Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Simple and Effective Content Encoder for Singing Voice Conversion via SSL-Embedding Dimension Reduction

Wangjin Zhou[1], *Tianjiao Du*[2], *Chenglin Xu*[3], *Sheng Li*[4], *Yi Zhao*[5], *Tatsuya Kawahara*[1]

[1]School of Informatics, Kyoto University, Japan; [2]Shenzhen International Graduate School, Tsinghua University, China; [3]Nanyang Technological University, Singapore; [4]Institute of Science Tokyo, Japan; [5]Department of Electronic Engineering, Tsinghua University, China

zhou@sap.ist.i.kyoto-u.ac.jp, dtj22@mails.tsinghua.edu.cn, cxu011@e.ntu.edu.sg, sheng.li@ieee.org, zy.2011@tsinghua.org.cn, kawahara@i.kyoto-u.ac.jp

Abstract

In any-to-any singing voice conversion (SVC), singing content can be encoded using either token-based or embedding-based approaches. Token-based methods often struggle with accurate content reconstruction, while embedding-based methods face significant timbre leakage. To address this trade-off, we propose a novel self-supervised learning (SSL)-based content representation method. By randomly selecting a subset of channels during training to serve as the new embedding and fixing them for subsequent SVC training, our approach achieves superior content modeling compared to token-based methods while mitigating timbre leakage typically observed in embedding-based approaches. We validate the effectiveness and generalizability of our method across SSL-based embeddings, SSL-based soft embeddings, and ContentVec.

Index Terms: Singing Voice Conversion, Zero-Shot, Content Representation, Cross Language Domain

1. Introduction

Any-to-any zero-shot singing voice conversion (SVC) aims to transform a singer's timbre to match that of any target singer while retaining the original melody and pitch. Typically, an SVC system is trained to reconstruct the original singing voice using disentangled content, pitch, and timbre representations [1–5]. During inference, the pitch is adjusted to fit the target singer's range, and the timbre representation is replaced with that of the target singer. A key challenge in SVC lies in effectively decoupling timbre information from the content representation. As shown in Figure 1, excessive timbre information in the content representation can cause the model to rely on content embeddings for timbre information instead of the singer condition, resulting in timbre leakage during inference.

Broadly speaking, there are two approaches to addressing this issue. One strategy involves optimizing the model to ensure that it learns timbre information exclusively from the singer condition [6,7]. The other, which is the focus of this paper, aims to reduce timbre leakage by directly minimizing the singer-related information in the content representation without modifying the model architecture.

Research has shown that discrete content representations are effective in removing timbre information from content embeddings [4, 8]. For any-to-any SVC, employing a large codebook for discretized tokens has been shown to preserve fine-grained singing content, such as articulation nuances, while eliminating singer-related information [3, 9]. This technique demonstrates strong potential in achieving accurate timbre conversion and expressive singing reconstruction.

However, our experiments reveal that this method fails to accurately reconstruct content for languages or accents not included in the training data. Even with expanded codebooks containing thousands of tokens, the limited representation space results in distortions in unseen pronunciations, which are often replaced with similar ones from the training data [10, 11]. In contrast, non-discretized embeddings with continuous content encoding spaces do not suffer from these distortions [12, 13]. To disentangle timbre from continuous embeddings, some studies [14, 15] employ supervised learning methods, but these approaches still exhibit significant timbre leakage.

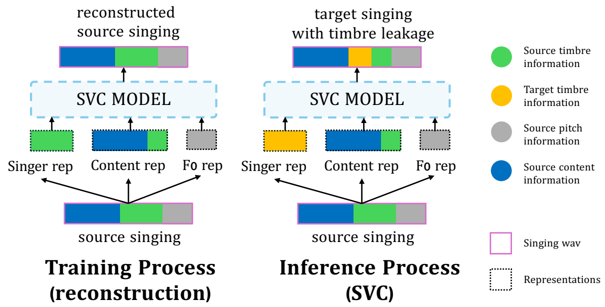

Figure 1: *How timbre leakage happens in SVC*

To address this, we investigate whether timbre information can be explicitly reduced in continuous embeddings. Assuming that timbre and content information are uniformly distributed across the embedding space, simple dimensionality reduction can proportionally reduce both types of information. While this raises concerns about whether the reduced content information can still support accurate singing reconstruction, prior studies on discrete representations suggest that even highly limited spaces can reliably reconstruct singing content within the training domain [4, 8]. We hypothesize that within a reasonable range of dimensionality reduction, timbre-related information can be significantly reduced while preserving sufficient content information for accurate reconstruction. This enables the model to rely on singer embeddings for timbre information, thereby enhancing SVC performance.

