Research on Speech Synthesis Technology Based on Rhythm Embedding

Tianxin Wu, Lasheng Zhao\textsuperscript{a}, Qiang Zhang

Key Laboratory of Advanced Design and Intelligent Computing Ministry of Education, Dalian University, Dalian 116622, China
wtx940907@163.com
\textsuperscript{a}zhaolasheng@dlu.com

Abstract. In recent years, Text-To-Speech (TTS) technology has developed rapidly. People have also been paying more attention to how to narrow the gap between synthetic speech and real speech, hoping that synthesized speech can be integrated with real rhythm. A rhythmic feature embedding method for Text-To-Speech was proposed in this thesis based on Tacotron2 model, which has arisen in the field of TTS in recent years. Firstly, rhythmic feature extraction through World vocoder can reduce redundant information in rhythmic features. Then, rhythmic feature fusion based on Variational Auto-Encoder (VAE) network can enhance rhythmic information. Experiments are carried out on the data set LJSpeech-1.0, and then subjective evaluation and objective evaluation are carried out on the synthesized speech respectively. Compared with the comparative literature, the subjective blind hearing test (ABX) score increased by 25%. At that same time, the objective Mel Cepstral Distortion value (MCD) declined to 12.77.

1. Introduction

To Speech Synthesis, also known as Text To Speech (TTS), is a technology that converts input text into corresponding speech[1]. Speech Synthesis, as an important part of man-machine communication, aims to produce a machine that can communicate with people in language and has a broad application market. Traditional speech synthesis methods mainly include statistical parameter speech synthesis and waveform splicing speech synthesis. Statistical parameter speech synthesis makes statistical modeling of acoustic features extracted by vocoder from the perspectives of digital signal processing and statistics, and then inputs the acoustic features predicted by the model into vocoder for speech synthesis[2]. However, the above processes all require researchers to have a wide range of knowledge reserves in professional fields. Waveform splicing speech synthesis is to select corresponding speech unit segments from corpus according to text information, and then splice the unit segments to synthesize speech[3][4]. This requires a very large speech library to select speech unit segments, and the synthesized speech lacks flexibility.

Due to the rapid development of in-depth learning, a new type of speech synthesis method is presented, which is different from the traditional speech synthesis method. Deep learning integrates internal modules into the same model, and does not need to establish parameter modules based on various specific domain knowledge, i.e. end-to-end speech synthesis system. End-to-end speech synthesis system has attracted extensive attention since its appearance, Wang et al.[5][6], who carried out speech synthesis by introducing a soft attention mechanism into seq2seq structure to reduce feature...
engineering and let the model learn features by itself. Char2wave[7][8][9] directly input English characters to predict acoustic parameters, and then performing speech synthesis through vocoder. Qiu et al.[10] made use of the end-to-end architecture to realize Chinese speech synthesis. However, the above speech synthesis method only starts from the intelligibility of the generated speech, improves and optimizes the model, and improves the intelligibility of the synthesized speech. The current mainstream Tacotron2[11] model can directly generate speech from graphics or phonemes. The model first predicts Mel spectrogram through a Seq2seq network with a position-sensitive attention mechanism, and then uses an improved Wavenet[12] network as a vocoder to synthesize the time domain waveforms of these spectrograms. Tacotron2 has great advantages in synthesis speed and intelligibility of synthesized speech, but in prosody prediction, Tacotron2 model does not add prosody module, and synthesized speech also lacks prosody sense. This is because Tacotron2 stores other information in the weight in order to obtain better convergence during the training process, allowing the model to randomly select and store itself, thus causing the synthesized speech to be too mechanized. Zhang et al.[13] introduced style recognition model into Tacotron2. This recognition model takes Mel spectrum as input and guides speech synthesis through style representation extracted from Mel spectrum. However, Mel spectrum focuses on several specific frequency ranges sensitive to human ears and their distribution ability, ignoring other characteristics of sound, such as emotion and rhythm.

Based on the above analysis, this article proposes a rhythmic embedding method based on Tacotron2 to guide speech synthesis. By embedding a rhythmic module in the model, the model can obtain the rhythmic information of the speaker and learn the speech mode at the same time, so as to reduce the error between the predicted Mel spectrum and the real Mel spectrum, thus making the synthesized speech closer to the target speaker's voice. The experimental results on the open source data set LJSpeech-1.0 show that it is effective to guide speech synthesis by introducing prosody.

