CA-VC: A Novel Zero-Shot Voice Conversion Method With Channel Attention

Ruitong Xiao*, Xiaofen Xing*‡†, Jichen Yang* and Xiangmin Xu*
* South China University of Technology, Guangzhou, China
‡ Zhongshan Institute of Modern Industrial Technology of SCUT, Zhongshan, China
E-mail: eerxiao@mail.scut.edu.cn, xfxing@scut.edu.cn, nisonyoung@gmail.com, xmxu@scut.edu.cn

***Abstract*—In recent years, more and more zero-shot voice conversion algorithms have been proposed. However, tasks about adaptation and disentanglement of speaker and content information in speech are still challenging. In this paper, we propose a novel zero-shot voice conversion method with channel attention (CA) under the framework of variational auto-encoder. In detail, the model consists of content encoder, speaker encoder and decoder. CA plays two different roles in our method. In content encoder, CA with channel width constraint forms a learnable bottleneck to reduce speaker information and retain content information simultaneously. In decoder, CA is used to combine speaker information and content information. Objective and subjective evaluations show that the proposed method can perform voice conversion well and generate high quality converted speech.**

I. Introduction

Voice conversion (VC) is to convert a source speaker's voice to a target speaker's voice without changing the linguistic content [1]. From the perspective of application, voice conversion is widely applied in many fields such as entertainment, creative industry, and spoofed speech generation [2].

In terms of learning methods, previous works of voice conversion can be roughly divided into supervised and unsupervised. Supervised voice conversion such as [3, 4] can achieve good performance, but has high requirements on data. Further speaking, it requires paired data or frame-level alignment between the source and the target speakers during training. In addition, if there is a large gap between the source and target domains, incorrect alignment may harm conversion performance. That's the reason why supervised voice conversion can't be widely used. On the contrary, parallel data is unnecessary in unsupervised voice conversion [5, 6]. Therefore, it becomes more easier to construct dataset for unsupervised voice conversion. Because unsupervised voice conversion has more relaxed requirements on data, it has attracted more and more attention in the community of voice conversion in recent years. In this work, we mainly focus on unsupervised voice conversion.

There are two kinds of unsupervised voice conversion: seen and unseen speaker. For the seen speaker voice conversion, although some previous proposed algorithms can only convert for seen speakers, such as StarGAN-VC [7] and CycleGAN-VC [8], more and more zero-shot voice conversion algorithms have been proposed for unseen speaker voice conversion recently [9, 10]. Generally speaking, voice signal often carries static information and dynamic information. Static information, such as the identity of the speaker and the environment of the voice recording device, is stable and unchanging in a segment of voice. However, dynamic information such as content and tone may change in each frame of speech. For voice conversion, it is required to change the static information while keep dynamic information unchanged. Therefore, the methods for the unseen speaker voice conversion usually focuses on how to disentangle the static information (speaker information) and dynamic information (content information). For example, vector quantization (VQ)-VC [11] was proposed to separate speaker and content information through discrete codes. And Lee et al. used the speaker identity classifier as discriminator against the content encoder to reduce the speaker identity information in content encoder [9]. AUTOVC [10] carefully designed proper width of bottleneck layer so that only content information is retained in content encoder. Chou et al. proposed a zero-shot voice conversion with instance normalization (AdaIN-VC) [12], which removes speaker information by instance normalization (IN) in the speaker encoder and adapts content representation with adaptive instance normalization (AdaIN) [13].

Although many effective voice conversion methods have been proposed, how to accurately separate speaker and content representations in speech and adapt voice to target speaker domain are still challenging tasks. When removing speaker information, content information may be removed at the same time. And the effect of conversion may be unstable. In this paper, we propose a novel channel attention (CA) based method for zero-shot voice conversion (CA-VC). As far as we know, channel attention has not been used in voice conversion task.