In this study, we compare various methods, including original continuous embeddings, ContentVEC [15], SSL-Soft [13] embeddings, discretized tokens, and our proposed dimension-reduction content encoder, all integrated within a modified baseline model based on So-VITS-SVC [1]. Our results show that by randomly selecting a subset of embedding dimensions and fixing them during training, we can achieve timbre leakage reduc-

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

10.21437/Interspeech.2025-1531

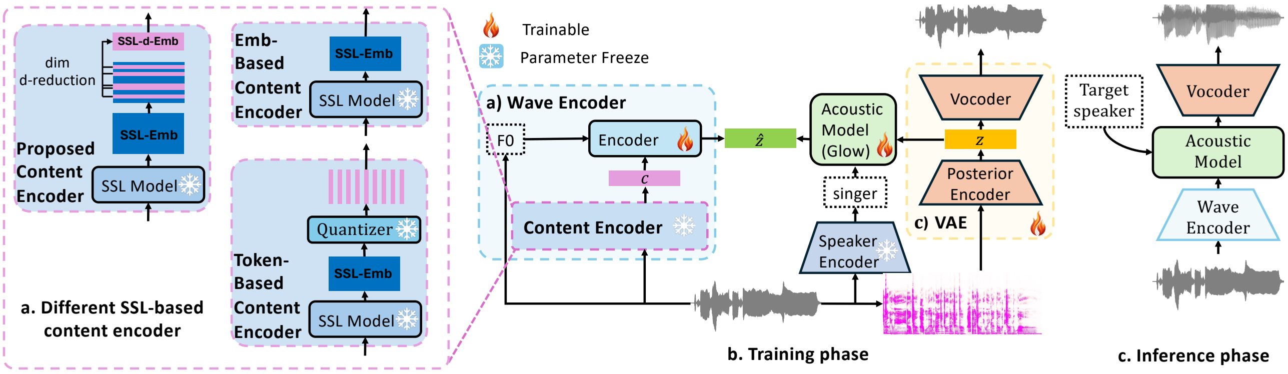

Figure 2: *Overview of our proposed content encoder and the experiment baseline system. To facilitate the comparison of different content encoders, we designed our baseline based on the widely used SVC open-source model, So-VITS-SVC.*

tion on par with token-based methods while improving content reconstruction. To evaluate this, we design an evaluation task where the SVC system is trained on a single-language dataset and tested on a multilingual dataset.

In summary, this paper makes the following key contributions:

- We demonstrate that using token-based methods for SVC leads to unnatural pronunciation, especially for languages outside the training set.
- We show that applying random dimensionality reduction to SSL embeddings and fixing the selected dimensions during training significantly mitigates timbre leakage while maintaining content reconstruction accuracy.
- We introduce a novel evaluation task in which SVC models are trained on a single-language dataset and tested on a multilingual dataset to assess content reconstruction quality.
- We apply dimensionality reduction to ContentVEC, achieving singer similarity on par with token-based systems while improving content reconstruction accuracy. Audio samples are available at our Demo Page [2].

2. Method

2.1. SVC Baseline

To comprehensively evaluate the performance of various embeddings in the SVC task, we adopted a widely used open-source model, So-VITS-SVC, as the baseline. The model comprises three modules: the wave encoder, the acoustic model, and the VAE.

Wave Encoder: The primary function of the Wave Encoder is to encode the source singing. First, the pitch information $F0$ is converted into an embedding, which is then combined with the content embedding c extracted by the content encoder. This combined embedding is processed through an attention mechanism to generate mean and variance vectors for constructing the prior distribution. By experimenting with different content encoders in the Wave Encoder, we aim to maintain content reconstruction accuracy while preventing timbre leakage.