2. Model Architecture

Based on Tacotron2, this article proposes a method of rhythm embedding for Speech Synthesis, so that the synthesized speech has a better sense of rhythm. The specific process is shown in Figure 1. The main components of the model are: (1) Rhythms Encoder, used to extract and generate the required rhythm information to participate in Speech Synthesis; (2) Based on the style of Tacotron2, the encoding module and the Seq2Seq model made through the position-sensitive attention mechanism are to predict the Mel spectrogram; (3) the vocoder model, which uses the Wavenet vocoder to receive the predicted Mel spectrogram and convert it into a speech waveform.

![Figure1 Speech Synthesis Model](image.png)

2.1. Rhythms Encoder

Different people have complex differences in physiology and psychology, so there are differences between each person's phonetic features, and even the phonetic features extracted from the same person's phonetic features in different contexts are different. The purpose of rhythmic coding module is to add rhythmic features of the target speaker to the synthesized speech so that the synthesized speech sounds closer to the target speaker's voice. Therefore, selecting which feature parameter to
represent the rhythmic features of the speaker is a key link in this article. The features of speech signals are divided into three categories: segmental features, suprasegmental features and speech features. Segment features mainly show rhythmic features of speech, including pitch frequency, spectrum features, spectrum envelope, etc. These characteristics are mainly related to people’s physiological conditions and sometimes influenced by the speaker’s current emotions. The pitch of a voice can be changed by changing the pitch frequency, and the spectral envelope can affect the atmosphere and timbre of the voice. Moreover, there are many prosody-related information in the spectral envelope of the voice, such as the semantic information of the voice and the identity information of the speaker. Therefore, the rhythmic coding module proposed in this method mainly fuses the pitch frequency and spectrum envelope of segment features to generate rhythmic features.

2.1.1. Prosodic Feature Extraction Based on World Vocoder

The World vocoder was created by Masanori et al [14] who proposed a high-quality speech analysis and synthesis model, which can extract clearer acoustic features and speed up the processing speed of the model. This article uses World vocoder to analyze speech and extract rhythmic features. As shown in Figure 2.

![Figure 2 Feature Extraction](image)

The feature extraction based on World vocoder includes three modules: Dio Algorithm, Cheap Trick Algorithm and Platinum Algorithm. The Dio Algorithm module extracts the fundamental frequency characteristics (i.e. F0) of the speech signal according to the input original waveform, corresponding to the periodic pulse sequence. The Cheap Trick Algorithm module calculates the Spectral Envelope according to the F0 extracted by the Dio Algorithm and the original waveform, which corresponds to the resonant part of the channel. The Platinum Algorithm module calculates an aperiodic parameter based on the original waveform, F0 and spectral envelope. Since the spectrum envelope and aperiodic signals are both of high-dimensional features, the pure fundamental frequency, spectrum envelope, and aperiodic signals will have higher dimensions after being fused, and the resulting Feature Sequences will carry a lot of pressure to the training of the model. Here, a convolution neural network is used to reduce the dimension of Feature Sequences. Finally, VAE network is used for feature fusion, and the obtained rhythmic features are input into Tacotron2 model.

2.1.2. Prosodic Feature Fusion Based on VAE Network

Variational Auto-Encoder (VAE) is one of the two main generative models in deep learning. This model was first proposed by Kingma et al. on ICLR 14 [15][16]. For a generative model, what we want to find is the distribution of the original data \(p(x)\). But we only have discrete samples of the true distribution \(\{x_1, x_2, ..., x_n\}\). Assume the type of distribution (e.g. Gaussian distribution), and the parameters are then calculated using the maximum likelihood function \(\theta = \arg\max_\theta \Sigma_i \log p(x_i)\) so that \(p_\theta(x)\) is obtained. But \(p(x)\) may be a very complicated distribution that is difficult to express, and it is difficult to choose an appropriate hypothetical distribution to calculate the maximum likelihood function. So that in addition to that visible variable \(x\), assume that there is an intermediate one that
obeys a simple distribution $p(z)$ potential variables of $z$. Then the distribution of the original data can be decomposed into:

$$p(x) = \int p(x|z)p(z)dz$$  \hspace{1cm} (1)

