**ABSTRACT**

WaveNet vocoder has seen its great advantage over traditional vocoders in voice quality. However, it usually requires a relatively large amount of speech data to train a speaker-dependent WaveNet vocoder. Therefore, it remains a challenge to build a high-quality WaveNet vocoder for low resource tasks, e.g. voice conversion, where speech samples are limited in real applications. We propose to use singular value decomposition (SVD) to reduce WaveNet parameters while maintaining its output voice quality. Specifically, we apply SVD on dilated convolution layers, and impose semi-orthogonal constraint to improve the performance. Experiments conducted on CMU-ARCTIC database show that as compared with the original WaveNet vocoder, the proposed method maintains similar performance, in terms of both quality and similarity, while using much less training data.

**Index Terms**— Voice Conversion (VC), WaveNet, Singular Value Decomposition (SVD)

1. **INTRODUCTION**

Voice conversion (VC) is a technique that aims to modify a speech signal of a source speaker to make it sound as if it is uttered by a target speaker without changing the content information. This technique has many applications, including voice morphing, emotion conversion, speech enhancement, movie dubbing as well as other entertainment applications.

One of the challenges is to generate high quality speech. Various statistical approaches have been proposed, e.g. Gaussian mixture model (GMM) [1–3], frequency warping [4–7] and exemplar based methods [8–10]. Recently, deep learning has become mainstream in this area, including deep neural network (DNN) [11–13], long short-term memory (LSTM) [14], variational auto-encoder (VAE) [15], and generative adversarial networks (GAN) [16–18].

Despite much progress, the quality of converted speech can not match that of the source speech. One major reason is that traditional vocoders with simplified assumption are generally adopted for speech reconstruction. To address this problem, several neural vocoders [19–22] are proposed to generate time-domain signal directly. The effectiveness of incorporating speaker dependent (SD) WaveNet [23, 24] into VC task for waveform generation has been investigated. Recently, an SD WaveNet [25] is first adopted to replace the traditional vocoder for converted speech generation; while another attempt [26] uses the WaveNet to learn a mapping between phonetic posteriorgram (PPG) and waveform samples to generate the speech for a specific target speaker. While SD WaveNet in general generates high quality speech, it requires a relatively large amount of data for training. According to a recent study [27], an SD WaveNet needs to be trained with at least 320 utterances, which is not always available in practice. To reduce the data requirement from target speaker, a speaker adapted WaveNet is proposed [28, 29], where a speaker independent WaveNet is first trained on a corpus consisting of various speakers, and then adapted with a small amount of target data.

In this paper, we propose to build an SD WaveNet with limited data from target speaker, where singular value decomposition (SVD) [30] technique is employed to reduce the number of WaveNet parameters. Specifically, SVD is applied to the dilated convolution layers by inserting a $1 \times 1$ convolution layer, where the original weight matrix is factorized into two small matrices. Inspired by [31], the semi-orthogonal constraint is used during the decomposed weight matrix updating, which leads to a rapid convergence for network training.

The contributions of this paper include,

1) We employ SVD to WaveNet which significantly reduces the number of parameters;

2) We significantly reduce the amount of data required for SD WaveNet training, while maintaining the quality of generated speech;

The rest of this paper is organized as follows. Section 2 briefly describes the overview of WaveNet vocoder. The details of our proposed method are discussed in Section 3. In
Sections 4 and 5, we describe the experimental setup and report the results. Finally, the conclusions are given in Section 6.