CA has two different tasks in our method. On one hand, we use CA with channel width constraint, which forms a learnable bottleneck, to reduce speaker information and retain content information in content encoder. On the other hand, inspired by MCCNet [14], CA is used for adapt voice to target speaker domain in decoder. In addition, the speaker encoder is followed by an auxiliary classifier that assists the speaker encoder to focus on speaker information. Because we do not use one-hot embedding for speakers and unseen speaker embedding can be extracted by the speaker encoder, our model has the potential to perform zero-shot voice conversion. Objective and

†Corresponding author. Email: xfxing@scut.edu.cn

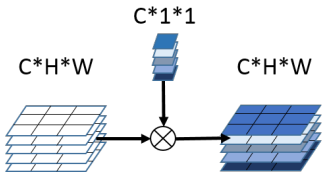


Fig. 1. Channel attention module for 2d convolution. Left set of square means input feature map, upper set of square means CA and right set of square means output feature map. $\bigotimes$ denotes element-wise product, C is the number of channel, meanwhile, H and W are spatial dimensions.

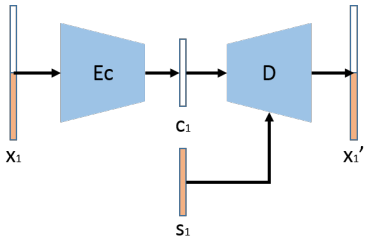


Fig. 2. The training phase of AUTOVC.

subjective evaluations conducted on CSTR VCTK corpus [15] show that our model can perform zero-shot voice conversion well and generate speech similar to the target speaker.

Contributions of our proposed method can be summarized as follows:

- We propose to use CA for disentangling speaker identity information and content information.
- We propose to use CA to adapt voice to target speaker domain.

The rest of the paper is organized as follows. Section 2 simply introduces related works. Section 3 introduces the proposed CA-VC in detail. Section 4 gives the experimental results and the corresponding analysis. Finally, the paper is concluded.

II. Related Works

In this section, we introduce related works of proposed method.

A. Channel attention

In the field of image classification, SENet [16] generates attention to the feature map in the channel dimension. The CA with size of $C \times 1 \times 1$ and the feature map with size of $C \times H \times W$ are made element-wise product as shown in Fig. 1 where C is the number of channel, meanwhile, H and W are spatial dimensions. Different channels are weighted respectively and important feature channels are labeled by weighting, so that the model focuses on important feature parts and forms attention. Similarly, CBAM [17] also uses similar CA to generate refined features. At the same time, CA is used for arbitrary video style transfer in MCCNet [14], and the method is available for Conv1d layers by generating channel attention of $C \times 1$ dimension.

Otherwise, the attention mechanism can be divided into soft attention and hard attention. Hard attention is equivalent to mask, and its characteristic is not differentiable. Soft attention is differentiable, it can calculate the gradient, and use the gradient back propagation to update and learn. Therefore, the application of soft attention in deep learning is more common.

B. AUTOVC

Qian et al.[10] propose a new style transfer scheme, which involves only a vanilla auto-encoder with a carefully designed bottleneck. Similar to CVAE, the proposed scheme only needs to be trained on the self-reconstruction loss, but it has a distribution matching property similar to GAN's. This is because the correctly-designed bottleneck will learn to remove the style information from the source and get the style-independent code by introducing carefully-tuned dimension reduction and temporal downsampling to constrain the information flow as shown in Fig. 2. This simple scheme leads to a significant performance gain. AUTOVC achieves superior performance on the traditional many-to-many conversion task, where all the speakers are seen in the training set, and perform zero-shot voice conversion with decent performance.

