MoEVC: A Mixture of Experts Voice Conversion System With Sparse Gating Mechanism for Online Computation Acceleration

Yu-Tao Chang[1], *Yuan-Hong Yang*[1], *Yu-Huai Peng*[3], *Syu-Siang Wang*[2], *Tai-Shih Chi* [1], *Yu Tsao*[2] *and Hsin-Min Wang*[3]

[1]Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[3]Institute of Information Science, Academia Sinica, Taiwan

changyutao.cm07g@nctu.edu.tw, abc1010102.cm07g@nctu.edu.tw

Abstract

Owing to the recent advancements in deep learning technology, the performance of voice conversion (VC) in terms of quality and similarity has significantly improved. However, complex computation is generally required for deep-learning-based VC systems. This can cause a notable latency, which limits the deployment of such VC systems in real-world applications. Therefore, increasing the efficiency of online computing has become an important task. In this study, we propose a novel mixture-of-experts (MoE) based VC system, termed MoEVC. The MoEVC system uses a gating mechanism to assign weights to feature maps to increase VC performance. In addition, applying sparse constraints on the gating mechanism can skip some convolution processes through elimination of redundant feature maps, thereby accelerating online computing. Experimental results show that by using proper sparse constraints, we can effectively reduce the FLOPs (floating-point operations) count by 70%, while improving VC performance in both objective evaluation and human subjective listening tests.

Index Terms: Voice conversion, variational autoencoder, non-parallel VC, fully convolutional network, mixture of experts

1. Introduction

Voice conversion (VC) aims to modify speech signals from one speaker to sound as if they came from another speaker without changing the linguistic content. There are a wide variety of VC applications, such as personalized text-to-speech system (TTS) [1, 2], emotion conversion [3, 4], speaking and hearing-aid devices [5, 6], and singing voice conversion [7, 8]. Numerous models have been proposed as fundamental tools for VC tasks. Well-known examples include the Gaussian mixture model (GMM) [9, 10], exemplar-based models [11], deep neural networks [12], variational autoencoder (VAE) [13, 14], and generative adversarial network (GAN) [15]. Traditionally, VC is performed in a frame-to-frame conversion manner. A notable limitation is that the temporal information cannot be modeled well. Recently, long short-term memory (LSTM) with high performance in modeling temporal information has been used in VC and achieved higher performance than other methods [16, 17, 18, 19]. Meanwhile, a fully convolutional network (FCN) has also been used to perform sequence-to-sequence conversion and obtain higher mean opinion score (MOS) scores than frame-to-frame conversion methods [20].

To develop VC for real-world applications, there are generally two requirements: (1) using a compressed model that requires lite storage, and (2) being able to generate converted voice with minimal computational cost and latency. For classification tasks, many algorithms have been proposed to meet these two requirements. However, for regression tasks (e.g., VC), there are only few studies in the literature. In [21] and [22], the authors proposed using parameter quantization and scaling to compress deep-learning models for speech enhancement (SE) tasks. The results show that when the compression rate is properly designed, a good balance can be achieved between model size and SE performance. In this study, we focus on the second requirement, of which the aim is to reduce the cost of online computation. To this end, we propose a novel mixture of experts voice conversion (MoEVC) system with sparse constraints on the gating mechanism.

The authors of [23] extended the conventional MoE model [24] to DeepMoEs, which utilizes an architecture that can reduce the computational complexity of deep-learning-based models with convolutional layers. The model combines a base convolutional network with an embedding network along with a sparse gating network [23]. The input passes through both the embedding network and the sparse gating network, and the resulting gating values are used to select the channels in each layer of the base convolutional network. Compared with other state-of-the-art channel pruning methods [25, 26, 27, 28], this model achieves comparable or even better performance. We believe that the DeepMoEs architecture can be suitably used in VC systems to reduce online computation costs without compromising the quality of speech.

The proposed MoEVC system uses an auxiliary classifier variational autoencoder (ACVAE) VC system [29] as the basic architecture and integrates the MoE module to increase the efficiency of online computing. Although the DeepMoEs model has the potential to be applied to any network architecture, three issues confine the direct combination of DeepMoEs and ACVAE. First, the DeepMoEs model is proposed for the classification of images with a fixed shape, while ACVAE supports speech conversion with arbitrary speech lengths. Second, all gating values in DeepMoEs are controlled by a single input, whereas ACVAE accommodates an additional speaker code as input. Third, DeepMoEs takes a single image as input, but the two inputs of ACVAE (i.e., the spectral features and the speaker code) are in different representation formats.

