Personalized End-to-End Mandarin Speech Synthesis using Small-sized Corpus

ChenHan Yuan[1], *Yi-Chin Huang*[2]

[1]Dept. of Computer Science, Virginia Tech, Blacksburg, VA 24060 USA
[2]Dept. of Computer Science, National Pingtung University, Pingtung, Taiwan
chris.yuan.ece@gmail.com, ychin.huang@gmail.com

Abstract

Conventionally, voice conversion techniques are based on the source-filter model, which extracts acoustic features and transforms the spectrum distribution from the source speaker to the target. Parallel corpora are usually required to learn the transformation and the alignment of phone units has to be done manually to obtain the optimal conversion. These requirements are hard to achieve in daily use. Therefore, we proposed an end-to-end method for personalized speech synthesis system by combining the ideas to tackle these problems and try to make the data collection task attainable.

We integrated the linguistic/acoustic feature extraction of the speech corpus by adopting suitable neural networks. In this way, the traditional linguistic feature extraction module which relies on the expert knowledge to build could be substituted. Then, for the personalized acoustic model, we adopted the variational auto-encoder, which focused on separating the speaker-related properties, such as timbre and speaker identity, from the underlying latent code, which assumed to be related to phoneme identity. Therefore, the requirement of manual alignment and parallel corpus could be overcome.

Finally, experimental results showed that the proposed system indeed useful for personalized speech synthesis and provides comparable performance with the conventional system while easier to build.

Index Terms: end-to-end speech synthesis, voice conversion, variational auto-encoder, word embedding

1. Introduction

The speech synthesis system consists of two main modules conventionally. One is the front-end text analysis module [1] and the other is the back-end vocoding module [2]. The front-end module usually is constructed by several sub-modules, such as text normalization, word segmentation, grapheme to phoneme, and part-of-speech tagging. After the text analysis, the linguistic representation of the input word sequence could be obtained and served as the model definition for the back-end vocoding module. The back-end vocoding module then extracts the acoustic features of the input speech and train the acoustic models based on the linguistic/acoustic features. Then, a speech synthesis system is constructed.

However, the problem of the front-end module is that there are several sub-modules and each of them could be not always perfect to predict the linguistic features. Therefore, there is inconsistent between training models and testing (e.g., synthesize speech). Besides, each of the sub-modules is not easily to build, which requires expert knowledge and quite patchy to construct.

Therefore, we attempt to integrate the front-end and back-end together as an unifying framework. In order to achieve this goal, we substituted the sub-modules of the front-end with several neural networks, such as a Seq2Seq model [3] for grapheme to phoneme mapping, a word2vec model [4] to capture the word-level class and attributes, and a RNN as the character-level language model [5]. The output vectors of these models will then be constructed as the character embeddings for the back-end acoustic model training. For the back-end vocoder, we adopt the wavenet vocoder [6, 7], which is also a neural network-based vocoder. By integrating the generated character embeddings for acoustic model training, the whole system should be capable to generate speech without a large amount of manually labeled corpus since the front-end modules such as Word2Vec and RNN-based language model are trained unsupervisedly.

For the personalized speech training, to alleviate the parallel corpora and phonemic alignment problem that usually required in the voice conversion training, the speaker-specific information and the underlying phone unit should be separated. This idea could be implemented by using the variational auto-encoder (VAE) [8, 9], which firstly encodes the phonemic representations which is speaker-independent. By combining the speaker-specific information and the underlying phonemic unit representation into the decoder process, the resultant speech should be able to perceived as uttered by the target speaker.

The organization of the paper is as follows. First, we describe the proposed method and related works in Sec. 2. Then, the performance of the proposed personalized speech synthesis system is evaluated via objective and subjective experiments, and the results are showed in Sec. 3. The concluding remarks are discussed in Sec. 4.

2. The Proposed End-to-End Personalized Synthesis System

The proposed method is focused on constructing a personalized speech synthesis system with a unified framework for Mandarin language. Fig. 1 shows the system framework for synthesizing the target speech with several sub-modules, which will be introduced in the section. The front-end block indicates the text analysis module and the back-end shows the vocoding process with the trained acoustic model and then the VAE decoding for transforming the spectrum features of the original acoustic feature to that of the target speaker using the codebook learnt from VAE training.

The text analysis of Mandarin is quite complicated since Mandarin sentences consist of sequences of characters and the grapheme-to-phoneme (G2P) process requires word segmentation done in advance in order to obtain correct G2P conversion. Therefore, we adopted several DNN frameworks that is suitable for different tasks and integrated them together for a unified

character embedding vector to cope with the linguistic analysis task. The details are depicted as below.