2. WAVENET VOCODER

2.1. WaveNet Vocoder

WaveNet vocoder [23–25] is an autoregressive generative model that can directly model raw audio waveforms from given acoustic features, e.g., spectral features, aperiodicity, and f0. Given a waveform \( x = [x_0, x_1, \ldots, x_{T-1}] \), the joint probability is modeled as the product of conditional distributions as follows.

\[
p(x) = \prod_{t=1}^{T} p(x_t|x_1, \cdots, x_{t-1}). \tag{1}
\]

In order to learn the long-range dependencies among temporal waveform samples, WaveNet utilizes an architecture based on a stack of dilated causal convolution layers and a gated activation unit. The calculation function at each dilated causal layer is expressed as follows.

\[
z = \tanh(W_f * i + V_f * h) \odot \sigma(W_g * i + V_g * h), \tag{2}
\]

where \( i \) is the input vector and \( h \) is the extra condition feature, e.g. spectral features. \( * \) and \( \odot \) represent the convolution and element-wise product operator respectively. \( f \) and \( g \) denote filter and gate, respectively. \( W \) and \( V \) indicate learnable convolution weights.

2.2. Advantages and Limitations

WaveNet vocoder directly establishes the mapping between acoustic features and time-domain waveforms in a data-driven manner. Therefore, some time-domain information which is usually discarded in the conventional vocoder, e.g. phase information, can be recovered, that allows for high fidelity voice reconstruction.

On the other hand, WaveNet vocoder is a deep autoregressive neural network which mainly utilizes stacked dilated convolutions. The dilation grows exponentially with the number of layers (e.g. 1, 2, ..., 512) and then repeated 3 or 4 times, which makes the network very deep. Hence, WaveNet generally consists of a large number of parameters, which requires a large number of data for training.

3. WAVENET FACTORIZATION WITH SINGULAR VALUE DECOMPOSITION

To reduce the data requirement of SD WaveNet training, in this section, we propose to factorize the WaveNet with singular value decomposition (SVD) technique.

3.1. SVD for WaveNet

Singular value decomposition (SVD) is a factorization technique in linear algebra, which has been successfully used to reduce model parameters in neural networks [31–34]. Generally, SVD is applied to decompose a learned weight matrix into two lower dimension matrices, and discard the smaller singular values [32–34]. With a reduced number of weight
parameters, the matrix is still able to achieve a close performance as the original large weight matrix.

Inspired by the success of [31], where an SVD is employed in time delay neural networks (TDNNs) for automatic speech recognition (ASR), we investigate the effectiveness of applying the SVD technique to adjust the network structure of WaveNet. Since the dilated convolution layers are the dominant component in the WaveNet model, which takes the majority number of weight parameters, we choose to apply SVD to these layers.

We use an example to show how it works. As shown in Fig. 1 (a), a 1 × 1 convolution layer is added after each dilated convolution layer, and we called it a SVD layer. DC is the output channel of dilated convolution layer and SC is the output channel of SVD layer. Then the original weight matrix is decomposed into two matrices, denoted as N and M respectively. Hence, the sizes of weight matrices N and M are controlled by DC and SC. For example, as shown in Fig. 1 (b), we set input dimension of dilated convolution layer to 200, filter size to 2 and DC to 256. The parameter number of original matrix is 200 × 2 × 256 = 102,400. By adding a 1 × 1 SVD layer with DC and SC of 32 and 256, shown in Fig. 1 (c), the parameter number of decomposed matrices N and M are 200 × 2 × 32 = 12,800 and 32 × 1 × 256 = 8,192. Consequently, the number of parameters are effectively reduced from 102,400 to 20,992.

3.2. Train with Semi-orthogonal Constraint

To prevent numerical instability and accelerate the convergence, the decomposed model is trained from a random initialization with a constraint of the matrix M to be semi-orthogonal. Specifically, we apply an update after every four time-steps of back-propagation to make the matrix to be close to a semi-orthogonal matrix. The deduced process of parameters update formula refers to [31], it is expressed as follows.

\[ M \leftarrow M - \frac{1}{2\alpha^2} (MM^T - \alpha^2 I) M, \]  \hspace{1cm} (3)

where \( \alpha \) is a scaling factor, and \( I \) is identity matrix. \( \alpha \) allows us to control how fast the parameters of different SVD layers change in a more consistent way. Let \( P = MM^T \), then we compute the scale \( \alpha \) as follows.

