Singer Adaptive Singing Voice Conversion

Haoyan Luo
School of Data Science
Chinese University of Hong Kong, Shenzhen
haoyanluo@link.cuhk.edu.cn

Xueyao Zhang
School of Data Science
Chinese University of Hong Kong, Shenzhen
xueyaozhang@link.cuhk.edu.cn

Zhizheng Wu
School of Data Science
Chinese University of Hong Kong, Shenzhen
wuzhizheng@cuhk.edu.cn

Abstract

Singing Voice Conversion (SVC) technology has seen significant advancements, transforming the potential of voice cloning in the music industry. However, the gap between the conversion of amateur voices and professional singing voices, which hold more commercial prospects, has not been adequately studied. This paper presents an in-depth analysis on the performance of So-VITS [1], a leading SVC system, when applied to both public and professional singing datasets. Our research introduces the ProSinger dataset, a collection of pure vocal singing corpora from multiple professional singers, to facilitate our investigation. Subjective evaluations based on Mean Opinion Score (MOS) reveal that the SoftVC-VITS model significantly outperforms a baseline SVC model across all datasets. Intriguingly, we found that the SVC model exhibited better performance when converting to public datasets than the ProSinger dataset. This discrepancy may be attributed to the challenges in extracting pure vocal voice from professional singing and the complexity of professional singing techniques. We hope the results will illuminate new research directions for enhancing the representation and modeling of professional singing voices. The code is available at https://github.com/clarenceluo78/singer-adaptive-svc.

1 Introduction

Singing Voice Conversion (SVC) is a burgeoning field in the domain of speech synthesis, which has the potential to revolutionize the music industry by enabling the transformation of one singing voice into another. By leveraging deep learning techniques, SVC aims to maintain the linguistic content and singing style of the source voice while altering the vocal characteristics to resemble a target singer. This technology, while promising, is fraught with numerous challenges that require detailed exploration and study.

Prior research has contributed substantially to the development of SVC technology, with a myriad of methods proposed to refine the accuracy and effectiveness of voice conversion. For example, [1] showed that soft units improve intelligibility and naturalness of the voice conversion with better linguistic content representation. Similarly, ContentVec [2] was proposed to remove speaker information while preventing loss of content information. With better speaker disentanglement, the content information in the songs is better captured. Speech synthesis technology has made rapid progress. Especially, VITS [3], an end to end speech synthesis system with additional function of many to many voice Conversion shows superior performance. To achieve realistic singing voice synthesis, many other approaches such as [4] which incorporates HiFi-GAN [5] a shallow diffusion mechanism is also proposed.

[1] https://github.com/svc-develop-team/so-vits-svc

So-VITS singing voice conversion system combines the above mentioned state-of-the-art voice manipulation modules and stands out as an advanced SVC system, showing superior performance over conventional methods. However, a significant challenge remains unaddressed: the disparity in performance when converting to amateur voices compared to professional singing voices.

In this paper, we undertake a comprehensive analysis of this issue, examining the performance gap of SVC systems when converting to public amateur voices versus professional singing voices. We introduce the ProSinger dataset, a unique corpus of professional singing voices, and compare its results with a range of public singing datasets. In addition, we scrutinize potential reasons for the observed discrepancy. The extraction of pure vocal voice from songs sang by professional singers and the complex nature of professional singing techniques may be identified as potential contributing factors.

Our main contributions can be summarized as follows:

- Introduction of the ProSinger dataset, a collection of vocal voices for professional singers, which serves as a tool for further research.
- A detailed investigation of the performance gap in SVC systems when converting amateur and professional singing voices.
- Comprehensive experiments and analysis, elucidating the challenges in modeling professional singing voices.

2 Related Work

The early studies for singing voice conversion include the use of spectral conversion techniques and the statistical generation architectures. These techniques were applied to convert singing voices across different genres and languages. However, they were limited by the inability to handle high-dimensional feature spaces, and required parallel training data, i.e. the source and the target speakers to sing the same songs during the training phase.

One of the earliest approaches for many-to-one singing voice conversion on non-parallel data was the use of deep neural networks (DNN) with an autoencoder-based approach. In the work of [6], a CNN-based encoder, a single WaveNet [7] decoder, and a data augmentation scheme were proposed to produce more natural and recognizable singing voices. An adversarially trained pitch regression network [8] is also introduced to enforce the encoder to network to learn singer-invariant and pitch-invariant representation. Moreover, some novel generation frameworks such as Gaussian mixture variational autoencoders (GM-VAEs) is also introduced to the SVC task.

