ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414427

PROSODY AND VOICE FACTORIZATION FOR FEW-SHOT SPEAKER ADAPTATION IN THE CHALLENGE M2VOC 2021

Tao Wang[1,2], *Ruibo Fu*[1,2], *Jiangyan Yi*[1], *Jianhua Tao*[1,2,3], *Zhengqi Wen*[1], *Chunyu Qiang*[1,2], *Shiming Wang*[1]

[1]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{tao.wang, ruibo.fu, jiangyan.yi, jhtao, zqwen, chunyu.qiang, shiming.wang}@nlpr.ia.ac.cn

ABSTRACT

The paper describes the CASIA speech synthesis system entry for challenge M2VoC 2021. The low similarity and naturalness of synthesized speech remains a challenging problem for speaker adaptation with few resources. Since the end-to-end acoustic model is too complex to interpret, overfitting will occur when training with few data. To prevent the model from overfitting, this paper proposes a novel speaker adaptation framework that decomposes the prosody and voice characteristics in the end-to-end model. A prosody control attention is proposed to control the phonemes' duration of different speakers. To make the attention controlled by the prosody information, a set of phoneme-level transition tokens is auto-learned from the prosody encoder in our framework and these transition tokens can determine the duration of phonemes in the attention mechanism. Secondly, when we need to use small data set for speaker adaptation, we just need to adapt the speaker related prosody model and decoder, which can prevent the model from overfitting. Further, we use a data puring model to automatically optimize the quality of datasets. Experiments demonstrate the effectiveness of speaker adaptation based on our method, and we (team identifier is **T03**) get the top three results in competition M2VoC by using this framework.

Index Terms— speech synthesis, end-to-end model, prosody and voice factorization, few-shot speaker adaptation, the M2VoC challenge

1. INTRODUCTION

With the development of deep neural networks (DNN) [1], a significant amout of efforts has been made to improve the naturalness of text-to-speech (TTS), such as Tacotron [2–4], wavenet [5]. Apart from the naturalness, people also expect a TTS system to be able to synthesize the voice of any speaker with few training data. To respond to this problem, many speaker adaptation methods are proposed and get effective results on several benchmark datasets [6–10]. The primary idea is to use a small amount of corpus to fine tune the model parameters. However, when there are fewer data, such as several of sentences, speaker adaptation remains a change.

The main challenge faced by few-shot speaker adaptation is that it is easy to overfit [11] when the dataset for training is too limited. Especially for the end-to-end acoustic models, such as Tacotron [2, 3], due to the complexity of the model structure, the encoder, the model parameters are larger, the loss function can be optimized to a very low level and the model will ignore the inherent laws of text pronunciation. To prevent the model from overfitting, there have been quite a few studies in machine learning [11]. A common method is to select a small hypothesis space for fine tuning which has small complexity, thus requiring fewer samples. Many researchers have tried to make the acoustic model to be more interpretable, so as to adapt new speaker with small hypothesis space [6–8]. However, due to the parameters of the encoder, decoder and attention mechanism are coupled with each other, it is difficult to find the hypothesis space that specifically controls the speaker's pronunciation characteristics. Therefore the factorization of the acoustic model is the key to make the model more interpretable. From a linguistic point of view, the prosody and voice are two dominant factors in TTS synthesis. Factorizing the prosody and voice may make the model more interpretable. Inspired by the forward attention [12] that can control the speed of speech, we can intergrate prosody control into the attention mechanism and factorizate the prosody and voice characteristics.

