A Phoneme Sequence Driven Lightweight End-To-End Speech Synthesis Approach

Zite Jiang, Feiwei Qin and Liaoying Zhao
College of Computer Science, HangZhou Dianzi University, Hangzhou 310018, China
zhaoly@hdu.edu.cn

Abstract. This paper develops an end-to-end neural network model for text-to-speech (TTS) system based on phoneme sequence. Inspired by the Tacotron-2, the proposed model adopts an encoder-decoder model with attention mechanism and applies mel-spectrogram to measure the intermediate acoustic feature. Phoneme sequence is used to replace the character sequence in order to overcome the shortage of the character feature used in Tacotron-2. Unlike the conventional concatenate methodology based TTS system, our model can generate waveform directly from phoneme sequence. In addition, analogue from text analysis, a new analysis methodology is proposed for phoneme analysis. Experiment result on LJ Speech dataset shows that, compared with char-based model, our model can get a comparative or better performance.

1. Introduction

Speech synthesis system has great potential in developing intelligent system like smart home devices, smart guided tours and human-like robots in the future. However, it is still a challenging work to synthesize an intelligibility and naturalness voice from text. Generally, a text-to-speech (TTS) system could be divided into two pieces: acoustic feature extraction and waveform generation (also called Vocoder).

Conventional TTS uses a concatenated methodology like unit selection synthesis [1] and statistical parametric based subband synthesize [2]. However, a concatenated speech synthesis requires complex model design and large database setting which is hard to be deployed in build-in systems. A seq-to-seq neural network model is proposed for machine translation and succeeds in speech synthesis [3].

How to efficiently extract expressive feature from the raw text is still a challenging issue in data-driven speech synthesis. Tacotron [4] and Tacotron-2 [5], two end-to-end models, are conditioned on character feature. A deep voice 3 [6] considers both character and phoneme feature.

In addition, acoustic feature extraction has a great influence on Vocoder process. STRAIGHT [7] is conditioned on fundamental frequency (F0), spectral envelope and aperiodicity features. WaveNet [8], a fully autoregressive model, is a robust generative model. However, it requires long training stage and is in low synthesis speed. A fast Griffin-Lim [9], an estimation of origin waveform, is based on spectrogram.

Inspired by the sequence-to-sequence char-based Tacotron-2, we develop a phoneme sequence mode that has more expressive feature than the character sequence in speech synthesis.

2. Related Work
There is no universal accepted linguistic feature as input and acoustic feature as output in neural network model currently. **Deep Voice 1** [10] uses phonemes, looking up from the phoneme dictionary, to predict the phoneme duration and F0. Tacotron proposes a character sequence as input to predict linear spectrogram. Deep Voice 3 considers both phoneme and character feature as input. In paper [11], a phonetic posteriorgrams is used in speech synthesis stage. In our model, we consider phoneme which have the same representability to text, as a sort of text.

Conventional acoustic features extractor is based on HMM [12]. A Bi-LSTM [13] and CNN [14] model performs well in sequence feature extraction in text processing. Attention mechanism performs well in text-based **machine translation** [15], **speech recognition** [16] and **speech synthesis** [17]. Inspired by text processing methodology, we adopt a model like text processing model and hope it work well in phoneme processing.

Generally, researchers conduct the process of short-time Fourier transformation (STFT) on the raw waveform to obtain the linear spectrogram. However, STFT is an irreversible operation, which causes an open issue to design a better **signal estimation** approach. A neural network model like WaveNet could generate smooth voice. However, it is tough to train the model. Griffin-Lim uses an iterative algorithm to generate voice. Hence in this paper, we adopt Griffin-Lim as Vocoder.

### 3. Overview of Approach

This section gives a brief introduction of the proposed model. As shown in Fig.1, the whole procedure can be divided into two stages: acoustic feature extractor and Vocoder. The phoneme sequence is converted from text (also called grapheme-to-phoneme). We will discuss how to obtain phoneme sequence from text in Section 3.1 and the details of acoustic features extractor in Section 3.2.

![Figure 1. Speech synthesis architecture. The Model uses phoneme sequence as input to predict spectrogram and then estimate the signal.](image)

**Grapheme-to-phoneme (G2P) transcription** [18] converts word to phoneme. There are two kinds of approaches respectively based on dictionary look-up and model. Since a model-based approach could not achieve hundred percentages to predict the phoneme, to reduce the unnecessary error causes in system, a dictionary look-up is used to convert grapheme to phoneme. **Fig. 2** shows an example of the dictionary look-up.

![Figure 2. A dictionary look-up example.](image)

Since a word could have multiple phonemes in different context and sound-link is an inevitable phenomenon in real human spoken, many sequence-to-sequence models are proposed to extract contextual information to synthesize more natural voice. Since an encoder-decoder model works well in generative processing, we consider the similar process. The acoustic feature extractor could be divided into three parts: encoder, decoder and postprocess net, as shown in **Fig. 3**.
3.1. Encoder and decoder
In the encoding stage, the model should be able to extract enough features that present the target spectrogram. The phoneme is embedded into a trainable vector, and then we pass the embedded phoneme into a bi-directional LSTM layer. Finally, encoded vector is generated after encoder.