In the proposed MoEVC, different structures are designed for the encoder and decoder to effectively combine DeepMoEs with ACVAE. In the experiments, we investigate the trade-off relationship between a reduction in FLOPs (floating-point operations) (the calculation of FLOPs followed [30]) and the quality of converted speech. To alleviate time-consuming listening tests, we first use the MOSNet [31], a deep-learning-based objective evaluator, to evaluate the VC results generated under

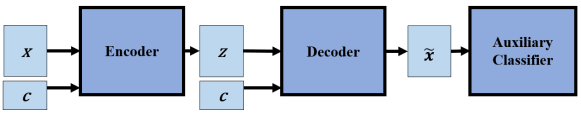

Figure 1: *The architecture of the ACVAE-VC system.*

various sparse constraints (i.e., computation acceleration rates). Subsequently, we validate the consistency of the human subjective listening test results and the MOSNet scores. Experimental results show that compared to the original ACVAE system, the MoEVC system can reduce the FLOPs count by 70% while improving speech quality.

The remainder of this paper is organized as follows. Section 2 reviews related works. Section 3 presents the MoEVC system. Section 4 exhibits the experimental setup and results. Finally, Section 5 provides the concluding remarks of this study.

2. Related works

This section presents related works on the MoEVC system.

2.1. VAE-VC

The VAE-VC is formed by an encoder-decoder architecture [13]. During training, given an input spectral feature $\boldsymbol{x}$, the encoder E_θ with the parameter set θ encodes $\boldsymbol{x}$ into a latent code: $\boldsymbol{z} = E_\theta(\boldsymbol{x})$. The speaker code, $\boldsymbol{c}$, of the input frame and the latent code, $\boldsymbol{z}$, are processed by the decoder G_ϕ with parameter set ϕ to reconstruct the input:

$$\bar{\boldsymbol{x}} = G_\phi(\boldsymbol{z}, \boldsymbol{c}) = G_\phi(E_\theta(\boldsymbol{x}), \boldsymbol{c}). \tag{1}$$

The model parameters can be learning from a speech corpus by maximizing the variational lower bound:

$$L_{vae}(\theta, \phi; \boldsymbol{x}, \boldsymbol{c}) = L_{recon}(\boldsymbol{x}, \boldsymbol{c}) + L_{lat}(\boldsymbol{x}), \tag{2}$$

where L_{lat} regularizes the encoder to align the approximate posterior with a prior distribution. Note that the speaker codes can be learned from the training data together with the model parameters θ and ϕ, or they can be defined in advance. For example, each speaker code can be assigned as a specific one-hot vector whose length is the number of training speakers.

In the conversion phase, given the source spectral frame $\boldsymbol{x}$ and the target speaker code, $\hat{\boldsymbol{c}}$, we can formulate the conversion function $f(.)$ and obtain the converted spectral feature, $\hat{\boldsymbol{x}}$, via:

$$\hat{\boldsymbol{x}} = f(\boldsymbol{x}, \hat{\boldsymbol{c}}) = G_\phi(\boldsymbol{z}, \hat{\boldsymbol{c}}) = G_\phi(E_\theta(\boldsymbol{x}), \hat{\boldsymbol{c}}). \tag{3}$$

Then, the converted spectral feature, $\hat{\boldsymbol{x}}$, is fed into the vocoder to generate the converted speech waveform.

Many studies have attempted to improve the VAE-VC framework. In [15], a GAN model was integrated with the VAE model to improve the VC performance. In [32], an auxiliary classifier was adopted to facilitate disentanglement for further improvement. Later, cross-domain VAE-VC was developed to jointly consider spectral features from different domains [14].