In this paper, we propose a prosody and voice factorization framework for few-shot speaker adaptation task. Firstly, we use a data puring model to automatic select the unmatching data pairs, which improves the efficiency of manual checking process. Secondly, a prosody control attention is proposed to control the phonemes' duration of different speakers. To make the attention mechanism controlled by the prosody information, a set of phoneme-level transition tokens is learned from the prosody encoder in our framework and these tokens determine the probability of each phoneme's transition in the attention mechanism. Thirdly, when we need

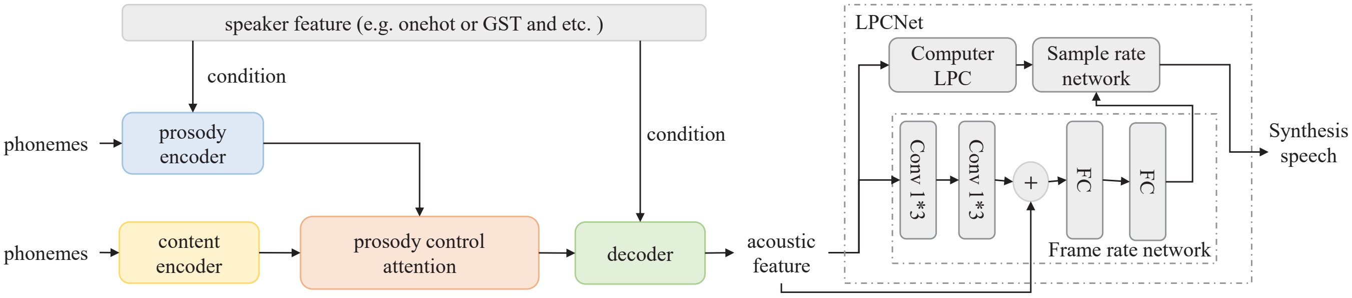

Fig. 1. An overview of our system. The acoustic model is based on the encoder and decoder model, and LPCNet is used as the vocoder.

to use a small amount of corpus to adapt new speakers, we only need to adapt the speaker related prosody model and decoder, which can prevent over fitting effectively. In summary, the main contributions are as follows:

- We use a data puring model to automatically select the unmatching data pairs, which improves the efficiency of manual checking process and shorts the time on the dataset building.
- A prosody control attention is proposed to control the duration of each phoneme of different speakers by using a set of phoneme-level transition token learned from prosody encoder.
- When facing speaker adaptation with few data, we only need to fine tune the prosody encoder and the decoder with smaller parameters, which can prevent the model from over fitting effectively.

2. PROPOSED METHOD

When the dataset is small, such as only several sentences, it is easy to overfit by adapting on the end-to-end acoustic model directly. In this section, we will introduce the prosody and voice factorization framework with prosody control attention for few-shot speaker adaptation. Firstly, we will give an introduction of the dataset processing.

2.1. Data selection and auto-labeling

We use an ASR model and silence detection model to automatic segmentation of the provided data. But there are still a lot of mistakes in the corpus. We use the force alignment by the ASR technology to further check the matching between the audio and the text. Furthermore, we also use a trained model to eliminate the unmatching audio-text data pairs. The model can find the mistakes of the text more thoroughly than the force alignment by ASR. Since some of the tracks (track 1b & 2b) are allowed to use additional open source data, we have added some additional open source data to improve the performance of our system. All the data are processed by the same technologies and checked by annotators.

The provided data contains some utterances that are too loud or too small. So we train a model by acoustic features to recognize these over expressive utterances. Experiments show that the system will be more stable after deleting some outlier utterances. In order to make sure that we really delete those over expressive utterances, we manually listen those deleted data and correct some mistakes. Then we use these manually checked data that need to be deleted as the negative samples to retrain the recognition model.

2.2. The factorization of prosody and voice

There are mainly two aspects of speaker characteristics. One is the prosody. For example, different people read the same text in different ways. The phonemes' durations of different speakers are different, which can reflect many characteristics of prosody, such as stress, speaking speed and so on. The other aspect is the voice of the speaker. This difference is mainly due to the different vocal organs of each person. To decompose the parameters of prosody part and voice part in the end-to-end acoustic model, our proposed framework is shown in the Fig. 1. The whole framework consists of four parts: the content encoder is mainly responsible for encoding text information into hidden features with contextual information. The prosody encoder learns phoneme-level transition tokens for each phoneme according to the input phonemes and speaker features automatically. The structure is shown in the Fig. 2. The phoneme-level transition tokens will determine the probability of transition of phonemes based on forward attention [12]. Suppose that in the process of attention decoding, the probability of the n-th phoneme transferring to the next is q_n, and the q_n is only determined by the features of phonemes and speakers. We can use the q_n as an indicator which describes the probability that the phoneme should move forward to the next one or keeping unmoved, which likes the transition agent in forward attention. Tacotron is used with the speaker features. It is worth noting that the speaker features here only need to be a vector

