VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-shot Voice Conversion

Disong Wang[1], Liqun Deng[2], Yu Ting Yeung[2], Xiao Chen[2], Xunying Liu[1], Helen Meng[1]*

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Huawei Noah's Ark Lab

{dswang,xyliu,hmmeng}@se.cuhk.edu.hk, {dengliqun.deng,yeung.yu.ting,chen.xiao2}@huawei.com

Abstract

One-shot voice conversion (VC), which performs conversion across arbitrary speakers with only a single target-speaker utterance for reference, can be effectively achieved by speech representation disentanglement. Existing work generally ignores the correlation between different speech representations during training, which causes leakage of content information into the speaker representation and thus degrades VC performance. To alleviate this issue, we employ vector quantization (VQ) for content encoding and introduce mutual information (MI) as the correlation metric during training, to achieve proper disentanglement of content, speaker and pitch representations, by reducing their inter-dependencies in an unsupervised manner. Experimental results reflect the superiority of the proposed method in learning effective disentangled speech representations for retaining source linguistic content and intonation variations, while capturing target speaker characteristics. In doing so, the proposed approach achieves higher speech naturalness and speaker similarity than current state-of-the-art one-shot VC systems. Our code, pre-trained models and demo are available at https://github.com/Wendison/VQMIVC.

Index Terms: vector quantization, mutual information, unsupervised disentanglement, one-shot, voice conversion

1. Introduction

Voice conversion (VC) is a technique used to modify para-linguistic factors of an utterance from a source speaker to sound like a target speaker. Para-linguistic factors include speaker identity [1], prosody [2] and accent [3], etc. In this paper, we focus on the conversion of speaker identity across arbitrary speakers under a one-shot scenario [4, 5], i.e., given only one target speaker's utterance for reference.

Previous work that use methods based on speech representation disentanglement (SRD) [6–8] attempted to address one-shot VC by decomposing the speech into speaker and content representations, and then the speaker identity can be converted by changing the source speaker's representation to that of the target speaker. However, it is difficult to measure the degree of SRD. Besides, previous approaches generally do not impose correlation constraints between speaker and content representations during training, which results in leakage of content information into the speaker representation, leading to VC performance degradation. To alleviate these issues, this paper proposes the vector quantization and mutual information-based VC (VQMIVC) approach, where mutual information (MI) measures the dependencies between different representations and can be effectively integrated into the training process to achieve

SRD in an *unsupervised manner*. Specifically, we first decompose an utterance into three factors: content, speaker and pitch, and then propose a VC system consisting of four components: (1) A content encoder using vector quantization with contrastive predictive coding (VQCPC) [9, 10] to extract frame-level content representations from acoustic features; (2) A speaker encoder that takes in acoustic features to generate a single fixed-dimensional vector as the speaker representation; (3) A pitch extractor that is used to compute normalized fundamental frequency (F_0) at the utterance level as the pitch representation; and (4) A decoder that maps content, speaker and pitch representations to acoustic features. During training, the VC system is optimized by minimizing VQCPC, reconstruction and MI losses. VQCPC aims to explore local structures of speech, and MI reduces the inter-dependencies of different speech representations. During inference, one-shot VC is achieved by only replacing the source speaker representation with the target speaker representation derived from a single target utterance. The main contribution of this work lies in applying the combination of VQCPC and MI to achieve SRD, without any requirements of supervision information such as text transcriptions or speaker labels. Extensive experiments have been conducted to thoroughly analyze the importance of MI, where information leakage issues can be significantly alleviated for enhanced SRD.

2. Related work

VC performance is critically dependent on the availability of the target speaker's voice data for training [11–17]. Hence, the challenge of one-shot VC is in performing conversion across arbitrary speakers that may be unseen during training, and with only one single target-speaker utterance for reference. Previous approaches for one-shot VC are based on SRD, which aims to separate speaker information from spoken content as far as possible. Related work include: tunable information-constraining bottleneck [6, 18, 19], instance normalization techniques [7, 20] and vector quantization (VQ) [8, 21]. We adopt VQCPC [9, 10], which is an improved version of VQ, to extract accurate content representations. Without explicit constraints between different speech representations, information leakage tends to occur, which degrades VC performance. We draw inspirations from information theory [22] in using MI as a regularizer to constrain the dependency between variables. As MI computation is challenging for variables with unknown distribution, various methods have been explored to estimate MI lower bound [23–25] in SRD-based speech tasks [26–28] . To guarantee the reduction of MI values, we propose to use variational contrastive log-ratio upper bound (vCLUB) [29]. While a recent effort [30] employs MI for VC by using speaker labels as the supervision for learning speaker representations, our proposed approach differs from [30] in terms of the combination of VQCPC and MI