2.2. ACVAE-VC

In [29], Kameoka et al., proposed an auxiliary classifier variational autoencoder (ACVAE) VC framework, as shown in Fig. 1. ACVAE-VC is based on vanilla VAE-VC, where a gated CNN is used for both the encoder and decoder. The input-output relation of the l-th layer can be formulated as:

$$\boldsymbol{h}'_{l-1} = [\boldsymbol{h}^T_{l-1}, \boldsymbol{c}^T_{l-1}]^T, \tag{4}$$

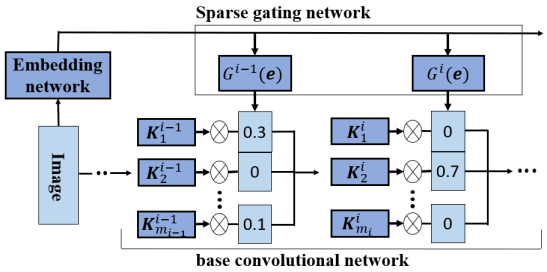

Figure 2: *The architecture of the DeepMoEs model.*

$$\boldsymbol{h}_l = (\boldsymbol{W}_l * \boldsymbol{h}'_{l-1} + \boldsymbol{b}_l) \odot \sigma(\boldsymbol{V}_l * \boldsymbol{h}'_{l-1} + \boldsymbol{d}_l), \tag{5}$$

where $\boldsymbol{h}_{l-1}$ and $\boldsymbol{c}_{l-1}$ represent the input and the speaker code of the $(l-1)$-th layer, respectively; $\boldsymbol{h}_l$ is the output; $\boldsymbol{W}_l$, $\boldsymbol{V}_l$, $\boldsymbol{b}_l$, and $\boldsymbol{d}_l$ are the weights and biases of the convolutional layer. $\boldsymbol{h}_{l-1}$ has the height and width of Q_{l-1} and N_{l-1}, and $\boldsymbol{c}_{l-1}$ is a 3D array consisting of a Q_{l-1}-by-N_{l-1} tiling of copies of the speaker code. Meanwhile, $\odot$ represents an element-wise multiplication, and σ denotes an element-wise sigmoid function.

The ACVAE-VC system uses an auxiliary classifier to encourage the model to generate output considering the speaker code. Specifically, the classifier approximates the posterior probability $p(\boldsymbol{c}|\boldsymbol{x})$ and is trained to maximize the lower bound of the mutual information between the decoder output and the speaker code. As derived in [29], the objective function is

$$J(\theta, \phi) + \lambda_L MI(\theta, \phi, \psi) + \lambda_I CE(\psi), \tag{6}$$

where J is the variational lower bound, MI is the mutual information lower bound, CE is the cross-entropy of the auxiliary classifier, and θ, ϕ, and ψ denote the parameters of the encoder, decoder, and auxiliary classifier, respectively.

2.3. Deep mixture of experts (DeepMoEs)

The work of [33] proved the advantages of stacking two layers of MoEs. Later, Wang et al. proposed the DeepMoEs architecture for image classification tasks, which can improve computational efficiency and may be applied to any model with convolutional layers [23]. Specifically, through the sparse gating mechanism, DeepMoEs allows each convolutional layer to dynamically select only a part of output feature maps to be activated during inference. Fig. 2 shows the architecture of DeepMoEs, which consists of three basic components: base convolutional network (BCN), embedding network (EMN), and sparse gating network (SGN). BCN and EMN share the same image input. EMN maps the input onto the embedding, $\boldsymbol{e}$, which is then transformed by SGN into a vector:

$$\boldsymbol{g}^l = G^l(\boldsymbol{e}) = \mathrm{ReLU}(\boldsymbol{W}^l_g \cdot \boldsymbol{e}), l = 1..., L, \tag{7}$$

where $\boldsymbol{g}^l$ is the gating vector for the l-th convolutional layer, and L is the total number of convolutional layers in BCN. As a result, DeepMoEs scales each feature map by the gating value generated from SGN. For a convolutional layer in BCN with tensor input $\boldsymbol{h}_i$ with spatial-resolution $W_i \times H_i$ and $\boldsymbol{C}_i$ input channels, a $\boldsymbol{C}_i \times K_W \times K_H \times \boldsymbol{C}_o$ convolutional kernel $\boldsymbol{K}$, and a set of gating values $[g_1, g_2, ..., g_{\boldsymbol{C}_i}]^T$ generated by SGN, the output $\boldsymbol{y}$ becomes

$$\boldsymbol{y} = \sum_{i=1}^{\boldsymbol{C}_i} g_i \boldsymbol{K}_i * \boldsymbol{h}_i. \tag{8}$$