C. Variational auto-encoder

Variational auto-encoder (VAE) [18] with speaker encoder, content encoder and decoder is a common choice for voice conversion [12]. Let x be the input speech segment and χ be the collection of training data. Let E_c be the content encoder, E_i be the speaker identity encoder and D be the decoder. In unsupervised learning, the reconstruction method is usually used for training. The reconstructed loss function is given as below.

$$L_{rec}(\theta_{E_c}, \theta_{E_i}, \theta_D) = \mathbb{E}_{x \epsilon \chi}\left[\|D(E_c(x), E_i(x)) - x\|_1^1\right] \quad (1)$$

In order to make the encoder output code obey the normal distribution, we use Kullback-Leibler (KL) loss to constrain the model. Let μ and σ be the mean and standard deviation of content code generated by E_c. The definition of KL loss is as

$$L_{KL}(\theta_{E_i}) = 0.5 \times \mathbb{E}\left[e^{\sigma} + \mu^2 - 1 - \sigma\right] \quad (2)$$

where e is natural constant. The objective function of VAE is defined as the weighted sum of the reconstruction loss function and KL loss, and the model is trained to minimize the objective function. The objective function of VAE is defined as follows, where λ_{rec} and λ_{KL} are the weights of L_{rec} and L_{KL}, respectively.

$$L_{VAE} = \lambda_{rec} L_{rec} + \lambda_{KL} L_{KL} \quad (3)$$

III. Proposed Approach

In this section, the proposed CA-VC will be introduced in detail. Our work is based on the VAE framework. Fig. 3 shows the architecture of CA-VC during training where AP,

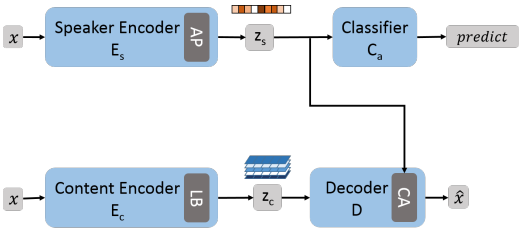


Fig. 3. The training phase of CA-VC where AP, LB and CA represent the modules of average pooling, learnable bottleneck and channel attention, respectively.

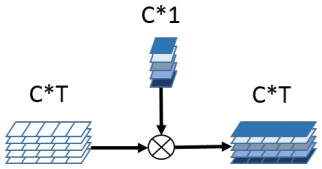


Fig. 4. Channel attention module for 1d convolution. $\otimes$ denotes element-wise product, C is the number of channel, meanwhile, T is time dimension.

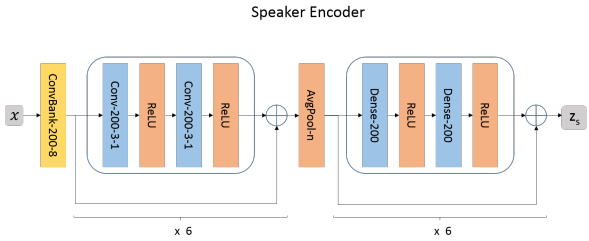


Fig. 5. The architecture of speaker encoder, where convolution layers are described as "Conv-output channel number-kernel size-stride" and dense layers are described as "Dense-output channel number".

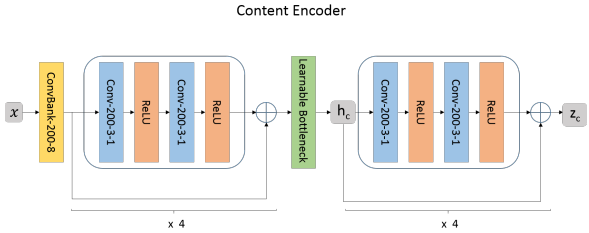


Fig. 6. The architecture of content encoder.

LB and CA represent the modules of average pooling, learnable bottleneck and channel attention module, respectively. In addition, x and $\hat{x}$ represent input feature and corresponding reconstructed feature, respectively. There are three modules in CA-VC, which are speaker encoder E_s, content encoder E_c and decoder D. Meanwhile, E_s is followed by an auxiliary classifier C_a for assisting the speaker encoder to focus on speaker information. They are trained at the same time without using pre-trained model. And we use Conv1d layers in encoders and decoder. Otherwise, channel attention for Conv1d layer is used and the process is shown in Fig. 4. Similar to channel attention for Conv2d, we generate attention to the feature map in the channel dimension. The CA with size of $C \times 1$ and the feature map with size of $C \times T$ dimension are made element-wise product. C is channel dimension and T is time dimension. During conversion, utterance of source speaker and utterance of target speaker are used as input of content encoder and speaker encoder, respectively. C_a will be removed if training is done. Next, we will introduce proposed approach and loss function.