Acoustic Model: We utilize generative flow (Glow) G as the acoustic model [16]. Its fundamental architecture consists of a stack of affine coupling layers, which are built from a series of WaveNet [17] residual blocks. During training, the model maps the posterior distribution $q(z)$, conditioned on the singer s, to the prior distribution $p(\hat{z})$ using G^{-1}, while KL divergence be used as the loss function. During inference, Glow G uses the prior distribution $p(\hat{z})$, conditioned on the singer s, to generate a predicted distribution $q(z^*)$.

VAE: We use a Variational Autoencoder (VAE) [18] for encoding and decoding singing. The process begins by converting the singing from waveform to spectrogram. The posterior encoder (the VAE's encoder) produces the mean and variance, which are used to sample the posterior distribution $q(z)$ from a Gaussian distribution. This sampled distribution $q(z)$ is then used to train the acoustic model. HiFiGAN [19] with NSF [20] is employed as the decoder to reconstruct waveform from $q(z)$. The decoder is trained using reconstruction loss, adversarial loss, and feature mapping loss, in conjunction with the acoustic model, wave encoder, and posterior encoder.

2.2. Proposed Content Encoder

In Figure 3, our idea can be intuitively illustrated using a three-dimensional sphere. By applying basic geometric principles, it is evident that the amount of information contained in a three-dimensional embedding space is greater than that in a dimension-reduced two-dimensional representation space.

Assuming that both timbre and content information are uniformly distributed within the embedding space, reducing dimensionality through random selection proportionally decreases the amount of both types of information. As Figure 3 shows, compared to the timbre information contained in the singer representation, the timbre information in the embedding space is non-negligible, which may lead to potential timbre leakage. However, after dimensionality reduction, the timbre information in the two-dimensional representation space becomes significantly smaller than that in the singer representation, effectively mitigating timbre leakage to some extent.

At the same time, we expect the content information retained in the dimension-reduced representation space to be greater than that in the discrete token representation space, thereby enabling better content reconstruction.

Specifically, by using the open-source HuBERT model [22], we extracted 768-dimensional embeddings $SSL\text{-}Emb$ and applied a dimensionality reduction by randomly selecting d dimensions from its original space. This reduced embedding, $SSL\text{-}d\text{-}Emb$, was used as content

[2] Demo Page: https://expdemos.github.io/Mono2PolySVC/

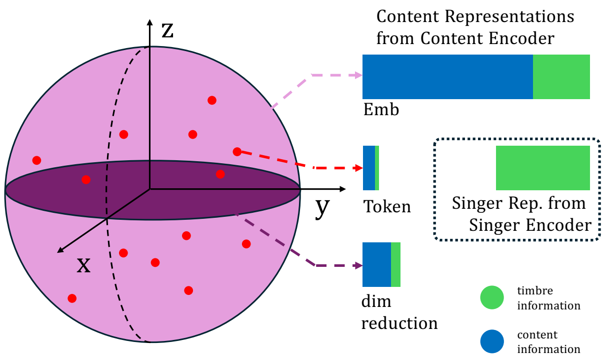

Figure 3: *A schematic illustration of our motivation in a simple three-dimensional space. On the left, the pink sphere represents the entire sample space of the embedding encoding, while the red dots indicate the locations of token representations obtained through the quantizer [21]. The purple region represents the representation space when only two dimensions (in this case, is XY plane) are selected. The right side illustrates the amount of information contained in different representation spaces, where green denotes the timbre information, and blue represents the content information.*

embeddings c during both the training and inference phases of the SVC task. We also explored how different values of d affected the accuracy of content reconstruction and similarity to the target timbre.

3. Experiments

3.1. Datasets

We employed an internal dataset that encompassed 200 hours of Chinese singing from a pool of 10,000 non-professional singers, with each individual contributing roughly one minute of recorded singing. To evaluate the model's performance, we selected sixteen out-of-domain singers as target singers, including eight males and eight females. Additionally, twenty songs in Chinese and twenty songs in other languages (English, Korean, Vietnamese, Japanese, and Cantonese) sung by different individuals were selected to act as source samples. A total of 640 paired SVC samples were used for evaluation. The singers in the training dataset do not appear in the test phase. All recordings in the training and evaluating dataset were captured with a sampling rate of 44.1 kHz.