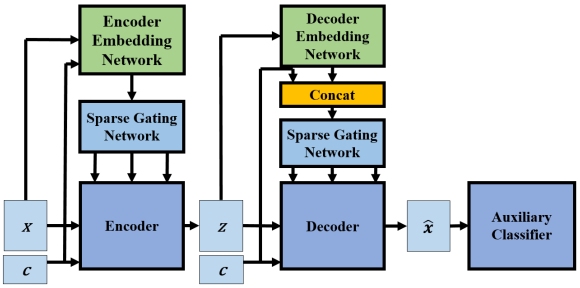

Figure 3: *The architecture of the proposed MoEVC system.*

Because the gating values are generated by the ReLU activation function, some elements may be zeros due to introducing the sparse constraint. To further control the sparsity of SGN and the diversity of EMN, an L1 regularization term can be added to the SGN outputs, along with the auxiliary classification loss. Accordingly, the overall loss function becomes

$$L_{all}(\boldsymbol{s};\boldsymbol{t}) = L_b(\boldsymbol{s};\boldsymbol{t}) + \lambda L_{\boldsymbol{g}}(\boldsymbol{s}) + \nu L_{\boldsymbol{e}}(\boldsymbol{s},\boldsymbol{t}), \qquad (9)$$

where $\boldsymbol{s}$ and $\boldsymbol{t}$ are the input image and the corresponding label, respectively, L_b is the loss from BCN, $L_{\boldsymbol{g}}$ is the L1 penalty term, and $L_{\boldsymbol{e}}$ is the auxiliary classification loss. Furthermore, λ and ν are coefficients for determining the weight between model accuracy, sparsity, and embedding diversity.

2.4. MOSNet: A deep-learning-based objective evaluator

Efficient and effective quality measurement of the generated speech has been a longstanding problem in TTS, SE, and VC tasks. In the VC community, objective metrics such as the mel-cepstral distance (MCD) [34] and global variance (GV) [35] are widely used to automatically measure the quality of converted speech. As reported in many previous studies [9, 36], such metrics do not always perfectly correlate with human perception as they mainly measure the distortion of acoustic features. Subjective listening tests, such as the mean opinion score (MOS) and similarity score, could represent the intrinsic naturalness and similarity of VC systems. Nevertheless, these human-involved evaluations are usually time-consuming and expensive because a large number of participants are required to perform listening tests and provide perceptual ratings. Recently, Lo et al. proposed a MOSNet, which uses a deep-learning-based model (formed by a CNN-BLSTM architecture) to perform speech naturalness and similarity assessments [31]. The results showed that the predicted MOS and similarity scores correlated well with human ratings. In this paper, we adopted MOSNet as a learning-based objective evaluator to evaluate the VC performance. Therefore, we could save the time of human listening and accelerate the development of the VC model.

3. Proposed MoEVC system

The proposed MoEVC system is shown in Fig. 3, which consists of BCN, SGN, and EMN. BCN was formed by ACVAE, which was designed and trained following [29]. To integrate DeepMoEs into the ACVAE-VC system, we mainly focus on the design of the EMN to meet the following requirements: (1) The decoder is not affected by the speaker code in the conversion phase. (2) The embedding network can process speech signals of any length. (3) The embedding network can accept multiple inputs from different domains.

Different from DeepMoEs that only uses one embedding network to control all gating values in the network, the proposed MoEVC system utilizes two embedding networks to control the gating values in the encoder and decoder separately. These two embedding networks are termed encoder/decoder embedding network (EEN/DEN), as shown in Fig. 3. To fulfil requirement (1), DEN only takes the latent code as input, and its output is concatenated with the speaker code to form the final embedding. Therefore, the decoder embedding is independent of the source speaker code in the conversion phase.

