

LEVERAGING DIVERSE SEMANTIC-BASED AUDIO PRETRAINED MODELS FOR SINGING VOICE CONVERSION

Xueyao Zhang¹ Zihao Fang¹ Yicheng Gu¹ Haopeng Chen¹
 Lexiao Zou² Junan Zhang¹ Liumeng Xue¹ Zhizheng Wu^{1,‡}

¹The Chinese University of Hong Kong, Shenzhen

²Shenzhen Research Institute of Big Data

ABSTRACT

Singing Voice Conversion (SVC) is a technique that enables any singer to perform any song. To achieve this, it is essential to obtain speaker-agnostic representations from the source audio, which poses a significant challenge. A common solution involves utilizing a semantic-based audio pretrained model as a feature extractor. However, the degree to which the extracted features can meet the SVC requirements remains an open question. This includes their capability to accurately model melody and lyrics, the speaker-independency of their underlying acoustic information, and their robustness for in-the-wild acoustic environments. In this study, we investigate the knowledge within classical semantic-based pretrained models in much detail. We discover that the knowledge of different models is diverse and can be complementary for SVC. Based on the above, we design a Singing Voice Conversion framework based on Diverse Semantic-based Feature Fusion (DSFF-SVC). Experimental results demonstrate that DSFF-SVC can be generalized and improve various existing SVC models, particularly in challenging real-world conversion tasks. Our demo website is available at <https://diversesemanticssvc.github.io/>.

Index Terms— Singing Voice Conversion, Semantic Features, Features Fusion, Robustness

1. INTRODUCTION

Singing Voice Conversion (SVC) aims to transform a singing signal into the voice of a target singer while maintaining the original lyrics and melody [1]. This allows any singer to perform any song. It has a wide range of applications, such as music entertainment, singing voice enhancement, vocal education, and artistic creation.

In recent years, generative models [2, 3] and conducting singing voice conversion with non-parallel data [4, 5] has attracted more attention. Figure 1 displays the classic pipeline for it. To empower the reference speaker to sing the source audio, the main idea is to extract the speaker-specific representations from the reference, extract the speaker-agnostic representations from the source, and then synthesize the converted audio using a decoder. Usually, speaker-specific representations can be just a one-hot speaker ID [6] or features extracted from a pretrained speaker verification model [6, 7]. For speaker-agnostic representations, common solutions involve using the intermediate output¹ from a semantic-based audio pretrained

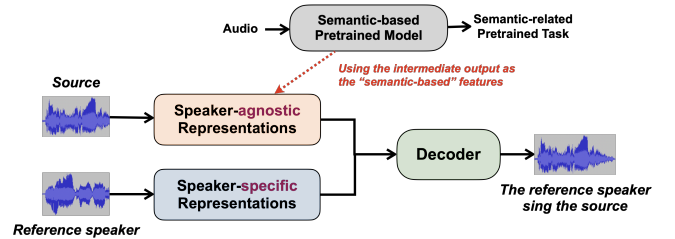


Fig. 1: The role of semantic-based pretrained model in the classic singing voice conversion pipeline.

Requirements of SVC	Capability of the Semantic-based Features
To model melody	Whether could or not remains unknown
To model lyrics	Could. But exactly how much remains unknown
To model auxiliary acoustic information	Could. But whether the information is speaker-agnostic or not remains unknown
To be robust for in-the-wild acoustic environment	Whether is robust or not remains unknown

Table 1: The extent to which the existing semantic-based features satisfy the requirements of singing voice conversion is still unclear.

model. The pretraining task of this model is typically designed to be semantic-related, such as Automatic Speech Recognition (ASR) [5, 8] or self-supervised learning guided by semantics [9, 10].

For SVC, high-quality speaker-agnostic representations should meet several requirements (Table 1). First, they should be capable of modeling melody and lyrics. Besides, they could also contain some auxiliary acoustic information (such as pronunciation, articulation, and prosody) to improve the naturalness and expressiveness of the synthesized audio. Last, they should be robust for varied acoustic environments of source audios (such as the in-the-wild singing voices separated from background music [11, 12]). However, for the semantic-based audio pretrained models, our understanding of their underlying knowledge remains limited, despite the considerable resources invested in pretraining [13, 14, 15]. It is still unclear to what extent the extracted semantic-based features can satisfy the requirements of SVC. For example, except for the semantic signals, there is also acoustic information including prosody in the semantic-based features [16, 17]. Is this sufficient, or how much additional information is required to model the melody? Furthermore, is this information speaker-agnostic, or could it cause the source’s timbre leakage [18] into the converted audio? Moreover, are these semantic-based pretrained models robust to diverse environments? How effective are these features when faced with in-the-wild audio data?