In the decoding stage, we use the same architecture described in Tacotron-2. Tacotron-2, a local-sensitive attention based autoregressive neural network, performs well in dynamically decoding the spectrogram information. The input of decoder each time step is concatenated by attention context vector and previous time step’s output. The attention context vector is calculated from the previous step using the same mechanism as the Tacotron-2 [5]. After pre-dense layer (fully connected neural network), the result is passed into uni-directional LSTM layer. Finally, the raw frame features are created.

3.2. postprocess net and audio generation
In the postprocess net, the output of the decoder, presenting the time features, are passed to a post convolutional residual network [19] to predict mel-spectrogram. The linear spectrogram, transformed from the mel-spectrogram, then convert to the waveform by Griffin-Lim algorithm.

4. Experimental Results

4.1. Experiment setup
The LJ Speech dataset was used in the experiment. There is totally about 23 hours human speech length. In the experiment, only half of the data (about 10 hours’ length) is used, the sample rate is 22,050Hz.

The training data occupies 95% in the whole data. This dataset has mappings between the text and waveform, but lacks the phoneme information. Carnegie Mellon university pronouncing dictionary (CMUdict) is used in grapheme-to-phoneme transcription. The dictionary has more than 130,000 words. Since the dictionary doesn’t include all vocabularies in LJ Speech dataset, these sentences are removed in our training and test stages.

In acoustic feature extractor, the parameter settings of our model are shown in Table. 1. In the post dense layer, a linear transform is used to fix the target length to 80-dimension, which is the same size
as mel-frequency cepstrum coefficients (MFCC) [20] dimension. During the training, we also predict the stop token in the dynamic LSTM to synthesize the reasonable length of the voice.

Table 1. Parameter settings in acoustic feature extractor.

| Module                  | Unit size | channels | depth | Kernel size |
|-------------------------|-----------|----------|-------|-------------|
| Embedding layer         | 512       | -        | -     | -           |
| Encode bi-LSTM          | 256       | -        | -     | -           |
| Pre-dense layer         | [256, 256]| -        | 2     | -           |
| uni-LSTM                | 1024      | 2        | -     | -           |
| Post convolutional residual network | -       | 512      | 5     | 5           |

At the signal estimation stage, Griffin-Lim algorithm is used to synthesize the final waveform presentation. In addition, both character sequence and phoneme sequence are used as input in the experiment for comparison.

4.2. Experiment Analysis

In English word, the length of the character is much longer than phoneme’s. When this phenomenon reflects in a sequence-to-sequence analysis, a phoneme-only based model will benefit from it since the length is shortened. We conduct a small evaluation on the length of the two different features in LJ Speech, the results are shown in Table 2. From which we can see that the length of the phonemes is shorter than characters.

Table 2. A brief comparison of two sequence features.

| Index                  | Value   |
|------------------------|---------|
| Number of Phonemes     | 730172  |
| Number of Characters   | 867121  |
| Ratio                  | 0.842   |
| Total words            | 191628  |

Character feature requires stronger contextual information to map the pronunciation. Here is an example, the ‘o’ in ‘hello’ and ‘world’ have different pronounce. The phoneme feature could eliminate the unnecessary contextual-dependence and will reduce the prediction error.

4.3. Experiment Results

In our experiment, we didn’t train our model for enough time to converge it. But it still takes about 70 hours training in Tesla K80 GPU.

The mel cepstral distortion (MCD) value and spectrum distortion value were used to evaluate the results objectively. Mel-scaled coefficient distortion can evaluate the human auditory-level perception distortion and spectrum distortion can evaluate the signal-level distortion. The formula of MCD is shown as (1), where T means the frames of the sentence, \( c_d^{target}(t) \) means the d-th coefficient of the targets, \( c_d^{predicted}(t) \) means the d-th coefficient of the predicted results, D means the number of coefficients. The formula of SD is shown as (2), where \( s_f^{target}(t) \) means the f-th frequency of the target results, \( s_f^{predicted}(t) \) means the f-th frequency of the predicted results.

\[
\text{MCD} = \alpha \cdot T^{-1} \cdot \sum_{t=0}^{T-1} \left( 2 \cdot \sum_{d=0}^{D-1} \left( c_d^{target}(t) - c_d^{predicted}(t) \right)^2 \right)^{1/2}, \quad \alpha = \frac{10^{0.2}}{\ln 10} \approx 6.14
\]

\[
\text{SD} = T^{-1} \cdot D^{-1} \cdot \sum_{f=0}^{D-1} \left( s_f^{target}(t) - s_f^{predicted}(t) \right)^2^{1/2}
\]

In the experiment, we also consider using the WaveNet [12] as our Vocoder. However, we trained the WaveNet more than 30 hours, the generated voice is still noisy. Instead, the Griffin-Lim algorithm can get a good performance without a long-time training. Fig.4 shows the evaluate result.
Figure 4. Result of mel spectrogram prediction.

