STFT Spectral Loss for Training a Neural Speech Waveform Model

Citation for published version:

Digital Object Identifier (DOI):
10.1109/ICASSP.2019.8683791

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
STFT SPECTRAL LOSS FOR TRAINING A NEURAL SPEECH WAVEFORM MODEL

Shinji Takaki\textsuperscript{1}, Toru Nakashika\textsuperscript{2}, Xin Wang\textsuperscript{1}, Junichi Yamagishi\textsuperscript{1,3}

\textsuperscript{1}National Institute of Informatics, Japan \textsuperscript{2}The University of Electro-Communications, Japan \textsuperscript{3}The University of Edinburgh, UK

takaki@nii.ac.jp, nakashika@uec.ac.jp, wangxin@nii.ac.jp, jyamagis@nii.ac.jp

ABSTRACT

This paper proposes a new loss using short-time Fourier transform (STFT) spectra for the aim of training a high-performance neural speech waveform model that predicts raw continuous speech waveform samples directly. Not only amplitude spectra but also phase spectra obtained from generated speech waveforms are used to calculate the proposed loss. We also mathematically show that training of the waveform model on the basis of the proposed loss can be interpreted as maximum likelihood training that assumes the amplitude and phase spectra of generated speech waveforms follow Gaussian and von Mises distributions, respectively. Furthermore, this paper presents a simple network architecture as the speech waveform model, which is composed of uni-directional long short-term memories (LSTMs) and an auto-regressive structure. Experimental results showed that the proposed neural model synthesized high-quality speech waveforms.

Index Terms— speech synthesis, neural waveform modeling, WaveNet

1. INTRODUCTION

Research on speech waveform modeling is advancing because of neural networks\textsuperscript{1,2}. The WaveNet\textsuperscript{2} directly models waveform signals and demonstrates excellent performance. The WaveNet can also be used as a vocoder, which converts acoustic features, e.g., a mel-spectrogram sequence, into speech waveform signals\textsuperscript{3,4}. Such a neural speech waveform model used as a vocoder can be integrated into a text-to-speech synthesis system, and it has been shown that it outperforms conventional signal-processing-based vocoders\textsuperscript{5}. Neural speech waveform models will be essential components for many speech synthesis applications.

There have been several investigations of neural speech waveform models about output distribution and training criteria. A categorical distribution for discrete waveform samples and cross entropy are introduced to train the original WaveNet\textsuperscript{2}. The mixture of logistic distribution and a discretized logistic mixture likelihood are also used to train the improved WaveNet\textsuperscript{5}. The parallel-WaveNet\textsuperscript{6} and the ClariNet\textsuperscript{7} use a distilling approach to transfer the knowledge of auto-regressive (AR) WaveNet to a simpler non-AR student model. The distilling approach typically introduces an amplitude spectral loss as the auxiliary loss in the distillation in order to avoid generating devoiced/whisper voices. We believe that losses using short-time Fourier transform (STFT) spectra can be used beyond the distillation, and they can be used for training speech waveform models themselves. Because it is known that (complex-valued) STFT spectra represent speech characteristics well, the spectra would be useful for efficient training of neural waveform models.

In this paper, we propose a new loss using the STFT spectra for the aim of training a high-performance neural speech waveform model that predicts raw continuous speech waveform samples directly (rather than the auxiliary loss used for the distillation process). Because not only amplitude spectra but also phase spectra represent speech characteristics\textsuperscript{8}, the proposed loss considers both amplitude and phase spectral losses obtained from generated speech waveforms. Also, we give interpretation based on probability distributions about training using the proposed loss. This interpretation provides us better understanding of the proposed method and its relationship to various spectral losses such as Kullback-Leibler divergence or Itakura–Saito divergence\textsuperscript{9,10}. This paper also presents a simple network architecture as a speech waveform model. The proposed simple network is composed of uni-directional long short-term memories (LSTMs) and an auto-regressive structure unlike the WaveNet, which uses a relatively complicated convolutional neural network (CNN) with stacked dilated convolution.

The rest of this paper is organized as follows. Section 2 of this paper presents a neural speech waveform and the proposed loss to train it. Section 3 describes the proposed network architecture. Experimental results are presented in Section 4. We conclude in Section 5 with a summary and mention of our future work.

2. THE PROPOSED LOSS FOR A WAVEFORM MODEL

A neural speech waveform model targeted in this paper is as follows.

\[
y = f^{(\lambda)}(x),
\]

where, \(y \in \mathbb{R}^{|t|}, x = [x_1, \ldots, x_i]^T, x_i \in \mathbb{R}^D, \) and \(\lambda\) represent a neural network’s outputs (i.e., speech waveform samples), an input sequence (e.g., a log-mel spectrogram), an input feature, and parameters of a neural network, respectively. A sample index, a frame index, and the dimension of an input feature are represented by \(m, i, \) and \(D, \) respectively. Back-propagation is generally used to get optimal parameters \(\lambda\) for a loss function. Because training the model is a regression task, a simple loss function is a square error between natural speech waveform samples and a neural network’s outputs \(y\) as follows.

\[
E^{(\text{wave})} = \sum_m (\hat{y}_m - y_m)^2,
\]

where, \(^\cdot\) denotes natural data. However, this criterion does not consider any information related to frequency characteristics of speech.