[‡]Correspondence to wuzhizheng@cuhk.edu.cn.

¹The intermediate output is a dense high-dimensional vector instead of just symbolic token. It is believed that such output contains not only semantic but also acoustic information, which can enhance the quality of synthesized audio. In this paper, we refer to it as *semantic-based* features.

School	Representative	Pretrained Task	Training Objective	Pretrained Data	Architecture
Supervised	WeNet [19]	ASR, <i>supervised</i> by linguistic labels	CTC, Next Token Prediction (character-level)	Text-only transcription from Gigaspeech/Wenet-speech (10k hours English /Chinese speech)	Encoder-decoder Conformer [23] or Transformer [21]
Weak-Supervised	Whisper [15]	Multitask including multilingual ASR, speech translation, spoken language identification, etc., large-scale <i>weak supervision</i>	Next Token Prediction (byte-level [24])	680k hours multilingual data, including both text-only and time-aligned transcription	Encoder-decoder Transformer [21]
Self-Supervised	ContentVec [20]	<i>Self-supervised</i> 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 [21]

Table 2: A systematic analysis for the three schools of the existing semantic-based pretrained models.

Motivated by this, by exploring the gap using only semantic-based features to conduct the SVC task, we investigate the underlying knowledge within the three classic semantic-based pretrained models respectively – WeNet [19], Whisper [15], and ContentVec [20]. Furthermore, we suppose that the capabilities of semantic-based features mainly depend on the pretraining tasks and data, and different pretraining ways will yield different underlying knowledge (Table 2). Based on this, we propose a concise and effective solution: by utilizing diverse semantic-based models that are pretrained distinctly, the capabilities of these jointly used semantic-based features will be enhanced for SVC. However, it is challenging to fuse the multiple features of diverse audio pretrained models. The reason is that the time resolutions of different models are usually mismatched, since their acoustic parameters like sampling rate and hop size can be different. To address it, we explored the efficacy and efficiency between a resampling strategy with cross attention [21, 22]. The experiments verify that such signal processing based method could achieve the comparable performance to cross attention but with a lower latency and no additional training cost (Section 4.4). Based on the above, we design a Singing Voice Conversion framework based on Diverse Semantic-based Features Fusion (DSFF-SVC). Both objective and subjective evaluation results verify the effectiveness, generalization, and robustness of our proposed framework.

2. RELATED WORK

The early singing voice conversion researches aim to design parametric statistical models such as HMM [25] or GMM [26, 27] 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 [4, 5], or recognition-synthesis SVC [9], is popular in recent years. The classic pipeline of the non-parallel SVC is displayed in Figure 1. To decouple speaker-agnostic and speaker-specific representations, some pioneering works leverage the information bottleneck [6] and the adversarial learning [4, 28] to design the end-to-end network. To introduce more explicit guidance instead of depending only on end-to-end learning, utilizing pretrained models as a feature extractor has become the trend.

For the speaker-agnostic representations, the most well-known solution is to leverage the phonetic posteriorgrams (PPG) from pretrained ASR models as the semantic-based features [29, 8]. In recent times, besides the supervised ASR models, an increasing number of semantic-based pretrained models have emerged under weak su-

pervision and self-supervised learning. Researchers have explored in conducting the SVC task based on HuBERT [14, 9], wav2vec 2.0 [13, 10], Whisper [15, 30] and more.

However, the specific role of the semantic-based pretrained models and the extent to which the extracted features can meet the SVC requirements is still an open question. For example, some researchers point out that there is also rich acoustic information like prosody in these features [17], which is beneficial to improve the naturalness and the speaker similarity [16]. On the contrary, others believe that this acoustic information could be speaker-specific, which will cause the source’s timbre leakage [18]. Motivated by this, we aim to investigate the specific knowledge within the semantic-based features by exploring to use them alone to conduct the SVC task. Besides, most existing works utilize a single source of semantic-based features. This study also researches whether different pretrained models can be complementary for SVC.

3. METHODOLOGY

In this section, we analyze classic semantic-based pretrained models to demonstrate the potential knowledge embedded in semantic-based features. We also explain why combining multiple pretrained models can be effective for SVC (Section 3.1). Additionally, we discuss the use of resampling and cross-attention strategies for addressing the feature fusion issue of multiple pretrained models with different time resolutions (Section 3.2). Building on this, we introduce the Diverse Semantic-based Features Fusion SVC framework (DSFF-SVC, Figure 2), which can integrate various models as its foundation.