Therefore, from the perspective of probability distribution, the variational self-encoder can be considered as a model to convert the probability distribution of the original data to the probability distribution of the target data, so that the original distribution and the target distribution are the same. Therefore, the essence of building a model is distribution transformation. According to formula (1), $p(x|z)$ is a model with generation from $Z$ to $X$ and we assume that $Z$ obeys the standard normal distribution, that is $p(z) = N(0,1)$. Then we can first sample one $Z$ from the standard normal distribution. And then calculate a $X$ according to the $Z$. The mean and variance of the original samples are used to reconstruct the data into the normal distribution of the potential space, then the normal distribution is randomly sampled, and the sampled results are decoded to generate the target data. Therefore, variational self-encoder is a reconstruction process and a sampling process.

In the process of refactoring, $X$ we want to refactor can minimize the error between the original distribution and the target distribution, but the reconstruction process is affected by many external factors, such as noise. Because $Z$ is resampled and not directly calculated by the encoder, thus increasing the difficulty of data reconstruction and the burden of computing resources. Since the mean and variance are calculated by neural networks, we can make the variance tend to 0 in the reconstruction process, which can not only reduce the difficulty of the data reconstruction process, but also ensure the randomness of the sampling process.

After the potential space is expressed as a normal distribution, there is a sampling process. We need to sample from $p(z|x)$, although it has been assumed that $p(z|x)$ is subject to normal distribution, but the mean value of the distribution $\mu$ and variance $\sigma$ still needs to be obtained by neural network, and then back propagation is carried out to optimize the mean and variance. However, the sampling process is not derivative, so a sampling technique of parameter reproduction is proposed in variational self-encoder. Assuming $z$ be subject to standard normal distribution $N(\mu, \sigma^2)$, then $\varepsilon = \frac{z-\mu}{\sigma}$ also obeys the standard normal distribution. And so $z = \varepsilon * \sigma + \mu$. Due to both of $\mu$ and $\sigma$ include other parameters in the training process, back propagation optimization parameters can be carried out, i.e. $z$ is sampled from $N(\mu, \sigma^2)$, which is equivalent to sample $\varepsilon$ from $N(0, 1)$ and then proceed $z = \varepsilon * \sigma + \mu$ transformation. The rhythmic fusion module proposed in this method adopts cyclic reference encoder and two fully connected layers to calculate the mean and variance of the distribution, the reference encoder consists of six two-dimensional convolution layers and a GRU layer. The feature coding of the reference audio is obtained by the reference encoder, and then the mean and standard deviation of the potential variables are generated through two independent fully connected layers (FC) with linear activation functions. Finally, the fused rhythmic features are obtained by normal distribution and re-parameterization.

2.2. Tacotron2 Model Based on Style Coding

Tacotron2 is currently the most widely used end-to-end speech synthesis model. It directly predicts Mel spectrum from character sequences and converts the predicted Mel spectrum into waveforms using Wavenet vocoder module. The Tacotron2 model based on style coding converts the text input into 512-dimensional phoneme sequence. At the same time, the Mel spectrum extracted from the speech signal deduces the potential style representation of the speech through VAE network, i.e. style coding. The style coding sequence and phoneme sequence are fused and processed by an encoder based on convolution neural network to obtain context features with style. Then the encoded context features are taken as the input of the position-sensitive attention module, so that the fully encoded sequence becomes a context vector of fixed length. Finally, through the decoder based on automatic regression cyclic neural network, the input sequence encoded in each step and each frame is predicted to be Mel spectrum. In order to reduce the complexity of the model and reduce the training time and prediction Mel spectrum time of the model at the same time, this method uses two residual GRU
structures with 256 GRU units in each layer in the decoder, and finally outputs the prediction Mel spectrum through a linear layer, so that multiple frames without overlap can be predicted during each decoding. And this article chooses Wavenet vocoder as the voice generator to improve the synthesized voice quality. The Mel spectrum predicted by Seq2Seq is input into the network, and the speech waveform is directly obtained by recovering the phase information in the signal by using the autoregressive characteristics of Wavenet vocoder.