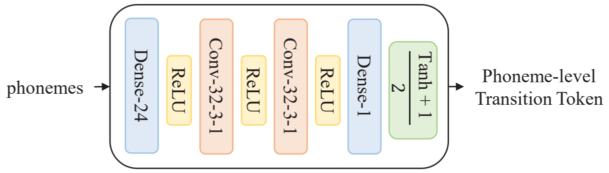

Fig. 2. The structure of prosody encoder.

with a fixed dimension, such as one-hot coding, global style token (GST) [13], ivector [14], dvector [15] and so on. The attention connects the two encoder modules and the decoder module, and controls the duration of each phoneme according to the phoneme-level transition tokens. The decoder generates acoustic features, which can be restored to speech by the vocoder. The prosody encoder's framework is shown in Fig. 2, and the purpose of the last layer is to make the output value between [0,1].

In the neural vocoder, we deploy the LPCNet [16], which significantly improve the efficiency of speech synthesis and remain high quality. In the frame rate network of LPCNet, we combine the trainable speaker embeddings from Tacotron with the output of convolution layer and the acoustic features that predicted by Tacotron .

2.3. Prosody and voice style adaptation

When we need to use a small amount of corpus for speaker adaptation, we only need to transfer the parameters related to speaker characteristics. As shown in the Fig. 1, since the parameters in the content encoder and attention mechanism are not related to speaker characteristics, these parameters have universal ability after we use the multi speaker dataset to pretrain the model parameters.

When we do speaker adaptation, we only need to fine tune the prosody encoder and the decoder module, which is shown in Fig. 3. Because the network structure of prosody encoder is composed of several convolutional networks, the parameters are small and the model is not easy to overfit, which can improve the stability of the fine-tuned model.

3. EXPERIMENTS

To verify the validity of the method, we conduct experiments to evaluate our proposed method. The multi speaker training datasets are AIShell-3 [17] and MST-Originbeat for track1a and track2a. Based on this, we add extended data (Data Baker [18], and DiDiSpeech [19]) to participate in the track1b and tracks2b. The testing set is divided into two cases, one is given three target speakers with different speech styles (game, chat, story), and each speaker has 100 speech samples, the other is given three target speakers with different speech styles (game, chat, story), each speaker has 5 speech samples. All the target speakers are not seen in the training dataset. Through these two testing sets, we can evaluate the ability of speaker adaptation in low resource and extremely low resource situations of our proposed model. All the wav files are sampled at 24KHz.

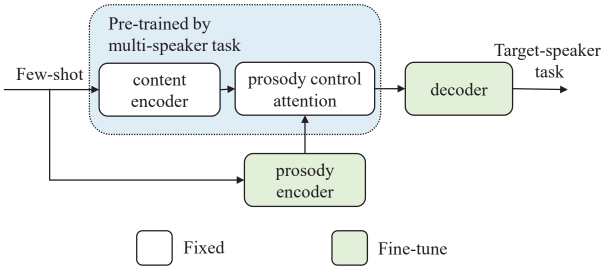

Fig. 3. The method of adapting the prosody and voice style.

3.1. Model details

The speaker feature which conditions the prosody encoder and the decoder is used as global style tokens (GST). The structures of the content encoder and the decoder module are the same as the encoder and decoder in tacotron2 [3]. Acoustic features are extracted with 10 ms window shift. LPCNet [16] is utilized to extract 32-dimension acoustic features. To evaluate the effect of prosody control attention and the ability of speaker adaptation based on few data, the following systems are established for comparison.