3.1. Analysis for Semantic-based Pretrained Models

In order for a model to capture the semantic signals of audio, we need to push its latent representations to align with semantic representations (such as characters or phonemes). In other words, the extent to which the latent representations contain semantic information depends on *the level of semantic supervision* with which the audio model is pretrained. Furthermore, the knowledge of these latent representations that *goes beyond semantic information* is also determined by the pretraining ways. Based on the above, we categorizes the existing semantic-based pretrained models into three schools (Table 2):

- **Supervised models:** The models are pretrained under the ASR task, whose representative is WeNet [19]. For such mod-

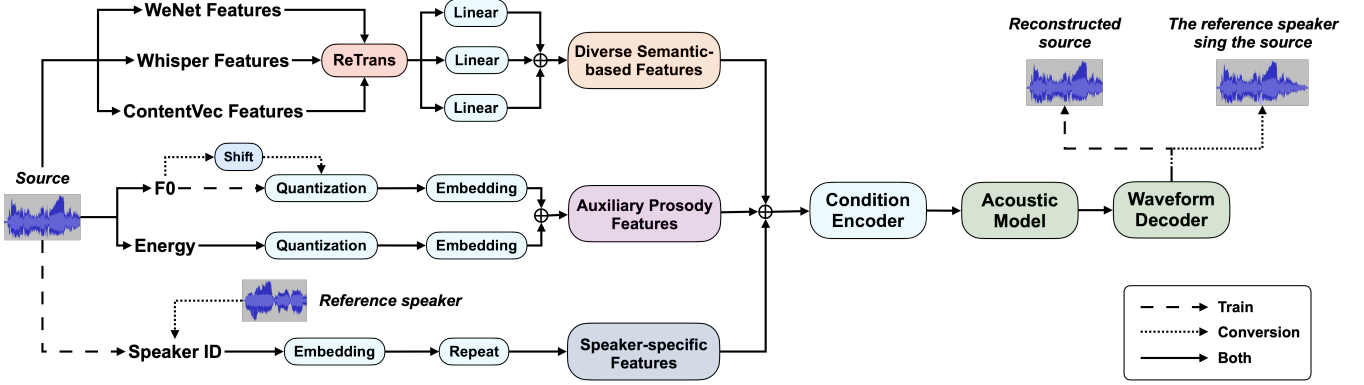


Fig. 2: The proposed Singing Voice Conversion framework based on Diverse Semantic-based Features Fusion (DSFF-SVC). It is capable of incorporating most existing models (i.e., acoustic model and waveform decoder) as a base.

els, the intermediate layers closer to the input (audio) contain more acoustic information, while layers closer to the output (character) contain more semantic information.

- **Weak-supervised models:** The models are pretrained under both ASR task and other auxiliary tasks, whose representative is Whisper [15]. Compared to supervised ones, such models are likely to contain additional knowledge related to the auxiliary tasks.
- **Self-supervised models:** The models conduct self-supervised learning under semantic (or speaker-agnostic) guidance, whose representative is ContentVec [20]. The knowledge of these models is highly related to the construction of pseudo labels during the initial audio tokenization, since these labels usually play a role of “teacher” [14]. Compared to the other two, the self-supervised ones contain more macro and general knowledge that is beneficial for modeling context.

In summary, the underlying knowledge of the three are likely to be different and diverse. In this study, we respectively select a representative for them – WeNet [19], Whisper [15], and ContentVec [20], to explore the gap when adopting them into the SVC task. We will also investigate whether they can be complementary to each other.

3.2. Features Fusion for Models of Mismatched Resolutions

When combining several pretrained audio models, a technical problem emerges because of the differing time resolutions of the models. In particular, for a specific utterance, the feature frame lengths extracted by different semantic-based pretrained models may vary. A classic signal processing based method for aligning features with different frame rates is to apply resampling. Besides of fusing the different semantic features, another common use case is to integrate a semantic-based model with a pretrained vocoder operating at different time resolutions.

3.3. Singing Voice Conversion Framework based on Diverse Semantic-based Features Fusion

The proposed Singing Voice Conversion framework based on Diverse Semantic-based Features Fusion (DSFF-SVC) is displayed in Figure 2. The fusion of multiple features derived from diverse semantic-based pretrained models serves as the speaker-agnostic representation. A trainable embedding layer is employed to model the speaker features derived from a one-hot speaker ID. Furthermore,

the incorporation of auxiliary prosody features, such as fundamental frequency (F0) and energy, facilitates the enhancement of melody modeling and the expressiveness of synthesized audio.

