an autoregressive and probabilistic generative model; it is therefore by nature perfectly suited to solving various complex problems in speech processing. It already achieves state-of-the-art performance in text-to-speech synthesis. It also constitutes a radically new and remarkably efficient tool to perform voice transformation, speech enhancement and speech compression. This paper presents a comprehensive review of the literature on WaveNet since its introduction in 2016. It identifies and discusses references related to its theoretical foundation, its application scope, and the possible optimization of its subjective quality and computational efficiency.

1 Introduction

Machine learning and artificial intelligence have recently contributed new signal and data processing tools with applications in many fields, including speech and audio processing. Among these tools is WaveNet [1], a deep artificial neural network adapted to the task of processing and generating raw audio waveforms autoregressively. WaveNet quickly emerged as a very promising tool for speech processing because its autoregressive model fittingly represents the speech generation process [2]. Many variations have already been published. Some of these variations improve the subjective quality of the generated audio, for example by presenting mel-spectrograms to the WaveNet network in Tacotron 2 [3]. Other variations significantly increase the computational efficiency of the neural network by introducing parallelization and distillation techniques [4][5], thus enabling WaveNet to be deployed on systems with more limited processing resources. Some further variations combine WaveNet with other architectures, such as the Glow architecture [6] in WaveGlow [7] or generative adversarial networks (GANs [8]) in [9][10][11], to increase the perceived quality of the output speech.

This paper presents a comprehensive review of the literature on WaveNet since its introduction in 2016. It identifies and discusses more than 60 references related to its theoretical background, its application scope, and its possible performance optimizations.

The paper is organized as follows. The principle and basic architecture of WaveNet are described in Section 2. The concepts of global and local conditioning, which are the basis for most applications of WaveNet, are presented in Section 3.

The application of WaveNet to solving selected problems in speech and audio processing is considered in Section 4. Papers dealing mainly with improving the subjective quality of the audio generated by WaveNet are summarized in Section 5. Papers focusing on improving computational efficiency are presented in Section 6. Finally, conclusions and perspectives are given in Section 7.

2 The WaveNet Architecture

WaveNet [1] is a convolutional neural network that performs autoregressive audio waveform generation. Its structure is inspired from PixelCNN [12], a network that auto-completes an occluded image according to its content by generating pixel predictions from a pixel's nearest neighbours. Unlike PixelCNN, which operates on two-dimensional RGB images, WaveNet processes audio waveforms, which are unidimensional time-series.

2.1 Fully Probabilistic and Autoregressive

The WaveNet generative model predicts a conditional probability distribution for sample x_t given a sequence of past generated samples $x = \{x_1, \ldots, x_{t-1}\}$. Thus, the probability of the audio sequence x is a chain rule of the conditional probabilities of every individual sample given their previous samples, as follows:

$$p(x) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}). \quad (1)$$

This points to two important aspects of WaveNet:

- It is a **fully probabilistic** model. To produce an output sample, it predicts a probability distribution function and selects the most probable discrete value from that distribution.
- It is an **autoregressive** model. It uses the new samples that it generates to determine the following samples.

2.2 μ-law Quantization

Since digital audio is usually represented using 16-bit integers, the fully probabilistic nature of WaveNet would require it to predict a probability distribution with 65,536 possible audio sample values. This would lead to an overly complex network. In the original WaveNet paper, this problem was partially solved by applying μ-law quantization [13] to the signal to be modeled:

$$f(x_t) = sign(x_t) * \frac{ln(1+\mu|x_t|)}{ln(1+\mu)}, \quad (2)$$

where $\mu = 255$ and $-1 < x_t < 1$. This reduces the dimensionality of the probability distribution to only 256 possible values, making it simpler for the network to learn a proper representation of the input audio samples.

Storing digital audio as 8-bit integers, even when using μ-law, induces a reduction in audio quality. However, as stated in [1], the reconstructed quantized signal sounds reasonably similar to the original unquantized one, especially for speech.

2.3 Dilated Causal Convolutions

Dilated causal convolutions are at the core of WaveNet. These forms of convolution can be compared to traditional linear filters, where the output is the weighted sum of some of the inputs. In the case of causal convolutions in deep learning, extra non-linear operations are applied after each convolution operation to enable non-linear expressivity. This makes it possible to learn input audio representations that cannot be captured using linear operations only. The causal convolution structure assures that only previous samples are used to generate the new ones.

Dilated causal convolutions are implemented in WaveNet by doubling the dilation factor until a set limit is reached, then repeated a set number of times {1, 2, 4, …, 512, 1, 2, 4, …, 512, … }. Stacking dilated convolution operations one over another as illustrated in Figure 1 exponentially increases the temporal support of the network, that is, the number of input samples required to calculate one output sample. The chosen architecture achieves a good compromise between the length of this temporal support, the computational complexity, and the modeling performance [14].

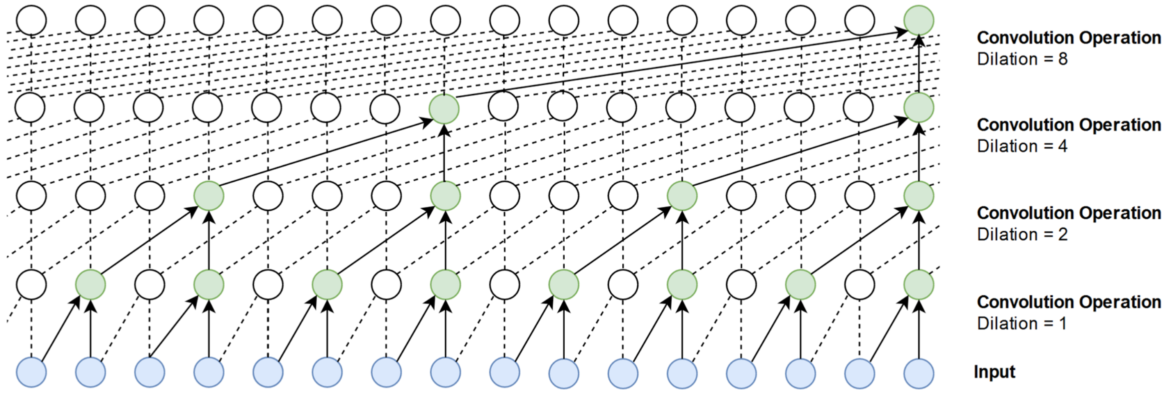

Figure 1. An illustration of dilated causal convolutional operations (figure inspired from [1]). Each green node performs a learned weighted sum of its inputs and applies a non-linearity. The arrows represent the current pipeline of operations and will shift to the right every time a new sample is generated.

