Leveraging Content-based Features from Multiple Acoustic Models for Singing Voice Conversion

Xueyao Zhang[1] **Yicheng Gu**[1] **Haopeng Chen**[1] **Zihao Fang**[1] **Lexiao Zou**[2]
Liumeng Xue[1] **Zhizheng Wu**[1]
[1]School of Data Science, The Chinese University of Hong Kong, Shenzhen
{xueyaozhang,yichenggu,haopengchen1,zihaofang}@link.cuhk.edu.cn
{xueliumeng, wuzhizheng}@cuhk.edu.cn
[2]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
200110803@stu.hit.edu.cn

Abstract

Singing voice conversion (SVC) is a technique to enable an arbitrary singer to sing an arbitrary song. To achieve that, it is important to obtain speaker-agnostic representations from source audio, which is a challenging task. A common solution is to extract content-based features (e.g., PPGs) from a pretrained acoustic model. However, the choices for acoustic models are vast and varied. It is yet to be explored what characteristics of content features from different acoustic models are, and whether integrating multiple content features can help each other. Motivated by that, this study investigates three distinct content features, sourcing from WeNet, Whisper, and ContentVec, respectively. We explore their complementary roles in intelligibility, prosody, and conversion similarity for SVC. By integrating the multiple content features with a diffusion-based SVC model, our SVC system achieves superior conversion performance on both objective and subjective evaluation in comparison to a single source of content features. Our demo page and code can be available here.

1 Introduction

Singing Voice Conversion (SVC) aims to convert the voice of a singing signal to the voice of a target singer without changing the underlying content and melody [8]. It can enable an arbitrary singer to sing an arbitrary song. SVC can be used for various applications including music entertainment, singing voice beautification, vocal education, and art creation.

In recent years, conducting singing voice conversion with non-parallel data [1, 10, 19, 22–24, 34] attracts more attention, since it does not rely on the scarce parallel singing voice corpus. Figure 1 displays the common pipeline for the non-parallel SVC. Its main idea is to first disentangle the speaker-agnostic representations from the source audio, and then inject the desired speaker information to synthesize the target (usually by an acoustic decoder and a subsequent waveform synthesizer). For modeling speaker-agnostic representations, the common solutions are to utilize content-based features from a pretrained acoustic model, such as extracting PPGs (or BNFs) from an Automatic Speech Recognition (ASR) model [1, 19, 22, 23] or a self-supervised model [9, 10].

With the rapid development of self-supervised learning and speech recognition oriented acoustic models, the choices for acoustic models are vast and varied. However, the characteristics of different content features during singing voice conversion are yet investigated. For example, in the voice conversion (VC), some researchers find that content features from the CE-loss-trained ASR model could be good at modeling prosody, while those from the CTC-loss-trained ASR model could be skilled in disentangling timbre [43]. Compared with speech, singing voice owns a wider range

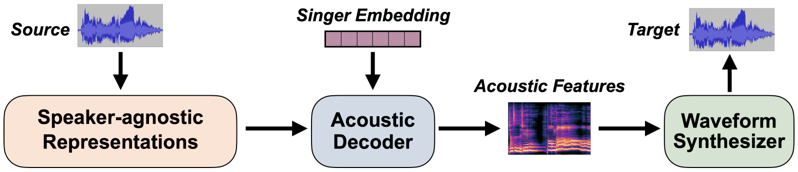


Figure 1: A classic pipeline of the non-parallel singing voice conversion system.

Table 1: A summary of WeNet, Whisper and ContentVec.

Model	Pretrained Task	Training Objective	Pretrained Data	Architecture
WeNet [39]	ASR, *supervised* by linguistic labels	CTC, Next Token Prediction (character-level)	Text-only transcription from Gigaspeech/Wenetspeech (10k hours English /Chinese speech)	Encoder-decoder Conformer [3] or Transformer [33]
Whisper [27]	Multitask including multilingual ASR, speech translation, spoken language identification, etc., large-scale *weak supervision*	Next Token Prediction (byte-level [29])	680k hours multilingual data, including both text-only and time-aligned transcription	Encoder-decoder Transformer [33]
ContentVec [26]	*Self-supervised* learning which conditions on disentangling speakers	Masked Token Prediction (frame-level)	Librispeech (1k hours English), using only speech but no any transcription	Encoder-only Transformer [33]

of pitch and energy variations and more diverse timbres (Section 4.3). However, there are some questions unclear: Which kind of content features do we need for SVC? Would integrating multiple content features help each other during conversion?

Motivated by that, this study aims to explore the specific roles of different content features. To obtain the content features, the existing acoustic model is designed to make the model's latent representations align with linguistic information (such as characters or phonemes). In other words, the content information of the features depends on *the level of linguistic supervision* which the acoustic models are pretrained with. Based on that, we investigate the three distinct content features (Table 1), sourcing respectively from:

- **WeNet** [39]: the acoustic model is by *supervised* trained on linguistic labels (i.e., characters). It is pretrained by text-only transcription data;
- **Whisper** [27]: the acoustic model is under *weak supervision* by large-scale multitask data. It is pretrained by both text-only and time-aligned transcription data;
- **ContentVec** [26]: the acoustic model conducts *self-supervised learning* which conditions on disentangling speakers. It is pretrained by only speech data and there is no explicit linguistic supervision.