2.1. End-to-End Speech Synthesis System

The proposed Mandarin speech synthesis system, which tried to substitute the conventional front-end text-to-label module and back-end label-to-feature module, is designed for Mandarin tonal language. The front-end module is constructed by combining three sub-modules, which are listed as follows:

- character-to-phoneme module C_P
- word-class and attribute module C_{lab}, and
- character-level sequence module C_{txt}.

The first character-to-phoneme module is used to predict the phonemic unit for each character, which consists of its unit identity and tone information based on the corresponding lexical word, in the input sentence. Here, we adopted the Seq2Seq model, which is useful for processing sequential data, such as machine translation [10], voice conversion [11] and speech recognition [12]. Here, to cope with homophone problem in Mandarin, the nearby characters (3 preceding/succeeding characters) of the current character are combined as the input sequence for the model training, and the output is the phonemic unit with tone information for the current character.

The second module, which deals with the word-class and attribute, is aimed to substitute the conventional POS tagging [13]. Here, the Word2Vec network is adopted, and is quite useful for clustering word with similar syntactic structure. After training with input continuous bag of word (CBOW) of the nearby words, the word-level embedding for each character could be obtained.

The third module is the substitute module for the conventional language model [14] that is served as the contextual linguistic information for the input sentence. Here, we adopted the RNNLM toolkit [15] in our system. However, the input features are the sequences of characters instead of the words. Since the RNN is capable to memorize the sequential information of the input features, the contextual linguistic features could be captured and the values of the hidden neurons are used as the linguistic information vector of the corresponding character.

The resultant three character-level vectors are then combined together and a vector quantization [16] process is applied to integrate the information of the three sub-modules and served as the character embeddings for the corresponding units. Then, for the back-end acoustic model training, we followed the wavenet vocoder [6, 7] and trained a source acoustic model with the character embeddings, and could obtained a speech synthesis system of the source speaker. In order to transform the acoustic model of the source speaker to the target one, we investigated the powerful variational auto-encoder (VAE), which has been adopted in several research domains, such as speech recognition [17], emotional speech generation [18], and voice conversion [9] and so on. Therefore, we tried to incorporate the VAE for its capability for training voice conversion model without the requirement of manually phonemic unit alignment.

2.2. Variational Auto-Encoder

The idea of the auto-encoder is to encode the input features to a sequence of the latent codes, and then decodes them for re-generating the original input features. For the VAE training, it hypothesizes that the encoder only make the input acoustic features into code sequence that is speaker-independent, with an additional speaker-specific information is added to the AE model. When decoding the code sequence to its original acoustic features, the additional speaker-specific information is concatenated and decoded as eq. 1:

$$\hat{x}_n = \hat{f}\left(x_n, y_n\right) = f_d\left(z_n, y_n\right) = \hat{f}\left(f_e\left(x_n\right), y_n\right) \quad (1)$$

where $\hat{x}_n$ is the transformed acoustic features x_n of the source speaker with length of n, $\hat{f}$ represents the entire VAE process, y_n is the target speaker information, z_n is the latent code, and f_e, f_e are the encoding and decoding process, respectively. Note that z_n only carries the phonemic information while the y_n is aligned to the length of x_n frames.

During model training, the VAE firstly decodes the input speech features and then restore to the original features, while optimizing the speaker-specific information based on the user defined code. Therefore, when applying this process to the corpora of multi-speakers, we could obtain a set of speaker-independent encoder, decoder and the speaker representation of each speaker. During the conversion step, the encoder encodes the input speech sequence to the latent code z_n, and we could substitute the speaker representation to the one that we desire to generate, which is denoted as y_n and combine with z_n to get the decoded results. In this way, the VAE process do not require the parallel data of each speaker and since it learns the distributions of speaker and phonemic unit separately, the manual alignment is also not necessary. The effectiveness of the VAE process is investigated in the experiment section.

3. Experiments

For evaluating the performance of the proposed system, we firstly evaluate the effectiveness of substituting the front-end text analysis modules. Here, we adopted the Mel-Cepstrum Distance (MCD) as the objective metric to confirm whether the proposed method could reproduce the characteristics of the original speech. Then, for the subjective evaluation, we held listening tests to evaluate the speech quality and speaker similarity of the generated speech by the proposed system with different settings.

3.1. Experimental settings

For training the front-end sub modules, we adopted several public text corpora. For the training of C_p, the TCC300 corpusc̃itechiang2012study is adopted, which consists of around $286k$ words. For C_{lab} and C_{txt}, the open-source corpus of the Wikipedia [19] is used here, which consist of around $2,800k$ text files. We used open-sourced CKIP toolkit [20] to obtained the Mandarin word segmentation. For the model training, we adopted the Seq2Seq tool [21] for training C_p, word2Vec for C_lab, and RNNLM toolkit [15] for C_txt. By applying VQ, the resultant character embedding size is set to 256.