2.4 Residual Learning Framework

To avoid the vanishing gradient problem associated with very deep neural networks [15] and thus to promote convergence and reduce training time, WaveNet adheres to a residual learning framework [16]. The structure of the residual units used in WaveNet is shown in Figure 2. Their integration and interconnection in the whole architecture of WaveNet is shown in Figure 3.

This architecture integrates shortcut connections, which are implemented under the form of skip-connections and identity mappings as shown in Figure 2. Skip-connections $F_2(r)$ bypass the following residual layers as shown in Figure 3. Identity mappings, on the other hand, consist in an element-wise addition between the residual input r and the non-linear output $F_1(r)$. The introduction of shortcut connections facilitates gradient propagation through all layers, which solves the vanishing gradient problem. Once optimized, a deep architecture implementing shortcut connections will be faster to train than architectures not implementing them [16].

As shown in Figure 2 and Figure 3, WaveNet implements several 1×1 convolution operations [17]. In the context of deep learning, a 1×1 convolution is an operation that takes as input a matrix composed of data points, each data point being a collection of a certain number of features (for example, a pixel and its three color channels). The 1×1 convolution operation computes a learned weighted sum of these features and outputs another data point with a possibly different number of features. This operation is repeated to cover the entire input matrix. It produces an output matrix with the same dimensions as the input matrix, but not necessarily the same number of features per data point. This operation is essential in the residual unit of Figure 2 to bring the number of features back to M, and before the SoftMax function in Figure 3, to bring the number of output non-normalized probabilities back to the number of discrete values to classify.

At the input of the network, a causal convolution without dilatations is applied to the μ-law audio samples.

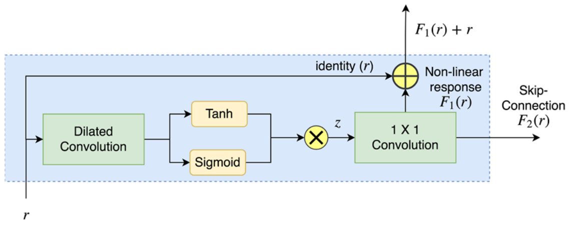

Figure 2. The residual unit used in WaveNet. Figure inspired from [1], with additional labels to better describe the residual network architecture.

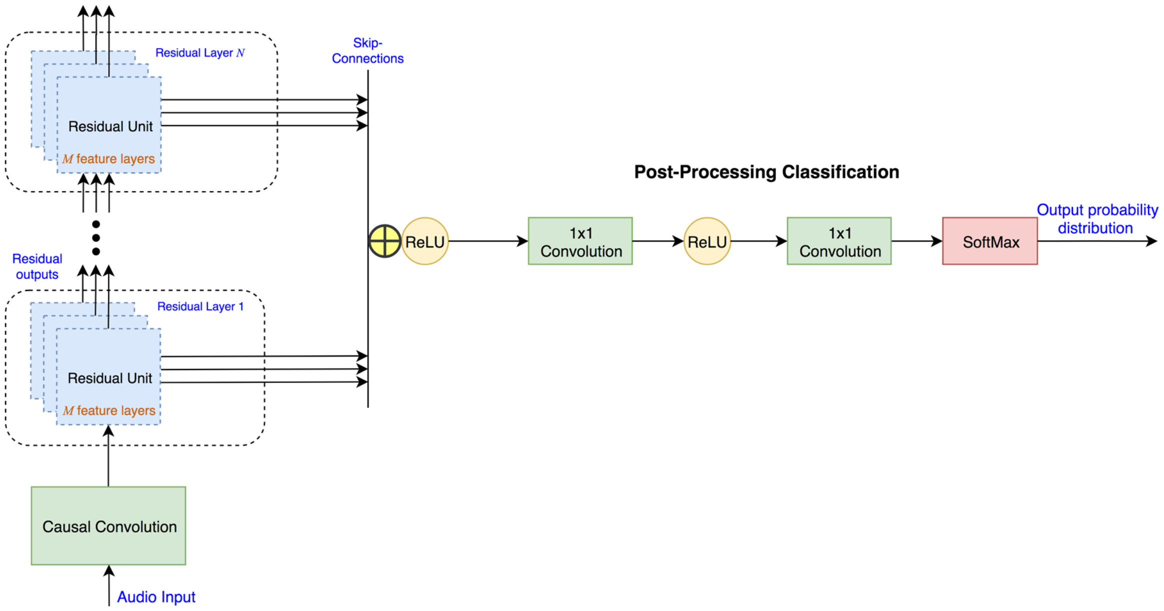

Figure 3. The architecture of WaveNet. Figure inspired from [1], with some additional blocks showing how N residual layers of M feature layers are stacked one over another. The dilation factor of the convolution operations changes from one layer to the other, as indicated in Figure 1.

2.5 Gated Activation Units

The main source of non-linearity in the WaveNet architecture comes from the activation units following the dilated causal convolution operation in the residual unit of Figure 2. They are the same gated activation units as in the PixelCNN network [12]:

$$z = \tanh\left(W_{f,k} * r\right) \odot S\left(W_{g,k} * r\right), \tag{3}$$

where r is the input of the residual unit, $*$ denotes a dilated causal convolution operation, $\odot$ is an element-wise multiplication, k is the layer index, f and g respectively denote filter (hyperbolic tangent activation tanh) and gate (sigmoid activation S), and W indicates learnable convolution filters. Experimental results suggest that this type of non-linearity works significantly better than a rectified linear unit (ReLU) activation function for modeling audio signals [1].

2.6 Softmax Distribution

WaveNet predicts a non-normalized probability distribution and transforms it into a proper probability distribution by using a normalized exponential function known as the SoftMax function:

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}\,. \tag{4}$$

The SoftMax function is a generalized logistic function that takes as input a vector z of K real values and outputs K positive real numbers that sum to 1.

3 Conditional WaveNet

Conditioners add parameters to the probability distribution function so that it depends not only on the previously generated samples, but also on some variables that describe the audio to be generated. Equation 1 then becomes:

$$p(x) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, h\,), \tag{5}$$

where h is a set of conditioners. Without h, the autoregressively generated sample x_t would have the most probable value depending on the previously generated samples, meaning that the final result will likely be a generalization of what WaveNet has learned to generate. For example, instead of generating a single-speaker sequence of speech that has the structure of a language, the result without conditioning would rather be a mixed-up sequence of phoneme shifting repeatedly between the voices of all the speakers on which WaveNet has been trained.