This work was partially supported by JST CREST Grant Number JP-MICR18A6, Japan and by MEXT KAKENHI Grant Numbers (16H06302, 16K16096, 17H04687, 18H04120, 18H04112, 18KT0051, 18K18069), Japan.
We then obtain a partial derivative of Eq. (7) w.r.t. \( n \).

### 2.2. Amplitude spectral loss

A loss function for an amplitude spectrum of frequency bin \( n \) at frame \( t \) is defined as a square error.

\[
E^{(amp)}_{t,n} = \frac{1}{2} (\hat{A}_{t,n} - A_{t,n})^2
\]

where, \( \hat{A}_{t,n} \) represents an amplitude spectrum of outputs (i.e., \( \theta_{t,n} \)), and \( A_{t,n} \) represents an amplitude spectrum of outputs (i.e., \( \theta_{t,n} \)). The amplitude and phase losses are related to each other through a neural network’s outputs during model training, although each loss function focuses on only amplitude or phase spectra.

In this paper, a combination of amplitude and phase spectral loss functions is used for training a neural speech waveform model.

\[
E^{(op)} = \sum_{t,n} (E^{(amp)}_{t,n} + \alpha_{t,n} E^{(ph)}_{t,n}),
\]

where, \( \alpha_{t,n} \) denotes a weight parameter. We use three types of \( \alpha_{t,n} \) for training a neural speech waveform model.

\( \alpha_{t,n} = 0 \): Using only the amplitude spectral loss function for model training.

\( \alpha_{t,n} = 1 \): Using a simple combination of the amplitude and phase spectral loss functions.

\( \alpha_{t,n} = \nu_{t} \): Here, \( \nu_{t} \) represents a voiced/unvoiced flag (1:voiced, 0:unvoiced). In this case, we assume that phase spectra in unvoiced parts are random values, and hence the phase spectral loss computed in unvoiced parts is omitted from model training.
Form samples. The output network with an auto-regressive structure is removed from feedback waveforms. CNNs with stacked dilated convolution are not used in our method. An output network is simply based on uni-directional LSTMs.

An overview of the proposed network is shown in Fig. 2. The processing below the dotted line in Fig. 2 is the same as that used in the WaveNet vocoder [12], in which input features are converted to hidden representations through an input network, and then they are duplicated to adjust time resolution. The processing above the dotted line is conceptually the same as that used in the WaveNet vocoder, in which feedback samples and hidden representations are fed back into an output network with an auto-regressive structure to output the next speech waveform sample. However, the following are two remarkable differences of the proposed network:

1: An output network is simply based on uni-directional LSTMs. CNNs with stacked dilated convolution are not used in our method.

2: Spectral amplitude information is removed from feedback waveform samples. The output network with an auto-regressive structure feeds natural waveform samples back to the network during teacher-forced training [13]. A network composed of uni-directional LSTMs tends to rely on the feedback waveform samples while ignoring the input features if natural waveform samples are directly fed into it. To solve this problem, feedback waveform samples are converted as follows.

\[ y'_{m-1 \rightarrow m-1} = F^{-1} \left( \frac{F(y_{m-1 \rightarrow m-1})}{F(y_{m-1 \rightarrow m-1})} \right) \]  

where, \( F \), \( F^{-1} \), the fraction bar, and \(|\cdot|\) denote FFT operation, inverse FFT operation, element-wise division, and the element-wise absolute, respectively. Speech waveform samples, whose amplitude spectra are 1, are obtained through the conversion. This conversion can be regarded as data dropout [15], although amplitude spectra are dropped and are replaced with 1 instead of 0.

4. EXPERIMENT

4.1. Experimental conditions

The proposed neural speech waveform models were evaluated as a vocoder5. We used a female speaker (slt) from the CMU-ARCTIC database [16], 1,032 and 50 utterances were used as training and test sets, respectively. Their speech waveforms have a sampling frequency of 16 kHz and a 16-bit PCM format.

As an input feature, an 80-dim log-mel spectrogram was used. Frame shift, frame length, and FFT size were 80, 400, and 512, respectively. Log-mel spectrograms and waveform samples were normalized to have 0 mean and 1 variance for training the proposed speech waveform model.