3. Test Results and Discussions

3.1. Description of specimens

In order to test the effect of speech synthesis models and the quality of synthesized speech, all models are trained and evaluated on the English open source data set LJSpeech-1.0 corpus. The data set was broadcast by a professional female announcer and contained 13,100 English audio clips and corresponding text tags. The total audio length was 24 hours, with an average of 15 words per sentence and an average duration of about 6.6 s. In the experiment, 13,000 pieces were selected for training and 100 pieces for testing. All the training is done on one machine using Ubuntu and NVIDIA 1080 graphics cards. According to the comparison [14] (hereinafter referred to as baseline), the batch size is 32, and the parameters are adaptively updated by the adaptive gradient descent algorithm (Adam). The exponential attenuation rate estimated by the first-order matrix of the parameters is set to be 0.9, the exponential attenuation rate estimated by the second-order matrix is set to be 0.999, and the initial learning rate is set to be 0.003. In order to improve the training efficiency, the model converges as soon as possible and is increased to 0.005 after 500K steps of training. The fitting degree of the model is measured by the error loss function of model training. As it is shown in Figure 3, the training error of our model gradually decreases, indicating that the model fits well and converges quickly.

![Figure3 The Loss of Our Speech Synthesis Model](image)

3.2. Experimental Evaluation

In order to evaluate the performance of the model, objective evaluation and subjective evaluation are adopted for evaluation:

(1) Subjective Evaluation

ABX blind listening test was carried out on the synthesized speech. ABX test mainly evaluates the synthesis effect according to the similarity of synthesized speech, and draws lessons from the principle of speaker recognition. During the test, the evaluator listened to 3 segments of speech A, B and X respectively, and judged whether speech A or B was closer to X in terms of rhythmic characteristics of speech. Among them, X is the speech from a real person, and A and B are the speech synthesized based on baseline and the speech synthesized by the method proposed in this article, respectively. Finally, the judgment results of all the appraisers are counted to calculate the percentage of similarity between the sound and the target voice. 100 pieces of the test set are selected for speech synthesis, and 100 pieces of speech synthesized based on the method proposed in this paper and 100 pieces of baseline speech synthesized based on the best model established in [13] are generated. Each synthetic speech in the ABX test is judged by 10 people who are familiar with the language of the training data,
and then the ABX average is calculated. The scoring standard is "which spoken rhythm is closer to the reference pronunciation", and there are three options: (1) the pronunciation synthesized by the method proposed in this paper is closer to the target pronunciation; (2) neutral; (3) Baseline speech is closer to the target speech.

![ABX Test Results](image)

As can be seen from Figure 4, the model proposed in this article is superior to the baseline-based model. The text of the test set is used for speech synthesis. After ABX scoring, 39% of the speech synthesized by the synthesis method proposed in this paper is closer to the reference speech, while only 24% of the speech synthesized based on baseline is 15% lower than the speech synthesized by the method proposed in this article. It is proved that the method proposed in this paper can make the loss information of predicting Mel spectrum less, thus synthesizing better speech.

(2) Objective Evaluation

The objective evaluation uses Mel Cepstral Distortion (MCD) values as evaluation criteria, and Mel Cepstrum Distortion (MCD) is used to measure the difference between Mel Cepstrum of two sequences. It is used to evaluate the quality of the speech synthesis system. If the MCD between the speech synthesized by the speech synthesis system and the natural speech is smaller, the synthesized speech is closer to the natural speech. In the experiment, the text corresponding to 100 sentences of speech in the test set is synthesized by Baseline model and the method proposed in this paper, and then the MCD value of each sentence is calculated. Finally, the Baseline model and the MCD average value of the model trained by the method proposed in this paper are calculated. The results are shown in Table 1:

| Model      | MCD Mean Value |
|------------|----------------|
| Baseline   | 13.15          |
| Our Method | 12.77          |

As can be seen from the table, the average MCD value calculated by the model synthesis speech proposed in this paper is 12.77 db, while the average MCD value calculated by the baseline synthesis speech is 13.15 db. It is obvious from the MCD value that the model trained by the method proposed in this article has advantages. This shows that the predicted Mel spectrum obtained by enhancing rhythmic features on the basis of baseline tends to be closer to the real Mel spectrum, and the synthesized speech obtained by the predicted Mel spectrum through vocoder is closer to the target speech. To be more precise, in terms of the degree of signal loss, the model proposed in this article has a lower amount of information loss.