Although the autoencoder-based models can obtain natural singing voices, redundant noise from input data may reduce the quality of the generated sounds. Many approaches such as adversarial training and phonetic posteriorgrams (PPG) have also been proposed to extract disentangled features to better convert the timbre. [9] propose a multilayer bidirectional LSTM (DBLSTM) network to map PPGs to Mel Cepstrals (MCEPs). The acoustic features together with the embeddings are them used to reconstruct the target singing voice through a vocoder. The authors of [10] implemented two separate encoders to generate mel spectrograms with PPGs as inputs. In an adversarial setting, [11] proposed a singing voice synthesis model with multi-task learning using the parametric vocoder features as auxiliary features. To address the problems such as pitch jitters and U/V errors casued by the instability of GAN, [12] suggested to feed harmonic signals to the SVC model in advance to enhance the audio generation. The experiments showed that the signals significantly improve the smoothness and continuity of harmonics in the generated audio and better match the target audio.

Among the above challenges, the speaker modeling in singing signals is non-trivial. One of the commonly adopted methods in the multi-singer system training is the speaker lookup table (LUT) [8, 13, 14]. Other approaches [12] leverage a speaker recognition network (SRN) to distill speaker information by extracting speaker embeddings from a target singer's reference audio. However, it is highly questionable that a single speaker embedding, extracted from either LUT or SRN, is adequate to capture the dynamically varying speaker characteristics of a singing utterance. In the work of [15], the authors applied a hierarchical speaker representation framework with an up-sampling streams and three down-sampling streams. In this work, we will explore more methods to capture fine-grained speaker characteristics at different granularity.

3 Approach

In the So-VITS system, pitch and intonation are preserved because to the system's usage of the SoftVC content encoder [2] to capture source audio speech elements, which are then fed straight into VITS [3] rather than being converted to a text-based intermediate. To address the issue of sound interruption, the vocoder was changed to NSF HiFiGAN [5]. The general architecture can found at Figure 1.

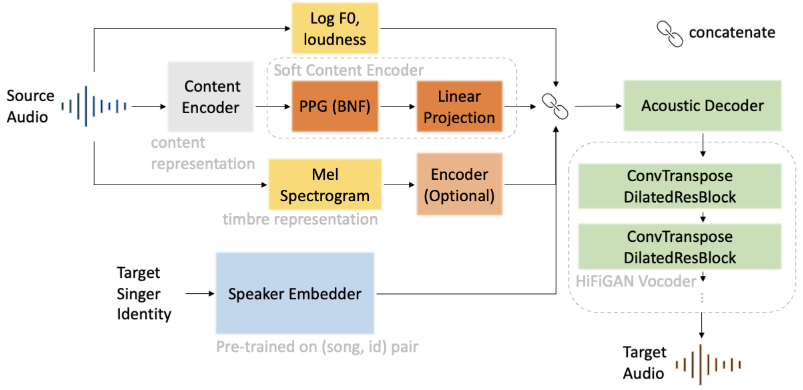

Figure 1: The overall architecture of the So-VITS model.

3.1 Soft Content Encoder

Soft speech units [1] are proposed to act as a bridge between discrete units and raw continuous characteristics. It models a distribution across discrete units to preserve more content information and improve comprehension. Discrete units cause an information bottleneck that forces speaker information out. In order to solve this problem, soft content encoder therefore acquire a speaker independent representation in order to accurately forecast the discrete units. On the other hand, speech sounds do not exist in a distinct place. As a result, some content information is lost during discretization.

3.2 VITS Synthesis

VITS [3] is an End-to-End (encoder -> vocoder together) TTS model that takes advantage of SOTA DL techniques like GANs, VAE, and Normalizing Flows. It is based on the Transformer architecture [16], which has been highly successful in various NLP tasks due to its ability to model long-range dependencies and its parallelizability. At the same time, VITS also incorporates elements from variational autoencoders [17], which are often used for their ability to learn meaningful and structured latent spaces in unsupervised learning tasks. The model architecture is a combination of GlowTTS encoder and HiFiGAN vocoder, which will be introduced later.

3.3 HiFi-GAN Vocoder

HiFi-GAN [5] is a type of vocoder developed for the task of high-fidelity speech synthesis. The vocoder's role in speech synthesis is to convert the intermediate mel-spectrogram representation into a waveform that can be played as sound. HiFi-GAN is trained to generate high-quality speech waveforms from mel-spectrograms. The GAN structure allows it to capture fine details in the speech signal, resulting in high-fidelity (HiFi) speech synthesis. In the case of So-VITS system, the use of HiFi-GAN can help to solve the problem of sound interruption once the generator gets better at producing realistic data, and the discriminator gets better at distinguishing real data from generated data.