3.2. Experiment Details

We trained the SVC baseline using different content encoders—SSL embeddings, clustered tokens, ContentVEC, SSL-Soft embeddings, dimension-reduced SSL embeddings (SSL-d-Emb), dimension-reduced SSL-Soft [3] embeddings (SSL-$Soft$-d-Emb), and dimension-reduced ContentVEC ($ContentVEC$-d-Emb)—on the same training dataset to ensure a fair comparison of model performance. Among them, The 'clustered tokens' encoder refers to tokens derived from a pre-trained K-means clustering model with $10,000$ clusters, trained on the same dataset used for training the SVC models. The remaining model inputs were standardized, including pitch F_0 and singer embedding s, as detailed in Section 2.1.

We employ `RMVPE` [23] to extract pitch information $F0$, which is then converted to Log-F0 and fed into a trainable embedding layer. During inference, we adjust the source singer's pitch to match the target singer's vocal range. Specifically, we first compute the mean $F0$ values for both the source and target singers, denoted as mean$(F0_{\text{src}})$ and mean$(F0_{\text{tgt}})$, respectively. The pitch of the source singer is then adjusted by scaling it according to the ratio of these mean values, producing the modified pitch $F0_{\text{src}}^{\text{tgt}}$, as illustrated in Equation 1:

$$F0_{\text{src}}^{\text{tgt}} = F0_{\text{src}} \times \frac{mean(F0_{\text{tgt}})}{mean(F0_{\text{src}})} \tag{1}$$

We directly use pre-trained Speaker Verification model Resnet34 [4] [24] implemented by WeSpeaker to extract a singer embedding s.

All models were trained with the AdamW [25] optimizer on eight A100 GPUs, using a mini-batch size of 80 for $100,000$ steps. During training, we applied a learning rate warm-up strategy for the first epoch. The maximum learning rate was set to 1×10^{-5}, and was reduced by weight decay of 1×10^{-3} each epoch.

3.3. Evaluation Metrics

Objective Metrics: We employ two types of objective metrics to assess the performance of the SVC models. We use Singer Similarity (SSIM) to evaluate the effectiveness of voice conversion. This is done by extracting speaker embeddings through a speaker verification (SV) model and computing the cosine similarity between these embeddings. For this purpose, we utilized two SV models: Wespeaker ResNet34 [26], which was used during the training phase of our model, and CAM++ [5], which was trained on both Chinese and English data.

Subjective Metrics: For testing on each content encoder, we selected 20 samples for subjective evaluation and invited 25 evaluators to assess them. The evaluation focused on two aspects: (1) Content Mean Opinion Score (CMOS 1-bad, 2-poor, 3-fair, 4-good, 5-excellen t), which measured the accuracy and naturalness of pronunciation referred to the source singing; and (2) a 5-point Similarity Mean Opinion Score (SMOS), which evaluate how closely the samples resembled the target singer's voice.

4. Results

The out-of-domain experimental results for different content encoders in the SVC baseline are presented in Table 1. To test the generalizability of our embedding extraction method, we independently chose two sets of low-dimensional representations for the d-dimensional reduction experiment: $SSL-d-Emb1$ and $SSL-d-Emb2$.

Initially, we evaluated the performance of the SSL-Token-based SVC system in both Chinese and other languages. Although the SVC model trained in Chinese showed impressive voice conversion abilities across various languages in terms of singer similarity, its effectiveness in accurately reconstructing content was not satisfactory. The case study on the demo page highlights that some pronunciations differ from the original singing, resulting in less clear articulation. This problem

[3] `https://github.com/bshall/soft-vc`

[4] `https://wespeaker-1256283475.cos.ap-shanghai.myqcloud.com/models/voxceleb/voxceleb_resnet34.zip`

[5] `https://www.modelscope.cn/models/iic/speech_campplus_sv_zh_en_16k-common_advanced`

Table 1: *Objective and subjective evaluation results of various content encoder-based SVC systems. Where dim indicates the channel dimensions of content representation. $SSIM_{wspk}$ refers to singer similarity calculated using WeSpeaker, while $SSIM_{CAM++}$ inferred as singer similarity calculated by CAM++.*