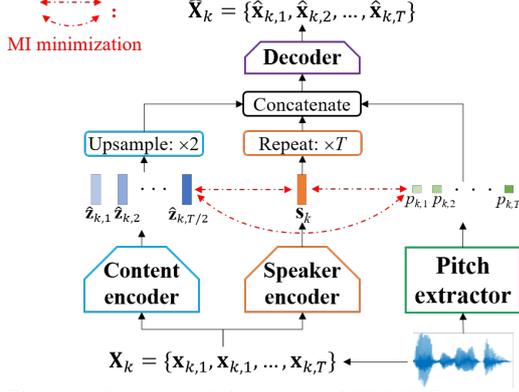*Work done during internship at Huawei Noah's Ark Lab

Figure 1: *Diagram of the proposed VQMIVC system.*

for fully unsupervised training, and the incorporation of pitch representations to maintain source intonation variations.

3. Proposed approach

This section first describes the system architecture of the VQMIVC approach, then elaborates on the integration of MI minimization into the training process, and finally shows how one-shot VC is achieved.

3.1. Architecture of the VQMIVC system

As shown in Figure 1, the proposed VQMIVC system includes four modules: content encoder, speaker encoder, pitch extractor and decoder. The first three modules respectively extract content, speaker and pitch representations from the input voice; and the fourth module, the decoder, maps these representations back to acoustic features. Assuming that there are K utterances, we use mel-spectrograms as acoustic features and randomly select T frames from each utterance for training. The kth mel-spectrogram is denoted as $\mathbf{X}_k = \{\mathbf{x}_{k,1}, \mathbf{x}_{k,2}, ..., \mathbf{x}_{k,T}\}$.

Content encoder θ_c: The content encoder strives to extract linguistic content information from $\mathbf{X}_k$ by using VQCPC as shown in Figure 2, which contains two networks h-net: $\mathbf{X}_k \rightarrow \mathbf{Z}_k$ and g-net: $\hat{\mathbf{Z}}_k \rightarrow \mathbf{R}_k$, and VQ operation q: $\mathbf{Z}_k \rightarrow \hat{\mathbf{Z}}_k$. h-net takes in $\mathbf{X}_k$ to derive a sequence of dense features $\mathbf{Z}_k = \{\mathbf{z}_{k,1}, \mathbf{z}_{k,2}, ..., \mathbf{z}_{k,T/2}\}$, where the length is reduced from T to $T/2$. Then the quantizer q discretizes $\mathbf{Z}_k$ with a trainable codebook B into $\hat{\mathbf{Z}}_k = \{\hat{\mathbf{z}}_{k,1}, \hat{\mathbf{z}}_{k,2}, ..., \hat{\mathbf{z}}_{k,T/2}\}$, where $\hat{\mathbf{z}}_{k,t} \in B$ is the vector closest to $\mathbf{z}_{k,t}$. VQ imposes an information bottleneck to remove non-essential details in $\mathbf{Z}_k$, making $\hat{\mathbf{Z}}_k$ to be related with underlying linguistic information. Then the content encoder θ_c is trained by minimizing the VQ loss [9]:

$$L_{VQ} = \frac{2}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T/2} \left\| \mathbf{z}_{k,t} - sg(\hat{\mathbf{z}}_{k,t}) \right\|_2^2 \quad (1)$$