4 Experiments

4.1 Data

We adopt two public datasets, Opencpop [18] and M4Singer [19], covering different singing styles, languages, and genres, and collect a professional singer dataset, ProSinger, to ensure a comprehensive evaluation. We select three professional singers Jian Li, Yijie Shi, and Adele from the ProSinger dataset to use in our experiments.

4.2 Evaluation method

In our experiments, we conduct subjective evaluation using Mean Opinion Score (MOS). MOS is a common evaluation metric used in the field of telecommunications and signal processing, particularly for assessing the quality of speech and audio signals. The MOS is essentially an average of numerical scores given by human evaluators. Evaluators are asked to rate the quality of audio on a scale from 1 to 5, where 1 is "bad" or "unacceptable" and 5 is "excellent". All of the converted samples for each system will be scored individually by 10 people.

4.3 Experimental Details

4.3.1 Training Setup

The model is trained iteratively over several epochs (capped at 1000 epochs) mainly using CUDA (RTX-2080). For the So-VITS model training, we can see notable improvement on singing voice conversion over 100 epochs, so we basically end the training at 100 to 200 epochs [2].

The model is evaluated on a validation dataset after each epoch, and the model state is saved periodically as a checkpoint. This allows for model recovery if the training process is interrupted, and also provides snapshots of the model state at different stages of training. The performance of the model is logged and monitored using TensorBoard, which provides insights into the model's learning over time and helps in identifying any issues or bottlenecks in the training process.

4.3.2 Model Version and Inference

We adopt the 4.0-Vec768-Layer12 version of So-VITS. The feature input is the Layer 12 Transformer output of the Content Vec [2], which ensures a more robust singer-disentangled representation. During the model inference stage, the mean filtering of F0 is enabled. It can effectively reduce the hoarse sound caused by the predicted fluctuation of pitch (the hoarse sound caused by reverb or harmony can not be eliminated temporarily). In our experiments, this function has been greatly improved on some songs.

4.3.3 Training and Evaluation Strategy

AdamW Optimizer: The AdamW optimizer [20] is an extension of the Adam optimizer. It decouples weight decay from the optimization steps, leading to improved performance and stability over the standard Adam optimizer. We employ AdamW for both the generator and discriminator to optimize their weights.

Exponential Learning Rate Scheduler: This method is used to adjust the learning rate during training. We start with a higher learning rate, which exponentially decays over the epochs. This approach is useful to converge quickly in the initial stages and fine-tune the model parameters in the later stages.

4.4 Results

WORLD-based SVC [9] is a baseline model which consists of two main modules, Content Extractor and Singer-specific Synthesizer. The main results of our experiments can be found at Table 1. The ablation study experiments can be found at Appendix A.1. From the results we can observe that the So-VITS system outperforms the WORLD-based baseline significantly. On the other hand, we can

[2] Model checkpoints are available at https://huggingface.co/clarenceluo/so-vits-svc-4.0-Vec768-pretrained

Table 1: MOS evaluation scores of the baseline models and our proposed model with additional module. GT means ground truth. BASE1 refers to [9]. ProSinger target 1, 2, and 3 refer to Jian Li, Yijie Shi, and Adele respectively. **Bold** means the best results between SVC model. Red means the best resulst among different target datasets.

	Model	Naturalness	Similarity
Source	GT	4.52 ± 0.18	2.69 ± 0.13
Opencpop target	GT	4.58 ± 0.09	4.33 ± 0.11
	BASE1	2.75 ± 0.12	2.44 ± 0.08
	So-VITS	**4.35 ± 0.11**	**3.98 ± 0.37**
ProSinger target 1	GT	4.55 ± 0.08	4.63 ± 0.12
	BASE1	2.71 ± 0.12	2.49 ± 0.05
	So-VITS	**4.25 ± 0.76**	**3.99 ± 0.92**
ProSinger target 2	GT	4.59 ± 0.07	4.60 ± 0.18
	BASE1	2.70 ± 0.59	2.48 ± 0.17
	So-VITS	**4.08 ± 0.38**	**3.73 ± 0.22**
ProSinger target 3	GT	4.64 ± 0.10	4.66 ± 0.09
	BASE1	2.63 ± 0.09	2.41 ± 0.15
	So-VITS	**4.38 ± 0.84**	**4.03 ± 0.20**

observe that the model performance on Opencpop target are generally better than the performance on ProSinger target.