For the semantic-based pretrained models, the WeNet, Whisper, and ContentVec are all Transformer-based [21] architecture. Here we utilize their encoder’s output as the semantic-based 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 frame length, and $D_{\mathcal{M}}$ is the latent representation dimension of the model \mathcal{M} . Using resampling, we transform their frame lengths as the same (i.e., $\hat{\mathbf{c}}_{\mathcal{M}} \in \mathbb{R}^{T \times D_{\mathcal{M}}}$), and adopt adding fusion to merge them:

$$\hat{\mathbf{c}} = \mathbf{Linear}(\hat{\mathbf{c}}_{\mathcal{M}_1}) \oplus \mathbf{Linear}(\hat{\mathbf{c}}_{\mathcal{M}_2}) \oplus \mathbf{Linear}(\hat{\mathbf{c}}_{\mathcal{M}_3}) \quad (1)$$

where $\hat{\mathbf{c}} \in \mathbb{R}^{T \times D}$, **Linear** means the linear layer, \mathcal{M}_1 , \mathcal{M}_2 , and \mathcal{M}_3 means WeNet, Whisper, and ContentVec, and \oplus means the element-wise adding operation.

For the auxiliary prosody features, we follow [31] 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 $\mathbf{e} \in \mathbb{R}^{T \times D}$. For the speaker-specific representations, a look-up table with a trainable embedding layer is adopted to learn the speaker embeddings $\mathbf{s} \in \mathbb{R}^D$. Finally, we adopt the adding fusion to merge all the features:

$$\mathbf{c} = \mathbf{CondEnc}(\hat{\mathbf{c}} \oplus \mathbf{f} \oplus \mathbf{e} \oplus \hat{\mathbf{s}}) \quad (2)$$

where **CondEnc** means a condition encoder that interacts with all the conditions of SVC. It is set as a simple linear layer in our experiments. $\hat{\mathbf{s}} \in \mathbb{R}^{T \times D}$ represents the frame-level speaker feature, being obtained by repeating \mathbf{s} T times.

During training, the DSFF-SVC conducts the reconstruction learning on the training corpus, and the speaker ID is just the speaker identity of every training sample. During conversion, given any source audio, we extract its semantic-based and energy features and stay them unchanged. To convert the speaker identity, we inject a reference speaker ID (which is seen in the training) to obtain the speaker-specific representations. For F0 features, we conduct the musical key transposition² to make the reference speaker sing the source song in his vocal range. Specifically, following [31, 1], we shift the source F0 features by multiplying a factor, which is

²[https://en.wikipedia.org/wiki/Transposition_\(music\)](https://en.wikipedia.org/wiki/Transposition_(music))

Semantic-based Features	MCD (↓)	F0CORR (↑)	F0RMSE (↓)	CER (↓)	SIM (↑)
Ground Truth	0.000	1.000	0.0	12.9%	1.000
WeNet	10.324	0.203	423.4	38.2%	0.912
Whisper	8.229	0.524	297.3	18.9%	0.914
ContentVec	8.972	0.491	361.0	22.1%	0.918
WeNet + Whisper	8.345	0.540	284.2	16.8%	0.911
WeNet + ContentVec	8.870	0.525	329.5	19.9%	0.912
Whisper + ContentVec	8.201	0.548	279.6	16.9%	0.912
WeNet + Whisper + ContentVec	8.249	0.572	278.5	16.1%	0.913

Table 3: Objective evaluation results of different semantic-based features and their integration. It can be observed that from WeNet to Wenet + Whisper, and then to Wenet + Whisper + ContentVec, the results of most metrics are promoted stage by stage.

computed as the ratio between the F0 medians among the reference speaker’s training corpus and the source audio.

Notably, the DSFF-SVC framework can support any architectures of acoustic models and waveform decoders (vocoders). During our experiments, we will investigate the generalization of DSFF-SVC for various base models including transformer-based, diffusion-based, and end-to-end models (Section 4.1).

4. EXPERIMENTS

We conduct experiments to answer the following questions:

- **EQ1:** How much the existing semantic-based pretrained models can meet the requirements of SVC? Could the multiple features from diverse semantic-based pretrained models be complementary?
- **EQ2:** How effective and generalized is the proposed singing voice conversion framework based on diverse semantic-based features fusion?
- **EQ3:** How effective is the resampling strategy compared to deep learning based method such as cross attention [21, 22] for fusing multiple features of mismatched time resolutions?

4.1. Experimental Setup

4.1.1. Evaluation Tasks

We adopt two conversion settings to evaluate: (1) **Recording Studio Setting:** following the most existing works, we utilize the high-quality singing corpus that is recorded in studio as the experimental data. The vocals are clean with virtually no noise or environmental interference. Specifically, we use Opencpop [32] as the target singer, whose training corpus is 5.2 hours of studio recorded singing voices. For source audios, we use M4Singer [33] and randomly 25 utterances for each timbre type respectively (including *Soprano*, *Alto*, *Tenor*, and *Bass*). (2) **In-the-Wild Setting:** in the real-world SVC application, usually the singing voices are separated from the background music, which will remain some artifacts or reverb in the vocals [11, 12]. We consider this as a more challenging conversion task to examine the robustness of the SVC systems. Specifically, we adopt a private corpus which contains 6.4 hours of singing voices of 15 professional singers (6 English singers and 9 Chinese singers). The vocals are separated by Ultimate Vocal Remover (UVR)³. We

adopt four singers as the targets (an English male, an English female, a Chinese male, and a Chinese female). For source audios, we randomly sample 100 utterances which cover multiple musical genres including *Pop*, *Rock*, *Folk* and *Soul*.