A. Speaker encoder

The architecture of speaker encoder E_s is shown in Fig. 5. We use convbank layer firstly to extract feature at different time scales from the input. Afterwards, a set of residual convolution blocks is used for extracting feature. Because speaker identity information is stable in a utterance, we perform global average pooling (AP) over time dimension, try to enforce the speaker encoder to learn global information only and remove dynamic information. Then the global information gets into a block of residual fully connection layers to generate z_s. Due to the role of average pooling, we assume that z_s only carries global information, including speaker information.

To assist speaker encoder focus on speaker information, we use a dense layer as auxiliary classifier C_a which will be removed after training. C_a takes z_s as input and predicts corresponding speaker. And the loss of classification is defined as follows.

$$L_{class} = \mathbb{E}_{x \epsilon \chi} \left[-log(C_a(c|z_s)) \right] \tag{4}$$

And L_{class} is the objective function with respect to C_a. It is also a part of objective function with respect to CA-VC model. With the constraints of the classification loss function, E_s tends to learn speaker identity information, hence we can assume that z_s is speaker representation. Because the speaker representation is learned and has the ability to expand, our model has the potential to perform zero-shot voice conversion.

B. Channel attention based disentangling method

In content encoder, we introduce a channel attention based module, which is helpful for disentangling.

Fig. 6 shows the architecture of content encoder. We also use convbank layer and a set of residual convolution blocks for extracting feature before channel attention layer, which provide the base for generating content representation.

After the front part of content encoder transforming the input feature, channel attention module is used to constrain the width of the channel. And we name the module with channel attention as learnable bottleneck (LB). The architecture of learnable bottleneck is shown as Fig. 7.

Different from CA used in SENet [16] or CBAM [17], CA at this part is not generated from the feature map, but a learnable parameter (LP) with size of $C \times 1$. We set zero as boundary of the parameter by using ReLU and get a soft mask $z_{CA} = ReLU(LP)$. Then z_{CA} and h_{pre}, which is extracted

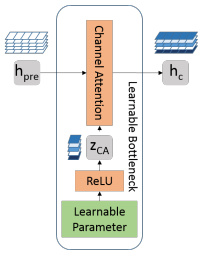


Fig. 7. The architecture of learnable bottleneck.

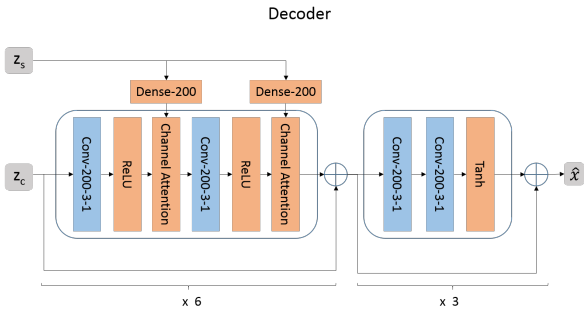


Fig. 8. The architecture of decoder.

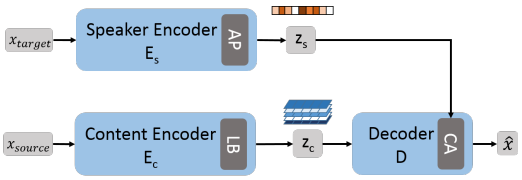


Fig. 9. The process of voice conversion using CA-VC.

by the front part of content encoder, are made element-wise product. z_{CA} is used as channel weights of h_{pre}. The channel with zero weight is shut down while the channel with positive weight is activated. When the training is done, z_{CA} is fixed during inference. After that, we get intermediate feature h_c.

