Deep neural network-based power spectrum reconstruction to improve quality of v/octed speech with limited acoustic parameters

Takuma Okamoto¹*, Kentaro Tachibana¹, Tomoki Toda²,¹, Yoshinori Shiga¹ and Hisashi Kawai¹

¹National Institute of Information and Communications Technology, 3–5, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619–0289 Japan
²Information Technology Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464–8601 Japan

(Received 1 June 2017, Accepted for publication 8 September 2017)

Keywords: Speech synthesis, Vocoder, Deep neural networks, Power spectrum

PACS number: 43.72.+q [doi:10.1250/ast.39.163]

1. Introduction

Text to speech synthesis (TTS) is an important technique in multilingual spoken language communications. Due to the flexibility and small footprint, statistical parametric speech synthesis (SPSS) has become mainstream in TTS. This paper investigates the improvement of the v/octed speech quality, which is a problem in SPSS.

The v/octer in SPSS is a module that converts acoustic features estimated from linguistic information by acoustic models into speech waveforms. Some v/ocoders have been investigated from a simple mel-log spectrum approximate (MLSA) filter with a simple pulse excitation and mel-cepstrum [1] to high-quality ones, such as STRAIGHT [2] and WORLD [3]. However, these high-quality vocoders are intended to analyze and convert high-quality speech and a number of acoustic parameters necessary to synthesize speech with the same quality as the original, but not for TTS. The number of parameters must be reduced to apply these high-quality vocoders to SPSS due to constraints on the number of parameters [4]. These constraints deteriorate synthesis quality even if the acoustic model perfectly estimates the acoustic parameters. In other words, speech quality in TTS will reach a peak due to the v/octer performance. Herein a method is investigated to improve this upper limit.

In SPSS, an acoustic model is trained from speech corpora and the maximum likelihood model parameters are estimated. Recently, deep neural networks (DNNs) have been introduced for acoustic model training in SPSS, DNN improves synthesis accuracy compared to the conventional hidden Markov model (HMM) [5,6]. Additionally corpus-dependent high-quality vocoders with DNNs have been investigated [7,8], whereas the conventional high-quality ones described above [2,3] are corpus-independent. Although corpus-dependent high-quality vocoders with DNNs improve the speech quality compared to the conventional STRAIGHT v/octer in both HMM- and DNN-based speech synthesis [7], the synthesis quality depends greatly on the estimation accuracy of the glottal closure instants [9]. Neural network-based vocoders such as WaveNet¹ and SampleRNN [8] require a huge amount of speech corpus for high-quality synthesis.

This paper provides a method to improve the v/octed speech quality and is applicable to arbitrary vocoders with limited acoustic parameters. The proposed method reconstructs high-quality speech signals from the v/octed ones with DNNs. As a first attempt, a neural network, which can reconstruct the power spectrum of the original speech waveform from that of the v/octed one, is trained. Experiments on both analysis-synthesis speech and SPSS with a Japanese female speech corpus are conducted. Compared to the conventional DNN-based postfiltering to reconstruct the mel-cepstral and STRAIGHT spectral coefficients from the over-smoothed acoustic parameters estimated by HMM-based acoustic models [10], the proposed method directly reconstructs the power spectrum of the v/octed speech waveform, improving the upper limit of vocoders with limited acoustic parameters.

2. Proposed approach

To recover the degradation of the v/octed speech, a neural network that estimates the power spectrum of the original speech from that of the v/octet one is trained. Figure 1 shows a block diagram of the proposed method.

In the training stage (Fig. 1(a)), the original and v/octet speech waveforms are extracted by a window function. Then the discrete Fourier transform (DFT) is performed to obtain each power spectrum. A neural network is constructed with the power spectrum of the v/octet \( P_v \) as the input and that of the original speech \( P_{org} \) as the output. Parameter \( \hat{\lambda} \) of the neural network is then trained as

\[
\hat{\lambda} = \arg \min_{\lambda} \frac{1}{2} \sum_t \| P_{org} - P_v \|^2, \tag{1}
\]

where \( t \) is the frame number. The network is trained to minimize the square error of Eq. (1) in all frames throughout the corpus. The network parameters are updated for each mini batch.