5 Analysis

5.1 Performance Comparison of SVC Models

Our experiments involved extensive evaluation of the So-VITS framework in comparison to the WORLD-based baseline SVC. In these tests, the performance of the So-VITS was found to be significantly superior. This superiority was quantified in terms of MOS. It suggests that So-VITS's approach, leveraging the power of variational Transformer models and HiFi-GAN vocoder, provides a robust and effective solution for SVC. This follows our intuition since So-VITS is a far more advanced system compared to the baseline.

Based on the experimental results in Table 1, we can infer that DBLSTM models from BASE1 obtains high accuracy in intonation by converting MCEP features and using pitch information from source singers. However, the similarity score is generally low, showing its shortage in singer-dependent decoder design. On the other hand, the ContentVec module in So-VITS system solve this problem accordingly, which obtain higher scores as additional singer-dependent information and invariant of the source audio are preserved.

5.2 Performance Comparison Between Opencpop and ProSinger

However, an interesting observation emerged when we compared the performance of the SVC model on different target datasets. The model found it relatively easy to convert voices from a public dataset (Opencpop) as opposed to the ProSinger dataset. The public dataset consisted of amateur voices, while the ProSinger dataset was made up of professional singer voices.

This discrepancy in performance could be attributed to several factors. First, the ProSinger dataset's vocal tracks were extracted from professional songs, which might have led to less clean, or 'impure', vocal data. This could have introduced challenges in the voice conversion process, affecting the overall performance. Second, professional singers tend to have a broader voice range and utilize more complex singing techniques compared to amateurs. These elements of professional singing might be harder to model and replicate accurately, posing additional challenges to the SVC system.

These findings suggest two possible directions for future research. First, there's a need for improved methods of vocal track extraction to obtain purer vocal data from professional songs. Second, advanced techniques are required for accurately modeling and replicating the broad voice range and complex singing techniques of professional singers.

5.3 Performance Comparison Among Professional Target Singers

The model demonstrated strong performance in converting the singing voice from one singer to another while maintaining the same lyrics and melody. However, it is worth noticing that among different professional singers presented in the experiments, we can observe that Adele generally has better performance than the other two professional singers.

The reason behind may be that Adele has a very special tone, so it may be conducive to voice transformation; Li Jian and Shi Yijie's songs usually include operatic singing style and the collaboration of the choir, so it may be possible to substitute some noise that does not belong to the singer's characteristics into the target-singer separation process, resulting in worse results. In addition, it was highly successful in scenarios where the source and target singers had similar vocal ranges (same gender) and styles (similar singing skills). This is likely due to the fact that the model could capture the subtle nuances of the source singer's voice and transpose them onto the target singer's vocal characteristics.

5.4 Limitations

The So-VITS model also faces challenges when the source and target singers had vastly different vocal ranges or singing styles. In such cases, the model sometimes produced outputs that were not as natural or convincing. This can be attributed to the inherent complexity of the task - transforming a singing voice while preserving the melody and lyrics is a non-trivial task, especially when the source and target voices are significantly different.

Furthermore, the model occasionally struggled with maintaining the precise timing and rhythm of the original song, particularly for songs with complex or rapid rhythmic patterns. This suggests that the model may need further tuning or additional training data to better capture the temporal aspects of singing voice conversion [3].

6 Conclusion

In the scope of this research, we delved into the realm of Singing Voice Conversion with a specific focus on the So-VITS, a cutting-edge system that has shown promising results in the field of SVC. Our investigation centered around examining the performance gap between conversion to Public Singing Datasets (amateur voices) and Professional Singing Datasets (professional voices). We collect ProSinger dataset, a unique corpus featuring pure vocal singing from multiple professional singers. In our experiments, we found that the So-VITS framework outperformes the baseline SVC model on both datasets. Notably, the gap in performance was more pronounced when converting to the ProSinger dataset. This performance discrepancy led us to several key observations. First, the extraction of pure vocal voices from professional songs posed a significant challenge. Second, the inherent complexity of professional singers' voices, which often employ sophisticated singing techniques, may be more challenging for the SVC model to learn the representations. The complexity of professional singing voices and the impurities in their vocal voice extraction are areas that need specific attention to bridge the gap in SVC performance.

[3] Demo is available at https://clarenceluo78.github.io/pages/svc.html

References

[1] Benjamin van Niekerk, Marc-Andre Carbonneau, Julian Zaidi, Matthew Baas, Hugo Seute, and Herman Kamper. A comparison of discrete and soft speech units for improved voice conversion. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2022.

[2] Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David Cox, Mark Hasegawa-Johnson, and Shiyu Chang. Contentvec: An improved self-supervised speech representation by disentangling speakers, 2022.

[3] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, 2021.

[4] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism, 2022.