According to AUTOVC [10], on the premise of preserving content information in intermediate feature, channel width of intermediate feature should be as small as possible. So that intermediate feature can reduce redundant information and only retains content information.

In order to make h_c retain necessary content information and reduce speaker information as much as possible, we add a penalty term p to reduce activate channels of h_c. The definition of p is as

$$p = \|z_{CA}\|_1^1 \tag{5}$$

where p is the L1 regularization of z_{CA}, which can make z_{CA} sparse. Therefore, the number of activated channel is reduced gradually during training and it is equal to constraining channel width of h_c. It also prevents the model from becoming an identity transformation.

As mentioned above, z_s tends to learn speaker feature and can not carry dynamic information. Because we use reconstruction loss as objective function and speaker information can be provided by z_s, complementary information of speaker information can only be passed from content encode E_c. In terms of getting content information, the limited channel resource of h_c preferentially tends to carry necessary dynamic content information for reconstructing. And speaker information will be abandoned if the limit of p is strong enough. Therefore, we assume that h_c with strong channel width limit is content representation and the channel attention based method is helpful for disentangling content and speaker representation by limiting the width of channel.

The subsequent set of residual blocks further adjusts h_c and generates μ_c and σ_c. We define μ_c and σ_c as the mean and standard deviation of z_c, respectively. And z_c is given as

$$z_c = \mu_c + \sigma_c \times \mathcal{N}(0, I) \tag{6}$$

where $\mathcal{N}(0, I)$ denotes standard Gaussian distribution.

C. Channel attention based adapting method

Inspired by MCCNet [14], channel attention module is used for adapting voice to target speaker domain. Channel attention module combines content information and speaker information in decoder. From Fig. 8, it can be seen that in the decoding process, speaker representation z_s gets into dense layer first and then it is provided to decoder (D) by channel attention module. CA at this part is generated from z_s and is made element-wise product with feature map. Size of CA and feature map are $C \times 1$ and $C \times T$, respectively. It can enhance or weaken feature based on speaker information for adaptation. Moreover, CA changes the standard deviation of features and ReLU implicitly changes mean of feature. These characteristics make channel attention module similar to AdaIN module. The two mechanisms may have some similarities. However, AdaIN operation, which normalizes the mean and variance of each feature map separately, potentially destroys information found in the magnitudes of the features relative to each other [19]. IN in AdaIN module may remove speaker information added by previous AdaIN module which may weaken the effect of adaptation. On the contrary, channel attention module does not need IN and have no such risk.

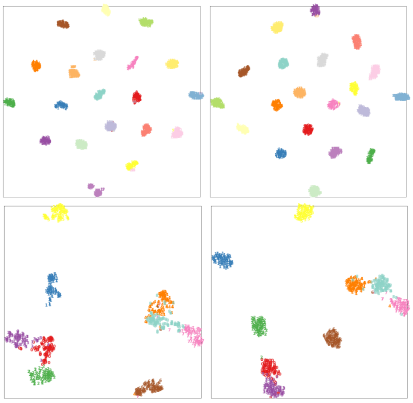
After adapting by blocks with channel attention module, a postnet consisting of Conv1d layers is used for refining.

Fig. 9 gives the conversion process of the proposed CA-VC. It can be observed that speaker representation z_s is obtained from the target speaker by speaker encoder. Because the speaker representation is learned and has the ability to expand, our model can be used for converting seen and unseen speaker's voice.

D. Loss function

Because our approach is unsupervised, we train the model by reconstruction. The reconstruction loss function is given as

$$L_{rec} = \mathbb{E}_{x \epsilon \chi} \left[\|D(z_c, z_s) - x\|_1^1 \right] \tag{7}$$



(a) (b)

Fig. 10. The visualization of speaker embedding. The upper and lower rows correspond to seen and unseen speakers, respectively. Columns from left to right correspond to (a) CA-VC without auxiliary classifier, (b) CA-VC with auxiliary classifier.