3.1 Global Conditioning

Global conditioning consists in using conditioners that describe the audio samples to be generated through all time steps, for example the speaker's identity. This enables WaveNet to model a conditional distribution that not only depends on previously generated samples, but also on the required global characteristics of the audio signal.

With the addition of global conditioning, Equation 3 describing the gated activation unit becomes:

$$z = \tanh\left(W_{f,k} * x + V_{f,k}^{T} h\right) \odot \sigma\left(W_{g,k} * x + V_{g,k}^{T} h\right), \quad (6)$$

where $V_{*,k}^{T}$ is a learnable transformation of the global conditioner h, projected to all time steps of the generated audio samples' discrete time resolution.

3.2 Local Conditioning

Local conditioners direct and organize the audio samples to be generated according to desired characteristics over windows of time. Text characters, encoded speech parameters, phonemes and musical notes are all valid local conditioners. A local conditioner usually has a lower sampling rate than the desired audio sequence. It thus need to be mapped into the same time resolution as the generated audio sequence. This can be done by using a transposed convolution operation that effectively applies learned upsampling. Conditioner h thus becomes $y = f(h)$, which is adjusted to the resolution of the audio sequence to be generated. It is then applied to the gated activation unit described by Equation 3 as follows:

$$z = \tanh\left(W_{f,k} * x + V_{f,k} * y\right) \odot \sigma\left(W_{g,k} * x + V_{g,k} * y\right), \quad (7)$$

where $V_{*,k} * y$ is a 1×1 convolution that upsamples the local conditioner.

In some implementations [3][18][19], a different approach is chosen in which $V_{*,k}$ includes an attention mechanism that effectively learns how to properly process the given and desired contextual information and learns the varying time dependencies between conditioners. Attention mechanisms enable WaveNet to consider current, previous, and possibly future conditioners for non-causal variations, to generate a sample.

4 Applications of WaveNet

In the original WaveNet paper [1], the model was adapted to different tasks. The first task was multi-speaker speech generation, where the model was only conditioned by a speaker identifier. In this case, WaveNet generates human speech that does not capture the linguistic syntax of the training data. This implementation served mainly as a proof of concept, showing the capacity of WaveNet to capture the identity of various speakers.

The second task was text-to-speech synthesis, where the WaveNet model was conditioned by linguistic features such as phonemes, their duration, and the fundamental frequency F0, all of which being provided by an additional acoustic model. This WaveNet model delivers better speech quality, expressed in terms of mean opinion scores (MOS), than conventional concatenative and parametric text-to-speech synthesis systems.

The third task was music generation. Both conditioned and unconditioned WaveNet models were experimented for this task. To increase the musicality of the generated samples, the temporal support of the network was increased. However, even with an increased temporal support, the dependencies captured between distant samples were not enough to prevent second-to-second variations in genre, musical instrument and volume. The resulting generated music was not formally evaluated by human subjects, but it is said to be often harmonic and aesthetically pleasing.

The final task was speech recognition. The suggested adaptation was non-causal. A mean-pooling layer used to group samples together into frames was added after the dilated causal convolutions of the residual unit from Figure 2. The mean-pooling layer was also followed by a few non-causal convolutional layers. Furthermore, it was trained with two loss terms, one to predict the next sample and the other to classify the current frame. This WaveNet model obtained state-of-the-art results compared to other deep learning methods based on recurrent neural networks.

In the following sections, many more applications of WaveNet are presented.

4.1 Text-to-speech

In Tacotron 2 [3], text embeddings are not used directly as local conditioners. Instead, they are first converted into a mel-spectrogram by a complementary deep neural network. This spectrogram generator uses location sensitive attention [20], introducing in Tacotron 2 a bi-directional recurrent neural network (BRNN). This BRNN allows WaveNet to learn dependencies between past and current local conditioners, and to make use of these dependencies when generating a new sample. Attention mechanisms are based on sequence-to-sequence models. An example of the usage of these models is in the task of language translation. In this context, an attention mechanism enables the model to be aware of a word's dependency with the other words in a sentence (the linguistic syntax, which can greatly differ from a language to another). This process is also useful for speech processing, where knowledge of what and how words in a sentence where pronounced can enhance the naturalness and flow of the generated sequence of samples. It is claimed that the Tacotron 2 model achieves a mean opinion score (MOS) of 4.53, which is comparable to the MOS of 4.58 obtained for professionally recorded speech.

A variant of Tacotron 2 is presented in [21], where the BRNN structure of [3] is replaced by a multi-head attention mechanism [22]. Instead of evaluating the local conditioners with a broad single attention mechanism, this particular attention mechanism integrates different learned representations of the local conditioners, separates them into different periods of time, and evaluates their respective dependencies. Multi-head attention can be parallelized, which reduces training and generation time, with a speed-up factor of 4.25 compared to the original Tacotron 2. It also achieves a MOS of 4.39, compared to a MOS of 4.44 for professional recordings.

A network similar to Tacotron 2, but specialized to generate voices it was not trained to generate, is presented in [18]. From a target reference waveform, speaker embeddings are generated. This process is learned using a specific loss function called "generalized end-to-end loss" [23]. This particular loss function ensures that the embeddings generated from utterances of similar speakers have a high degree of similarity, while the embeddings for dissimilar speakers are farther apart in the embedding space. The proposed model takes as input certain linguistic features such as a sequence of graphemes or phonemes, passes them through a neural synthesizer of log-mel spectrogram, and finally provides the output of this synthesizer to a WaveNet decoder. Authors report a speech naturalness MOS of 4.2, but a speech similarity MOS of 3.28, on unseen speakers when trained on the VCTK speech corpus [24].

Deep Voice 2 [25] is a modification of its predecessor Deep Voice [26], where the neural network decoder is replaced by a WaveNet decoder that takes for conditioning a sequence of phonemes. The duration and fundamental frequency of the phonemes are afterwards both predicted from the phoneme labels extracted from text and from a given speaker identity. Deep Voice 2 also includes some minor optimizations. Deep Voice 3 [27] proposes a similar architecture but with a different mel-spectrogram prediction network, and is adapted to other generation techniques than WaveNet.