We perform a systematic analysis with the three acoustic models and assess their performance in terms of prosody, intelligibility, and conversion similarity for SVC. Moreover, we find that integrating different content features could be compatible and complementary, which can further improve the singing voice conversion performance.

In addition to the exploration of content features, this paper also presents our efforts in improving the waveform synthesizer (i.e., Vocoder) for a high-quality singing voice. Specifically, we conduct a comprehensive analysis of the characteristics between speech and singing voice (Section 4.3). By using a pretrained speech BigVGAN [17] vocoder as our foundation model, we research why fine-tuning it by singing voice data can be helpful, and illustrate what the specific benefits are. Based on the above, we present an SVC system that contains a diffusion-based [23] acoustic decoder that uses multiple content features and our fine-tuned BigVGAN singing voice vocoder. The experimental results of objective and subjective evaluation verify our proposed system is superior to the existing state-of-the-art SVC methods.

2 Related Work

2.1 Singing Voice Conversion

The early singing voice conversion researches aim to design parametric statistical models such as HMM [32] or GMM [13, 14] to learn the spectral features mapping of the parallel data. Since the parallel singing voice corpus is hard to collect on a large scale, the non-parallel SVC [1, 24], or recognition-synthesis [9] SVC, becomes popular in recent years. To obtain speaker-agnostic features for the non-parallel SVC, extracting PPGs from pretrained ASR models as content-based features is one of the most common methods [1, 19, 22, 23]. Some researchers also propose to use the low quefrencies of MFCC as the content features [42]. Besides obtaining content features from supervised acoustic models, extracting BNFs from the self-supervised model [9, 10] has also been adopted. Researchers of [34] point out that self-supervised features like HuBERT [5] can also supply the acoustic reference to improve the prosody for SVC. However, most existing works utilize a single source of content features. This study aims to research whether integrating multiple content features would be complementary to singing voice conversion.

2.2 Singing Voice Vocoder

The singing voice has broader and more diverse frequency bands than speech, making it hard to adopt a speech vocoder in the singing voice domain explicitly. The existing works of singing voice vocoders devote to tackling the synthesis challenges in both time and frequency domains. On the one hand, to reconstruct the spectrogram precisely in the frequency domain, [6] proposes to generate audio in a multi-band process, which allows the generator to pay attention to different frequency bands adaptively. HPG [31] adopts anti-aliasing techniques to improve the reconstruction of high-frequency parts for vocoders. On the other hand, to keep the synthesized audios consistent in the time domain and avoid the "glitches", SingGAN [7] and NSF-HiFiGAN [35] introduce the sine excitation modeling for F0, which can stabilize the auditory quality. SawSing [37] suggests modeling phase continuities explicitly. In a word, the existing works focus more on improving the representation or the architecture of vocoders. However, this study makes efforts from the data perspective, which explores how singing voice data can improve a speech vocoder and what the specific benefits are.

3 Methodology

The proposed SVC framework that can leverage multiple content features is illustrated in Figure 2. Note that the multiple content features can be plugged into conversion model of any architecture. The diffusion-based conversion model here is just to exemplify, whose architecture is same as [23]. After obtaining the predicted mel spectrograms, we utilize a waveform synthezier, which is a BigVGAN vocoder [17] finetuned by the singing voice data (Section 3.3), to get the synthesized audio.

3.1 Features Extraction and Multiple Content Features Fusion

We utilize the multiple content features (WeNet, Whisper, and ContentVec), the acoustic/musical features (F0 and Energy), and the speaker features (Speaker ID) for singing voice conversion. The key challenge is how to extract and then fuse them, especially for the fusion of multiple content features sourcing from different pretrained models. We will describe it in details in this section.

3.1.1 Features Extraction

For the content features, the WeNet, Whisper, and ContentVec are all designed as the Transformer-like [33] architecture. We utilize the encoder's output as the content features. Formally, given the utterance u and the acoustic model $\mathcal{M}$, the extracted content features $\mathbf{c}_{\mathcal{M}} = \mathcal{M}(u) \in \mathbb{R}^{T_{\mathcal{M}} \times d_{\mathcal{M}}}$, where $\mathcal{M}$ can be either WeNet, Whisper, or ContentVec, $T_{\mathcal{M}}$ is the length of frames, and $d_{\mathcal{M}}$ is the latent representation dimension of the model $\mathcal{M}$.