where $sg(\cdot)$ denotes the stop-gradient operator. To further encourage $\hat{\mathbf{Z}}_k$ to capture local structures, contrastive predictive coding (CPC) is employed by adding RNN based g-net taking in $\hat{\mathbf{Z}}_k$ to obtain aggregation $\mathbf{R}_k = \{\mathbf{r}_{k,1}, \mathbf{r}_{k,2}, ..., \mathbf{r}_{k,T/2}\}$. Given $\mathbf{r}_{k,t}$, the model is trained to distinguish a positive sample $\hat{\mathbf{z}}_{k,t+m}$ that is m steps in the future from negative samples drawn from the set $\Omega_{k,t,m}$ by minimizing the InfoNCE loss [31]:

$$L_{CPC} = -\frac{1}{KT'M} \sum_{k=1}^{K} \sum_{t=1}^{T'} \sum_{m=1}^{M} \log \left[\frac{\exp(\hat{\mathbf{z}}_{k,t+m}^T \mathbf{W}_m \mathbf{r}_{k,t})}{\sum_{\tilde{\mathbf{z}} \in \Omega_{k,t,m}} \exp(\tilde{\mathbf{z}}^T \mathbf{W}_m \mathbf{r}_{k,t})} \right] \quad (2)$$

where $T' = T/2 - M$, $\mathbf{W}_m$ (m=1,2,...,M) is trainable projection matrix. By predicting future samples with probabilistic contrastive loss (2), local features (e.g., phonemes) spanning many
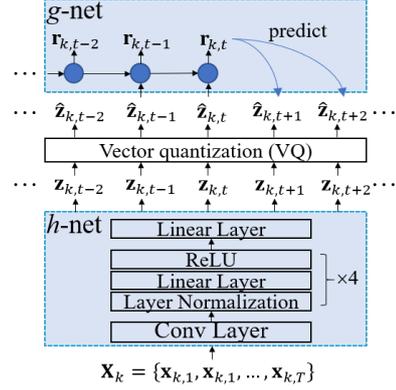

Figure 2: *Details of the VQCPC based content encoder.*

time steps are encoded into $\hat{\mathbf{Z}}_k = f(\mathbf{X}_k; \theta_c)$, which is the content representation used to accurately reconstruct the linguistic content. During training, the negative samples set $\Omega_{k,t,m}$ is formed by randomly selecting samples from the current utterance.

Speaker encoder θ_s: The speaker encoder takes in $\mathbf{X}_k$ to generate a vector $\mathbf{s}_k = f(\mathbf{X}_k; \theta_s)$, which is used as the speaker representation. $\mathbf{s}_k$ captures global speech characteristics to control the speaker identity of the generated speech.

Pitch extractor: The pitch representation is expected to contain intonation variations but exclude content and speaker information, so we extract F_0 from the waveform and perform z-normalization for each utterance independently. In our experiments, we adopt log-normalized F_0 (log-F_0) as $\mathbf{p}_k = (p_{k,1}, p_{k,2}, ..., p_{k,T})$, which is speaker-independent so that speaker encoder is forced to provide the speaker information, e.g., vocal ranges.

Decoder θ_d: The decoder is used to map the content, speaker and pitch representations to mel-spectrograms. Linear interpolation based upsampling ($\times 2$) and repetition ($\times T$) are performed on $\hat{\mathbf{Z}}_k$ and $\mathbf{s}_k$ respectively to align with $\mathbf{p}_k$ as inputs of decoder to generate mel-spectrograms $\hat{\mathbf{X}}_k = \{\hat{\mathbf{x}}_{k,1}, \hat{\mathbf{x}}_{k,2}, ..., \hat{\mathbf{x}}_{k,T}\}$. Decoder is jointly trained with content and speaker encoders by minimizing a reconstruction loss:

$$L_{REC} = \frac{1}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T} [\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_1 + \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2] \quad (3)$$

3.2. MI minimization integrated into VQMIVC training

Given the random variables $\mathbf{u}$ and $\mathbf{v}$, the MI is Kullback-Leibler (KL) divergence between their joint and marginal distributions as $\mathbf{I}(\mathbf{u}, \mathbf{v}) = D_{KL}(P(\mathbf{u}, \mathbf{v}); P(\mathbf{u})P(\mathbf{v}))$. We adopt vCLUB [29] to compute the upper bound of MI as:

