IMPROVE FEW-SHOT VOICE CLONING USING MULTI-MODAL LEARNING

Haitong Zhang, Yue Lin

NetEase Games AI Lab, China
{zhanghaitong01, gzlinyue}@corp.netease.com

ABSTRACT

Recently, few-shot voice cloning has achieved a significant improvement. However, most models for few-shot voice cloning are single-modal, and multi-modal few-shot voice cloning has been understudied. In this paper, we propose to use multi-modal learning to improve the few-shot voice cloning performance. Inspired by the recent works on unsupervised speech representation, the proposed multi-modal system is built by extending Tacotron2 with an unsupervised speech representation module. We evaluate our proposed system in two few-shot voice cloning scenarios, namely few-shot text-to-speech (TTS) and voice conversion (VC). Experimental results demonstrate that the proposed multi-modal learning can significantly improve the few-shot voice cloning performance over their counterpart single-modal systems.

Index Terms: few-shot voice cloning, text-to-speech, voice conversion, multi-modal

1. INTRODUCTION

Recently, text-to-speech has witnessed a great improvement due to the development of the sequence-to-sequence models [1, 2] and high-fidelity neural vocoders [3, 4]. Meanwhile, voice conversion has gained rapid development using various techniques [5].

With the rapid development of TTS and VC, few-shot voice cloning, namely cloning new voices with a small amount of data, has become an active research topic. For few-shot TTS, previous works have shown that fine-tuning either parts of the multi-speaker initial model or the whole model can provide high-quality results [6–8]. Another approach is to train a speaker-adaptive model conditioned on a speaker embedding extracted from a pre-trained speaker recognition model [9, 10]. This approach is useful when a few seconds of data is available and requires no fine-tune process. However, this approach has a drawback that the speaker similarity stops improving as more data is available [7]. Meanwhile, there are some works on building a voice conversion system for target speakers using a limited amount of data [5].

Although few-shot voice cloning has achieved a significant improvement, most systems for few-shot voice cloning are single-modal, and multi-modal few-shot voice cloning has been understudied. There are only a few previous works on multi-modal voice cloning. [11–13] proposed to receive multi-modal inputs using separate encoders so that the system can tackle TTS and VC simultaneously. A random masker or XOR training operator or the KL divergence between two encoder outputs is used to train the model for two tasks simultaneously. However, as found in [11, 12], the TTS performance deteriorates while the VC results get better, indicating multi-modal learning for few-shot voice cloning is challenging while it may be helpful.

Inspired by previous works on unsupervised speech representation [14–16], this paper extends Tacotron2 (a state-of-the-art TTS system) with an unsupervised speech representation module to achieve multi-modal learning. To our best knowledge, this is the first work to leverage unsupervised speech representation to achieve multi-modal voice cloning. In addition, experimental results demonstrate that the proposed multi-modal system can significantly improve the few-shot voice cloning performance over its counterpart single-modal systems.

2. THE PROPOSED SYSTEM

2.1. Motivation

As found in [14], the relationship between the unsupervised quantized speech representations matches the relationship between phonemes, indicating that the unsupervised quantized speech representations are similar to phonemes. Similarly, [15] found that the quantized speech representations learned by VQ-VAE are consistently associated with phonemes. Thus, the quantized speech representations have been used in text-to-speech, voice conversion, and automatic speech recognition [14–17]. Inspired by the previous works, this paper aims to improve few-shot voice cloning by extending the single-modal TTS model Tacotron2 with an unsupervised quantized speech representation module.