For the acoustic/musical features, we follow [23] to obtain the quantized F0 and energy features. We adopt the trainable embedding layers to get the F0 embeddings $\mathbf{f} \in \mathbb{R}^{T \times d}$ and energy embeddings $\mathbf{e} \in \mathbb{R}^{T \times d}$, where d is set as 384 in our experiments.

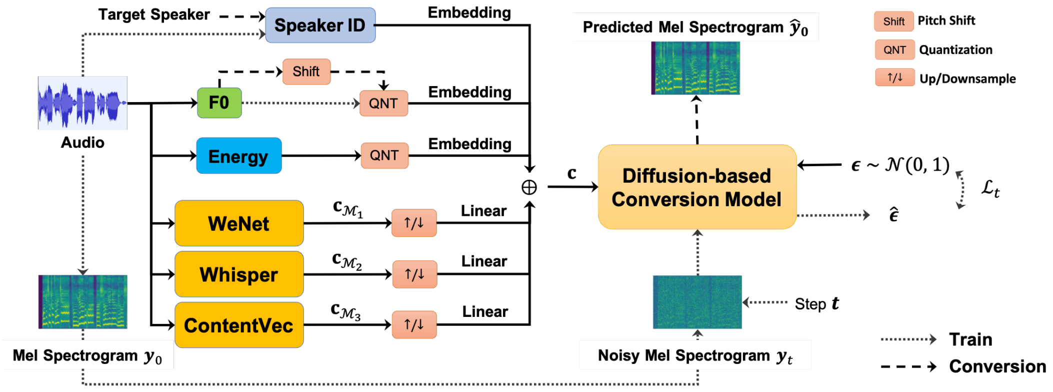


Figure 2: Diffusion-based SVC framework that leverages multiple content features.

For the speaker identity, we adopt a look-up table and use a trainable embedding layer to learn the speaker embeddings $\mathbf{s} \in \mathbb{R}^d$.

3.1.2 Features Fusion

Usually, the different pretrained model that could produce content features owns different acoustic parameters (such as sampling rate and frameshift) [26, 27, 39]. As a result, the content feature $\mathbf{c}_{\mathcal{M}}$ sourcing from the different model $\mathcal{M}$ will have the different frame length $T_{\mathcal{M}}$, which makes it challenging to fuse the multiple content features of different sources.

To address it, when fusing multiple content features, we first downsample (or upsample) the content features $\mathbf{c}_{\mathcal{M}}$ to $\tilde{\mathbf{c}}_{\mathcal{M}} \in \mathbb{R}^{T \times d_{\mathcal{M}}}$ based on their frameshift parameters. Subsequently, we use the linear layer to project $\tilde{\mathbf{c}}_{\mathcal{M}}$ into $\hat{\mathbf{c}}_{\mathcal{M}} \in \mathbb{R}^{T \times d}$, whose dimension is same to the acoustic/musical features. Finally, we adopt the adding fusion strategy to merge all kinds of features:

$$\mathbf{c} = \hat{\mathbf{c}}_{\mathcal{M}_1} \oplus \hat{\mathbf{c}}_{\mathcal{M}_2} \oplus \hat{\mathbf{c}}_{\mathcal{M}_3} \oplus \mathbf{f} \oplus \mathbf{e} \oplus \hat{\mathbf{s}} \tag{1}$$

where $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ means WeNet, Whisper, and ContentVec. $\hat{\mathbf{s}} \in \mathbb{R}^{T \times d}$ represents the frame-level speaker feature, which is obtained by repeating $\mathbf{s}$ as T times. $\oplus$ means the element-wise adding operation.

3.2 Diffusion-based Conversion Model

The diffusion-based model has been proven effective in singing voice generation related tasks [20, 23, 31]. To investigate the effectiveness of different content features and their integration for SVC, we employ a diffusion model similar to DiffSVC [23] to be the conversion model.

During training, the main idea of diffusion model is to learn a noise predictor based on specific conditions [4, 23]. In this study, we adopt the fused features $\mathbf{c}$ as the conditions, and consider the mel spectrogram as the objective ($\mathbf{y}_0$) to train the diffusion-based conversion model. Following [23], we adopt a noise predictor like WaveNet [25] architecture. At every diffusion step t, given the noise predictor F, we first sample a Gaussian noise ϵ and then minimize the MSE loss:

$$\mathcal{L}_t = \mathbf{MSE}(\hat{\epsilon}, \epsilon), \text{ where } \hat{\epsilon} = F(t, \mathbf{c}, \mathbf{y}_t), \tag{2}$$

where $\mathbf{y}_t$ is the noisy acoustic features at the step t.

During conversion, we obtain the predicted mel spectrogram by giving the diffusion model a random Gaussian noise ϵ and the conversion conditions $\mathbf{c}$. To obatain $\mathbf{c}$, given any source audio, we extract its content and energy features and leave them unchanged. To convert the speaker identity, we inject a target speaker ID (which is seen in the training) to obtain the speaker features. For F0 features, we conduct the musical key transposition[1] to make the target singer could singing the source song in his vocal range. Specifically, following the baseline B01 of [8], we shift the source F0 features by

[1] https://en.wikipedia.org/wiki/Transposition_(music)

multiplying a factor, which is computed as the ratio between the F0 median of the target speaker's training data and the median of the source audio.