$$\mathbf{I}(\mathbf{u}, \mathbf{v}) = E_{P(\mathbf{u}, \mathbf{v})} \left[\log Q_{\theta_{\mathbf{u}, \mathbf{v}}}(\mathbf{u}|\mathbf{v}) \right] - E_{P(\mathbf{u})} E_{P(\mathbf{v})} \left[\log Q_{\theta_{\mathbf{u}, \mathbf{v}}}(\mathbf{u}|\mathbf{v}) \right] \quad (4)$$

where $\mathbf{u}, \mathbf{v} \in \{\hat{\mathbf{Z}}, \mathbf{s}, \mathbf{p}\}$, $\hat{\mathbf{Z}}$, $\mathbf{s}$ and $\mathbf{p}$ are content, speaker and pitch representations respectively, $Q_{\theta_{\mathbf{u}, \mathbf{v}}}(\mathbf{u}|\mathbf{v})$ is the variational approximation of ground-truth posterior of $\mathbf{u}$ given $\mathbf{v}$ and can be parameterized by a network $\theta_{\mathbf{u}, \mathbf{v}}$. The unbiased estimation for vCLUB between different speech representations is given by:

$$\hat{\mathbf{I}}(\hat{\mathbf{Z}}, \mathbf{s}) = \frac{2}{K^2 T} \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{t=1}^{T/2} \left[\log Q_{\theta_{\hat{\mathbf{Z}}, \mathbf{s}}}(\hat{\mathbf{z}}_{k,t}|\mathbf{s}_k) - \log Q_{\theta_{\hat{\mathbf{Z}}, \mathbf{s}}}(\hat{\mathbf{z}}_{l,t}|\mathbf{s}_k) \right] \quad (5)$$

$$\hat{\mathbf{I}}(\mathbf{p}, \mathbf{s}) = \frac{1}{K^2 T} \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{t=1}^{T} \left[\log Q_{\theta_{\mathbf{p}, \mathbf{s}}}(\mathbf{p}_{k,t}|\mathbf{s}_k) - \log Q_{\theta_{\mathbf{p}, \mathbf{s}}}(\mathbf{p}_{l,t}|\mathbf{s}_k) \right] \quad (6)$$

$$\hat{\mathbf{I}}(\hat{\mathbf{Z}}, \mathbf{p}) = \frac{2}{K^2 T} \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{t=1}^{T/2} \left[\log Q_{\theta_{\hat{\mathbf{Z}}, \mathbf{p}}}(\hat{\mathbf{z}}_{k,t}|\hat{\mathbf{p}}_{k,t}) - \log Q_{\theta_{\hat{\mathbf{Z}}, \mathbf{p}}}(\hat{\mathbf{z}}_{l,t}|\hat{\mathbf{p}}_{k,t}) \right] \quad (7)$$

where $\hat{\mathbf{p}}_{k,t} = (\mathbf{p}_{k,2t-1} + \mathbf{p}_{k,2t})/2$. With a good variational approximation, (4) provides a reliable MI upper bound. Therefore,

Algorithm 1. Training process of the proposed VQMIVC

Input: mel-spectrograms $\{\mathbf{X}_k\}$, normalized log-F_0 $\{\mathbf{p}_k\}$, learning rate α and β
1. **for** each training iteration **do**
2. $\hat{\mathbf{Z}}_k \leftarrow f(\mathbf{X}_k; \theta_c)$, $\mathbf{s}_k \leftarrow f(\mathbf{X}_k; \theta_s)$
3.   Calculate log-likelihood $L_{\mathbf{u},\mathbf{v}}$ (9), then update:
4. $\theta_{\mathbf{u},\mathbf{v}} \leftarrow \theta_{\mathbf{u},\mathbf{v}} + \alpha \nabla_{\theta_{\mathbf{u},\mathbf{v}}} L_{\mathbf{u},\mathbf{v}}$, $\mathbf{u},\mathbf{v} \in \{\hat{\mathbf{Z}}, \mathbf{s}, \mathbf{p}\}$
5.   Calculate VC loss L_{VC} (10), then update:
      $\theta \leftarrow \theta - \beta \nabla_\theta L_{VC}$, $\theta \in \{\theta_c, \theta_s, \theta_d\}$
6. **end for**
7. **return** $\theta_c, \theta_s, \theta_d$