For comparison, we adopted the same architecture of the phoneme-based sequence model, while the only difference is the character feature to phoneme feature. Models with character sequence inputs take a bit more time than phoneme sequence in training under the same training conditions, and same thing happens in synthesis stage, however, it didn’t get a better performance. In addition, we also train the Tacotron-2 model before the WaveNet, and test in the same condition. The result is summarized in Table. 3. The phoneme-based model have slightly overperform than the char-based model. The Griffin-Lim distortion is calculated to estimate the Griffin-Lim algorithm’s error.

Table 3. Comparison of char-based and phoneme-based model

| System                        | MCD(dB)         | SD(dB)             |
|-------------------------------|-----------------|--------------------|
| Char-based (our model)        | 11.4083294573   | 0.45788899644      |
| **Phoneme-based (our model)** |                 |                    |
| Tacotron-2 (char-based)       | 11.1987140227   | 0.45884787579      |
| Tacotron-2 (phoneme)          | 11.2795017916   | 0.457523534425     |
| Griffin-Lim distortion        | 11.8089875978   | 0.459087886076     |

5. Conclusion and Future Work
A new model using a phoneme sequence feature is proposed in this paper. The new model can achieve comparative or better performance to the character sequence-based model. Compared with the character-based model Tacotron-2, our model is faster in the training and get a lower MCD value. The result shows that the phoneme can be analysed as the text in speech signal processing and this feature is better than character in the speech synthesis.

By eliminating the error caused by the text prediction from phoneme, using phoneme as intermedia feature will make a progress in voice conversion system.

Acknowledgments
This work was supported in part by the Key Research & Development Program of Zhejiang Province (No. 2019C03127). The authors would like to thank the reviewers in advance for their comments and suggestions.

References
[1] Merritt T, Clark RA, Wu Z, Yamagishi J, King S. Deep neural network-guided unit selection synthesis. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2016), 2016 Mar 20 (pp. 5145-5149).
[2] Okamoto T, Tachibana K, Toda T, Shiga Y, Kawai H. An investigation of subband WaveNet vocoder covering entire audible frequency range with limited acoustic features. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2018), 2018 Apr 15 (pp. 5654-5658).
[3] Fan Y, Qian Y, Xie FL, Soong FK. TTS synthesis with bidirectional LSTM based recurrent neural networks. In Fifteenth Annual Conference of the International Speech Communication Association 2014.

[4] Wang Y, Skerry-Ryan RJ, Stanton D, Wu Y, Weiss RJ, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S, Le Q. Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135. 2017 Mar 29.

[5] Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, Wang Y, Skerrv-Ryan R, Saurous RA. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2018), 2018 Apr 15 (pp. 4779-4783).

[6] Ping W, Peng K, Gibiansky A, Arik SO, Kannan A, Narang S, Raiman J, Miller J. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. arXiv preprint arXiv:1710.07654. 2017 Oct 20.

[7] Kawahara H. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. Acoustical science and technology. 2006;7(6):349-53.

[8] Oord AV, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499. 2016 Sep 12.

[9] Griffin D, Lim J. Signal estimation from modified short-time Fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1984 Apr;32(2):236-43.

[10] Arik SO, Chrzanowski M, Coates A, Diamos G, Gibiansky A, Kang Y, Li X, Miller J, Ng A, Raiman J, Sengupta S. Deep voice: Real-time neural text-to-speech. arXiv preprint arXiv:1702.07825. 2017 Feb 25.

[11] Sun L, Li K, Wang H, Kang S, Meng H. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In 2016 IEEE International Conference on Multimedia and Expo (ICME2016), 2016 Jul 11 (pp. 1-6).

[12] Wen Z, Tao J, Pan S, Wang Y. Pitch-scaled spectrum based excitation model for HMM-based speech synthesis. Journal of Signal Processing Systems. 2014 Mar 1;74(3):423-35.

[13] Zhou C, Sun C, Liu Z, Lau F. C-LSTM neural network for text classification. arXiv preprint arXiv:1511.08650. 2015 Nov 27.

[14] Abdel-Hamid O, Mohamed AR, Jiang H, Penn G. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2012), 2012 Mar 25 (pp. 4277-4280).

[15] Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025. 2015 Aug 17.

[16] Chorowski JK, Bahdanau D, Serdyuk D, Cho K, Bengio Y. Attention-based models for speech recognition. In Advances in Neural Information Processing systems 2015 (pp. 577-585).

[17] Bulyko I, Ostendorf M. Joint prosody prediction and unit selection for concatenative speech synthesis. In ICASSP 2001 May 7 (pp. 781-784).

[18] Bisani M, Ney H. Joint-sequence models for grapheme-to-phoneme conversion. Speech communication. 2008 May 1;50(5):434-51.

[19] Laina I, Rupprecht C, Belagiannis V, Tombari F, Navab N. Deeper depth prediction with fully convolutional residual networks. In 2016 Fourth International Conference on 3D Vision (3DV), 2016 Oct 25 (pp. 239-248). IEEE.

[20] Zheng F, Zhang G, Song Z. Comparison of different implementations of MFCC. Journal of Computer science and Technology. 2001 Nov 1;16(6):582-9.