3.3 Waveform Synthesizer

To train a high-quality singing voice vocoder, we leverage a pretrained BigVGAN [17] speech vocoder and fine-tune it with singing voice data. The reasons why we can develop a singing voice vocoder from a pretrained speech one are two folds. On the one hand, speech and singing voice are both human sounds, whose waveform structures are homogeneous [30]. Besides, compared with speech, high-quality and copyrighted singing voice data is scarcer and harder to access. Therefore, conducting transfer learning from a pretrained speech vocoder to a singing voice vocoder is viable and will save financial and computational costs. On the other hand, compared with speech, there is a wider range of pitch and energy and more diverse timbres in the singing voice (Section 4.3). A vocoder trained by only speech data will produce problems like glitches and poor frequency reconstruction when synthesizing for singing voices. In Section 4.3, we will display the specific improvements after fine-tuning BigVGAN with singing voice data.

4 Experiments

We conduct experiments to answer the following evaluation questions:

- **EQ1**: How effective are the different content features? Could they help each other after integrating?
- **EQ2**: Could singing voice data improve a pretrained speech vocoder and why? If it does promote, what are the specific benefits after fine-tuning?
- **EQ3**: How effective is the proposed system compared to the existing SVC approaches?

4.1 Experimental Setup

4.1.1 Evaluation Tasks

We adopt three conversion tasks to evaluate: one is the conversion from M4Singer [41] to Opencpop [36]. We randomly sample 100 utterances from M4Singer [41] as source audios. The other two are provided by the Singing Voice Conversion Challenge (SVCC) 2023[2] [8], including an in-domain and a cross-domain conversion. There are 48 source utterances for each task. For the training data, we utilize five datasets: Opencpop [36], SVCC training data [8], VCTK [38], OpenSinger [6], and M4Singer [41]. There are 83.1 hours speech and 87.2 hour singing data in total.

4.1.2 Evaluation Metrics

For objective evaluation, following [8], we adopt Mel-cepstral distortion (MCD) [16], F0 Pearson correlation coefficient (F0CORR), F0 Root Mean Square Error (F0RMSE), Character Error Rate (CER) which is obtained with the recognition results of Whisper [27], and D_{embed} which estimates the speaker similarity by using RawNet-3 [11] as the objective metrics. For subjectivce evalution, we invite 12 volunteers who are experienced in the audio generation areas to conduct the Mean Opinion Score (MOS) evaluation in terms of naturalness and similarity. The naturalness score ranks from 1 ("Bad") to 5 ("Excellent"), and the similarity score ranks from 1 ("Different speaker, sure") to 4 ("Same speaker, sure").

4.1.3 Baselines

We adopt two existing SVC approaches as the baselines:

- **FastSVC** [22]: It is an official baseline provided by SVCC 2023. It utilizes a pretrained conformer [3] to extract content features, which is similar to WeNet. We adopt the official pretrained model to conduct conversion.
- **SoVITS**: It is one of the most popular open-source project in the SVC area. It utilizes ContentVec as the content features and uses a flow-based conversion model.

[2] http://vc-challenge.org/

Table 2: Statistics of the fine-tuned datasets for our singing voice vocoder. VCTK and SVCC (cross-domain) are two speech datasets, while the others are singing voice datasets.

Dataset	**Language**	**#Hours**	**#Utterances**	**#Speakers**
VCTK [38]	English	82.9	33,971	109
SVCC (cross-domain) [8]	English	0.2	311	2
OpenSinger [6]	Chinese	51.9	43,075	74
Popbutfy [21]	English	30.7	28,971	24
M4Singer [41]	Chinese	29.7	20,896	19
PopCS [20]	Chinese	5.9	1,651	1
Opencpop [36]	Chinese	5.2	3,756	1
CSD [2]	English, Korean	4.1	2,864	1
SVCC (in-domain) [8]	English	0.4	309	2
PJS [15]	Japanese	0.4	291	1
Total (Speech / Singing Voice)		83.1 / 128.3	34.2k / 101.8k	111 / 123

Table 3: Objective evaluation results of different content features and their integration. D_{embed} represents an objective metric for speaker similarity. The best and the second best results of every column (except those from Ground Truth) are **bold** and *italic*.