---

we can decrease the correlation among different speech representations by minimizing (5)-(7), and the total MI loss is:

$$L_{MI} = \hat{\mathbf{I}}(\hat{\mathbf{Z}}, \mathbf{s}) + \hat{\mathbf{I}}(\hat{\mathbf{Z}}, \mathbf{p}) + \hat{\mathbf{I}}(\mathbf{p}, \mathbf{s}) \tag{8}$$

During training, variational approximation networks and VC network are optimized alternatively. The variational approximation networks are trained to maximize the log-likelihood:

$$L_{\mathbf{u},\mathbf{v}} = \log Q_{\theta_{\mathbf{u},\mathbf{v}}}(\mathbf{u}|\mathbf{v}), \quad \mathbf{u},\mathbf{v} \in \{\hat{\mathbf{Z}}, \mathbf{s}, \mathbf{p}\} \tag{9}$$

while the VC network is trained to minimize VC loss:

$$L_{VC} = L_{VQ} + L_{CPC} + L_{REC} + \lambda_{MI} L_{MI} \tag{10}$$

where λ_{MI} is a constant weight to control how MI loss enhances the disentanglement. The final training process is summarized in Algorithm 1. We note that no text transcriptions or speaker labels are used during training, so the proposed approach achieves disentanglement in a fully unsupervised way.

3.3. One-shot VC

During conversion, the content and pitch representations are first extracted from source speaker's utterance $\mathbf{X}_{src}$ as $\hat{\mathbf{Z}}_{src} = f(\mathbf{X}_{src}; \theta_c)$ and $\mathbf{p}_{src}$ respectively, the speaker representation is extracted from only one target speaker's utterance $\mathbf{X}_{tgt}$ as $\mathbf{s}_{tgt} = f(\mathbf{X}_{tgt}; \theta_s)$, then the decoder generates the converted mel-spectrograms as $f(\hat{\mathbf{Z}}_{src}, \mathbf{s}_{tgt}, \mathbf{p}_{src}; \theta_d)$.

4. Experiments

4.1. Experimental setup

All experiments are conducted on the VCTK corpus [32] with 110 English speakers, which are randomly split into 90 and 20 speakers as training and testing sets respectively. The testing speakers are treated as unseen speakers that are used to perform one-shot VC. For acoustic features extraction, all audio recordings are downsampled to 16kHz, 80-dim mel-spectrograms and F_0 are both calculated with 25ms Hanning window, 10ms frame shift and 400-point fast Fourier transform.

The proposed VC network consists of the content encoder, speaker encoder and decoder. The content encoder contains a h-net, a quantizer q and a g-net. The h-net is composed of a convolutional layer with stride of 2, four blocks with layer normalization, 512-dim linear layer and ReLU activation function for each block. The quantizer contains a codebook with 512 64-dim learnable vectors. The g-net is a 256-dim unidirectional RNN layer. For CPC, the future prediction step M is 6 and the number of negative samples $|\Omega_{k,t,m}|$ is 10. The speaker encoder follows [7], which contains 8 ConvBank layers to encode the long-term information, 12 convolutional layers with 1 average-pooling layer, and 4 linear layers to derive the 256-dim speaker representation. The decoder follows [6] with a 1024-dim LSTM layer, three convolutional layers,

Table 1: *MI among content, speaker and pitch representations, where MI is estimated on all testing speakers for 10 rounds, and mean ± standard-variance are reported.*

λ_{MI}	$\hat{\mathbf{I}}(\hat{\mathbf{Z}}, \mathbf{s})$	$\hat{\mathbf{I}}(\hat{\mathbf{Z}}, \mathbf{p})$	$\hat{\mathbf{I}}(\mathbf{p}, \mathbf{s})$
0	24.65 ± 1.79	29.47 ± 0.31	0.12 ± 0.01
1e-3	8.55 ± 1.26	2.59 ± 0.03	0.06 ± 0.01
1e-2	8.00 ± 1.07	0.30 ± 0.01	**0.02 ± 0.01**
1e-1	**5.52 ± 0.58**	**0.09 ± 0.01**	0.03 ± 0.01

Table 2: *CER/WER for generated speech using speaker representations from Same and Mixed utterances by varying λ_{MI}.*