To fulfil requirements (2) and (3), we used convolutional layers in EEN because the CNN structure has been proven to be an effective feature extractor in previous works [37, 38]. Moreover, we utilized a temporal pooling layer proposed by [39] to average the input along the time dimension, so that we could obtain a fixed-sized vector; then, the final embedding was generated by several fully-connected layers for non-linear transformations. As for the input of the EEN, we first created a tensor, which is a Q-by-N tiles of $\boldsymbol{c}$, where Q and N are the spatial-resolution of $\boldsymbol{x}$. For the architecture of DEN, we combined a sequential autoencoder [40] with several fully-connected non-linear transformations. Specifically, we jointly trained a sequential autoencoder to reconstruct $\boldsymbol{z}$, and used the encoder part of the sequential autoencoder to perform feature extraction. The feature was concatenated with $\boldsymbol{c}$ and passed through fully-connected transformations to generate the final embedding.

The number of CNN kernels and the fully connected units in EEN and DEN were carefully designed so that the total parameters increased by less than 20% while the total FLOPs increased by less than 1%. The reason why the increases in parameters and FLOPs are different is that a CNN layer has fewer parameters than a fully-connected layer, but requires higher FLOPs. The increase in parameters is mainly due to the increase in fully connected units that require fewer FLOPs. Finally, the loss function for training the MoEVC model is:

$$L_{all}(\boldsymbol{x};\boldsymbol{c}) = L_{base}(\boldsymbol{x};\boldsymbol{c}) + \alpha L_{ae}(\boldsymbol{z}) + \beta L_{spc}(\boldsymbol{x}), \qquad (10)$$

where L_{base}, L_{ae}, and L_{spc} are the losses of BCN, DEN sequential autoencoder, and L1 norm of SGN, respectively.

4. Experiments

4.1. Experimental setup

We performed evaluation of the proposed MoEVC system using the Voice Conversion Challenge 2016 dataset. The speech signals were recorded by US English speakers in a professional recording studio without noticeable noise effects. The dataset consisted of 162 utterances for training and 54 utterances for evaluation from each of the 5 source and 5 target speakers. We used all training and evaluation data of two female speakers 'SF2' and 'SF3' and two male speakers 'SM1' and 'SM2' in the experiments. Thus, there were four types of conversions (female to female, male to male, female to male, and male to female). We report the average score for each conversion type and the average score for all types.

All speech signals were sampled at 16kHz. The WORLD vocoder [41] was used to extract 513-dimensional spectra (SPs), 513-dimensional aperiodic components (APs), and a fundamental frequency (F0) every 5ms. Then, 36-dimensional mel-cepstral coefficients (MCCs) were extracted from SPs and further normalized per pixel dimension based on the corresponding mean and standard deviation; these two statistics were saved and utilized in the conversion phase. All models were trained

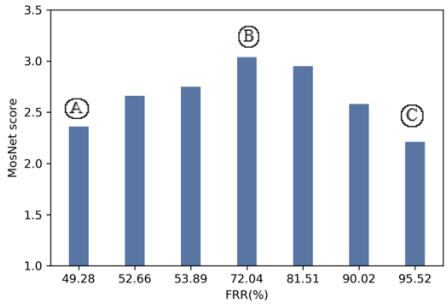

Figure 4: *The MOSNet scores of converted speech by the MoEVC system with different online computation costs (FRRs).*

for 5000 epochs using the Adam optimizer [42] with the initial learning rate, β_1, and β_2 set to 0.001, 0.9, 0.999 and a batch size of 64.

4.2. Experiment results

In this study, we focused our attention on a comparison of online computation cost and the naturalness of the converted speech. We first used MOSNet to evaluate the converted utterances with different sparse constraints. Setting a larger sparse constraint value causes the gating values to be sparser. More feature maps will then be pruned to achieve the purpose of significantly reducing the computational cost. As we investigated several spare constraint parameters, it was difficult to conduct listening tests for all converted speech results. Therefore, we first applied MOSNet to objectively evaluate the converted speech. Then we conducted listening tests to further confirm trends observed in the MOSNet scores.

4.3. MOSNet results

Fig. 4 shows a correlation between the MOSNet score and the computational cost, which is expressed as the FLOPs reduction rate (FRR). Since we intended to observe the relation between FRR and the MOSNet score, we did not apply any post-filtering methods to the converted speech, such as GV adjustment and maximum likelihood parameter generation. F0 was converted using linear mean-variance transformation in the log domain, while the APs were kept unmodified.