- **FA** forward attention and without prosody encoder. When adapting new target speaker, we only fine tune the decoder module which is conditioned by the speaker features.
- **PCA** our proposed framework with prosody control attention, when adapting new target speaker, all parameters of the model are fine tuned.
- **Proposed** our proposed framework with prosody control attention, when adapting new target speaker, only adapt the prosody encoder and decoder modules.
- For vocoder, the LPCNet vocoder is denoted as **L**, and the world vocoder [20] is denoted as **W**. For example, **FA+W** is expressed as forward attention based acoustic model plus world vocoder.

3.2. Evaluations for quality with different systems

Firstly, we compare different vocoders and different acoustic features with multi-speaker speech synthesis. Eight Chinese native speakers participate in the evaluations. In each experimental group, 20 parallel sentences are selected randomly from testing sets of each system. By observing the scores of subjective evaluations in the Table 1, it can be found that the LPCNet can significantly improve the quality of synthetic speech. Besides, since the dimension of acoustic

Table 1. Average perference score (%) on speech quality among different acoustic models and different vocoders, where N/P stands for "no perference". The p-values <0.01.

System A	Scores A(%)	N/P Neural(%)	Scores B(%)	System B
FA+W	16.73	13.01	70.26	**FA+L**
FA+W	28.65	43.64	27.71	**Proposed+W**
FA+L	25.64	47.64	26.72	**Proposed+L**
Proposed+W	12.57	8.83	78.60	**Proposed+L**

Table 2. Average perference score (%) on speech similar among different systems, where N/P stands for "no perference". The p-values <0.01.

	FA	PCA	Proposed	N/P
FA vs **PCA**	38.52	54.32	–	7.16
FA vs **Proposed**	31.47	–	58.83	9.7
PCA vs **Proposed**	–	25.64	36.52	37.84

features for LPCNet is 32, while the dimension of acoustic features for WORLD is 187. The decreasing of dimension of predicted acoustic features can reduce the complexity of Tacotron model. Therefore, in the following sections we mainly concentrate on LPCNet vocoder.

3.3. Evaluations for few-shot speaker adaptation

We also conduct an ABX subjective perference test for similarity of speech. The results are listed in Table 2. In each experimental group, 20 parallel sentences are selected randomly from testing sets of each system. It is obvious that speaker adaptation based on our framework can achieve better similarity, this is because the prosody control attention can decompose prosody information from voice information. By comparing **PCA** and **proposed** system, we can find that only adapting the prosody encoder and decoder can achieve better results, because fine tuning on less parameter can effectively prevent model from over fitting.

3.4. Evaluations in the challenge M2VoC

The method in this paper is used to participate in the 2021 multi-speaker multi-style voice cloning (M2VoC) challenge [21]. The M2VoC challenge aims to provide a common sizable dataset as well as a fair testbed for benchmarking the voice cloning task. Using the provided dataset, we took part in the track1a (100 sentences and 18 teams, only using provided datasets), track1b (100 sentences and 22 teams, can use other open source datasets, track2a (5 sentences and 17 teams, only using provided datasets, and track2b (5 sentences and 19 teams, can use other open source datasets) respectively.

The results are shown in Table 3 and Table 4. Our team identifier is **T03**. It can be found that we have achieved the top three results in each track on speech quality, style and speaker similarity, ranking at the leading level, which can also show the effectiveness of our proposed model.

Table 3. The final score (average MOS score on speech quality, style and speaker similarity) of challenge M2VoC by using the data provided by the competition only (track 1a and track 2a). Our team is shown in **bold**.

	First	Second	Third	Fourth
100 sentences (track1a)				
SpeakerSimilarity	4.2484	**4.1455**	4.1370	3.8832
SpeechQuality	4.2373	**4.1741**	4.0623	4.0214
StyleSimilarity	4.1488	4.1212	**3.9348**	3.8027
5 sentences (track2a)				
SpeechQuality	**4.0905**	3.9568	3.8941	3.8768
SpeakerSimilarity	3.2250	3.2205	**3.2168**	3.1368

Table 4. The final score (average MOS score on speech quality, style and speaker similarity) of challenge M2VoC by using the data provided by the competition and other open source datasets (track 1b and track 2b). Our team is shown in **bold**.