Content Features	**MCD** ($\downarrow$)	**F0CORR** ($\uparrow$)	**F0RMSE** ($\downarrow$)	**CER** ($\downarrow$)	D_{embed} ($\uparrow$)
Ground Truth	0.000	1.000	0.0	12.9%	1.000
WeNet	10.324	0.203	423.4	38.2%	0.743
Whisper	*8.229*	0.524	297.3	18.9%	0.858
ContentVec	8.972	0.491	361.0	22.1%	0.809
WeNet + Whisper	8.345	0.540	284.2	*16.8%*	0.856
WeNet + ContentVec	8.870	0.525	329.5	19.9%	0.758
Whisper + ContentVec	**8.201**	*0.548*	*279.6*	16.9%	**0.886**
WeNet + Whisper + ContentVec	8.249	**0.572**	**278.5**	**16.1%**	*0.865*

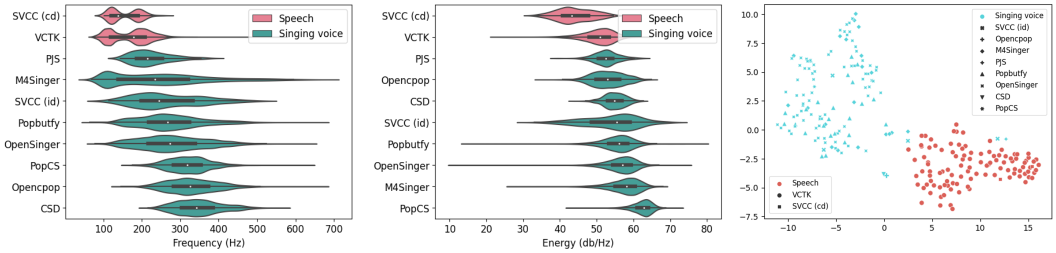
4.1.4 Implementation Details

For WeNet, we use the official models pretrained by 10k hours Wenetspeech and 10k hours Gigaspeech to extract content features. For Whisper, we use the multilingual MEDIUM model. For ContentVec, we use the official 500-CLASS model preatrained by 1k hours Librispeech. For BigVGAN vocoder, we adopt the offical 24KHZ_100BAND model that pretrained by 585 hours LibriTTS [40]. We fine-tune it with the five training datasets and additional four datasets, in total 83.1 hours of speech and 128.3 hours of singing voice data (Table 2). The sampling rate is 24kHz, the mel bands are 100, and the hop size is 256.

4.2 Performance of Different Content Features (EQ1)

To evaluate the effectiveness of different content features *alone* for singing voice conversion, in this section, we only use the content features as the conditions of the diffusion-based conversion model. We choose Opencpop [36] as the training data (target speaker) and use M4Singer [41] as source audios. The experimental results are illustrated in Table 3.

It can be observed that: (1) When using single source of content features, Whisper outperforms the others among all the metrics. The multitask training and the large-scale weak supervision make it contain comprehensive speaker-agnostic information. Besides, ContentVec behaves better than WeNet although it uses less training data. It indicates that under speech data pertaining, self-supervised representations could be more robust and generalized to other domains like singing voice. (2) During integration, sometimes introducing WeNet will be harmful to the spectrogram reconstruction and speaker similarity. We speculate that the WeNet is only pretrained on speech, which is not well robust for the singing voice. (3) Integrating multiple content features could be helpful in most cases (see the representative samples here). It reveals their complementary roles in spectrogram reconstruction, prosody, intelligibility, and speaker disentanglement.



(a) F0

(b) Energy

(c) LTAS (t-SNE)

Figure 3: The distributions of F0, energy, and LTAS (t-SNE results) for speech and singing voice. Compared with speeches, the singing voices own a larger and wider range of F0 and energy. The singing timbres are also distinct and more varied.

Table 4: Improvements after fine-tuning a pretrained speech vocoder by singing voice data. The "before/after" means the results before and after fine-tuning.

Dataset	**MCD (↓)**	**F0RMSE (↓)**	**PESQ (↑)**	**FAD (↓)**
VCTK	2.11 / 0.96	355.6 / 169.0	4.16 / 4.19	1.27 / 0.01
SVCC (cd)	1.90 / 1.33	96.1 / 21.1	3.81 / 3.87	1.46 / 0.28
OpenSinger	2.27 / 1.15	69.5 / 18.5	3.90 / 4.03	0.53 / 0.03
Popbutfy	2.11 / 0.96	56.3 / 52.5 °	4.01 / 4.12	0.41 / 0.18
M4Singer	2.32 / 1.01	62.9 / 17.1	3.97 / 4.15	0.64 / 0.03
PopCS	2.38 / 1.34	143.8 / 37.3	3.98 / 4.12	0.53 / 0.03
Opencpop	3.14 / 1.36	48.1 / 8.8	3.83 / 4.02	0.36 / 0.02
CSD	2.96 / 1.37	97.8 / 33.3	3.71 / 3.86	0.43 / 0.05
SVCC (id)	2.37 / 1.24	82.7 / 21.1	3.84 / 3.95	0.33 / 0.24
PJS	2.09 / 0.86	176.8 / 47.1	3.88 / 4.00	0.31 / 0.01

[*] All the improvements are significant (p-value is less 0.01) except for °

4.3 Effect of Singing Voice Data for Vocoder (EQ2)

To investigate the similarity or distinction between singing voice and speech, we conduct a preliminary data analysis from the aspects of F0, energy, and speaker timbre. Specifically, we analyze the differences between singing voice and speech in the range of F0 and energy, and the distribution of the Long-Term Average Spectrum (LTAS). Note that LTAS represents the formant and resonance features of the human vocal tract, which can reveal the timbre characteristics [18]. We use the ten datasets of Table 2 for analysis. In Figure 3, we observe that there are larger and wider ranges of F0 and energy in singing voices. Besides, singing timbres distribute distinctly from speaking ones and are more diverse. Given these specific features of the singing voice, it is necessary to explore whether fine-tuning a pretrained speech vocoder with singing voice data can be helpful.