Input features, i.e., the log-mel spectrogram sequence, were first converted into hidden representations through an input network composed of an 80-unit bi-directional LSTM and a CNN with 80 filters whose size is 5 (time direction) × 80 (frequency direction). Then, hidden representations were duplicated to adjust time resolution. The previous 400 samples and hidden representations obtained from the input network were fed into an output network. The output network was composed of three 256-unit Uni-directional LSTMs.

Frame shift, frame length, and FFT size used to obtain the STFT spectra of a neural network’s outputs were 1, 400, and 512, respectively. Voiced/Unvoiced flags were extracted by RAPT [17]. Mini-batches were created, each from 120 randomly selected speech segments. Each mini-batch contained a total of 15 s speech waveform samples (each speech segment equaled 0.125 s). We used the Adam optimizer [13], and the number of updating iteration was 100k.

We trained four speech waveform models with the proposed network architecture. The difference among these four models was the loss function used for model training. Eq. 2 and Eq. 13 with three types of \( a_{t,n} \) were used as loss functions for training the four models. For comparison, the WaveNet was trained by using 80-dim log-mel spectrograms and 1,024 mu-law discrete waveform samples as input and output training data. The network architecture of the WaveNet was the same as that used in [12]. The WORLD [19] was also used as a baseline signal-processing-based vocoder. Spectral envelopes and aperiodicity measurements obtained by utilizing the WORLD were converted to 59-dim mel-cepstrum and 21-dim band aperiodicity. The total dimensions of a WORLD acoustic feature was 82 (60 (mel-cepstrum) + 1 (voiced/unvoiced flag) + 1 (f0) + 21 (band aperiodicity)). In total, we used six vocoders (the four proposed models, the WaveNet, the WORLD) in the experiment.

5 Synthetical speech samples and codes for model training can be found at https://nii-yamagishilab.github.io/TSNetVocoder/index.html and https://github.com/nii-yamagishilab/TSNetVocoder respectively.
4.2. Experimental results

4.2.1. Spectrogram

Fig. 3 shows spectrograms of analysis-by-synthesis speech waveform segments. There is an unvoiced part around 0.2 seconds. We can see from Fig. 3(d) and (e) that the proposed models trained by using $E^{(wav)}$ and $E^{(sp)}$ without the phase spectral loss function (i.e., $\alpha = 0$) generated noisy spectrograms. Also, it can be seen from Fig. 3(f) that large value amplitude spectra are observed around 7 kHz in the unvoiced part, whereas they are not observed in the natural spectrogram. Such artifacts are not observed in the spectrogram shown as Fig. 3(g). These results indicate that training a proposed model by using the amplitude and phase spectral loss functions is adequate and adjusting the weight parameter $\alpha$ by using the voiced/unvoiced flag further improves the performance.

Both spectrograms of the WaveNet and the proposed model trained by adjusting $\alpha_{t,n}$ by using the voiced/unvoiced flags (i.e., Fig. 3(c) and (g)) are similar to the natural one, although the detailed spectral structures are changed.

4.2.2. Subjective evaluation result

The subjective evaluation was conducted using 158 crowd-sourced listeners. Natural samples and samples synthesized from the 12 systems were evaluated. The number of synthetic samples was 650 (13 systems x 50 test sentences). Participants evaluated speech naturalness on a five-point mean opinion score (MOS) scale. Each synthetic sample was evaluated 40 times, giving a total of 26,000 data points. Thanks to the large number of data points, differences between all combinations of evaluated systems are statistically significant (i.e., $p < 0.05$).

First, among the proposed models, those trained together with the phase spectral loss outperformed the others. Using not only amplitude spectra but also phase spectra to calculate loss is useful for training a neural speech waveform model. The proposed model trained by adjusting $\alpha_{t,n}$ by using the voiced/unvoiced flags obtained the best score among the proposed models.

Second, it can be seen from Fig. 3 that the best performance of the proposed model (i.e., $\alpha = v$) is better than that of the WORLD. In the WORLD, the TTS result is drastically decreased from the AbS result. On the other hand, we can see that the proposed model with the best configuration (i.e., $\alpha = v$) is robust against acoustic features predicted from the acoustic model.

Finally, compared with the WaveNet, the best performance of the proposed models (i.e., $\alpha = v$) is slightly worse. But note that the number of parameters of the proposed model (1.8M) is less than that of the WaveNet (2.3M). Although we need to improve loss functions and network architecture to achieve comparable performance with the WaveNet, it is obvious that the proposed model achieved high-performance speech waveform modeling.

5. CONCLUSION

We proposed a new STFT spectral loss to train a high-performance speech waveform model directly. We also presented a simple network architecture for the model, which is composed of uni-directional LSTMs and an auto-regressive structure. Experimental results showed that the proposed model can synthesize high-quality speech waveforms.

This is a part of our sequential work. In our next paper [13], we will show that the above model can be improved further to achieve performance comparable with the WaveNet. Our other future work includes training based on other time-frequency analysis such as modified discrete cosine transform.
6. REFERENCES