The proposed system mainly consists of two modules, which is illustrated as Figure 1. The first module is the unsupervised quantized speech representation module, which is used to extract unsupervised linguistic units given speech. The second module includes a Tacotron2-like module to generate mel-spectrogram conditioned on either unsupervised

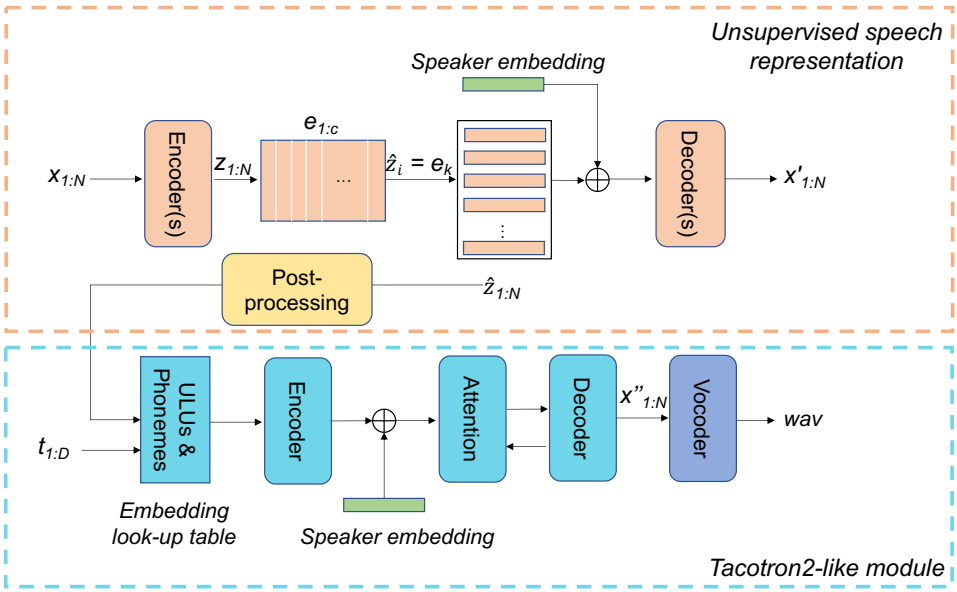


Fig. 1. The proposed model structure.

linguistic units or phonemes and a neural vocoder to reconstruct waveform based on predicted mel-spectrogram. We call the proposed multi-modal system MD-Tacotron in this paper.

2.2. Unsupervised speech representation Module

In this paper, we use VQ-VAE model [18] as the extractor of discretized linguistic units due to its promising performance [18]. VQ-VAE is an auto-encoder model with a codebook dictionary $e = C * D$, where C refers to the number of embeddings, and D is the size of each embedding. The model also contains an encoder and decoder. The encoder takes speech features $x_{1:N}$ as inputs, and produces higher-level hidden representation $z_{1:N}$ before mapping into discretized embedding $\hat{z}_{1:N}$ in the dictionary by finding the nearest one as $\hat{z}_i = e_k$, where $k = argmin_j \ ||z_i - e_j||$, and $j \in 1, 2, \cdots, C$. Finally, the discretized embedding $\hat{z}_{1:N}$ and the speaker embedding s are concatenated and passed into the decoder to reconstruct speech feature $x^{'}_{1:N}$. We use straight-through gradient estimation to approximate the gradients from the argmin operation [18]. The loss function of the model is

$$\begin{aligned} L &= ||x_{1:N} - x^{'}_{1:N}||_2^2 \\ &+ ||sg(z(x)) - e_j)||_2^2 \\ &+ \beta * ||z(x) - sg(e_j)||_2^2 \end{aligned} \tag{1}$$

where the first term is L2 loss for reconstructing the speech feature. The second term updates the codebook dictionary, with sg denotes stop-gradient operation. The third term, as the commitment loss, encourages the encoder output z to get close to the codebook embeddings, and the hyper-parameter β balances the loss terms.

2.3. Sequence-to-sequence module

The second module is a seq-to-seq model to generate mel-spectrogram and a neural vocoder to reconstruct the waveform conditioned on the predicted mel-spectrogram. We use Tacotron2 as our seq-to-seq module framework due to its outstanding performance in TTS [2]. To support multi-speaker training, we concatenated the encoder output with the speaker embedding extracted from the speaker embedding loop-up table. We also replaced the original attention with GMM attention for robust sequence modeling. In original Tacotron2, the model usually receives text representation as inputs, such as phonemes. In the proposed system, the Tacotron2-like module takes either unsupervised linguistic units (extracted from the VQ-VAE module) or phonemes as inputs. Unsupervised linguistic units (ULUs) and phonemes are embedded using an embedding lookup table. Specifically, the embedding lookup table is a $(N_{ULUs} + N_{Phone}) * N_{Dim}$ matrix, where N_{ULUs}, N_{Phone}, and N_{Dim} refer to the number of unsupervised linguistic units, phonemes, and the embedding dimension, respectively. The module architecture is illustrated in the bottom part of Figure 1.