4.1.2. Evaluation Metrics

For objective evaluation, following [1], we adopt Mel-cepstral distortion (MCD) [34], F0 Pearson correlation coefficient (F0CORR), F0 Root Mean Square Error (F0RMSE), Character Error Rate (CER) which is obtained with the recognition results of whisper-large ASR model [15]. Besides, following [35], we use the WavLM model [36] finetuned for speaker verification⁴ to compute the cosine speaker similarity score (SIM). For subjective evaluation, 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. Base Models

We select three base models to verify the generalization ability of DSFF-SVC framework:

- **TransformerSVC:** It adopts a vanilla encoder-only transformer model[21] as the acoustic model. Its output is the mel-spectrogram.
- **VitsSVC:** It is a VITS-based [37] model which is similar to the SoftVC-VITS⁵. It is an end-to-end framework and can directly produce waveform.
- **DiffWaveNetSVC:** It adopts a diffusion-based acoustic model and could generate mel-spectrogram, which is proposed by [31]. The internal encoder of the diffusion framework is based on Bidirectional Non-Causal Dilated CNN[38], which is similar to WaveNet[39].

4.1.4. Implementation Details

For WeNet, we use the official models pretrained by 10k hours Wenetspeech⁶ to extract semantic-based features. For Whisper, we use the multilingual MEDIUM model⁷. For ContentVec, we use the

³<https://github.com/Anjok07/ultimatevocalremovergui>

⁴<https://huggingface.co/microsoft/wavlm-base-plus-sv>

⁵<https://github.com/svc-develop-team/so-vits-svc>

⁶<https://github.com/wenet-e2e/wenet>

⁷<https://github.com/openai/whisper>

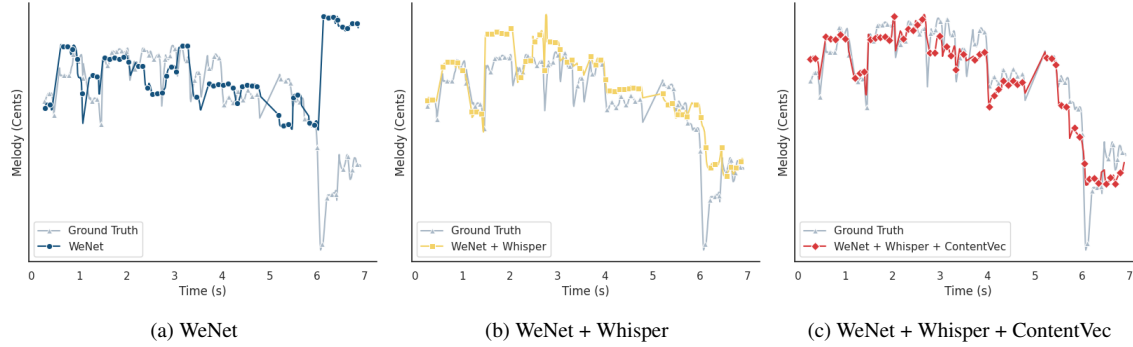


Fig. 3: The complementary role of diverse semantic-based features in melody modeling. More benefits of the joint usage of diverse semantic-based features (including spectrogram reconstruction, lyrics modeling, etc.) can be seen at our [demo website](#).

official 500-CLASS model pretrained by 1k hours Librispeech⁸. The latent space dimension D is set as 384. We adopt Parselmouth [40] to extract F0 and compute L2-norm of the amplitude of each short-time Fourier transform frame as energy features [41]. Following [31], we set their numbers of bins for quantization as 256. For the three base models, we adopt the implementations of Amphion⁹ [42]. For DiffWaveNetSVC and TransformerSVC, we use the pretrained Amphion Singing BigVGAN¹⁰ as the vocoder to decode waveform from mel-spectrogram.

4.2. Performance of Different Semantic-based Features (EQ1)

In this section, we aim to explore the gap using semantic-based features *alone* to conduct the SVC task. We select DiffWaveNetSVC as the base model and conduct the task under the Recording Studio Setting. The experimental results are illustrated in Table 3.