Furthermore, we adopt MCD, F0RMSE, PESQ [28], and FAD [12] as metrics to explore the performance change after fine-tuning speech vocoder with the singing voice data. For VCTK, OpenSinger, Popbutfy, and M4Singer, we select a random speaker for each dataset as the unseen speakers for testing. For SVCC and the other four single-speaker datasets, we randomly sample 5% utterances for each dataset for testing. All the other data is used for training. The results is illustrarted in Table 4. It indicates that the fine-tuned vocoder gets promoted among spectrogram reconstruction (MCD), pitch modeling (F0RMSE), auditory quality (PESQ), and perceptual audio quality (FAD).

Besides of the quantitative results, we also conduct a qualitative analysis to investigate the specific benefits. For example, in Figure 4, the reconstruction for high-frequency bands is improved after fine-tuning. We provide more cases on our demo page, which displays the problems of the original speech vocoder (such as glitches) and how much fine-tuning can alleviate them.

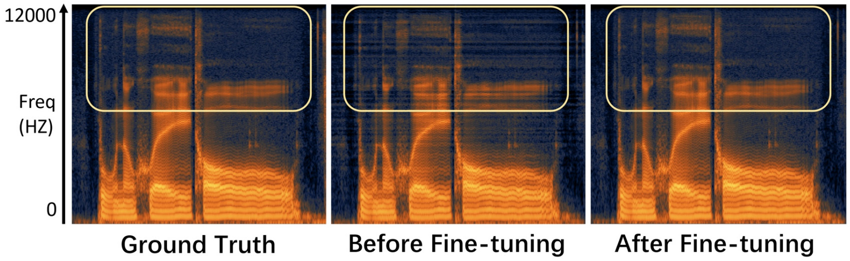


Figure 4: The mel spectrograms before and after fine-tuning. It can be observed that the fine-tuned vocoder can reconstruct the energies of the high-frequency bands better.

Table 5: Objective evaluation results of different systems. Ours (W+C) and Ours (W+W+C) represent our systems which use Whisper + ContentVec and WeNet + Whisper + ContentVec as content features repesctively.

Task	System	F0CORR (↑)	F0RMSE (↓)	CER (↓)	D_{embed} (↑)
M4Singer ↓ **Opencpop**	Ground Truth	1.000	0.0	12.9%	1.000
	FastSVC	0.636	273.9	49.2%	0.410
	SoVITS	0.801	**166.4**	17.5%	0.622
	Ours (W+C)	**0.804**	175.7	**13.8%**	**0.744**
	Ours (W+W+C)	0.801	167.7	16.6%	0.714
SVCC **(in-domain)**	Ground Truth	1.000	0.0	10.5%	1.000
	FastSVC	0.799	232.2	38.0%	0.448
	SoVITS	0.966	**57.6**	17.0%	0.725
	Ours (W+C)	**0.968**	59.2	17.7%	**0.891**
	Ours (W+W+C)	0.954	60.4	**16.3%**	0.869
SVCC **(cross-domain)**	Ground Truth	1.000	0.0	7.2%	1.000
	FastSVC	0.526	441.2	47.2%	0.341
	SoVITS	0.966	**52.1**	17.1%	0.746
	Ours (W+C)	0.967	57.1	15.1%	0.795
	Ours (W+W+C)	**0.970**	52.7	**14.5%**	**0.799**

4.4 Performance of Our Proposed System (EQ3)

Based on the explorations for multiple content features and singing voice fine-tuned vocoder, we further research the effectiveness of our whole SVC system.

4.4.1 Objective Evaluation

We adopt F0CORR, F0RMSE, CER, and D_{embed} as the metrics to objectively evaluate the prosody, intelligibility, and conversion similarity of different systems.