We use Parallel WaveNet to generate waveform conditioned on mel-spectrogram. Inspired by [19], we use a single Gaussian as the output distribution. We train a universal neural vocoder for all experiments in this paper.

Table 1. Results of objective and subjective evaluations for the few-shot TTS scenario.

Model	MCD	WER	NAT	SIM
GT	-	6.17	$4.30 \pm .06$	$4.09 \pm .09$
Tacotron	6.41	8.03	$3.80 \pm .07$	$\mathbf{3.66 \pm .09}$
MD-Tacotron	**5.95**	**6.93**	$\mathbf{3.90 \pm .07}$	$3.64 \pm .08$

2.4. Training and inference

Although the VQ-VAE and seq-to-seq module can be trained in an end-to-end mode, we use a two-stage training mode in this paper and left joint training for future investigation. In the first stage, we train the VQ-VAE module using the training data and use the trained VQ-VAE module to extract unsupervised linguistic units for the whole training data. We then apply the post-processing method to remove the consecutively repetitive unsupervised linguistic units as in [16] to get a <unsupervised linguistic units (ULUs), audio> paired data. We combine this dataset with the original <phonemes, audio> paired data into the final dataset. In the second stage, we train the seq-to-seq module using this final dataset. At each training step, a batch of <ULUs, audio> and <phonemes, audio> is randomly sampled from the final dataset.

During inference, for VC, we first extract the unsupervised linguistic units for the input utterance using trained VQ-VAE module, and apply the post-processing method, then take the unsupervised units as the inputs of the Tacotron2-like module to generate waveform; for TTS, we use phonemes as the inputs of the seq-to-seq module to synthesize waveform.

3. EXPERIMENTAL SETUP

3.1. Datasets and Training setup

We evaluate the proposed method on voice cloning (including TTS and VC) using only 2-minute data from the target speakers. We implement experiments using VCTK data [20], which contains recordings from 108 speakers. We re-sample all waveforms into 16kHz.

Following [16], we set the number of embedding C and the size of each embedding D in the VQ-VAE module into 256 and 64, respectively. The encoder contains 3 convolution layers with a kernel size of 5 and channel size of 512, and 2 BLSTM layers with a hidden size of 256. The decoder includes 3 BLSTM layers with a hidden size is 256. We used the Adam algorithm for optimization. The initial learning rate is 1e-3, and we started decaying the learning rate from the 10k step with the decay step and rate is 15k and 0.5, respectively. We set β into 0.25 as in [18]. We set the output-per-step in the seq-to-seq module into 1. We strongly encourage readers to refer to [2] for other training setups.

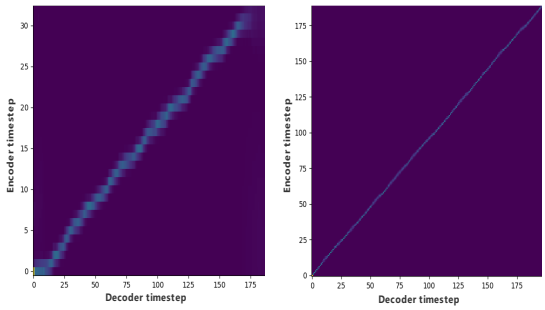


Fig. 2. An attention alignment example, the left sub-figure comes from Tacontron2 which uses phonemes as the model's inputs, while the right one comes from MD-Tacotron while using unsupervised linguistic units as inputs.