On the one hand, the underlying knowledge of different semantic-based models can be distinct and diverse: (1) For modeling melody (F0CORR and F0RMSE), Whisper and ContentVec perform better than WeNet. This is mainly because the auxiliary tasks of the weak-supervised Whisper and the self-supervised ContentVec are beneficial for modeling more prosody-related signals. (2) For modeling lyrics (CER), Whisper performs best compared to the other two. It is assumed that the multilingual ASR tasks and the large-scale pretraining data allow Whisper to model more valid and robust semantic information. Moreover, ContentVec will provide better semantic signals than WeNet, although both are pretrained only on speech. It reveals that the self-supervised pretrained models could be more robust than the classical supervised ASR model. (3) When measuring speaker independence (SIM), the speaker similarity results of the three are comparable and all above 0.9, which is difficult to rank from human perception (see our demo page). This means that when using only semantic-based features for SVC, the three can all be considered speaker-agnostic.

On the other hand, the diverse semantic-based features can be complementary in most cases. For example, from WeNet to WeNet + Whisper, and then to WeNet + Whisper + ContentVec, the results of most metrics are promoted stage by stage. We display a case study of melody modeling in Figure 3. It illustrates that after integrating diverse semantic-based features, the trajectories of the melody

between converted audios and ground truth are closer. However, we can also find that using only semantic-based features is hard to model melody adequately (the highest F0CORR in Table 3 is 0.572), appearing the “out of tune” for human hearing. Therefore, introducing explicit melody modeling (such as F0 features) for SVC remains necessary in the present technology context.

4.3. Performance of the DSFF-SVC framework (EQ2)

To verify the generalization and robustness of the proposed DSFF-SVC framework, we conduct experiments based on both the recording studio and the in-the-wild settings for the three base models. Here we also use the auxiliary prosody features (including F0 and energy), to improve the melody modeling and the expressiveness of the system. The objective and subjective evaluation results can be seen in Table 4 and Table 5. It reveals that: (1) Conducting SVC under in-the-wild setting is more difficult. Facing to the more complex acoustic environment, only using the supervised WeNet produces a very poor performance. Integrating with the weak-supervised Whisper and the self-supervised ContentVec are usually beneficial, especially for Whisper. (2) For the generalization ability, we can observe that for all three models, the proposed idea of diversion semantic-based features fusion is effective in most cases under both settings. (3) For the robustness, particularly, under the in-the-wild setting, the improvement of integrating diverse semantic-based features is more obvious, especially for the objective CER and the subjective metrics.

Compared with Table 3 and 4, there are also many interesting observations about auxiliary prosody features (F0 and energy): (1) After injecting such prosody signals into SVC models, the intelligibility (CER) has been improved a lot. In our opinion, this is because the prosody in singing voice is expressive and characteristic, which is hard to be modeled without the explicit assistant of prosody features like F0. Therefore, using only semantic-based features, some tone- and intonation-related signals will not be learned well, leading to the worse intelligibility. (2) Except for the intelligibility, it is also notable about the change of conversion speaker similarity (SIM). After introducing F0 features, the melody-related metrics have been improved a lot and the converted singing voices are not out of tune any more. In the meantime, however, the speaker similarity is also harmed. We suppose such auxiliary prosody features own some speaker-specific signals, e.g. the personalized *vibrato* patterns within F0 features. Therefore, directly copying the original prosody features from the source audio could not be the best choice for SVC, although it is a common way in the most SVC works.

⁸<https://github.com/auspicious3000/contentvec>

⁹<https://github.com/open-mmlab/Amphion/tree/main/egs/svc>

¹⁰https://huggingface.co/amphion/BigVGAN_singing_bigdata

Base Model	Semantic-based Features	Recording Studio Setting				In-the-Wild Setting			
		F0CORR (↑)	F0RMSE (↓)	CER (↓)	SIM (↑)	F0CORR (↑)	F0RMSE (↓)	CER (↓)	SIM (↑)
TransformerSVC	WeNet	0.849	149.3	15.6%	0.878	0.871	210.0	40.0%	0.865
	+ Whisper	0.924	77.2	14.9%	0.881	0.848	183.8	18.7%	0.867
	+ Whisper + ContentVec	0.931	75.5	16.2%	0.883	0.857	186.7	23.3%	0.868
VitsSVC	WeNet	0.937	175.3	19.1%	0.890	0.919	91.3	57.7%	0.869
	+ Whisper	0.945	144.4	17.8%	0.890	0.920	86.9	35.2%	0.869
	+ Whisper + ContentVec	0.946	112.9	17.7%	0.886	0.921	79.5	32.3%	0.870
DiffWaveNetSVC	WeNet	0.936	55.5	15.8%	0.875	0.901	87.8	60.8%	0.855
	+ Whisper	0.943	49.5	15.2%	0.884	0.921	73.6	21.1%	0.865
	+ Whisper + ContentVec	0.940	55.2	15.7%	0.884	0.919	79.9	23.3%	0.867

Table 4: Objective Evaluation Results of the proposed DSFF-SVC framework.