λ_{MI}	*Same*: CER / WER	*Mixed*: CER($\triangle_C$) / WER($\triangle_W$)
0	23.4% / 32.1%	35.9%(12.5%) / 59.9%(27.8%)
1e-3	12.8% / 25.6%	13.9%(1.1%) / 27.7%(2.1%)
1e-2	**12.3% / 24.9%**	**12.9%**(0.6%) / **25.9%**(1.0%)
1e-1	12.7% / 25.6%	**12.9%**(0.2%) / **25.9%**(0.3%)

two 1024-dim LSTM layers and a 80-dim linear layer. Besides, a 5-layer convolutional based Postnet is added to refine predicted mel-spectrograms, which are converted to waveform by Parallel WaveGAN vocoder [33] that is trained by VCTK corpus. The variational approximation $Q_{\theta_{\mathbf{u},\mathbf{v}}}(\mathbf{u}|\mathbf{v})$ for all MI is parameterized in Gaussian distribution as $Q_{\theta_{\mathbf{u},\mathbf{v}}}(\mathbf{u}|\mathbf{v}) = \mathcal{N}(\mathbf{u}|\mu(\mathbf{v}), diag(\sigma^2(\mathbf{v})))$ with mean $\mu(\mathbf{v})$ and variance $\sigma^2(\mathbf{v})$ inferred by a two-way fully-connected network $\theta_{\mathbf{u},\mathbf{v}}$ that is composed of four 256-dim hidden layers. The VC network is trained using Adam optimizer [34] with 15-epoch warmup increasing the learning rate from 1e-6 to 1e-3, which is halved every 100 epochs after 200 epochs until 500 epochs in total. Batch size is 256 and 128 frames are randomly selected from each utterance for training per iteration. Variational approximation networks are also trained with the Adam optimizer with a learning rate of 3e-4. We compare our proposed VQMIVC method with AutoVC [6], AdaIN-VC [7] and VQVC+ [8], which are among the state-of-the-art one-shot VC methods.

4.2. Experimental results and analysis

4.2.1. Speech representation disentanglement performance

In the VC loss (10), λ_{MI} determines the capacity of MI to enable SRD, we first vary λ_{MI} to evaluate disentanglement degrees between different speech representations extracted from all testing utterances by computing vCLUB as shown in Table 1. We can see that when λ_{MI} increases, MI tends to decrease to reduce the correlation among different speech representations.

To measure how much content information is entangled with speaker representation, we adopt two ways to generate the speech, i.e., (1) *Same*, i.e., the content, speaker and pitch representations of the same utterance are used to generate the speech; (2) *Mixed*, i.e., the content and pitch representations of one utterance and speaker representation of another utterance are used to generate the speech, both utterances belong to the same speaker. Then an automatic speech recognition (ASR) system is used to obtain character/word error rate (CER/WER) of the generated speech. The increased CER and WER from '*Same*' to '*Mixed*' are denoted as $\triangle_C$ and $\triangle_W$ respectively. As the only difference in inputs for speech generation is that of speaker representation, we can conclude that larger values of $\triangle_C$ and $\triangle_W$ reflect that more content information is leaked to speaker representation. All testing speakers are used for speech generation, and the publicly released Jasper-based ASR system [35] is used. The results are shown in Table 2, we can see that when MI is

Table 3: *Speaker classification accuracy on $\hat{Z}$ (content) and s (speaker), and prediction loss for p (pitch) inferred by $\hat{Z}$.*

λ_{MI}	$\hat{Z}$-accuracy	s-accuracy	p-loss
0	9.7%	**100%**	0.279
1e-3	9.5%	99.7%	0.284
1e-2	9.4%	99.5%	0.287
1e-1	**9.3%**	98.1%	**0.289**

Table 4: *ASR and F_0-PCC results for one-shot VC.*

Methods	CER	WER	F_0-PCC
Source (Oracle)	3.5%	9.0%	1.0
AutoVC	15.7%	30.5%	0.455
AdaIN-VC	27.1%	47.1%	0.346
VQVC+	35.5%	59.5%	0.237
VQMIVC (proposed)	**14.9%**	**29.3%**	**0.781**
w/o MI (proposed)	38.0%	62.1%	**0.781**

not used (λ_{MI}=0), the generated speech is severely contaminated by the undesired content information that resides in the speaker representations as indicated by the largest values of $\triangle_C$ and $\triangle_W$. However, when MI is used (λ_{MI}>0), significant reductions of $\triangle_C$ and $\triangle_W$ can be obtained. As λ_{MI} increases, both $\triangle_C$ and $\triangle_W$ decrease, showing that higher λ_{MI} can, to a larger degree, alleviate leakage of content information into the speaker representation.