3.2. Evaluation metrics

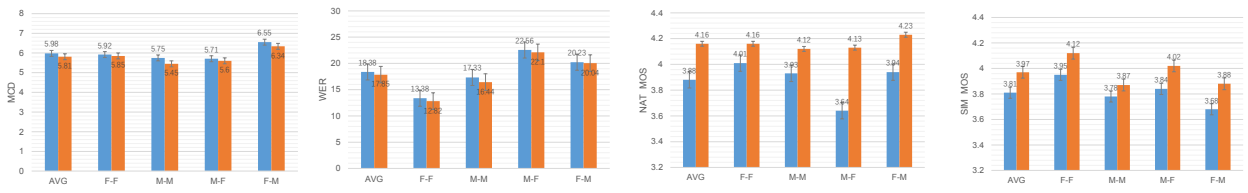
We carry out both objective and subjective evaluations. For the objective evaluations, we compute the dynamic-time-warping (DTW) mel cepstrum distortion (MCD) [21], a commonly used measure of spectral distortion in TTS and VC, and the word error rate (WER), which reflects the intelligibility of the synthesized or converted speech. For the subjective evaluation, we design listening tests for evaluating the naturalness (NAT) and speaker similarity (SIM) of the synthesized speech. We evaluate the proposed system in two scenarios. In the few-shot TTS scenario, 20 utterances are randomly selected for evaluations. In the few-shot VC scenario, 10 utterances from each source speaker are randomly selected for evaluation. Each utterance is rated by 15 listeners using the mean opinion score (MOS) on a five-point scale. Speech demos are available at https://haitongzhang.github.io/Orchestra/ .

4. RESULTS AND DISCUSSION

4.1. Few-shot TTS

We first investigate the effectiveness of our proposed system in the few-shot TTS scenario, where we use speakers p225 and p226 as the target speakers. In this scenario, the evaluation data includes 100 utterances from each target speaker. Since the proposed method takes Tacotron2 as the seq-to-seq module, it is natural to compare it with the Tacotron2 model (the single-modal counterpart) in the few-shot TTS scenario. Thus, the models investigated in this scenario include:

- Tacotron: Tacotron2 pre-trained using $<$ phonemes, audio$>$ paired data from 106 speakers (excluding p225 and p226), then fine-tuned using 2-minute $<$ phonemes, audio$>$ paired data from the target speakers (namely p225 and p226);



(a) (b) (c) (d)

Fig. 3. Results of the objective and subjective evaluations for few-shot any-to-one voice conversion; (a) DTW-MCD results; (b) WER results; (c) Naturalness MOS; (d) Speaker similarity MOS. For DTW-MCD and WER, the lower, the better; for naturalness and speaker similarity MOS, the higher, the better. Blue denotes the VQ-VAE model and orange denotes MD-tacotron.

- MD-Tacotron: The proposed system pre-trained using < phonemes, audio> and < ULUs, audio> paired data from 106 speakers (excluding p225 and p226), then fine-tuned using 2-minute < phonemes, audio> and < ULUs, audio> paired data from the target speakers (namely p225 and p226);

To prevent overfitting, we fixed the encoder when fine-tuning the models [7]. The results are given in Table 1. As shown in Table 1, the proposed system out-performs Tacotron2 in terms of MCD, WER, and naturalness and achieves a comparable performance to Tacotron2 on speaker similarity.

We attribute the results to two possible reasons. (1) Since ULUs are more fine-grained than phonemes (see Figure 2), the attention mechanism could possibly learn more robust alignment when it takes both fine-grained ULUs and coarse phoneme sequence as inputs. This speculation is confirmed by MD-Tacotron producing less typical alignment errors (i.e., skipping and repeating) than tacotron2 since MD-Tacotron achieves a smaller number of insertion and deletion (while computing the WER) than Tacotron2. (2) We can regard using the unsupervised linguistic units as a way of data augmentation on the textual representation side, which leads to a performance improvement when a limited amount of data is available.