In Table 5, it indicates that: (1) In general, our proposed multiple content features based diffusion model outperforms FastSVC and SoVITS, especially in the aspects of intelligibility (CER) and conversion similarity (D_{embed}); (2) Comparing between the two in-domain tasks of different languages – SVCC (in-domain) in English, and M4Singer → Opencpop in Mandarin, we observe that the CER performances are comparable, while the F0 modeling and speaker disentanglement are worse for M4Singer → Opencpop. We speculate that modeling pitches for the tone language like Mandarin could be harder than the intonation language like English for SVC; (3) Comparing between SVCC in-domain and cross-domain tasks, on the one hand, the CER of cross-domain conversion is lower. The reason could be that content features are pretrained with more speech data than the singing voice, resulting in better intelligibility. On the other hand, the conversion similarity of cross-domain is worse, which is caused by the distinct distribution between singing voice timbres and speech timbres just like shown in Figure 3c; (4) In comparison between Table 3 and Table 5, we observe that although introducing the auxiliary F0 features improves prosody significantly, it can also make the speaking style of the converted audio near to the source, resulting in a worse conversion similarity to the target.

Table 6: MOS evaluation results (with 95% confidence interval) on SVCC in-domain task. The full scores of Naturalness and Similarity are 5 and 4.

System	**Naturalness** (↑)	**Similarity** (↑)
Ground Truth	4.67 ± 0.18	3.17 ± 0.29
FastSVC	1.02 ± 0.04	1.23 ± 0.15
SoVITS	2.98 ± 0.31	2.74 ± 0.31
Ours (Whisper + ContentVec)	**3.52 ± 0.26**	**2.95 ± 0.30**
Ours (WeNet + Whisper + ContentVec)	3.48 ± 0.29	2.56 ± 0.35

4.4.2 Subjective Evaluation

As is described in Section 4.1, we invite 12 volunteers who are experienced in the audio generation areas to conduct MOS evaluation on SVCC in-domain task. For naturalness (or similarity) evaluation, every subject rates 4 randomly sampled utterances (or utterance pairs) for every system. Given the five systems of Table 6, a total of 20 tests will be displayed to the subject in random order. The specific source of every sample is anonymous to subjects.

In Table 6, we can see that: (1) our systems which leverage multiple content features are better in terms of naturalness and conversion speaker similarity than FastSVC and SoVITS; (2) introducing WeNet could even damage the performance. We speculate that WeNet is pretrained by only speech, which is not well robust when modeling the singing voice; (3) The integration of Whisper and ContentVec outperforms the others, especially in the conversion similarity, which is close to the ground truth.

5 Conclusion

In this paper, we investigate three distinct content features sourcing respectively from WeNet, Whisper, and ContentVec. We perform a systematic analysis of the three and explore their complementary roles in terms of prosody, intelligibility, and conversion similarity for singing voice conversion. Moreover, we analyze the differences between the singing voice and speech, and explain why fine-tuning a speech vocoder with singing voice data could improve. The experimental results on the SVCC in-domain evaluation data suggest that more sources of content features can be complementary, but it will degrade the similarity when WeNet is included. We will leave this to future research.

References

[1] Xin Chen, Wei Chu, Jinxi Guo, and Ning Xu. Singing voice conversion with non-parallel data. In *MIPR*, pages 292–296. IEEE, 2019.

[2] Soonbeom Choi, Wonil Kim, Saebyul Park, Sangeon Yong, and Juhan Nam. Children's song dataset for singing voice research. In *ISMIR*, 2020.

[3] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *INTERSPEECH*, pages 5036–5040. ISCA, 2020.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[5] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3451–3460, 2021.

[6] Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. Multi-singer: Fast multi-singer singing voice vocoder with A large-scale corpus. In *ACM Multimedia*, pages 3945–3954. ACM, 2021.

[7] Rongjie Huang, Chenye Cui, Feiyang Chen, Yi Ren, Jinglin Liu, Zhou Zhao, Baoxing Huai, and Zhefeng Wang. Singgan: Generative adversarial network for high-fidelity singing voice generation. In *ACM Multimedia*, pages 2525–2535. ACM, 2022.

[8] Wen-Chin Huang, Lester Phillip Violeta, Songxiang Liu, Jiatong Shi, Yusuke Yasuda, and Tomoki Toda. The singing voice conversion challenge 2023. *arXiv*, abs/2306.14422, 2023.

[9] Wen-Chin Huang, Shu-Wen Yang, Tomoki Hayashi, and Tomoki Toda. A comparative study of self-supervised speech representation based voice conversion. *IEEE J. Sel. Top. Signal Process.*, 16(6):1308–1318, 2022.

[10] Tejas Jayashankar, Jilong Wu, Leda Sari, David Kant, Vimal Manohar, and Qing He. Self-supervised representations for singing voice conversion. In *ICASSP*, pages 1–5, 2023.

[11] Jee-weon Jung, You Jin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. Pushing the limits of raw waveform speaker recognition. In *INTERSPEECH*, pages 2228–2232. ISCA, 2022.

[12] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *INTERSPEECH*, pages 2350–2354. ISCA, 2019.

[13] Kazuhiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura. Statistical singing voice conversion with direct waveform modification based on the spectrum differential. In *INTERSPEECH*, pages 2514–2518. ISCA, 2014.

[14] Kazuhiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura. Statistical singing voice conversion based on direct waveform modification with global variance. In *INTERSPEECH*, pages 2754–2758. ISCA, 2015.