A slight modification to the original WaveNet text-to-speech system from [1] is presented in [28], where a conditioner, the fundamental frequency (F0), is predicted from the linguistic features by some pre-processing layers instead of given. During training, this architecture influences the gradient of the pre-processing layers to learn a solution optimizing two objectives, potentially varying their weights into a more optimal solution. This is found to improve the quality of the resulting audio compared to the original WaveNet architecture.

4.2 Speech Compression

Another application of WaveNet is as a decoder in a speech compression application [29]. Because of parametrization and data compression, some speech quality and naturalness is normally lost when using vocoders such as Codec 2 [30]. WaveNet can learn to compensate for this degradation, and can be used in place of the decoder. In the model proposed in [29], the input audio is encoded into parameters using Codec 2, and these parameters are used as conditioners for a WaveNet decoder. A quality-check classifier is also included, which enables the network to switch between WaveNet decoding and predictive decoding. During training, the classifier learns to switch between modes by evaluating the speech degradation of the WaveNet decoder output with respect to the original input. During inference, the classifier chooses between modes independently using only the coded parameters. Transitions between these two modes are smoothed by a classical synchronization mechanism known as waveform similarity overlap-add (WSOLA) [31]. The resulting system compresses speech at a bitrate of 2400 bits/s. The MOS score of the decoded speech signal is estimated at 2.9, compared to 2.7 for the decoder of Codec 2. At 42 kbit/s, the MOS score is estimated at 4.7, very close to the original signal's MOS.

4.3 Voice Conversion

Voice conversion solutions can take advantage of WaveNet's ability to generate speech by modifying its conditioners in different ways. One implementation extracts various acoustic features from the input speech, and then adapts these features to the desired voice [32]. These new features are then passed to a WaveNet as conditioners. This voice conversion technique achieves higher conversion accuracy compared to conventional voice conversion techniques based on Gaussian mixture models.

Another proposed model aims at directly learning how to transform a voice into another by using only the input acoustic features [33]. During training, the acoustic features of the input-voice audio sequence are fitted to the target-voice audio sequence using dynamic time warping. During training, the model learns how to replicate the target waveform using the given transformed acoustic features. During inference, the model is expected to transform the acoustic features of the initial source waveform without applying dynamic time warping into a converted waveform.

A solution based on a variational autoencoder is also presented in [34]. An autoencoder is an end-to-end deep-learning encoder-decoder structure with a bottleneck between the encoder and the decoder. This bottleneck structure allows for learning a compressed low-dimensional latent-variable space [35] of the input data that is called an embedding [36]. Variational autoencoders further improves the embedding by forcing it to follow a Gaussian distribution. In [34], during training, the autoencoder learns to transform any speech signal into the desired voice. During inference, the input waveform passes through the encoding part of the autoencoder, which outputs embeddings representing features of the target speech. These embeddings are then used as conditioners for WaveNet. This solution provides better results for voice conversion compared to conventional methods.

An unconventional WaveNet model demonstrating the feasibility of shifting the tone of the generated voice so that it sounds like someone singing is presented in [37]. This model closely resembles the original WaveNet model, but with some key differences. Instead using past raw unidimensional waveforms, data with time and frequency dimensions are processed. Phonetic features are also presented as local conditioners to dictate what to sing. Furthermore, instead of using a SoftMax output, a mixture of logistics inspired from PixelCNN++ [38] is used. To make the model more robust, a denoising objective is also added, meaning that Gaussian noise is added to the training data. This works precisely like data augmentation by noise injection, where the network learns to generate an uncorrupted prediction. Listening tests showed preference for this synthesizer over parametric and concatenative systems.

4.4 Noise Reduction

WaveNet can generate clean audio from noisy audio by removing all but the parts that it has been trained to model [39]. This WaveNet is non-causal and uses the speaker's identity for global conditioning, making it so that the learned voice is distinguished from the rest of the audio. This enables the extraction of only the desired speaker's voice. This approach is shown to produce better denoising results than conventional Wiener filtering [40].

4.5 Music Generation

Some experiments have explored music synthesis with WaveNet. A proposed architecture [41] uses an LSTM recurrent neural network that learns different styles of music and that is used to generate a symbolic musical representation. This representation is provided as a conditioner to WaveNet so that it generates the melody of an instrument. The proposed architecture combines the strength of raw audio models with that of symbolic models.

In [42], WaveNet is combined with an autoencoder to synthesise musical instruments, in a similar way to what is described in Section 4.3 for voice conversion. This paper has inspired the "NSynth Super" project [43], which is a neural music synthesizer that manipulates generated embeddings learned from a large selection of instruments to create entirely synthetic, but absolutely natural sounding, instrument notes.

4.6 Musical Timbre Transfer

With an autoencoder latent variable structure similar to that discussed in Section 4.5 for music generation, an experiment to convert music from one domain to another is presented in [44]. This paper shows that WaveNet can be used to convert the timbre of raw audio from an instrument to another without altering the inherent melody. To make training of the encoder reliable, a specialized loss function named "confusion loss" [45] is used, enabling domain transfer by discriminating on features that are not domain-independent. A WaveNet decoder then uses these latent variables as conditioners to generate the converted audio.

The task of musical instrument synthesis with timbre control is considered in [46]. The proposed system involves a recurrent neural network architecture that learns to generate spectrograms from symbolic music representations and instrument embeddings.

TimbreTron [10] adapts image-based style transfer techniques to audio. The proposed method is based on CycleGan [47], which learns how to translate an image from a domain to another when no paired training data is available. One popular demonstration of this architecture consists in transforming the subject of a video, a horse, into a zebra. The training technique based on generative adversarial networks [8] is adapted to audio with TimbreTron, which transforms the melody of an instrument into another by a modified architecture using constant Q transform (CQT) spectrograms that are used as conditioners for WaveNet. Based on human perceptual evaluations, TimbreTron's ability to transfer timbre while preserving the musical content has been proved.

4.7 Source Separation

Another possible application of WaveNet is source separation, where a sequence of raw audio containing multiple overlapping sounds is separated into individual sounds. A proposed model for this task is named Wave-U-Net [48], which is based on the U-Net model [49] originally developed for image segmentation. U-Net is a fully convolutional network for image segmentation, where an image is processed into a segmented image. The proposed architecture, Wave-U-Net, adapts the U-Net model to WaveNet, enabling it to extract different sources from the input raw audio.