Because the framework of our model is VAE, we apply KL loss.

Finally, the objective function is defined as follows, where λ_{rec}, λ_{KL}, λ_p and λ_{class} are the weights of L_{rec}, L_{KL}, L_p and L_{class}, respectively.

$$L = \lambda_{rec}L_{rec} + \lambda_{KL}L_{KL} + \lambda_p p + \lambda_{class}L_{class} \quad (8)$$

IV. Experiment and Evaluation

A. Dataset and experiment setup

The proposed method is evaluated on CSTR VCTK corpus [15], which includes speech data uttered by 109 native English speakers with various accents. Each speaker has about 400 sentences. Although there are some parallel data in the dataset, we do not use the characteristic of parallel data during training. The corpus is randomly spilt into training set with 90 speakers and validation set with 19 speakers. In addition, following [12], 512-dimension mel-spectrogram is extracted as the input acoustic feature. And input frame length is 180. So the size of input x is 512×180. Furthermore, h_{pre}, h_c, μ_c, σ_c and z_c are all with the size of 200×180. We omit batch size dimension here.

Because vocoder is not the key point of our research and Griffin-Lim with 512-dimension mel-spectrogram can also get acceptable generation quality, following [12], inverse linear transformation and Griffin-Lim algorithm [20] are used for converting the mel-spectrograms back to waveform.

B. Speaker embedding visualization

In order to visualize the speaker embedding in the 2D space by using t-SNE [21], we randomly chose 20 seen speakers and 9 unseen speakers, each with 100 utterances, which are used as input to obtain z_s. We compared the visualization of two train settings which were CA-VC with auxiliary classifier and CA-VC without auxiliary classifier.

As shown by the left column of Fig. 10, the points representing the utterances of the same speaker are clustered together and the utterances of different speakers can be separated. We found that the speaker representation z_s generated by the model without auxiliary classifier can correspond to the speaker well even the speaker identity label is not used for training. It is probably because when channel width of content encoder is limited, speaker information tends to pass from speaker encoder even without classify supervision. And similar conclusions are also mentioned in [12, 22]. But z_s generated by the model with auxiliary classifier can be separated more easily, especially for unseen speakers as shown in the right lower part of Fig. 10. It illustrates that auxiliary classifier is helpful for speaker embedding.

C. Evaluation of disentanglement

To verify the disentangling effect of LB module, we performed an ablation study and compared the speaker identity prediction accuracy on z_c from AdaIN-VC, CA-VC and CA-VC without LB module. For comparing prediction accuracy, we trained three speaker classifiers, which is consisting of average pooling and 3 dense layers, with z_c from three models as input respectively and then calculate speaker prediction accuracy.

Otherwise, we used each model to generate 20 converted utterances from unseen speaker utterances, transcribed utterances by using google speech recognition API. Then we used transcription of source utterances as label to calculate the word error rate (WER) of converted utterances.

TABLE I
THE ACCURACY FOR SPEAKER IDENTITY PREDICTION ON CONTENT REPRESENTATION AND THE WORD ERROR RATE (WER).

Methods	***Accuracy***	***WER***
AdaIN-VC [12]	32%	44%
CA-VC	35%	30%
CA-VC **without LB**	63%	26%

Table I shows the results, the lower the better for both accuracy and WER. The accuracy of CA-VC with LB is lower than the result of ablated model and is close to that of AdaIN-VC. It proves that LB module can reduce speaker information in content representation. Meanwhile, our method performed better than AdaIN-VC in term of WER, which shows that our method can retain content information better. Because IN in AdaIN-VC may cause slight loss of content, the loss may be obvious as the number of IN layers increases. But our method only uses one LB module so there won't be such cumulative effect.

D. Evaluation of voice conversion

Fig. 11 shows a conversion phase mel-spectrogram result sample of our method. Converted mel-spectrogram has similar characteristic of target mel-spectrogram, such as peak inteval, and trend of curve changed in time domain is similar to source mel-spectrogram.