Fig. 4 shows that when we slightly increase the sparse constraint to achieve a higher FRR (from point Ⓐ to point Ⓑ), the MOSNet score will increase. However, when further increasing the FRR, the MOSNet score begins to decline (from point Ⓑ to point Ⓒ). This trend is often observed in model compression and acceleration tasks. After removing the redundant components to properly compress the model, the deep learning model can produce better performance. In contrast, further compression or acceleration may cause performance degradation. The MOSNet score reached the highest point at Ⓑ, for which the FRR and the MOSNet score were 72% and 3.2, respectively.

4.4. Human listening test results

Next, we conducted human listening tests to verify the findings from the evaluation results based on the MOSNet score. Twenty subjects were recruited to listen to samples of converted speech generated by the MoEVC system associated with points Ⓐ, Ⓑ, and Ⓒ in Fig. 4 and samples of converted speech by the original ACVAE-VC system. Each listener was asked to score 12 utterances, which were also evaluated by MOSNet, in terms of naturalness and similarity to the target speaker. Tables 1 and 2 show the MOS scores (for naturalness) and the similarity scores, where **X2Y** denotes the conversion of speaker type **X** to speaker type **Y**, **F** and **M** represent female speakers and male speakers, and **AVE** denotes the average score of all conversion pairs.

Table 1: *The MOS scores of converted speech produced by MoEVC with different FRRs and ACVAE.*

	FRR	F2F	F2M	M2F	M2M	AVE
Ⓐ	49%	2.97	2.18	1.95	1.58	2.13
Ⓑ	72%	**3.75**	**2.86**	**3.12**	**3.60**	**3.22**
Ⓒ	95%	1.85	1.63	1.69	1.94	1.74
ACVAE	0%	3.43	2.58	2.90	3.50	2.98

Table 2: *The similarity scores of converted speech produced by MoEVC with different FRRs and ACVAE.*

	FRR	F2F	F2M	M2F	M2M	AVE
Ⓐ	49%	2.83	2.05	1.78	1.85	2.06
Ⓑ	72%	**3.27**	2.10	2.05	3.32	2.48
Ⓒ	95%	2.41	1.63	1.5	2.06	1.79
ACVAE	0%	2.95	**2.13**	**2.20**	**3.36**	**2.50**

From Table 1, we note that the trend of MOS results is the same as in Fig. 4, i.e., case Ⓑ (fully compressed) is better than case Ⓐ (under compressed) and case Ⓒ (over compressed). Moreover, case Ⓑ even obtains better MOS scores than the original ACVAE-VC system, where no sparse constraints were applied, i.e., no acceleration. From Table 2, we can also see that case Ⓑ is better than case Ⓐ, case Ⓒ, and the original ACVAE-VC system. However, the improvement in the similarity score seems to be less significant than the improvement in the MOS score. There may be two reasons. First, the MoEVC model was adjusted by the MOSNet score, which was highly correlated with the MOS score. Second, model compression may be more harmful to similarity than naturalness.

5. Conclusions

The main contribution of this study is twofold. First, we confirmed the effectiveness of DeepMoEs for the acceleration of online computation for deep-learning-based VC tasks. According to our experimental results, under the condition of reducing the FLOPs count by 70%, the proposed MoEVC system will not harm the performance, and even improve the naturalness and similarity of the converted speech. Second, we proved that MOSNet can be used as an effective learning-based objective evaluator for VC tasks. We confirmed that the predicted scores are consistent to the results of human listening tests. Because it is difficult to conduct extensive human listening tests, using MOSNet to predict MOS scores during the model development phase can greatly speed up model development. We hope that our work can promote the research on model compression and online computation acceleration for VC.

Acknowledgement. This work was supported in part by the MOST-Taiwan Grants MOST 107-2221-E-001-008-MY3, MOST 109-2221-E-001-016, and MOST 109-2634-F-001-009.

6. References

[1] 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," *arXiv preprint arXiv:1903.12389*, 2019.

[2] H.-T. Luong and J. Yamagishi, "Bootstrapping non-parallel voice conversion from speaker-adaptive text-to-speech," *arXiv preprint arXiv:1909.06532*, 2019.