\[ \alpha = \sqrt{tr(PPT)/tr(P)}, \]  \hspace{1cm} (4)

4. EXPERIMENTAL SETUPS

4.1. Training of WaveNet Vocoder

A WaveNet vocoder takes acoustic features as input [24] and generates speech waveforms as output. WORLD vocoder [35] was employed for speech analysis to prepare the acoustic features for WaveNet training. We extracted 513-dimensional spectrum, 1-dimensional aperiodicity coefficient and 1 dimensional F0 with 5 ms frame shift. Then 40-dimensional mel cepstral coefficients (MCCs) were calculated from spectrum using speech signal processing toolkit (SPTK) 1. The 40-dimensional MCCs, 3-dimensional logarithmic fundamental frequency (static, delta and delta delta) together with the aperiodicity and the unvoiced/voiced (U/V) flag, form a 45-dimensional acoustic feature for a speech frame.

The WaveNet vocoder consists of 3 blocks, with 10 dilution layers in each block. The dilation in each block starts from 1 and exponentially increases by a factor of 2 until 512. The filter size of causal dilated convolution is 2. The hidden units of residual connection, skip connection and the SVD layer before the softmax output layer are set to 256. We denote the output channel of dilated convolution layer as DC, the output channel of SVD layer as SC, as shown in Fig. 1. The network is optimized by Adam optimization method with a constant learning rate of 0.0001, and a mini-batch size of 14,000. The speech waveforms are encoded in form of 8 bits µ-law.

We compare three WaveNet configurations that are shown in Table 1. Different number of utterances, e.g. 100, 300 and 500, from a specific speaker, are used for SD-WaveNet vocoder training. To avoid over-fitting, we vary the iteration steps according to the training data size, e.g. 400,000 iteration steps for 500 utterances, 300,000 and 100,000 steps for 300 and 100 utterances respectively.

4.2. Voice Conversion Experiments

Experiments were conducted on CMU-ARCTIC dataset [36]. We selected two source speakers (clb and rms) and two target speakers (bdl and slt). Intra-gender and inter-gender conversions were conducted between following pairs: rms to bdl (M2M), clb to slt (F2F), clb to bdl (F2M) and rms to slt (M2F). 100 utterances were used for training, another 20 non-overlap utterances of each speaker were used for evaluation. In total, 80 converted sentences were generated for the 4 speaker pairs.

We implemented the PPG approach to voice conversion [37] in the experiments. 42 dimensional PPG features of CMU-ARCTIC dataset were extracted with the PPG extractor trained with Kaldi [38] on Wall Street Journal corpus [39]. All audio files were downsampled to 16 kHz. The PPG conversion model consists of one feed-forward layer, two long short-term memory (LSTM) layers and one linear output layer. Each hidden layer consists of 128 units. Adam optimization method is used for model training with a learning rate of 0.002. The PPG conversion model takes the PPG features as input and generate MCC features for WaveNet vocoder.

1https://sourceforge.net/projects/sp-tk/
Table 1. WaveNet vocoder configurations, where \textit{DC} and \textit{SC} denote the output channel of dilated convolution layer and SVD layer, respectively. WaveNet-SVD64 and WaveNet-SVD32 are WaveNets with 64 and 32 output channels in the dilated convolution layer.

<table>
<thead>
<tr>
<th>Vocoder</th>
<th>Dilation Layers</th>
<th>Filter Size</th>
<th>Residual Channel</th>
<th>Skip Channel</th>
<th>DC</th>
<th>SC</th>
<th>Model Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>WaveNet</td>
<td>30</td>
<td>2</td>
<td>256</td>
<td>256</td>
<td>256</td>
<td>N/A</td>
<td>49.9MB</td>
</tr>
<tr>
<td>WaveNet-SVD64</td>
<td>30</td>
<td>2</td>
<td>256</td>
<td>64</td>
<td>64</td>
<td>256</td>
<td>32.4MB</td>
</tr>
<tr>
<td>WaveNet-SVD32</td>
<td>30</td>
<td>2</td>
<td>256</td>
<td>32</td>
<td>32</td>
<td>256</td>
<td>16.7MB</td>
</tr>
</tbody>
</table>

5. EVALUATIONS

We validated our proposed factorized WaveNet with two sets of experiments, that cover the training of WaveNet vocoder and WaveNet vocoder in voice conversion. Both objective and subjective evaluation results were reported.