In addition, we design two speaker classifiers, taking $\hat{Z}$ and s as inputs respectively; and one predictor, taking in $\hat{Z}$ to infer p. The classifiers and predictor are all 4-layer fully-connected network with 256-dim hidden size. Higher speaker classification accuracy denotes more speaker information in $\hat{Z}$ or s, while higher prediction loss (mean square error) for p denotes less pitch information in $\hat{Z}$. The results are shown in Table 3. We can observe that $\hat{Z}$ contains less speaker and pitch information when λ_{MI} increases to achieve lower accuracy and higher pitch loss. Speaker classification accuracy on s is high for all λ_{MI}, while the accuracy decreases when λ_{MI} increases, showing that s contains abundant speaker information, but higher λ_{MI} tends to make s lose speaker information. To ensure proper disentanglement, we set λ_{MI} to 1e-2 for the following experiments.

4.2.2. Content preservation and F_0 variation consistency

To evaluate whether the converted voice maintains linguistic content and intonation variations of the source voice, we test the CER/WER of the converted speech and calculate the Pearson correlation coefficient (PCC) [36] between F_0 of source and converted voice. PCC ranges from -1 to 1 and can be effectively used to measure the correlation between two variables, where a higher F_0-PCC denotes that the converted voice has higher F_0 variation consistency with the source voice. 10 testing speakers are randomly selected as the source speakers, and the remaining 10 testing speakers are treated as target speakers, which leads to 100 conversion pairs where all source utterances are used for conversion. The results for different methods are shown in Table 4, where the results for source speech are also reported as the performance upper bound. It can be seen that VQMIVC achieves the lowest CER and WER among all methods, which shows the robustness of the proposed VQMIVC method to preserve the source linguistic content. Besides, we observe that ASR performance drops significantly without using MI (w/o MI), as the converted voice is contaminated by undesired content information entangled with speaker representations. In addition, by providing source pitch representations, we can explic-
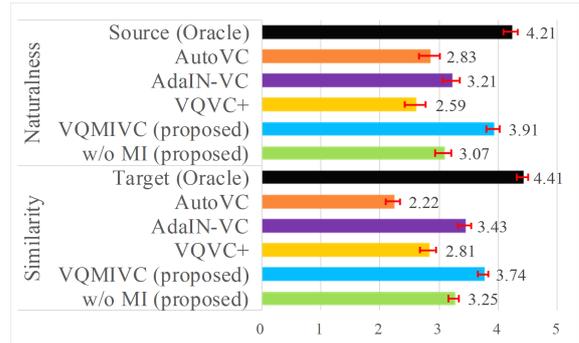

Figure 3: *Comparison results of MOS with 95% confidence intervals for speech naturalness and speaker similarity.*

itly and effectively control intonation variations of the converted voice to achieve high F_0 variation consistency, as indicated by largest F_0-PCC of 0.781 obtained by the proposed methods.

4.2.3. Speech naturalness and speaker similarity

Subjective tests are conducted by 15 subjects to evaluate the speech naturalness and speaker similarity, in terms of 5-point mean opinion score (MOS), i.e, 1-bad, 2-poor, 3-fair, 4-good, 5-excellent. We randomly select two source speakers and two target speakers from the testing speakers, each source or target set contains one male and one female speaker, which results in 4 conversion pairs, where 18 converted utterances from each pair are evaluated by each subject. The scores are averaged across all pairs and reported in Figure 3. Source (Oracle) and Target (Oracle) denote that the speech is synthesized with ground-truth mel-spectrograms of source and target utterances by Parallel WaveGAN respectively. We observe that the proposed method, denoted as 'w/o MI', outperforms AutoVC and VQVC+, but is inferior to AdaIN-VC. Pronunciation errors are frequently detected in the converted voice by 'w/o MI' via our official listening tests, which can be reflected by the high CER/WER of 'w/o MI' in Table 4. These issues can be greatly alleviated by the proposed MI minimization, where improved speech naturalness and speaker similarity are achieved. This indicates that MI minimization facilitates proper SRD to derive accurate content representation and effective speaker representation, which can be used to generate the natural speech with high voice similarity to the target speaker.