[3] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," *arXiv preprint arXiv:2002.00198*, 2020.

[4] K. Qian, Y. Zhang, S. Chang, D. Cox, and M. Hasegawa-Johnson, "Unsupervised speech decomposition via triple information bottleneck," *arXiv preprint arXiv:2004.11284*, 2020.

[5] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 172–183, 2013.

[6] D. Wang, J. Yu, X. Wu, S. Liu, L. Sun, X. Liu, and H. Meng, "End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction," in *Proc. ICASSP 2020*.

[7] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[8] K. Vijayan, H. Li, and T. Toda, "Speech-to-singing voice conversion: The challenges and strategies for improving vocal conversion processes," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 95–102, 2018.

[9] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[10] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Speech and Audio Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[11] Y.-C. Wu, H.-T. Hwang, C.-C. Hsu, Y. Tsao, and H.-M. Wang, "Locally linear embedding for exemplar-based spectral conversion," in *Proc. Interspeech 2016*.

[12] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[13] 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 *Proc. APSIPA ASC 2016*.

[14] W.-C. Huang, H.-T. Hwang, Y.-H. Peng, Y. Tsao, and H.-M. Wang, "Voice conversion based on cross-domain features using variational auto-encoders," in *Proc. ISCSLP 2018*.

[15] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. Interspeech 2017*.

[16] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. ICASSP 2015*.

[17] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "Atts2svc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *Proc. ICASSP 2019*.

[18] M. Zhang, B. Sisman, S. S. Rallabandi, H. Li, and L. Zhao, "Error reduction network for dblstm-based voice conversion," in *Proc. APSIPA ASC 2018*.

[19] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, 2019.

[20] W.-C. Huang, Y.-C. Wu, C.-C. Lo, P. L. Tobing, T. Hayashi, K. Kobayashi, T. Toda, Y. Tsao, and H.-M. Wang, "Investigation of f0 conditioning and fully convolutional networks in variational autoencoder based voice conversion," in *Proc. Interspeech 2019*.

[21] Y.-T. Hsu, Y.-C. Lin, S.-W. Fu, Y. Tsao, and T.-W. Kuo, "A study on speech enhancement using exponent-only floating point quantized neural network (eofp-qnn)," in *Proc. SLT 2018*.

[22] J. H. Ko, J. Fromm, M. Philipose, I. Tashev, and S. Zarar, "Precision scaling of neural networks for efficient audio processing," *arXiv preprint arXiv:1712.01340*, 2017.

[23] X. Wang, F. Yu, L. Dunlap, Y.-A. Ma, R. Wang, A. Mirhoseini, T. Darrell, and J. E. Gonzalez, "Deep mixture of experts via shallow embedding," *arXiv preprint arXiv:1806.01531*, 2018.

[24] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton *et al.*, "Adaptive mixtures of local experts." *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.

[25] J.-H. Luo, J. Wu, and W. Lin, "Thinet: A filter level pruning method for deep neural network compression," in *Proc. ICCV 2017*.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR 2017*.

[27] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *arXiv preprint arXiv:1608.08710*, 2016.

[28] Z. Huang and N. Wang, "Data-driven sparse structure selection for deep neural networks," in *Proc. ECCV 2018*.

[29] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Acvae-vc: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder," *arXiv preprint arXiv:1808.05092*, 2018.

[30] *https://github.com/Lyken17/pytorch-OpCounter/tree/master/thop*.

[31] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "Mosnet: Deep learning based objective assessment for voice conversion," *arXiv preprint arXiv:1904.08352*, 2019.

[32] J.-c. Chou, C.-c. Yeh, H.-y. Lee, and L.-s. Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," *arXiv preprint arXiv:1804.02812*, 2018.

[33] D. Eigen, M. Ranzato, and I. Sutskever, "Learning factored representations in a deep mixture of experts," *arXiv preprint arXiv:1312.4314*, 2013.

[34] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. PACRIM 1993*.

[35] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, "Ways to implement global variance in statistical speech synthesis," in *Proc. Interspeech 2012*.

[36] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *Proc. ICML 2019*.

[37] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[38] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[39] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP 2018*.

[40] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *Proc. ICML 2015*.

[41] 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.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.