[15] Junya Koguchi, Shinnosuke Takamichi, and Masanori Morise. PJS: phoneme-balanced japanese singing-voice corpus. In *APSIPA*, pages 487–491. IEEE, 2020.

[16] Robert Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, volume 1, pages 125–128. IEEE, 1993.

[17] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *ICLR*. OpenReview.net, 2023.

[18] Sang-Hyuk Lee, Hee-Jun Kwon, Hyun-Jin Choi, Nam-Hun Lee, Sung-Jin Lee, and Sung-Min Jin. The singer's formant and speaker's ring resonance: a long-term average spectrum analysis. *Clinical and experimental otorhinolaryngology*, 1(2):92–96, 2008.

[19] Zhonghao Li, Benlai Tang, Xiang Yin, Yuan Wan, Ling Xu, Chen Shen, and Zejun Ma. Ppg-based singing voice conversion with adversarial representation learning. In *ICASSP*, pages 7073–7077. IEEE, 2021.

[20] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *AAAI*, pages 11020–11028. AAAI Press, 2022.

[21] Jinglin Liu, Chengxi Li, Yi Ren, Zhiying Zhu, and Zhou Zhao. Learning the beauty in songs: Neural singing voice beautifier. In *ACL (1)*, pages 7970–7983. Association for Computational Linguistics, 2022.

[22] Songxiang Liu, Yuewen Cao, Na Hu, Dan Su, and Helen Meng. Fastsvc: Fast cross-domain singing voice conversion with feature-wise linear modulation. In *ICME*, pages 1–6. IEEE, 2021.

[23] Songxiang Liu, Yuewen Cao, Dan Su, and Helen Meng. Diffsvc: A diffusion probabilistic model for singing voice conversion. In *ASRU*, pages 741–748. IEEE, 2021.

[24] Eliya Nachmani and Lior Wolf. Unsupervised singing voice conversion. In *INTERSPEECH*, pages 2583–2587. ISCA, 2019.

[25] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[26] Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David D. Cox, Mark Hasegawa-Johnson, and Shiyu Chang. Contentvec: An improved self-supervised speech representation by disentangling speakers. In *ICML*, volume 162, pages 18003–18017. PMLR, 2022.

[27] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv*, abs/2212.04356, 2022.

[28] Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, pages 749–752. IEEE, 2001.

[29] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL (1)*. The Association for Computer Linguistics, 2016.

[30] Johan Sundberg and Thomas D Rossing. The science of singing voice, 1990.

[31] Naoya Takahashi, Mayank Kumar, Singh, and Yuki Mitsufuji. Hierarchical diffusion models for singing voice neural vocoder. In *ICASSP*, pages 1–5, 2023.

[32] Oytun Türk, Osman Büyük, Ali Haznedaroglu, and Levent Mustafa Arslan. Application of voice conversion for cross-language rap singing transformation. In *ICASSP*, pages 3597–3600. IEEE, 2009.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[34] Chao Wang, Zhonghao Li, Benlai Tang, Xiang Yin, Yuan Wan, Yibiao Yu, and Zejun Ma. Towards high-fidelity singing voice conversion with acoustic reference and contrastive predictive coding. In *INTERSPEECH*, pages 4287–4291. ISCA, 2022.

[35] Xin Wang, Shinji Takaki, and Junichi Yamagishi. Neural source-filter waveform models for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:402–415, 2019.

[36] Yu Wang, Xinsheng Wang, Pengcheng Zhu, Jie Wu, Hanzhao Li, Heyang Xue, Yongmao Zhang, Lei Xie, and Mengxiao Bi. Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis. In *INTERSPEECH*, pages 4242–4246. ISCA, 2022.

[37] Da-Yi Wu, Wen-Yi Hsiao, Fu-Rong Yang, Oscar Friedman, Warren Jackson, Scott Bruzenak, Yi-Wen Liu, and Yi-Hsuan Yang. Ddsp-based singing vocoders: A new subtractive-based synthesizer and A comprehensive evaluation. In *ISMIR*, pages 76–83, 2022.

[38] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019.

[39] Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. In *INTERSPEECH*, pages 4054–4058. ISCA, 2021.

[40] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. In *INTERSPEECH*, pages 1526–1530. ISCA, 2019.

[41] Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, and Zhou Zhao. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. In *NeurIPS*, 2022.

[42] Ying Zhang, Peng Yang, Jinba Xiao, Ye Bai, Hao Che, and Xiaorui Wang. K-converter: An unsupervised singing voice conversion system. In *ICASSP*, pages 6662–6666. IEEE, 2022.

[43] Xintao Zhao, Feng Liu, Changhe Song, Zhiyong Wu, Shiyin Kang, Deyi Tuo, and Helen Meng. Disentangling content and fine-grained prosody information via hybrid ASR bottleneck features for voice conversion. In *ICASSP*, pages 7022–7026. IEEE, 2022.