Semantic-based Features	Recording Studio		In-the-Wild	
	Nat. (↑)	Sim. (↑)	Nat. (↑)	Sim. (↑)
WeNet	2.72 ± 0.22	2.64 ± 0.21	2.85 ± 0.21	2.34 ± 0.20
+ Whisper	4.02 ± 0.18	3.13 ± 0.17	3.70 ± 0.18	2.86 ± 0.23
+ Whisper + ContentVec	4.14 ± 0.19	3.25 ± 0.18	3.71 ± 0.18	2.82 ± 0.23

Table 5: Subjective MOS evaluation results (with 95% confidence interval) of DiffWaveNetSVC. The full scores of Naturalness (Nat.) and Similarity (Sim.) are 5 and 4.

4.4. Performance of the Resolution Transformation based Features Fusion (EQ3)

Fusion Strategy	Computational Cost		Conversion Quality		
	RTX (↑)	RTF (↓)	F0CORR (↑)	CER (↓)	SIM (↑)
Cross attention	191.9	2.57	0.896	16.4%	0.871
Resampling	415.0	0.86	0.936	15.8%	0.875

Table 6: The comparison between resampling and cross attention for features fusion.

To verify the performance and efficiency of the resampling for feature fusing, we select the cross attention [21] as the compared method. Specifically, we utilize only WeNet to extract semantic-based features and try different strategies to align its resolution to that of F0 and energy features. We conduct the experiments on DiffWaveNetSVC and follow NaturalSpeech2 [22] to adopt the cross attention for features fusion within the diffusion model.

Except for the quality evaluation for the conversion results, we also measure the computational cost of the both. Assume the duration of the whole training corpus is d_{train} , the duration of the evaluation samples to be converted is d_{infer} , the time to train an epoch is t_{train} , and the time to infer the evaluation samples is t_{infer} . We define RTX as d_{train}/t_{train} , which can measure the training efficiency. We define RTF as t_{infer}/d_{infer} to measure the inference latency. The computational platform is a single NVIDIA Tesla V100.

The comparison results between resampling and cross attention are displayed in Table 6. We can see that when fusing the features of mismatched time resolutions for the SVC task, resampling is even slightly better than cross attention in terms of conversion quality. Moreover, as a non-learning method, resampling does not introduce any additional parameters and greatly accelerates both training efficiency (RTX) and inference speed (RTF) much.

5. CONCLUSION AND FUTURE WORK

This paper investigates three semantic-based pretrained models of supervised, weak-supervised, and self-supervised fashions for SVC. We discover their capabilities including modeling melody and lyrics are diverse and can be complementary. Based on these findings, we propose the SVC framework based on Diverse Semantic-based Features Fusion. The experimental results confirm the effectiveness of our proposed framework, particularly for real-world applications involving in-the-wild data. In addition, this paper raises other questions worth exploring. Is it possible to pretrain one model that can capture both melody and lyrics signals while being speaker-agnostic at the same time? How should we select the pretraining method and construct the pretraining corpus to improve the robustness? We will leave these to our future research.

6. REFERENCES

- [1] Wen-Chin Huang, Lester Phillip Violeta, Songxiang Liu, Jiatong Shi, Yusuke Yasuda, and Tomoki Toda, “The singing voice conversion challenge 2023,” in *ASRU*. 2023, IEEE.
- [2] Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu, “Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation,” in *Proc. of SLT*, 2024.
- [3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, “High-resolution image synthesis with latent diffusion models,” in *CVPR*. 2022, pp. 10674–10685, IEEE.
- [4] Eliya Nachmani and Lior Wolf, “Unsupervised singing voice conversion,” in *INTERSPEECH*. 2019, pp. 2583–2587, ISCA.
- [5] Xin Chen, Wei Chu, Jinxi Guo, and Ning Xu, “Singing voice conversion with non-parallel data,” in *MIPR*. 2019, pp. 292–296, IEEE.
- [6] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, “Autovc: Zero-shot voice style transfer with only autoencoder loss,” in *ICML*. 2019, vol. 97, pp. 5210–5219, PMLR.
- [7] Shahan Nercessian, “Zero-shot singing voice conversion,” in *ISMIR*, 2020, pp. 70–76.
- [8] Zhonghao Li, Benlai Tang, Xiang Yin, Yuan Wan, Ling Xu, Chen Shen, and Zejun Ma, “Ppg-based singing voice con-