5.1. Evaluation Metrics

5.1.1. Objective Metrics

For objective evaluation, we employed signal-to-noise ratio (SNR) and root mean square error (RMSE) to evaluate distortion between the original speech and synthesized speech in time-domain and frequency-domain respectively.

\[
SNR = 10 \log_{10} \left( \frac{\sum_{n=1}^{N} (x(n) - y(n))^2}{\sum_{n=1}^{N} y(n)^2} \right), \quad (5)
\]

\[
RMSE = \sqrt{\frac{1}{F} \sum_{f=1}^{F} \left(20 \log_{10} \frac{|Y(f)|}{|X(f)|} \right)^2}, \quad (6)
\]

where, \(x(n)\) and \(y(n)\) represent the windowed signals of synthesized speech and natural speech, respectively. \(n\) denotes the sample position in time domain. \(N\) is the frame length. \(|X(f)|\) and \(|Y(f)|\) represent the magnitude of synthesized speech and natural speech at \(f\)-th frequency bin, respectively. \(F\) is the number of frequency bins.

For both \(SNR\) and \(RMSE\) calculation, a time shift between \(\pm200\) points is used to maximize the correlation between original and synthesized speech. Lower \(SNR\) and \(RMSE\) indicate the smaller distortion.

5.1.2. Subjective Metrics

For subjective evaluation, we conducted AB and XAB preference tests to assess speech quality and speaker similarity respectively. During AB tests, we randomly selected samples A and B from the proposed method and the reference method respectively. Each listener was required to select a sample with better quality. For XAB tests, X represented the reference sample of the target, A and B referred to the converted samples randomly selected from the comparison methods. Noted that the language content of X, A and B were the same. Then listeners were asked to choose the sample more close to the reference sample or no preference.

We conducted the mean opinion score (MOS) tests to assess the naturalness of the converted speech. Each listener was asked to rate opinion score on a five-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad)

For each system, 20 sentences, randomly selected from the 80 converted samples, were used for listening tests. 10 proficient English listeners participated in all listening tests.

5.2. Evaluations of WaveNet Vocoder

5.2.1. Factorized WaveNet with different SVD configurations

We start by studying WaveNet as a vocoder presented in Section 4.1, with and without SVD.

First, we validated the model sizes with different configurations. As shown in Table 1, the model size of original WaveNet is 49.9MB. While, by applying SVD to factorize the WaveNet, the model size decreases to 32.4MB and 16.7MB with \textit{DC} of 64 and 32 respectively. We observe that the model size is effectively reduced through SVD.

Table 2. Comparison of average \(SNR\) and \(RMSE\) with and without semi-orthogonal constraint for different WaveNet training.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Constraint</th>
<th>SNR (dB)</th>
<th>RMSE (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WaveNet</td>
<td>w</td>
<td>3.597</td>
<td>9.499</td>
</tr>
<tr>
<td>WaveNet-SVD64</td>
<td>w</td>
<td>3.598</td>
<td>8.545</td>
</tr>
<tr>
<td>WaveNet-SVD32</td>
<td>w</td>
<td>3.601</td>
<td>8.345</td>
</tr>
<tr>
<td>WaveNet-SVD32</td>
<td>w/o</td>
<td>3.605</td>
<td>8.465</td>
</tr>
</tbody>
</table>

Then, we validated the performance by varying the output channel size of dilation layer using 100 training utterances. As shown in Table 2, it is observed that the \(RMSE\) decreases when SVD is applied. The best performance is achieved with the \textit{DC} of 32. We don’t observe that SVD has a significant impact on the \(SNR\) among the various configurations.