4.8 Speech Recognition

Keyword spotting is an example of how WaveNet can be used as a speech recognition model [50]. Adapted to this task, WaveNet is trained to recognize the "Hey Snips" keyword. This model is shown to outperform networks based on recurrent neural networks, reducing the false detection rate significantly.

4.9 Speech Enhancement

In the speech compression application built around the Codec 2 vocoder already presented in Section 4.2, WaveNet is used as the decoder [29]. The speech signal is encoded using Codec 2 at the 8 kHz sampling rate, then decoded using WaveNet at the 16 kHz sampling rate. This shows that WaveNet is effectively capable of performing blind artificial audio bandwidth extension.

5 Subjective Quality Optimizations

5.1 Noise Shaping

Considering that audio samples are usually recorded with a resolution of 16 bits per sample while WaveNet generates samples with a resolution of only 8 bits, there is a risk that quantizing with μ-law as proposed in the original WaveNet paper introduces some quantization noise. A WaveNet that uses noise-shaping [51] has been proposed in order to evaluate the benefits of shaping the quantization noise [52]. The principle of this WaveNet is to generate from acoustic features an excitation signal rather than raw audio. During training, a time-invariant noise weighting filter is calculated from the speech training database. The training speech data is then filtered to get an excitation signal, which is then quantized. The resulting signal is presented to WaveNet as the data to replicate during training according to the extracted acoustic features of the input speech sequence. During inference, the excitation signal generated by WaveNet from acoustic features is dequantized and filtered by the noise-shaping filter to obtain the reconstructed speech signal. The paper presenting this variant of WaveNet showed that, since the prediction error is much larger than the quantization error, this structure yields marginal results to mitigate quantization noise. It is however useful to mitigate the spectral distortion generated by WaveNet in higher frequency bands.

Instead of using a fixed filter calculated using the entire training database, it is possible to use an adaptive filter derived from mel-cepstral coefficients [53]. Unlike the previously-cited noise-shaping paper, this one suggests that mel-cepstrum noise-shaping significantly improves speech quality by shaping the quantization noise.

Another solution to reduce the noisy nature of the decoded speech signal is to implement a linear predictive (LPC) filter as in the ExcitNet [54] and LP-WaveNet [55] models. These models can be seen as cross-overs between a traditional LPC vocoder and WaveNet. The WaveNet part predicts an excitation signal that is then passed through an LPC filter to reconstruct the speech signal. This strategy makes it possible to outperform both traditional linear prediction vocoders and basic WaveNet vocoders.

5.2 Training on Small Multi-Speaker Datasets

Instead of generating speech directly, reference [56] proposes a WaveNet capable of generating a glottal excitation signal from acoustic features. The glottal signal is then passed through a vocal tract filter to obtain a speech signal. The WaveNet model is thus more speaker independent because most of the information about the speaker is supposedly contained in the vocal tract filter. This means that WaveNet only has to learn a generic excitation signal instead of speaker specific waveforms. When only small datasets are available, the suggested architecture provides better performance than generating speech directly.

5.3 Elimination of Collapsed Segments

A particular kind of voice degradation in WaveNet vocoders named "collapsed segments", and a solution to that problem, are discussed in [57]. Features extracted by the encoder of a WORLD vocoder [58] are decoded in two ways, first by the original WORLD decoder, then by a WaveNet decoder in which the extracted features are used for local conditioning. The envelopes of both decoded signals is calculated and compared. When the difference between the two envelopes exceeds a certain threshold, a collapsed segment in WaveNet is detected. In that situation, the speech signal is regenerated using linear predictive coding (LPC). The resulting signal gets a 77% user preference in subjective quality compared to the signal containing collapsed segments.

5.4 Improvement of Speech Naturalness

A common problematic situation for WaveNet is when local conditioners such as acoustic features that are delivered by a complementary acoustic model are mismatched with the desired audio. This causes significant audio degradation. A proposed solution [9] is to introduce generative adversarial networks (GANs) [8] in order to further train the acoustic model to increase speech naturalness by reducing the rate of mismatched characteristics. GANs consist of two competing networks: a generator, which creates a sequence of samples, and a discriminator, which classify real sequences from generated sequences. The two networks compete against each other during training, both getting progressively better at their respective tasks. The use of GANs for training the complementary acoustic models of WaveNet resulted in very high mean opinion scores with a broad number of GAN variants.

Attempts were made to train a WaveNet to generate speech by conditioning the network on parameters normally used for speech compression. An investigation has been published about the best local conditioners to use for a WaveNet-based statistical vocoder [59]. The best results were obtained using mel filter banks with the fundamental frequency as conditioning. Another investigation was made showing that training a multi-speaker WaveNet vocoder yields better results if it has been trained with the same voices as those generated during inference, showing difficulty when generating voices it has never been trained on [60].

Distillation optimizations, as will be shown in Section 6.3, offer substantial improvements in computational efficiency, making possible real-time audio generation with Parallel WaveNet [4] and ClariNet [5]. However, reference [11] has shown that audio sequences generated by student networks do not reach the quality of natural waveforms. The student model is also difficult to adapt to new speakers, first necessitating the adaptation of the teacher network. To work around these drawbacks, a post-distillation framework using generative adversarial networks (GAN) is proposed [11]. First, a WaveNet student network is distilled from a WaveNet teacher as explained in [4][5]. Then, the student network is trained adversarially with a discriminator from a GAN. This process can be used to adapt the student network to other speaker identities. It can also be used to increase the perceptual quality of the student's generated sequence of samples by tricking the discriminator into predicting that the generated waveform is a natural waveform. The proposed method was shown to reduce the quality gap between sequences generated by the distilled WaveNet model and natural waveforms. It also demonstrated the ability to directly adapt the student network to new speakers.

5.5 Language-Specific Optimizations

Tacotron 2 [3] showed promising results, outperforming both neural text-to-speech systems and conventional systems based on text analyzers and duration models, from the point of view of the naturalness of synthesized speech. However, its applicability to other languages than English was not thoroughly evaluated in [3]. A new Tacotron 2 model that is adapted to the task of end-to-end neural Japanese speech synthesis is proposed in [61]. Three models are experimented in order to identify the best parameters and processing architectures for the Japanese language. All three generate different conditioners that are then used in WaveNet. Although the results obtained with this approach do not match those obtained with traditional pipeline systems using text analyzers and duration models, the experiments reported in this paper are presented as being a step forward for end-to-end Japanese speech synthesis.