For training the acoustic model, we used the speech corpus collected from our lab, which consists of a $1,927$ utterances by a professional female speaker, is recorded in recording studio, the audio format is $48kHz$ and $16bits$. The content of the utterances includes news transcripts and child stories. In order to evaluate the effectiveness of the VAE, we also collected a small-sized parallel corpora, which consists of 72 utterance pairs and 4 speakers, to compare the training results.

For feature extraction, we used open-sourced Wavenet vocoder [6] to extract Mel-Cepstral feature, aperiod feature, fundamental frequency f_0, and voiced/unvoiced features. The

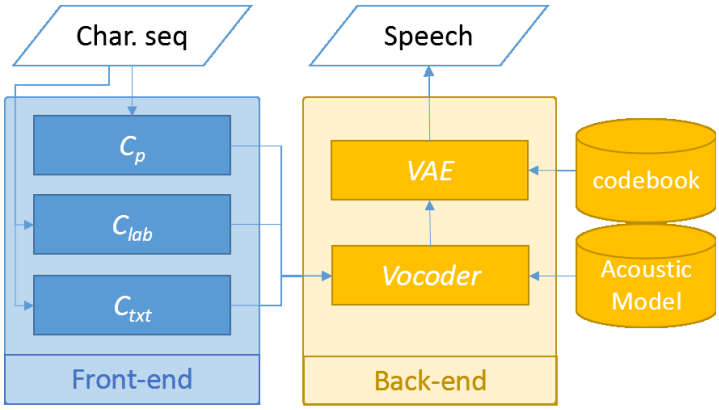

Figure 1: *The proposed system framework for synthesizing personalized speech.*

hyper parameter settings are based on the suggested settings in the tool.

3.2. Objective Evaluation

To evaluate the distortion between the generated speech and the synthesized speech, we compared the Mel-Cepstrum Distance (MCD) as follows:

$$MCD = \frac{10}{log10}\sqrt{2\sum_{m=1}^{M}\left(c_o\left(m\right)-c_s\left(m\right)\right)^2} \qquad (2)$$

where c_o and c_s is the cepstral feature of the original speech and the synthesized one, respectively. M is the number of dimension of the Mel-Cepstrum. Here, the baseline system is the conventional text analysis module that our lab constructed for Mandarin [22], which is also trained using the same set of corpus as the proposed neural network-based module. The back-end are the same for both system. The MCD of both systems compared to the original speech utterances and the results are shown in table. 1. Note that the inside test is held by using the same set of utterances for acoustic model training, while the outside test is by using a set of 72 utterances not included for training. The results showed that by unsupervised learning using NNs, the proposed method is capable to generated speech with similar distribution of distortion (no significant differences for both inside and outside tests). Therefore, the proposed unified framework is effective for substituting the conventional front-end modules.

Data	baseline	proposed
inside	4.7 ± 0.056	4.3 ± 0.053
outside	5.5 ± 0.047	5.6 ± 0.049

Table 1: *The averaged MCD between original speech and synthesized speech (in dB)*

After applying the VAE module for generating the speech utterances of the target speaker, it is interesting to evaluate whether the spectrum transformation could result in decreasing the cepstral features. Here, the VAE is trained between cepstral features extracted from the synthesized speech of the source speaker and the natural speech of the target speaker. We adopted the speech data of a male speaker as the target speaker, and compare the MCD between synthesized speech and original speech of the target speaker. Note that we only collected 72 utterances of the target speaker for VAE training, therefore only inside test is reported, which is 8.3 ± 0.072 before spectral transformation and 6.2 ± 0.65 after transformation. The results suggests that the VAE is indeed helpful for transforming the cepstrum features from source to target. However, MCD is not a perception metric for speech, therefore, we conducted subjective evaluations for further performance investigation.

3.3. Subjective Evaluation

For subjective subjective evaluation, we conducted two experiments to validate whether VAE training could be done without parallel corpora. First, we synthesized the same set of speech utterances as collected from target speaker to simulate the parallel data (denoted as VAE_{pseu}). The second set is randomly generated speech utterances that are different from the target speaker (denoted as VAE_{non}). We also have the third set of speech utterances that are the natural speech utterances of the source speaker and is parallel to that of the target speaker, which serves as the golden standard (denoted as VAE_{par}). We asked 10 native Mandarin speaker to evaluate the speech quality and speaker similarity. For the speech quality test, 5-point Mean Opinion Score (MOS) test is held, while the ABX preference test is conducted for evaluating the speaker similarity. The 20 testing utterances are all outside sentences, which are not in the training set nor parallel set.