4.2. Few-shot voice conversion

In this scenario, we study the performance of the proposed method on the few-shot VC task. We set p225 and p226 as the target speakers and p227 and p228 as the source speakers. Since the proposed method takes VQ-VAE as the extractor of unsupervised linguistic units, it is natural to compare with VQ-VAE Model (the single-modal counterpart) in the few-shot VC scenario. The models investigated in this scenario include:

- VQ-VAE: The single-modal VQ-VAE model pre-trained with speech data from 104 speakers (excluding the source and target speakers), then fine-tuned using 2-minute speech data from the target speakers.
- MD-Tacotron: The proposed multi-modal system pre-trained with speech data from 104 speakers (excluding the source and target speakers); then only the seq-to-seq module fine-tuned using 2-minute < ULUs, audio> paired data from the target speakers;

The results are provided in Fig. 3. From the objective evaluations (see Fig. 3 (a) and (b)), we found that the proposed multi-modal system outperforms the single-modal VQ-VAE VC model in MCD and WER. While computing the WER of synthesized speech, we found that the number of substitutions of the proposed system is smaller than that of the VQ-VAE model, which indicates that MD-Tacotron generates more intelligible speech than the VQ-VAE model.

From the subjective evaluations (see Fig. 3 (c) and (d)), we found that the proposed system significantly outperforms the baseline VQ-VAE model in naturalness and similarity, which indicates multi-modal learning can improve the few-shot VC performance by a large margin. In addition, we found that the proposed system provides a stable performance on both inter-gender and intra-gender conversion, while VQ-VAE's inter-gender performance is significantly worse than the intra-gender performance.

5. CONCLUSION

This paper proposes a multi-modal voice cloning system by extending Tacotron2 with an unsupervised speech representation module. We verify the effectiveness of the proposed system by comparing it with its single-modal counterparts in both few-shot TTS and VC scenarios. Experimental results reveal that (1) in few-shot TTS, the proposed multi-modal system out-performs its single-modal counterpart (Tacotron2) on MCD, WER, and naturalness, and achieves a comparable performance to Tacotron2 on speaker similarity; (2) in few-shot VC, the proposed multi-modal system out-performs its single-modal counterpart (VQ-VAE) on all evaluation metrics.

6. REFERENCES

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

[2] 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.

[3] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, pp. 125–125.

[4] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 3918–3926.

[5] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," *arXiv preprint arXiv:2008.12527*, 2020.

[6] S. Ö. Arık, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 10 040–10 050.

[7] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie *et al.*, "Sample efficient adaptive text-to-speech," in *International Conference on Learning Representations*, 2018.

[8] H. B. Moss, V. Aggarwal, N. Prateek, J. González, and R. Barra-Chicote, "BOFFIN TTS: Few-shot speaker adaptation by bayesian optimization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7639–7643.

[9] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez-Moreno *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *NeurIPS*, 2018.

[10] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," in *International Conference on Machine Learning*, 2018, pp. 3683–3691.

[11] T.-H. Kim, S. Cho, S. Choi, S. Park, and S.-Y. Lee, "Emotional voice conversion using multitask learning with text-to-speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7774–7778.

[12] M. Zhang, X. Wang, F. Fang, H. Li, and J. Yamagishi, "Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet," *Proc. Interspeech 2019*, pp. 1298–1302, 2019.

[13] H.-T. Luong and J. Yamagishi, "Nautilus: a versatile voice cloning system," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2967–2981, 2020.

[14] A. H. Liu, T. Tu, H.-y. Lee, and L.-s. Lee, "Towards unsupervised speech recognition and synthesis with quantized speech representation learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7259–7263.

[15] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[16] H. Zhang and Y. Lin, "Unsupervised learning for sequence-to-sequence text-to-speech for low-resource languages," *Proc. Interspeech 2020*, pp. 3161–3165, 2020.

[17] H. Zhang, "The neteasegames system for voice conversion challenge 2020 with vector-quantization variational autoencoder and wavenet," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 175–179.

[18] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *NIPS*, 2017.

[19] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," in *International Conference on Learning Representations*, 2018.

[20] C. Veaux, J. Yamagishi, and K. MacDonald, "Superseded - CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017. [Online]. Available: http://datashare.is.ed.ac.uk/handle/10283/2651

[21] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1. IEEE, 1993, pp. 125–128.