We used the global variance (GV) [23] firstly. GV is a method to visualize spectral variance distribution of speaker

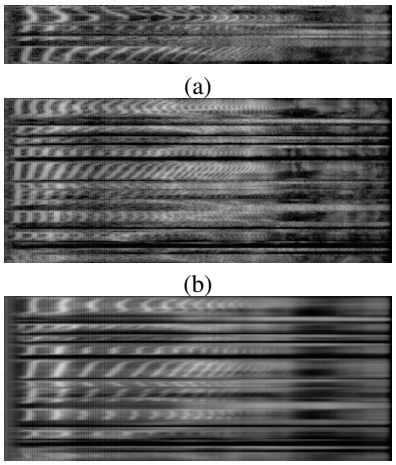


Fig. 11. Mel-spectrogram of utterances. The horizontal axis represents the frequency domain, and the vertical axis represents time. Rows from top to bottom correspond to (a) target utterance, (b) source utterance and (c) converted utterance.

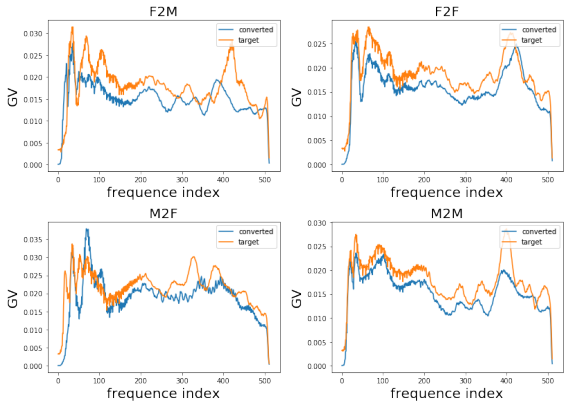


Fig. 12. The global variance of converted result and target speaker utterance.

and can be used to measure conversion effect [24]. We generated 100 converted utterances (including male to male (M2M), male to female (M2F), female to male (F2M) and female to female (F2F), 4 pairs of unseen speakers) and chose 100 utterances of target speaker randomly. As shown in Fig. 12, we evaluated the GV for each of the frequency index for 4 speaker pairs and found that the distribution of converted result is similar to the distribution of target speaker's utterances. It declared that channel attention module can convert the voice and our model is effective for voice conversion.

TABLE II
RESULTS COMPARISON AMONG DIFFERENT METHODS. THE RESULT OF SIMILARITY TEST IS WITH THE FORM AS TARGET/ SOURCE/ NO ONE.

Methods	**Metric**		
	MCD	***Naturalness***	***Similarity***
AdaIN-VC [12]	3.4	3.1	59%/ 11%/ 30%
AUTOVC [10]	3.7	3.4	46%/ 24%/ 30%
CA-VC	**3.3**	3.6	**75%**/ 15%/ 10%
CA-VC **without LB**	4.9	**3.7**	20%/ 38%/ 42%

Table II shows the experiment results comparison among AdaIN-VC, AUTOVC (with WaveNet [25]), CA-VC and CA-VC without LB module. Mel cepstral distortion (MCD) [26] is used for objective test. Mean opinion score (MOS) of naturalness and similarity test are used for subjective test.

For Objective evaluation, we calculated MCD between converted speech and target speech. There are parallel data in CSTR VCTK corpus, though we do not use the characteristic of parallel data during training, but data are not strict aligned. So we used dynamic time warping before calculating MCD. The lower MCD the better. Table II shows the results for different systems. It can be observed that the proposed method can give a comparable performance and achieve the best result.

For Subjective evaluation, by using unseen speaker as both source and target speaker, each model generated 20 converted utterances (including M2M, M2F, F2M and F2F, 4 pairs of unseen speakers). Speaker pairs of different methods are same. In MOS test of naturalness, 18 subjects evaluated the naturalness of utterances by scoring from 1 to 5 and the higher MOS score the better. In similarity test, subjects chose which speaker is more similar to converted speech, target speaker or source speaker or no one. We define conversion success rate as chosen rate of target speaker.