	First	Second	Third	Fourth
100 sentences (track1b)				
SpeakerSimilarity	4.2466	4.1427	**4.1027**	4.0305
SpeechQuality	4.3132	4.3005	**4.2636**	4.2486
StyleSimilarity	4.1056	4.0624	3.9574	3.9438
5 sentences (track2b)				
SpeechQuality	4.1650	**4.1086**	3.8718	3.4818
SpeakerSimilarity	3.2409	**3.1964**	3.1445	3.0255

4. CONCLUSION

In this paper, the multi-speaker end-to-end speech synthesis system built for the challenge M2VoC is introduced. There are several improvements based on the end-to-end speech synthesis. The first one is the use of data puring model. Secondly, we present an end-to-end TTS model with prosody control attention mechanism for prosody and voice factorization, aiming to improve the similarity of speaker adaptation based on few data. A set of auto-learned phoneme-level transition tokens are learned from prosody encoder to help the attention mechanism to control prosody information. Thirdly, when adapting speaker with small dataset, only the prosody encoder and the decoder is needed to fine tune. Experiments demonstrate that the effectiveness of our proposed model, and we achieve the top three results in the challenge M2VoC. In the future, we will continue to explore the ways to decompose prosody and voice characteristics and the strategy of speaker adaptation.

5. ACKNOWLEDGEMENTS

This work is supported by the National Key Research and Development Plan of China (No.2020AAA0140002), the National Natural Science Foundation of China (NSFC) (No.61831022, No.61901473, No.61771472, No.61773379), a Inria-CAS Joint Research Project (No. 173211KYSB 20190049) and the CCF-Tencent Open Research Fund.

6. REFERENCES

[1] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio, *Deep learning*, vol. 1, MIT press Cambridge, 2016.

[2] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[3] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj 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.

[4] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, and Ming Zhou, "Close to human quality TTS with transformer," *CoRR*, vol. abs/1809.08895, 2018.

[5] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[6] Zhizheng Wu, Pawel Swietojanski, Christophe Veaux, Steve Renals, and Simon King, "A study of speaker adaptation for dnn-based speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] Yi Zhao, Daisuke Saito, and Nobuaki Minematsu, "Speaker representations for speaker adaptation in multiple speakers blstm-rnn-based speech synthesis," *space*, vol. 5, no. 6, pp. 7, 2016.

[8] Pawel Swietojanski and Steve Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 171–176.

[9] Yuchen Fan, Yao Qian, Frank K Soong, and Lei He, "Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4475–4479.

[10] Quanjie Yu, Peng Liu, Zhiyong Wu, Shiyin K Ang, Helen Meng, and Lianhong Cai, "Learning cross-lingual information with multilingual blstm for speech synthesis of low-resource languages," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5545–5549.

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] J. Zhang, Z. Ling, and L. Dai, "Forward attention in sequence- to-sequence acoustic modeling for speech synthesis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4789–4793.

[13] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[14] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19 – 41, 2000.

[15] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.

[16] Jean-Marc Valin and Jan Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[17] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," *arXiv preprint arXiv:2010.11567*, 2020.

[18] "https://www.data-baker.com/open_source.html," .

[19] Tingwei Guo, Cheng Wen, Dongwei Jiang, Ne Luo, Ruixiong Zhang, Shuaijiang Zhao, Wubo Li, Cheng Gong, Wei Zou, Kun Han, and Xiangang Li, "Didispeech: A large scale mandarin speech corpus," 2020.

[20] Masanori Morise, Fumiya Yokomori, and Kenji 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.

[21] Qicong Xie, Xiaohai Tian, Guanghou Liu, Kun Song, Lei Xie, Zhiyong Wu, Hai Li, Song Shi, Haizhou Li, Fen Hong, Hui Bu, and Xin Xu, "The multi-speaker multi-style voice cloning challenge 2021," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.