4. Conclusion

This article introduces a method based on World vocoder and VAE network to embed rhythmic features to optimize Tacotron2 model. Firstly, World vocoder is used to extract rhythmic features of speech signals, then VAE network is used to fuse the obtained high-dimensional rhythmic feature sequences, and finally low-dimensional feature information is obtained. The rhythmic information generated by Tacotron2 model embedded with rhythmic features to predict Mel spectrum is enhanced. Experiments show that based on Tacotron2 model, the rhythmic feature embedding module proposed in this paper can participate in guiding speech synthesis, which can make the decoder output less information to predict Mel spectrum loss, and the synthesized speech has advantages in subjective evaluation of ABX and objective evaluation of MCD value, which is better than baseline. In the later research, we will further consider how to further refine other potential feature information, and study
how to distinguish and fuse the feature information of many people, in order to establish a speech synthesis system with multi-speaker and multi-style conversion.

Acknowledgments
This work was supported by the Liaoning Natural Science Foundation Guidance Plan: Research on Enterprise Patent Litigation Forecast and Patent Value Mining Methods Based on Potential Semantic Analysis (2019-ZD-0569).

References
[1] Mohammadi, Seyed Hamidreza, et al. An overview of voice conversion systems[J]. Speech Communication, 2017, 88: 65-82(in Sweden).
[2] Cai Wenbin, Wei Yunlong, Xu Haihua, et al. Hybrid unit selection speech synthesis system target cost construction[J]. Computer Engineering and Applications, 2018, 54(24): 20-25(in America).
[3] Gonzalvo X, Tazari S, Chan C A, et al. Recent advances in google real-time HMM-driven unit selection synthesizer[C]. Interspeech, 2016: 2238-2242(in California).
[4] Takaki S, Yamagishi J. A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, 2016: 5535-5539(in Chinese).
[5] Wang Y, Skerry-ryan R J, Stanton D, et al. Tacotron: towards end-to-end speech synthesis[C]. Interspeech, 2017: 4006-4010. (in Stockholm)
[6] Ping W, Peng K, Chen J. Clarinet: parallel wave generation in end-to-end text-to-speech[J]. arXiv preprint arXiv:1807.07281, 2018.
[7] Chrzanowski M, Coates A, Diamos G, et al. Deep voice: real-time neural text-to-speech[C]. International Conference on Machine Learning, 2017: 195-204. (in Tuscany)
[8] Gibiansky A, Arik S, Diamos G, et al. Deep voice2: multi-speaker neural text-to-speech[C]. Advances in Neural Information Processing Systems, 2017: 2962-2970(in California).
[9] Sotelo J, Mehri S, Kumar K, et al. Char2wav: end-to-end speech synthesis[J]. International Conference on Learning Representations, 2017(in Toulon).
[10] Qiu Zeyu, Qu Dan, Zhang Lianhai. End-to-end speech synthesis based on WaveNet[J]. International Journal of Computer Applications, 2019, 39(05): 1325-1329(in Chinese).
[11] Shen J, Pan R, Weiss R.J, et al. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions[C].Conference on Acoustics, Speech, and Signal Processing. 2018: 4779-4783 (in Canada).
[12] Oord A, Dieleman S, Zen H, et al. WaveNet: a generative model for raw audio[C]. International Speech Communication Association Speech Synthesis Workshop, 2016: 125(in Sunnyvale).
[13] Zhang Y, Pan S, He L, et al. Learning latent representations for style control and transfer in end-to-end speech synthesis[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, 2019: 6945-6949(in Aberdeen).
[14] Morise M, Yokomori F, Ozawa K, et al. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications[J]. IEICE Transactions on Information and Systems, 2016, 99(7): 1877-1884(in Japan).
[15] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv: Machine Learning, 2013.
[16] Rezende D J, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models[J]. arXiv preprint arXiv:1401.4082, 2014.
[17] E. Shelhamer, J. Long, and T. Darrell, Fully convolutional networks for semantic segmentation[J]. IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440(in America).
[18] Ioffe S, Szegedy C. Batch Normalization: Accelerating deep network training by reducing internal covariate shift[C]. International Conference on Machine Learning, 2015: 448-456(in France).