In MOS test of naturalness, CA-VC produced high quality speech and got the second highest score. Although CA-VC without LB got the highest score, it does not mean that it worked well. Because without LB, content representation contains too much redundant information including speaker information, which improves naturalness but leads to poor performance during conversion as shown in similarity test. In similarity test, the result shows that our model is more effective and got the highest conversion success rate. It is probably because the reason mentioned in section III-C. It verifies the disentangling effect of LB module and proves that channel attention module can adapt source data to target domain for voice conversion.

V. CONCLUSIONS

In this work, a novel zero-shot unsupervised voice conversion method with channel attention is proposed. As far as we know, channel attention has not been used in voice conversion task. CA has two different tasks in our method. We use channel attention for disentangling speaker and content information in content encoder. Meanwhile, we combine speaker and content information in decoder by channel attention. Objective and subjective evaluations reveal that channel attention has significant effect during disentangling and combining speaker and content information. Thus, the proposed method is able to generate high quality converted speech.

VI. ACKNOWLEDGEMENTS

The work is supported by the National Natural Science Foundation of China U1801262, Key-Area Research and Development Program of Guangdong 2019B010154003, the Science and Technology Project of Guangzhou 202103010002, Fundamental Research Funds for the Central Universities

(2019PY21) and Science and Technology Project of Zhongshan (2019AG024).

REFERENCES

[1] B. Sisman, J. Yamagishi, S. King, and H. LI, "An eoverview of voice conversion and its challeng:from statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2021.

[2] Z. Wu, P. L. D. Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z.-H. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagishi, "Anti-spoofing for text-independent speaker verification: an initial database,comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 768–783, 2016.

[3] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[4] X. Tian, S. Lee, Z. Wu, E. S. Chng, and H. Li, "An example-based approach to frequency warping for voice conversation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1863–1875, 2017.

[5] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2016.

[6] X. Tian, J. Wang, H. Xu, E. S. Chng, and H. Li, "Average modeling approach to voice conversion with non-parallel data," in *The Speaker and Language Recognition Workshop (Odyssey)*, 2018, pp. 227–232.

[7] T. Kaneko, H. Kameoka, T. Kou, and N. Hojo, "Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion," in *INTERSPEECH*, 2019.

[8] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6820–6824.

[9] S. Lee, B. Ko, K. Lee, I.-c. Yoo, and D. Yook, "Many-to-many voice conversion using conditional cycle-consistent adversarial networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6279–6283.

[10] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning (ICML)*, 2019, pp. 5210–5219.

[11] D.-y. Wu and H.-y. Lee, "One-shot voice conversion by vector quantization," in *45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7734–7738.

[12] J.-c. Chou, C.-c. Yeh, and H.-y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," in *INTERSPEECH*, 2019.

[13] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1510–1519.

[14] Y. Deng, F. Tang, W. Dong, H. Huang, C. Ma, and C. Xu, "Arbitrary video style transfer via multi-channel correlation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1210–1217.

[15] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[17] S. Woo, J. Park, J.-Y. Lee, and I.-S. Kweon, "CBAM: Convolutional block attention module," in *14th European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

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

[19] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116, 2020.

[20] D. Griffin and J. Kim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[21] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 2605, pp. 2579–2605, 2008.

[22] M. Ravanelli and Y. Bengio, "Learning speaker representations with mutual information," *arXiv preprint arXiv:1812.00271*, 2018.

[23] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *ICASSP 2005-2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, vol. 1. IEEE, 2005, pp. I–9.

[24] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for hmm-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. E90D, no. 5, pp. 1–6, 2007.

[25] A. V. D. 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," *ArXiv preprint arXiv:1609.03499v2*, 2016.

[26] R. F. 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, 1993, pp. 125–128.