Finally, we examined the performance of semi-orthogonal constraint with WaveNet-SVD32 and 100 training utterances. As shown in Table 2, by using the semi-orthogonal constraint, the \(RMSE\) decreases from 8.465 dB to 8.345 dB, while \(SNR\) remains almost the same. This confirms the need for semi-orthogonal constraint in decomposed weight matrix updating.
Table 3. Comparison of the SNR and RMSE between WaveNet and WaveNet-SVD vocoders.

<table>
<thead>
<tr>
<th>Vocoder</th>
<th>Data Size</th>
<th>SNR (dB)</th>
<th>RMSE (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100utts</td>
<td>300utts</td>
<td>500utts</td>
</tr>
</tbody>
</table>

With the empirical observations above, we set the output channel of dilated layer as 32 with the semi-orthogonal constraint in the rest of the experiments.

5.2.2. Objective Evaluation

Table 3 shows the objective evaluation results of WaveNet vocoders with and without applying SVD. It is observed that all the systems achieve similar performance in terms of SNR. According to RMSE, it is observed that system performance improves as training data increase for both WaveNet and WaveNet-SVD. In addition, we found that WaveNet-SVD consistently outperforms WaveNet for different amounts of training data. Moreover, we also observe that WaveNet-SVD (100 utts) achieves results that are similar to WaveNet (300 utts) and close to WaveNet (500 utts). Fig. 2 gives an example of spectrum generated from (a) original, (b) WaveNet (500 utts), (c) WaveNet (100 utts) and (d) WaveNet-SVD (100 utts). The spectrum generated by WaveNet-SVD (100 utts) is clearly sharper than that of WaveNet (100 utts). And it is also closer to that of WaveNet (500 utts) and the original speech.

Fig. 2. Example of spectrums of the same utterance ‘b0436’ from bdl, which are synthesized with different systems. (a) original, (b) WaveNet (500 utts), (c) WaveNet (100 utts), (d) WaveNet-SVD (100 utts).

5.2.3. Subjective Evaluation

We further conducted AB preference and MOS tests to assess the speech quality generated from different systems. The subjective results of AB tests are presented in Fig. 3. We first validated the effect of the amount of training data. As shown in Fig. 3 (a), WaveNet (500 utts) outperforms WaveNet (100 utts) significantly. Then, we compared various WaveNet configurations with different amounts of training data. Fig. 3 (b) suggests that WaveNet-SVD (100 utts) achieves a significantly better performance than that of WaveNet (100 utts). Results showed in Fig. 3 (c) indicates that WaveNet-SVD (100 utts) slightly outperforms WaveNet (300 utts). In Fig. 3 (d), the preference scores of WaveNet-SVD (100 utts) and WaveNet (500 utts) fall into each others confidence intervals, which means they are not statistically significant. The results of MOS test in Table 4 are consistent with those in Fig. 3. Hence, we conclude that WaveNet-SVD (100 utts) significantly outperforms WaveNet (100 utts), and achieves similar results as WaveNet (500 utts).

Fig. 3. Results of speech quality preference tests with 95% confidence intervals for (a) WaveNet (100 utts) vs. WaveNet (500 utts), (b) WaveNet (100 utts) vs. WaveNet-SVD (100 utts), (c) WaveNet (300 utts) vs. WaveNet-SVD (100 utts), (d) WaveNet (500 utts) vs. WaveNet-SVD (100 utts).

Table 4. Mean opinion scores on speech quality of waveforms reconstructed using different vocoders.