5. Conclusions

We propose a novel approach by combining VQCPC and MI for unsupervised SRD-based one-shot VC. To achieve proper disentanglement of content, speaker and pitch representations, VC model is not only trained to minimize the reconstruction loss, but also VQCPC loss to explore local structures of speech for content, and MI loss to reduce the correlation between different speech representations. Experiments verify the efficacy of proposed methods to mitigate information leakage issues by learning accurate content representation to preserve source linguistic content, speaker representation to capture desired speaker characteristics, and pitch representation to retain source intonation variations, which results in high-quality converted voice.

6. Acknowledgements

7. References

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

[2] D. Rentzos, S. Vaseghi, E. Turajlic, Q. Yan, and C.-H. Ho, "Transformation of speaker characteristics for voice conversion," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*. IEEE, 2003, pp. 706–711.

[3] K. Oyamada, H. Kameoka, T. Kaneko, H. Ando, K. Hiramatsu, and K. Kashino, "Non-native speech conversion with consistency-aware recursive network and generative adversarial network," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 182–188.

[4] S. Liu, J. Zhong, L. Sun, X. Wu, X. Liu, and H. Meng, "Voice conversion across arbitrary speakers based on a single target-speaker utterance." in *Interspeech*, 2018, pp. 496–500.

[5] H. Lu, Z. Wu, D. Dai, R. Li, S. Kang, J. Jia, and H. Meng, "One-shot voice conversion with global speaker embeddings." in *Interspeech*, 2019, pp. 669–673.

[6] 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*. PMLR, 2019, pp. 5210–5219.

[7] J.-c. Chou and H.-Y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *Interspeech*, pp. 664–668, 2019.

[8] D.-Y. Wu, Y.-H. Chen, and H.-y. Lee, "Vqvc+: One-shot voice conversion by vector quantization and u-net architecture," *Interspeech*, pp. 4691–4695, 2020.

[9] B. van Niekerk, L. Nortje, and H. Kamper, "Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge," *Interspeech*, pp. 4836–4840, 2020.

[10] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

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

[12] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[13] D. Erro, A. Moreno, and A. Bonafonte, "Inca algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2009.

[14] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.

[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," *arXiv preprint arXiv:1704.00849*, 2017.

[16] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[17] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[18] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6284–6288.

[19] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7836–7846.

[20] Y.-H. Chen, D.-Y. Wu, T.-H. Wu, and H.-y. Lee, "Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5954–5958.

[21] 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.

[22] B. Gierlichs, L. Batina, P. Tuyls, and B. Preneel, "Mutual information analysis," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2008, pp. 426–442.

[23] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[24] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 297–304.

[25] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *International Conference on Machine Learning*, 2018, pp. 531–540.

[26] M. Ravanelli and Y. Bengio, "Learning speaker representations with mutual information," *Interspeech*, pp. 1153–1157, 2019.

[27] Y. Kwon, S.-W. Chung, and H.-G. Kang, "Intra-class variation reduction of speaker representation in disentanglement framework," *Interspeech*, pp. 3231–3235, 2020.

[28] T.-Y. Hu, A. Shrivastava, O. Tuzel, and C. Dhir, "Unsupervised style and content separation by minimizing mutual information for speech synthesis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3267–3271.

[29] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "Club: A contrastive log-ratio upper bound of mutual information," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1779–1788.

[30] S. Yuan, P. Cheng, R. Zhang, W. Hao, Z. Gan, and L. Carin, "Improving zero-shot voice style transfer via disentangled representation learning," *arXiv preprint arXiv:2103.09420*, 2021.

[31] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[32] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[33] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

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

[35] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," *Interspeech*, pp. 71–75, 2019.

[36] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.