The speech quality results are shown in Fig. 2. Note that "Source" and "Natural" indicates the synthesized and natural speech of the source speaker, respectively. The speech quality of the synthesized utterances is slightly inferior than the natural speech. However, the VAE module integration does not decrease speech quality significantly, since the speech quality is similar to the one that does not perform the VAE conversion.

The results of the speaker similarity is shown in Fig. 3. The gray bar indicates the no-preference between two comparative systems. The VAE-based methods all generated speech utterances perceived much similar to the target speaker compared to the source utterances. However, by investigating the results of the three VAE-based systems, there is no preference shown even though VAE_{par} achieved slightly better preferences. This result shows that the parallel corpora is not required for training the VAE model for significantly better voice conversion re-

sults. However, there is still some buzz sounds in the generated speech utterances, which could be caused by the vocoder settings. We could further conducts different experiments using other vocoders [2, 23].

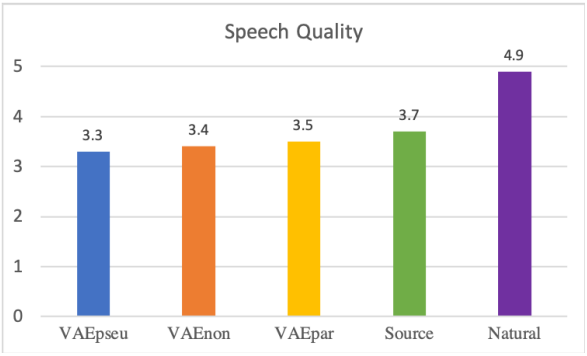

Figure 2: *The MOS test results of speech quality.*

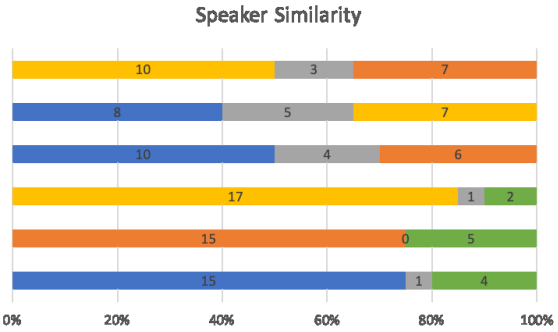

Figure 3: *The ABX test results of speaker similarity.*

4. Conclusions

In this paper, we have introduced a unified end-to-end personalized speech synthesis system for Mandarin language. By integrating several unsupervised neural networks to the front-end text analysis module, it showed the similar performance comparing with the conventional text analysis modules. By adding the VAE module, the parallel corpora is not required to generate the speech utterance of an arbitrary speaker. Future work will focus on improving the speech quality and also incorporate the linguistic information for VAE training to enhance the transformation results.

5. References

[1] E. Brill, "A simple rule-based part of speech tagger," in *Proceedings of the third conference on Applied natural language processing*. Association for Computational Linguistics, 1992, pp. 152–155.

[2] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[3] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[5] T. Mikolov, S. Kombrink, L. Burget, J. Černockỳ, and S. Khudanpur, "Extensions of recurrent neural network language model," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5528–5531.

[6] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder." in *Interspeech*, 2017, pp. 1118–1122.

[7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[8] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *arXiv preprint arXiv:1611.07308*, 2016.

[9] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[10] M.-T. Luong, E. Brevdo, and R. Zhao, "Neural machine translation (seq2seq) tutorial," *https://github. com/tensorflow/nmt*, 2017.

[11] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using sequence-to-sequence learning of context posterior probabilities," *arXiv preprint arXiv:1704.02360*, 2017.

[12] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4845–4849.

[13] X. Zheng, H. Chen, and T. Xu, "Deep learning for chinese word segmentation and pos tagging," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 647–657.

[14] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[15] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, and J. Cernocky, "Rnnlm-recurrent neural network language modeling toolkit," in *Proc. of the 2011 ASRU Workshop*, 2011, pp. 196–201.

[16] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Springer Science & Business Media, 2012, vol. 159.

[17] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 16–23.

[18] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," *arXiv preprint arXiv:1804.02135*, 2018.

[19] S. Reese, G. Boleda, M. Cuadros, and G. Rigau, "Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus," 2010.

[20] P.-H. Li, T.-J. Fu, and W.-Y. Ma, "Remedying bilstm-cnn deficiency in modeling cross-context for ner," *arXiv preprint arXiv:1908.11046*, 2019.

[21] "Cmu sphinx sequence-to-sequence g2p toolkit," https://github.com/cmusphinx/g2p-seq2seq.

[22] C.-C. Hsia, C.-H. Wu, and J.-Y. Wu, "Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in hmm-based speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 1994 –2003, nov. 2010.

[23] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.