6 Performance Optimizations

6.1 Training Optimizations

In the original WaveNet implementation, 8 bit μ-law quantization is applied to the audio samples. The architecture is also adapted to generating 8 bit samples by generating a probability distribution used to classify between 256 discrete values. The generated audio quality would be increased if 16 bit samples were generated instead of 8 bit samples, thus classifying between 65,536 discrete values instead of 256. However, with the architecture suggested in the original WaveNet paper, the classification task would be complicated by several orders of magnitude. One solution proposed in various papers is to introduce a logistic mixture model in WaveNet [4][3][37]. This technique is highly inspired by a similar improvement introduced in the PixelCNN++ network [38]. This optimization shows that the output distribution can be simplified into logistic functions aggregated together into the same probability distribution. The use of logistic functions also imposes that adjacent values have a very similar probability value. This reduces the number of parameters required to model WaveNet's probability distribution. Only the parameters of each chosen logistic function can be modified, which speeds-up training and simplifies the classification task.

In the case of ClariNet [5], instead of using a mixture of logistic distributions to model raw waveforms, a single Gaussian is shown to suffice without degradation of the audio quality.

A lot of data is necessary to train WaveNet for a specific application. A solution proposed in [62] is to use a WaveNet that is first pre-trained on a large dataset, then specialized to another dataset that can be much more limited. This WaveNet uses for conditioning the speaker identity, a sequence of phonemes extracted from the input text, as well as the logarithm of the fundamental frequency. Initially, the weights of the WaveNet core and an embedding vector representing the speaker's identity are trained. Then, three different WaveNet data specialization methods are presented in the paper. The first one allows the modification of all weights in the model, which is effective but prone to overfitting. The second one aims at adjusting only the embedding vector corresponding to the learned speaker identity, adapting it to the already functioning WaveNet core. This way, the WaveNet core does not lose its ability to generalize based on the knowledge it gained from the much larger initial dataset. The third adaptation method modifies the architecture, so that the speaker identity embedding is not predicted from a given ID, but directly from the raw data. The third suggested architecture is trained without applying any dataset specialization process, but instead directly trained on a limited dataset. As shown in the paper, good speech similarity and quality results are obtained with only 10 seconds of speech from a new speaker, and quality increases when more data is available.

A similar experiment proposes a model that is initially trained before fine-tuning it into the desired model [19]. It is able to convert a person's voice into Lombard speech, which is the result of an involuntary tendency of humans to vocalize with extra effort in noisy environments. There are very few Lombard speech datasets available, making such model adaptations necessary for the task of specific high-performance speech synthesis. The proposed model uses a sequence-to-sequence model with attention as well as a WaveNet model to generate the samples. The proposed architecture is shown to perform better than WaveNet trained exclusively with a Lombard speech dataset.

6.2 Faster Sequential Generation

A simple way to accelerate WaveNet consists in eliminating the redundant operations performed in a naive implementation. Buffers containing the results of the dilated causal convolution operations with different lengths on every layer are kept, in a way that the same operation does not have to be repeated [63]. The proposed generation algorithm is faster than the naïve implementation and exponentially faster when WaveNet has about 10 dilated causal convolution layers.

Reference [64] proposes to increase the sampling frequency of WaveNet up to 48 kHz, covering the whole audible frequency spectrum using a subband architecture. Considering that a high number of dilated convolution layers in WaveNet is exponentially slower to execute than a WaveNet having a low number of layers, an increase in sampling rate increases the model size significantly, which in turn increases the computational load. The proposed subband solution separates the input into 8 kHz subbands, demodulates and downsamples these subbands, and passes them as input to their respective subband WaveNet. The subband WaveNet outputs are modulated back together to create a full band waveform. The subband WaveNets can be parallelized, increasing computational gains even further. With this proposed architecture, it is claimed that synthesis speed is increased by a factor of 4.

6.3 Distillation Optimizations

Parallel WaveNet [4] and ClariNet [5] both use a similar optimization technique that involves the use of a pre-trained teacher network and a student network. The student attempts to learn the teacher's probability distribution with a technique named "probability density distillation loss" introduced in Parallel WaveNet [4]. With this loss term, the student is trained to match the distribution of its generated samples to the distribution of the teacher. Some extra loss terms are also added to increase the final student's quality and proximity to the desired output. This technique enables the use of a smaller network, the student, that can learn efficiently from the teacher instead of converging extremely slowly with the training data. It also considerably speeds up the audio generation process.

6.4 Parallelization Optimizations

The original WaveNet architecture is sequential due to its autoregressive structure and can only be parallelized during training. Some techniques have been introduced that enable some interesting parallelization optimizations during generation. A normalizing flow [65] is a procedure that starts with a simple posterior distribution and iteratively applies invertible transformations resulting in a more flexible distribution. This transformation can be applied to audio to reduce the number of variables needed to represent a sequence of samples. WaveNet can then be trained to generate these variables instead of the raw waveforms. This is the idea behind inverse autoregressive flows [66], which consist in using a generative network such as WaveNet that generates a compressed latent variable representation autoregressively. These variables can then be transformed in parallel into the raw audio. Parallel WaveNet [4] implements this solution and ClariNet [5] implements a similar solution using Gaussian autoregressive flows. Parallel WaveNet has been shown to be fast enough to be used as the voice generator behind Google assistant.

While Parallel WaveNet [4] and ClariNet [5] both propose a highly computationally optimized implementation of the network, these techniques necessitate complex loss terms because of the implemented distillation optimizations. FloWaveNet [67] and WaveGlow [7] both propose a normalizing flow-based model that only needs a single loss term due to the absence of a student-teacher distillation process, which greatly simplifies the training procedure. They are shown to have comparable quality to the original WaveNet implementation while being both computationally efficient.

7 Conclusions

This paper has presented a review of the literature on WaveNet, an autoregressive deep generative artificial neural network for raw audio waveforms, since its introduction in 2016. This review has been made as complete as possible up to December 2018. With several new papers being published each month, it may, however, not remain complete for a very long time.

The basic principle of WaveNet, as well as the concepts of global and local conditioning, have been summarized in Sections 2 and 3. A summary of papers presenting various applications of WaveNet in the field of speech and audio processing and various improvements in terms of subjective quality or computational efficiency has been presented in Sections 4 to 6.