Chinese					
Content Encoder Type	**dim**	**Objective Metrics**		**Subjective Metrics**	
		SSIM$_{wspk}$↑	SSIM$_{CAM++}$↑	CMOS↑	SMOS↑
SSL-Token	768	0.850	0.735	4.086 ± 0.132	$\mathbf{4.122} \pm 0.120$
SSL-Emb	768	0.752	0.521	$\mathbf{4.366} \pm 0.108$	2.750 ± 0.067
SSL-Soft-Emb	256	0.785	0.582	4.298 ± 0.112	3.262 ± 0.103
ContentVEC-Emb	768	0.815	0.656	4.284 ± 0.112	3.752 ± 0.125
SSL-256-Emb1	256	0.813	0.642	4.184 ± 0.108	3.686 ± 0.122
SSL-256-Emb2	256	0.822	0.664	4.258 ± 0.120	3.738 ± 0.127
SSL-128-Emb1	128	0.837	0.681	4.244 ± 0.101	3.920 ± 0.124
SSL-128-Emb2	128	0.839	0.695	4.162 ± 0.113	3.908 ± 0.125
SSL-Soft-128-Emb	128	0.801	0.629	4.280 ± 0.133	3.472 ± 0.112
ContentVEC-256-Emb	256	0.842	0.717	4.224 ± 0.097	4.034 ± 0.121
Other Langauges					
Content Encoder Type	**dim**	**Objective Metrics**		**Subjective Metrics**	
		SSIM$_{wspk}$↑	SSIM$_{CAM++}$↑	CMOS↑	SMOS↑
SSL-Token	768	0.837	0.641	3.210 ± 0.148	4.006 ± 0.117
SSL-Emb	768	0.755	0.462	$\mathbf{4.326} \pm 0.121$	2.494 ± 0.070
SSL-Soft-Emb	256	0.774	0.484	4.266 ± 0.123	3.154 ± 0.091
ContentVEC-Emb	768	0.799	0.554	4.250 ± 0.135	3.512 ± 0.126
SSL-256-Emb1	256	0.805	0.544	4.186 ± 0.134	3.532 ± 0.120
SSL-256-Emb2	256	0.814	0.560	4.204 ± 0.129	3.570 ± 0.128
SSL-128-Emb1	128	0.821	0.591	4.220 ± 0.120	3.726 ± 0.116
SSL-128-Emb2	128	0.821	0.588	4.178 ± 0.131	3.700 ± 0.121
SSL-Soft-128-Emb	128	0.799	0.549	4.230 ± 0.130	3.476 ± 0.116
ContentVEC-256-Emb	256	0.835	0.634	4.256 ± 0.125	$\mathbf{4.038} \pm 0.128$

arises from the overly compressed information by discrete tokens, causing inaccuracies in content reconstruction for languages not included in the training set. These insights highlight the significance of this study.

Both the objective metric SSIM and the subjective metric SMOS indicate that as the number of selected dimensions d decreases, the similarity between the SVC samples and the target timbre increases. The subjective metric CMOS shows that reducing the number of dimensions d does not significantly affect content reconstruction accuracy, demonstrating the feasibility of our proposed method.

Moreover, SSL-256-Emb surpasses the more complexly trained SSL-Soft-Emb, despite having the same number of representation dimensions, illustrating that our system is not only simpler but also more effective. Notably, SSL-128-Emb even outperforms the previous state-of-the-art ContentVEC, highlighting that our straightforward random dimension selection approach can achieve better timbre similarity compared to methods relying on more complex supervised training processes.

Furthermore, our dimension reduction method outperforms each baseline embedding when applied to all three types (SSL-Emb, SSL-Soft-Emb, ContentVEC), demonstrating strong generalization capability. When we applied our proposed dimensionality reduction method directly to ContentVEC, by randomly selecting 256 dimensions from the 768-dimensional ContentVEC embeddings to use as content embeddings, the resulting SVC model achieved comparable similarity metrics to those based on SSL-Token, while significantly improving content reconstruction accuracy. Based on these findings, we propose the Mono2PolySVC model, which allows for training in a single language and inference across multiple languages. We used colored bars in the table to highlight the improvements of this new model over token-based SVC.

5. Conclusion

We introduce a straightforward yet innovative SVC content encoding strategy that effectively reduces timbre leakage by randomly selecting d dimensions from the original content embedding to create a lower-dimensional representation. The result indicates that decreasing the number of selected dimensions enhances control over timbre leakage without sacrificing content reconstruction accuracy. This method's versatility is further confirmed by its successful application across three different types of embeddings. When integrated with ContentVEC, our approach enables an SVC model to achieve singing voice conversion performance comparable to token-based methods in terms of singer similarity, while surpassing them in content reconstruction. To further demonstrate its effectiveness, we conducted experiments where the model was trained on a single-language dataset and evaluated on a multilingual dataset, showcasing its robust performance across different linguistic contexts.