- version with adversarial representation learning,” in *ICASSP*. 2021, pp. 7073–7077, IEEE.
- [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.*, vol. 16, no. 6, pp. 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*, 2023, pp. 1–5.
 - [11] Naoya Takahashi, Mayank Kumar Singh, and Yuki Mitsufuji, “Robust one-shot singing voice conversion,” *arXiv*, vol. abs/2210.11096, 2022.
 - [12] Yongyi Zang, You Zhang, Mojtaba Heydari, and Zhiyao Duan, “Singfake: Singing voice deepfake detection,” in *ICASSP*. 2024, IEEE.
 - [13] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *NeurIPS*, 2020.
 - [14] 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.*, vol. 29, pp. 3451–3460, 2021.
 - [15] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” in *ICML*. 2023, vol. 202, pp. 28492–28518, PMLR.
 - [16] 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*. 2022, pp. 4287–4291, ISCA.
 - [17] Chang Liu, Zhen-Hua Ling, and Ling-Hui Chen, “Pronunciation dictionary-free multilingual speech synthesis using learned phonetic representations,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3706–3716, 2023.
 - [18] Binzhu Sha, Xu Li, Zhiyong Wu, Ying Shan, and Helen Meng, “Neural concatenative singing voice conversion: rethinking concatenation-based approach for one-shot singing voice conversion,” in *ICASSP*. 2024, IEEE.
 - [19] 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*. 2021, pp. 4054–4058, ISCA.
 - [20] 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*. 2022, vol. 162, pp. 18003–18017, PMLR.
 - [21] 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*, 2017, pp. 5998–6008.
 - [22] Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian, “Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers,” in *ICLR*. 2024, OpenReview.net.
 - [23] 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*. 2020, pp. 5036–5040, ISCA.
 - [24] Rico Sennrich, Barry Haddow, and Alexandra Birch, “Neural machine translation of rare words with subword units,” in *ACL (1)*. 2016, The Association for Computer Linguistics.
 - [25] Oytun Türk, Osman Büyük, Ali Haznedaroglu, and Levent Mustafa Arslan, “Application of voice conversion for cross-language rap singing transformation,” in *ICASSP*. 2009, pp. 3597–3600, IEEE.
 - [26] 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*. 2014, pp. 2514–2518, ISCA.
 - [27] 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*. 2015, pp. 2754–2758, ISCA.
 - [28] Chengqi Deng, Chengzhu Yu, Heng Lu, Chao Weng, and Dong Yu, “Pitchnet: Unsupervised singing voice conversion with pitch adversarial network,” in *ICASSP*. 2020, pp. 7749–7753, IEEE.
 - [29] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen M. Meng, “Phonetic posteriorgrams for many-to-one voice conversion without parallel data training,” in *ICME*. 2016, pp. 1–6, IEEE Computer Society.
 - [30] Ziqian Ning, Yuepeng Jiang, Zhichao Wang, Bin Zhang, and Lei Xie, “Vits-based singing voice conversion leveraging whisper and multi-scale f0 modeling,” in *ASRU*. 2023, IEEE.
 - [31] Songxiang Liu, Yuewen Cao, Dan Su, and Helen Meng, “Diffsvc: A diffusion probabilistic model for singing voice conversion,” in *ASRU*. 2021, pp. 741–748, IEEE.
 - [32] 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*. 2022, pp. 4242–4246, ISCA.
 - [33] 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.
 - [34] Robert Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in *Proceedings of IEEE pacific rim conference on communications computers and signal processing*. IEEE, 1993, vol. 1, pp. 125–128.
 - [35] Ziyue Jiang, Yi Ren, Zhenhui Ye, Jinglin Liu, Chen Zhang, Qian Yang, Shengpeng Ji, Rongjie Huang, Chunfeng Wang, Xiang Yin, Zejun Ma, and Zhou Zhao, “Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias,” *arXiv*, vol. abs/2306.03509, 2023.
 - [36] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and

- Furu Wei, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [37] Jaehyeon Kim, Jungil Kong, and Juhee Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *ICML*. 2021, vol. 139, pp. 5530–5540, PMLR.
- [38] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro, “Diffwave: A versatile diffusion model for audio synthesis,” in *ICLR*. 2021, OpenReview.net.
- [39] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, “Wavenet: A generative model for raw audio,” in *SSW*. 2016, p. 125, ISCA.
- [40] Yannick Jadoul, Bill Thompson, and Bart de Boer, “Introducing Parselmouth: A Python interface to Praat,” *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.
- [41] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” in *ICLR*. 2021, OpenReview.net.
- [42] Xueyao Zhang, Liumeng Xue, Yuancheng Wang, Yicheng Gu, Xi Chen, Zihao Fang, Haopeng Chen, Lexiao Zou, Chaoren Wang, Jun Han, Kai Chen, Haizhou Li, and Zhizheng Wu, “Amphion: An open-source audio, music and speech generation toolkit,” *arXiv*, vol. abs/2312.09911, 2023.