Future papers on WaveNet are likely to present more applications in the field of speech and audio processing together with further refinements and improvements. Papers presenting variants of WaveNet with application to multidimensional audio, and even to autoregressive sequential data other than audio, are also likely to be published.

References

[1] A. van den Oord et al., “WaveNet: A Generative Model for Raw Audio,” [demo] *9th ISCA Workshop on Speech Synthesis*, 13-15 September 2016, Sunnyvale, USA. Full paper available as arXiv preprint arXiv:1609.03499, September 2016.

[2] X. Wang et al., “A Comparison of Recent Waveform Generation and Acoustic Modeling Methods for Neural-Network-Based Speech Synthesis,” *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018. arXiv preprint arXiv:1804.02549, April 2018.

[3] J. Shen et al., “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 4779-4783. arXiv preprint arXiv:1712.05884, December 2017.

[4] A. van den Oord et al., “Parallel WaveNet: Fast High-Fidelity Speech Synthesis,” *35th International Conference on Machine Learning (ICML 2018)*, 10-15 July 2018, Stockholm, Sweden. arXiv preprint arXiv:1711.10433, November 2017.

[5] W. Ping, K. Peng and J. Chen, “ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech,” arXiv preprint arXiv:1807.07281, July 2018.

[6] D. P. Kingma and P. Dhariwal, “Glow: Generative Flow with Invertible 1x1 Convolutions,” *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, 2-8 December 2018, Montreal, Canada. arXiv preprint arXiv:1807.03039, July 2018.

[7] R. Prenger, R. Valle and B. Catanzaro, “WaveGlow: A Flow-based Generative Network for Speech Synthesis,” arXiv preprint arXiv:1811.00002, October 2018.

[8] I. J. Goodfellow et al., “Generative Adversarial Nets,” *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 8-13 December 2014, Montreal, Canada. arXiv preprint arXiv:1406.2661, June 2014.

[9] Y. Zhao et al., “Wasserstein GAN and Waveform Loss-Based Acoustic Model Training for Multi-Speaker Text-to-Speech Synthesis Systems Using a WaveNet Vocoder,” in *IEEE Access*, vol. 6, pp. 60478-60488, 2018. arXiv preprint arXiv:1807.11679, July 2018.

[10] S. Huang et al., “TimbreTron: A WaveNet (CycleGAN (CQT (Audio))) Pipeline for Musical Timbre Transfer,” arXiv preprint arXiv:1811.09620, November 2018.

[11] Q. Tian et al. “Generative Adversarial Network based Speaker Adaptation for High Fidelity WaveNet Vocoder,” arXiv preprint arXiv:1812.02339, December 2018.

[12] A. van den Oord et al., “Conditional Image Generation with PixelCNN Decoders,” *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 5-10 December 2016, Barcelona, Spain. arXiv preprint arXiv:1606.05328, June 2016.

[13] *ITU-T Recommendation G.711: Pulse code modulation (PCM) of voice frequencies*, ITU-T, November 1988.

[14] F. Yu and V. Koltun, “Multi-Scale Context Aggregation by Dilated Convolutions,” *4th International Conference on Learning Representations (ICLR 2016)*, 2-4 May 2016, San Juan, Puerto Rico. arXiv preprint arXiv:1511.07122, April 2016.

[15] R. K. Srivastava, K. Greff and J. Schmidhuber, “Highway Networks,” *34th International Conference on Machine Learning (ICML 2017)*, 6-11 August 2017,

Sydney Australia. arXiv preprint arXiv:1505.00387, November 2015.

[16] K. He et al., "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 2016, pp. 770-778. arXiv preprint arXiv:1512.03385, December 2015.

[17] A. Prakash, "One-by-One convolutions," internet: https://iamaaditya.github.io/2016/03/one-by-one-convolution/, March 25, 2016 [Nov. 11, 2018].

[18] Y. Jia, et al, "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis," *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, 2-8 December 2018, Montreal, Canada. arXiv preprint arXiv:1806.04558, November 2018.

[19] B. Bollepalli, L. Juvela and P Alku, "Speaking Style Adaptation in Text-To-Speech Synthesis using Sequence-to-Sequence Models with Attention," arXiv preprint arXiv:1810.12051, October 2018.

[20] J. Chorowski et al., "Attention-Based Models for Speech Recognition," *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 7-12 December 2015, Montreal, Canada. arXiv preprint arXiv:1506.07503, June 2015.

[21] N. Li et al., "Close to Human Quality TTS with Transformer," arXiv preprint arXiv:1809.08895, September 2018.

[22] A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 4-9 December 2017, Long Beach, USA. arXiv preprint arXiv:1706.03762, June 2017.

[23] L. Wan et al., "Generalized End-to-End Loss for Speaker Verification," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 4879-4883. arXiv preprint arXiv:1710.10467, January 2018.

[24] C. Veaux, J. Yamagishi and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," March 2017. [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). http://dx.doi.org/10.7488/ds/1994.

[25] S. O. Arik et al., "Deep Voice 2: Multi-Speaker Neural Text-to-Speech," *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 4-9 December 2017, Long Beach, USA. arXiv preprint arXiv:1705.08947, September 2017.

[26] S. O. Arik et al, "Deep Voice: Real-time Neural Text-to-Speech," *34th International Conference on Machine Learning (ICML 2017)*, 6-11 August 2017, Sydney Australia. arXiv preprint arXiv:1702.07825.

[27] W. Ping et al., "Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning," *6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, Canada, 30 April - 3 May 2018. arXiv preprint arXiv:1710.07654, October 2017.

[28] Y. Gu and Y. Kang, "Multi-task WaveNet: A Multi-task Generative Model for Statistical Parametric Speech Synthesis without Fundamental Frequency Conditions," *Interspeech 2018*, 2-6 September 2018, Hyderabad, India. arXiv preprint arXiv:1806.08619, June 2018.

[29] W. B. Kleijn et al., "Wavenet Based Low Rate Speech Coding," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 676-680. arXiv preprint arXiv:1712.01120, December 2017.

[30] D. Rowe, "Codec 2 - Open Source speech coding at 2400 bit/s and below," *TAPR and ARRL 30th Digital Communications Conference 2011 Proceedings*, 23-25 September 2011, Baltimore, USA, pp. 80-84.

[31] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," *1993 IEEE International Conference on Acoustics,*

Speech, and Signal Processing (ICASSP), Minneapolis, USA, 1993, pp. 554-557.