[5] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, 2020.

[6] Eliya Nachmani and Lior Wolf. Unsupervised Singing Voice Conversion. In *Proc. Interspeech 2019*, pages 2583–2587, 2019.

[7] Aäron 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. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, page 125, 2016.

[8] Chengqi Deng, Chengzhu Yu, Heng Lu, Chao Weng, and Dong Yu. Pitchnet: Unsupervised singing voice conversion with pitch adversarial network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7749–7753, 2020.

[9] Xin Chen, Wei Chu, Jinxi Guo, and Ning Xu. Singing voice conversion with non-parallel data. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 292–296, 2019.

[10] 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 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7073–7077, 2021.

[11] Tae-Woo Kim, Min-Su Kang, and Gyeong-Hoon Lee. Adversarial Multi-Task Learning for Disentangling Timbre and Pitch in Singing Voice Synthesis. In *Proc. Interspeech 2022*, pages 3008–3012, 2022.

[12] Haohan Guo, Zhiping Zhou, Fanbo Meng, and Kai Liu. Improving adversarial waveform generation based singing voice conversion with harmonic signals. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6657–6661, 2022.

[13] Naoya Takahashi, Mayank Kumar Singh, and Yuki Mitsufuji. Hierarchical disentangled representation learning for singing voice conversion, 2021.

[14] Songxiang Liu, Yuewen Cao, Na Hu, Dan Su, and Helen Meng. Fastsvc: Fast cross-domain singing voice conversion with feature-wise linear modulation, 2021.

[15] Xu Li, Shansong Liu, and Ying Shan. A Hierarchical Speaker Representation Framework for One-shot Singing Voice Conversion. In *Proc. Interspeech 2022*, pages 4307–4311, 2022.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[17] Yin-Jyun Luo, Chin-Chen Hsu, Kat Agres, and Dorien Herremans. Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders, 2019.

[18] 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, 2022.

[19] 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 *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[21] Yi Ren, Xu Tan, Tao Qin, Jian Luan, Zhou Zhao, and Tie-Yan Liu. Deepsinger: Singing voice synthesis with data mined from the web, 2020.

A Appendix

A.1 More Experiment Results

Ablation Study Ablation studies are conducted to demonstrate contributions of different proposed modules for the So-VITS system. See the results in Table 2.

Table 2: Results of the ablation study. w/o means without. HV denotes the content encoder using ContentVec [2], HG denotes the adversarial generative module (use HiFi-GAN [7] as baseline vocoder), and Vec denotes the Transformer output of the 12 layer of ContentVec. ProSinger target 1 refers to Jian Li.

	Model	**Naturalness**	**Similarity**
Opencpop target	So-VITS w/o HV	4.15 ± 0.13	3.72 ± 0.17
	So-VITS w/o HG	4.03 ± 0.26	$\mathbf{3.99 \pm 0.20}$
	So-VITS w/o Vec	4.28 ± 0.44	3.602 ± 0.09
	So-VITS	$\mathbf{4.35 \pm 0.11}$	3.98 ± 0.37
ProSinger target 1	So-VITS w/o HV	4.18 ± 0.59	3.91 ± 0.25
	So-VITS w/o HG	4.24 ± 0.91	3.88 ± 0.98
	So-VITS w/o Vec	4.14 ± 0.33	3.38 ± 0.16
	So-VITS	$\mathbf{4.25 \pm 0.76}$	$\mathbf{3.99 \pm 0.92}$

A.2 Architecture

The singer-adaptive encoder for target singer representation follows the design in [21], tentatively. The encoder consists of 1) a pre-net to preprocess the linear-spectrograms of the target singing voice, 2) several Transformer blocks to generate a hidden sequence, and 3) an average pooling module to compress the hidden sequence into the singer embedding which contains the timbre information of the targetsinger. See the encoder architecture in Figure 2.

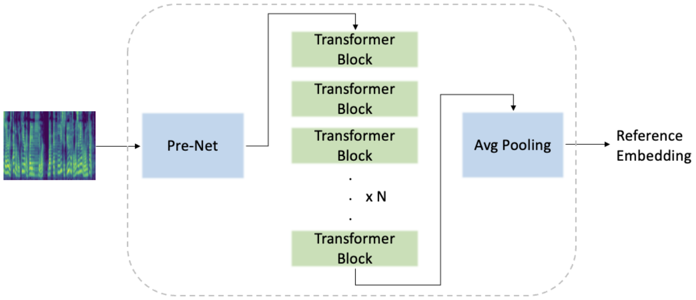

Figure 2: Singer-adaptive encoder for target singer representation