When evaluating the distortion in the speech spectrum, spectral distortion (SD) is often introduced and defined as

\[
SD = 10 \log_{10} \frac{1}{FT} \sum_{f=1}^{F} \sum_{t=1}^{T} | P_{org} - P_v |, \tag{2}
\]

where \( F, T \) are the numbers of frequency bins and frames, respectively. In the proposed method, Eqs. (1) and (2)
indicate that the neural network is just trained to minimize the
SD of the vocoded speech.
In typical DNN-based acoustic modeling for SPSS, a
maximum likelihood parameter generation (MLPG) algorithm
[11] is often introduced, and both static and dynamic acoustic
features are used as the neural network output to obtain
smooth parameter trajectories [5]. Preliminary experiments
using the proposed method with or without dynamic features
did not yield significant differences. Therefore, only static
features are used in this paper.
In the synthesis stage, the vocoded speech is generated
from the acoustic parameters of the original speech signal in
the analysis-synthesis case and those estimated by the acoustic
models in SPSS (Fig. 1(b)). The power spectrum of the
vocoded speech is sequentially transferred to the reconstruct-
ed one by the trained network.
3. Experiments
To evaluate the effectiveness of the proposed method, two
experiments were conducted using a Japanese female speech
corpus recorded with a sampling frequency 48 kHz and
downsampled to 16 kHz. The first one employed analysis-
synthesis speech to confirm that the proposed approach
improves the vocoder when the acoustic features are perfectly
estimated in SPSS. The other one evaluated the proposed
method in SPSS using DNN-based SPSS.
3.1. Experiment with analysis-synthesis speech
The experiment with analysis-synthesis speech employed
7,000 (about 4.7 hours) and 20 utterances as the training and
test sets, respectively. The analysis window function was a
Hann window with a frame length of 256 samples and frame
shift of 16 samples (= 1 ms). The input and output of the
proposed neural network were 129-dimensional vectors of the
power spectra (0 to 8 kHz) of the vocoded and original speech
waveforms. There were 3 hidden layers with 1,024 nodes.
The activation function was a sigmoid function. An Adam
optimization algorithm [13] was used to update the param-
eters. The mini batch size was 256, and the network
parameters were updated up to 20 epochs. The input and
output vectors were normalized to have a zero-mean and
unit-variance.
The experiment evaluated two types of vocoders:
MLSA: A simple vocoder with the fundamental frequen-
cy ($ f_o $) analyzed by STRAIGHT, and the 0-th to 24-th mel-
cestral coefficients (25 dimensions) obtained by SPTK 3.9
and simple pulse excitations.
STRAIGHT: Similar to a previous work [4,5], $ f_o $, the
0-th to 24-th mel-cepstral and the 5-band aperiodic coef-
ficients extracted from the smoothed spectrum analyzed by
STRAIGHT.
The fundamental frequency was initially analyzed at 1 ms
and downsampled to 5 ms intervals. The other parameters
were analyzed every 5 ms based on the standard parameters in
HTS 2.1.1.
The proposed method was evaluated by the reconstructed
power spectrum and its SD. In the evaluation, the power
spectrum re-analyzed from the waveform generated by the
reconstructed power spectrum and the vocoded phase com-
ponent was used instead of that directly reconstructed by the
trained DNN.
Figure 2 shows examples of spectrograms for the original
speech waveform in the test set, the analysis-synthesis one by
each vocoder, and that reconstructed by the proposed method.
In addition, SD defined in Eq. (2) was introduced as an
objective evaluation criterion. Table 1 shows the results of the
averaged test set SD for each vocoder, while Fig. 3 plots the
averaged results at each frequency band for the STRAIGHT
vocoder. These results suggest that although the harmonic
structure and fluctuation components of the vocoded spectro-
grams at high-frequency bands deteriorate, the proposed
neural network can successfully restore the power spectrum of
the analysis-synthesis vocoded speech.
Furthermore, a paired comparison listening test between
the analysis-synthesis vocoded speech by STRAIGHT and
that reconstructed by the proposed method was conducted for
\[ http://sp-tk.sourceforge.net \]
\[ http://hts.sp.nitech.ac.jp/archives/2.1.1/ \]
a subjective evaluation. All 20 utterances of the test set were used as the evaluation speech set and presented by headphones. The listening subjects were two adult females and five adult males without hearing loss. The simple vocoded speech and the reconstructed one by the proposed method of a test set utterance were continuously presented in random order in 1 s intervals. Subjects could freely re-listen to them many times.

The subjects compared and judged the quality of two stimuli. An additional answer of neutral was added to denote cases where the subject could not detect a quality difference between the two stimuli. Table 2 shows the results of the listening test. The statistical test of the results indicates that the proposed method can significantly improve the vocoded speech quality.