[32] K. Kobayashi et al., "Statistical Voice Conversion with WaveNet-Based Waveform Generation," *Interspeech 2017*, 20-24 August 2017, Stockholm, Sweden, pp. 1138-1142.

[33] J. Niwa et al., "Statistical Voice Conversion Based on Wavenet," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 5289-5293.

[34] W. C. Huang et al., "Refined WaveNet Vocoder for Variational Autoencoder Based Voice Conversion," arXiv preprint arXiv:1811.11078, November 2018.

[35] C. Bishop, "Latent Variable Models," in *Learning in Graphical Models*, pp. 371-403, MIT Press, January 1999.

[36] Y. Bengio, A. Courville and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798-1828, August 2013. arXiv preprint arXiv:1206.5538, April 2014.

[37] M. Blaauw and J. Bonada, "A Neural Parametric Singing Synthesizer," *Interspeech 2017*, 20-24 August 2017, Stockholm, Sweden. arXiv preprint arXiv:1704.03809, August 2017.

[38] T. Salimans et al., "PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications," *5th International Conference on Learning Representations (ICLR 2017)*, 24-26 April 2017, Toulon, France. arXiv preprint arXiv:1701.05517, January 2017.

[39] D. Rethage, J. Pons and X. Serra, "A WaveNet for Speech Denoising," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 5069-5073. arXiv preprint arXiv:1706.07162, June 2017.

[40] P. Scalart and J. Vieira Filho, "Speech Enhancement Based on A Priori Signal to Noise Estimation," *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Atlanta, USA, 1996, pp. 629-632.

[41] R. Manzelli et al., "Conditioning Deep Generative Raw Audio Models for Structured Automatic Music," *19th International Society for Music Information Retrieval Conference (ISMIR 2018)*, 23-27 September, Paris, France. arXiv preprint arXiv:1806.09905, June 2018.

[42] J. Engel et al., "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders," *34th International Conference on Machine Learning (ICML 2017)*, 6-11 August 2017, Sydney Australia. arXiv preprint arXiv:1704.01279, April 2017.

[43] Magenta. "NSynth Super". Internet: https://nsynthsuper.withgoogle.com/ [23 November 2018].

[44] N. Mor, L. Wolf, A. Polyak and Y. Taigman, "A Universal Music Translation Network," arXiv preprint arXiv:1805.07848, May 2018.

[45] E. Tzeng et al., "Deep domain confusion: Maximizing for domain invariance," arXiv preprint arXiv:1412.3474, December 2014.

[46] J. W. Kim, R. Bittner, A. Kumar and J. P. Bello, "Neural Music Synthesis for Flexible Timbre Control," arXiv preprint arXiv:1811.00223, November 2018.

[47] J.-Y. Zhu et al. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," arXiv preprint arXiv:1703.10593, March 2017.

[48] D. Stoller, S. Ewert and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," *19th International Society for Music Information Retrieval Conference (ISMIR 2018)*, 23-27 September, Paris, France. arXiv preprint arXiv:1806.03185, June 2018.

[49] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, LNCS, vol. 9351, pp. 234-241, 2015. arXiv preprint arXiv:1505.04597, May 2015.

[50] A. Coucke et al., "Efficient Keyword Spotting using Dilated Convolutions and Gating," arXiv preprint arXiv:1811.07684, November 2018.

[51] B. S. Atal and M. R. Schroeder, "Adaptive Predictive Coding of Speech Signals," *The Bell System Technical Journal*, vol. 49, no. 8, pp. 1973-1986, October 1970.

[52] K. Tachibana et al., "An Investigation of Noise Shaping with Perceptual Weighting for Wavenet-Based Speech Generation," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 5664-5668.

[53] T. Yoshimura et al., "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, July 2018.

[54] E. Song, K. Byun and H. G. Kang, "ExcitNet Vocoder: a Neural Excitation Model for Parametric Speech Synthesis Systems," arXiv preprint arXiv:1811.04769, November 2018.

[55] M. J. Hwang et al., "LP-WaveNet: Linear Prediction-based WaveNet Speech Synthesis," arXiv preprint arXiv:1811.11913, November 2018.

[56] L. Juvela et al., "Speaker-Independent Raw Waveform Model for Glottal Excitation," *Interspeech 2018*, 2-6 September 2018, Hyderabad, India. arXiv preprint arXiv:1804.09593.

[57] Y.-C. Wu et al., "Collapsed speech segment detection and suppression for WaveNet vocoder," *Interspeech 2018*, 2-6 September 2018, Hyderabad, India. arXiv preprint arXiv:1804.11055, April 2018.

[58] M. Morise, F. Yokomori and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions on Information and Systems*, vol. 99-D, no. 7, pp. 1877-1884, July 2016.

[59] N. Adiga, V. Tsiaras and Y. Stylianou, "On the use of Wavenet as a Statistical Vocoder," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 5674-5678.

[60] T. Hayashi et al., "An Investigation of Multi-Speaker Training for Wavenet Vocoder," *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, 2017, pp. 712-718.

[61] Y. Yasuda, X. Wang, S. Takaki and J. Yamagishi, "Investigation of Enhanced Tacotron Text-to-Speech Synthesis Systems with Self-Attention for Pitch Accent Language," arXiv preprint arXiv:1810.11960, October 2018.

[62] Y. Chen et al., "Sample Efficient Adaptive Text-to-Speech," arXiv preprint arXiv:1809.10460, September 2018.

[63] T. Le Paine et al, "Fast Wavenet Generation Algorithm," arXiv preprint arXiv:1611.09482, November 2016.

[64] T. Okamoto et al., "An Investigation of Subband Wavenet Vocoder Covering Entire Audible Frequency Range with Limited Acoustic Features," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 5654-5658.

[65] D. Jimenez Rezende and S. Mohamed, "Variational Inference with Normalizing Flows," *32nd International Conference on Machine Learning (ICML 2015)*, 06-11 July 2015, Lille, France. arXiv preprint arXiv:1505.05770, June 2016.

[66] D. P. Kingma et al., "Improved Variational Inference with Inverse Autoregressive Flow," *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 5-10 December 2016, Barcelona, Spain. arXiv preprint arXiv:1606.04934; January 2017.

[67] S. Kim, S. G. Lee, J. Song and S. Yoon, "FloWaveNet: A Generative Flow for Raw Audio," arXiv preprint arXiv:1811.02155, November 2018.