6. References

[1] Y.-J. Luo, C.-C. Hsu, K. Agres, and D. Herremans, "Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders," in *Proc. IEEE-ICASSP*. IEEE, 2020, pp. 3277–3281.

[2] S. Liu, Y. Cao, N. Hu, D. Su, and H. Meng, "Fastsvc: Fast cross-domain singing voice conversion with feature-wise linear modulation," in *Proc. IEEE-ICME*. IEEE, 2021, pp. 1–6.

[3] W.-C. Huang, L. P. Violeta, S. Liu, J. Shi, and T. Toda, "The singing voice conversion challenge 2023," in *Proc. IEEE-ASRu*. IEEE, 2023, pp. 1–8.

[4] W. Zhou, F. Zhang, Y. Liu, W. Guan, Y. Zhao, and H. Qu, "Zero-shot singing voice conversion: Built upon clustering-based phoneme representations," *arXiv preprint arXiv:2407.07728*, 2024.

[5] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[6] S. Liu, Y. Cao, D. Su, and H. Meng, "Diffsvc: A diffusion probabilistic model for singing voice conversion," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 741–748.

[7] S. Chen, Y. Gu, J. Zhang, N. Li, R. Chen, L. Chen, and L. Dai, "Ldm-svc: Latent diffusion model based zero-shot any-to-any singing voice conversion with singer guidance," *arXiv preprint arXiv:2406.05325*, 2024.

[8] D.-Y. Wu, Y.-H. Chen, and H.-Y. Lee, "Vqvc+: One-shot voice conversion by vector quantization and u-net architecture," *arXiv preprint arXiv:2006.04154*, 2020.

[9] Q. Wang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Drvc: A framework of any-to-any voice conversion with self-supervised learning," in *Proc. IEEE-ICASSP*. IEEE, 2022, pp. 3184–3188.

[10] M. Łajszczak, G. Cámbara, Y. Li, F. Beyhan, A. van Korlaar, F. Yang, A. Joly, Á. Martín-Cortinas, A. Abbas, A. Michalski *et al.*, "Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data," *arXiv preprint arXiv:2402.08093*, 2024.

[11] Y. Yang, Y. Kartynnik, Y. Li, J. Tang, X. Li, G. Sung, and M. Grundmann, "Streamvc: Real-time low-latency voice conversion," in *Proc. IEEE-ICASSP*. IEEE, 2024, pp. 11 016–11 020.

[12] T. Jayashankar, J. Wu, L. Sari, D. Kant, V. Manohar, and Q. He, "Self-supervised representations for singing voice conversion," in *Proc. IEEE-ICASSP*. IEEE, 2023, pp. 1–5.

[13] B. Van Niekerk, M.-A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper, "A comparison of discrete and soft speech units for improved voice conversion," in *Proc. IEEE-ICASSP*. IEEE, 2022, pp. 6562–6566.

[14] C. Wang, Z. Li, B. Tang, X. Yin, Y. Wan, Y. Yu, and Z. Ma, "Towards high-fidelity singing voice conversion with acoustic reference and contrastive predictive coding," *arXiv preprint arXiv:2110.04754*, 2021.

[15] K. Qian, Y. Zhang, H. Gao, J. Ni, C.-I. Lai, D. Cox, M. Hasegawa-Johnson, and S. Chang, "Contentvec: An improved self-supervised speech representation by disentangling speakers," in *International Conference on Machine Learning*. PMLR, 2022, pp. 18 003–18 017.

[16] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[17] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, vol. 12, 2016.

[18] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[19] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[20] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. IEEE-ICASSP*. IEEE, 2019, pp. 5916–5920.

[21] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 1177–1178.

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[23] H. Wei, X. Cao, T. Dan, and Y. Chen, "Rmvpe: A robust model for vocal pitch estimation in polyphonic music," *arXiv preprint arXiv:2306.15412*, 2023.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[26] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *Proc. IEEE-ICASSP*. IEEE, 2023, pp. 1–5.