### Table 1 Results of the averaged test set spectral distortion for analysis-synthesis speech.

<table>
<thead>
<tr>
<th></th>
<th>Default</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLSA</td>
<td>−33.9</td>
<td>−35.6</td>
</tr>
<tr>
<td>STRAIGHT</td>
<td>−36.3</td>
<td>−37.1</td>
</tr>
</tbody>
</table>

### Table 2 Results of the paired comparison listening test. S, P, and N are the answer number of simple STRAIGHT, STRAIGHT with proposal, and neutral.

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>P</th>
<th>N</th>
<th>p-value</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>22</td>
<td>99</td>
<td>19</td>
<td>&lt;10^{-10}</td>
<td>−7.0</td>
</tr>
</tbody>
</table>

3.2 Applying to SPSS

To evaluate the proposed method applicability to SPSS, a DNN-based acoustic model was trained from 6,000 of the 7,000 utterances used in the experiment for analysis-synthesis speech. The input parameter of the acoustic model of DNN was the 438-dimensional context label used in [6]. The outputs for both the MLSA and STRAIGHT vocoders were $26 \times 3 + 1 = 79$-dimensional and $46 \times 3 + 1 = 139$-dimensional vectors constructed from the static and dynamic features, and the voice/unvoice flag, $f_o$ was logarithmically converted and continuously interpolated. The input and output vectors were normalized to have a zero-mean and unit-variance as in Sect. 3.1. The DNN configuration of the acoustic model was the same as that for the proposed method. A smooth static parameter trajectory was generated from the

Fig. 2 Spectrograms of the original, vocoded and proposed speech waveforms.

Fig. 3 Results of the averaged test set spectral distortion at each frequency band for the STRAIGHT vocoder.
Table 3 Results of the averaged test set spectral distortion for DNN-based speech synthesis. (a): default, (b): proposal trained from analysis-synthesis power spectrum, (c) proposal trained from vocoded SPSS power spectrum, (d) proposal trained from vocoded SPSS power spectrum with correct \( f_0 \).

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLSA</td>
<td>-29.5</td>
<td>-29.9</td>
<td>-21.8</td>
<td>-25.2</td>
</tr>
<tr>
<td>STRAIGHT</td>
<td>-29.8</td>
<td>-30.0</td>
<td>-22.5</td>
<td>-26.3</td>
</tr>
</tbody>
</table>

estimated static and dynamic features by the MLPG algorithm [11]. The speech waveform was generated by a vocoder where the acoustic parameters were estimated by the trained acoustic model. The vocoded SPSS waveform was then converted to the reconstructed one by the proposed method where the reconstruction network trained in subsection 3.1 was used. In addition, the reconstruction network was also learned from the vocoded SPSS power spectrum of the training set. Furthermore, to exclude the harmonic structure error between the input and output, that with the correct \( f_0 \) was also investigated.

Table 3 shows the results of SD. Although there are slight improvements by the proposed method with the analysis-synthesis network (Table 3(b)), the quality improvements are unfortunately not found in preliminary listening tests. This is because the power spectra of the vocoded SPSS speech waveforms do not match the training ones. However, the results in Table 3(c) and (d) suggest that the proposal with the vocoded SPSS power spectrum not reconstructs but degrades the estimated SPSS acoustic features, and the reconstruction network cannot be well trained. To apply the proposal to SPSS, high-quality acoustic models which can reduce the over-smoothing and discontinuity in estimated acoustic features are required. In addition, although this paper introduced a simple feedforward DNN with frame-wise training, the proposed method with multiple frame input and recurrent neural networks will be investigated as in the conventional postfiltering approach [10].

Consequently, the results of the experiments confirm that the proposed method can significantly improve vocoded speech quality for analysis-synthesis speech. This implies that the proposed method is useful for SPSS when the acoustic parameters are perfectly estimated. Future work includes spectral phase modeling, applications to SPSS and speech corpus with high sampling frequency such as 48 kHz.

4. Conclusions

A method to improve the vocoded speech quality applicable to arbitrary vocoder with limited acoustic parameters is proposed based on deep neural networks. First, a neural network is trained using the input and output as the power spectra of the vocoded and original speech. Then the trained network reconstructs the vocoded speech. Both objective and subjective experiments using a Japanese female speech corpus validate the effectiveness of the proposed method for analysis-synthesis speech.

References