<table>
<thead>
<tr>
<th>MOS</th>
<th>WORLD</th>
<th>WaveNet-SVD (100 utts)</th>
<th>WaveNet (100 utts)</th>
<th>WaveNet (500 utts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.44</td>
<td>3.70</td>
<td>3.28</td>
<td>3.74</td>
<td></td>
</tr>
</tbody>
</table>

5.3. Evaluation of Voice Conversion

5.3.1. Objective Evaluation

We further studied the effect of WaveNet vocoder in PPG voice conversion systems as discussed in Section 4.2. RMSE was used as evaluation metric.

Table 5 shows the objective results for different systems. It is observed that the average RMSE of WaveNet decreases from 13.515 dB to 13.313 dB, as training data increases from 100 to 500. Moreover, we also observe that the average RMSE of WaveNet-SVD outperforms WaveNet (100 utts)
Table 5. Comparison of RMSE (dB) between VC results generated with WaveNet and WaveNet-SVD vocoders.

<table>
<thead>
<tr>
<th>Vocoder</th>
<th>Utts</th>
<th>Intra</th>
<th>Inter</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>WaveNet</td>
<td>500</td>
<td>13.304</td>
<td>13.321</td>
<td>13.313</td>
</tr>
<tr>
<td>WaveNet</td>
<td>100</td>
<td>13.436</td>
<td>13.594</td>
<td>13.515</td>
</tr>
<tr>
<td>WaveNet-SVD</td>
<td>100</td>
<td>13.394</td>
<td>13.385</td>
<td>13.390</td>
</tr>
</tbody>
</table>

and is close to that of WaveNet (500 utts). Similar trends are also observed in Intra-gender and Inter-gender conversion.

5.3.2. Subjective Evaluation

We then conducted AB, XAB and MOS tests to assess the generated speech quality and speaker similarity between WaveNet and WaveNet-SVD.

Fig. 4. Quality preference tests of converted speech samples with 95% confidence intervals of different vocoders for (a) WaveNet (100 utts) vs. WaveNet (500 utts), (b) WaveNet (100 utts) vs. WaveNet-SVD (100 utts), (c) WaveNet (500 utts) vs. WaveNet-SVD (100 utts).

The results of AB test are presented in Fig. 4. We first examined the effect of the amount of training data in terms of speech quality. As shown in Fig. 4 (a), WaveNet (500 utts) outperforms WaveNet (100 utts) significantly. Then, we compared the speech quality generated by proposed WaveNet-SVD and WaveNet with the same 100 training utterances. Fig. 4 (b) suggests that WaveNet-SVD (100 utts) achieves a significantly better performance than WaveNet (100 utts). Finally, we compared the speech quality generated by proposed WaveNet-SVD and WaveNet with different amounts of training utterances. Results showed in Fig. 4 (c) indicate that the preference scores of WaveNet-SVD (100 utts) and WaveNet (500 utts) fall into each other’s confidence intervals, which means they are not significantly different. The results of MOS test in Table 6 are consistent with those in Fig. 4.

Fig. 5. Similarity preference tests of converted speech samples with 95% confidence intervals of different vocoders for (a) WaveNet (100 utts) vs. WaveNet (500 utts), (b) WaveNet (100 utts) vs. WaveNet-SVD (100 utts), (c) WaveNet (500 utts) vs. WaveNet-SVD (100 utts).

6. CONCLUSIONS

In this study, we proposed a factorized WaveNet using singular value decomposition. Specifically, a $1 \times 1$ SVD layer was added after each dilated convolution layer to decompose one parameter matrix into two smaller matrices. The experimental results demonstrated that 1) the proposed method reduced the model parameters effectively; 2) it maintained similar performance with much less training data (100 utts) than the original WaveNet (500 utts), in terms of both quality and similarity.

7. ACKNOWLEDGEMENTS

This research is supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-100E-2018-006), and by Programmatic grant no. A18A2b0046 from the Singapore government’s Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain). It is also partially supported by the National Key Research and Development Program of China (No.2017YFB1002102).

https://dhqadg.github.io/ASRU-WaveNet-SVD/
8. REFERENCES


