SKQVC: One-Shot Voice Conversion by K-Means Quantization with Self-Supervised Speech Representations

Youngjun Sim*
Graduate School of Artificial Intelligence
POSTECH
Pohang, South Korea
youngjunsim@postech.ac.kr

Jinsung Yoon*
Graduate School of Artificial Intelligence
POSTECH
Pohang, South Korea
truestar2001@postech.ac.kr

Young-Joo Suh†
Graduate School of Artificial Intelligence
POSTECH
Pohang, South Korea
yjsuh@postech.ac.kr

***Abstract*—One-shot voice conversion (VC) is a method that enables the transformation between any two speakers using only a single target speaker utterance. Existing methods often rely on complex architectures and pre-trained speaker verification (SV) models to improve the fidelity of converted speech. Recent works utilizing K-means quantization (KQ) with self-supervised learning (SSL) features have proven capable of capturing content information from speech. However, they often struggle to preserve speaking variation, such as prosodic detail and phonetic variation, particularly with smaller codebooks. In this work, we propose a simple yet effective one-shot VC model that utilizes the characteristics of SSL features and speech attributes. Our approach addresses the issue of losing speaking variation, enabling high-fidelity voice conversion trained with only reconstruction losses, without requiring external speaker embeddings. We demonstrate the performance of our model across 6 evaluation metrics, with results highlighting the benefits of the speaking variation compensation method.**

***Index Terms*—voice conversion, self-supervised learning, disentanglement, quantization**

I. Introduction

One-shot voice conversion (VC) aims to transform the speaker identity of a source into that of an arbitrary target using only a single utterance. This process typically employs disentanglement-based methods to separate content and speaker information, replacing the source speaker's information with that of the target speaker. The key challenge lies in effectively disentangling content and speaker information while preserving both. To address this, various strategies have been proposed, including information bottlenecks [1, 2], additional loss functions [3, 4], normalization techniques [5, 6], and vector quantization (VQ) methods [7–9]. VQ methods capture content information by replacing the input embedding with the nearest vectors from a discrete codebook, which primarily represents phonetic features within the continuous content space, thus removing speaker information. As a result, speaker identity is excluded, and a discrete content embedding is extracted. VQVC [7] disentangles speech attributes from mel-spectrograms using VQ with a randomly initialized codebook and codebook loss [10]. The quantized output is used as the content embedding, and the speaker embedding is extracted from the residual of VQ. However, using a randomly initialized codebook leads to loss of content information during the quantization process, negatively affecting intelligibility and quality. VQMIVC [9] addresses this issue by applying mutual information loss and vector quantization with contrastive predictive coding (VQCPC), enhancing the quality of the converted speech.

Recently, many studies have increasingly employed complex and sophisticated methods to enhance VC performance [11, 12]. Additionally, external speaker embedding derived from pre-trained speaker verification (SV) models has been widely used to extract more accurate speaker information [1, 3]. YourTTS [11], based on a variational autoencoder (VAE) architecture, utilizes a conditional normalizing flow method and a monotonic alignment search (MAS) module [13], while using external speaker embedding as a condition for the posterior encoder and a duration predictor. FreeVC [12] captures content information using SSL features, combined with data perturbation, a bottleneck network, and a conditional normalizing flow method, while employing external speaker embedding to achieve high naturalness and similarity in voice conversion.

SSL features, which are speech representations derived from self-supervised learning (SSL) models such as HuBERT [14] and WavLM [15], have demonstrated the ability to linearly predict various speech attributes [16]. These features are encoded such that instances of the same phone are closer together than different phones, meaning that nearby features share similar phonetic content [17, 18]. Due to this inherent characteristic, SSL features have been increasingly used in recent voice conversion models [12, 19–21]. The models [19–22] use K-means quantization (KQ) with SSL features, a VQ method where the codebook is initialized by applying K-means clustering to SSL features. This codebook effectively

*Equal contribution
†Corresponding author

represents the continuous content space, ensuring that content information from the source speech is well captured while removing speaker information. However, when using a small codebook, the discrete content embedding derived from KQ may lose speaking variation information, including prosodic detail and phonetic variation. This negatively impacts the performance of VC, which relies on precisely following the prosody and content of the source speech. To address this issue, [21] used residual vector quantization (RVQ) [23] with a codebook size of 1024, and [20] employed a codebook size of 2000, along with an energy predictor and a corresponding loss function.

In this paper, we propose a simple disentanglement architecture that leverages the characteristics of SSL features and speech attributes. Specifically, we utilize SSL features containing rich acoustic information to extract not only content-related information but also speaker-related information. Discrete content embedding is captured through KQ, while speaker and speaking variation embeddings are extracted from the residual, through information bottleneck layers without additional constraint losses. This approach enables high-fidelity one-shot VC using only reconstruction losses, eliminating the need for external speaker embedding. Furthermore, by compensating for speaking variation, it preserves the prosody and content accuracy of the source speech even with a small codebook size of 256. Our model demonstrates superior performance across 6 objective and subjective evaluation metrics. The experiments highlight the advantage of the proposed compensation method across different codebook sizes, and ablation studies confirm the critical role of the key components. Audio samples are available at `https://simyoungjun.github.io/SKQVC-DEMO/`

II. Method

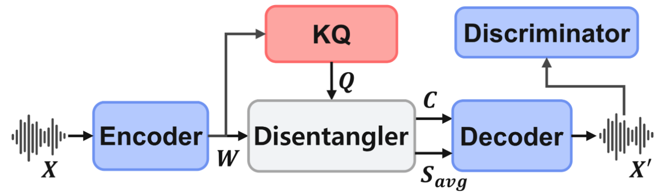
The proposed model architecture is illustrated in Fig. 1(a). It includes a WavLM encoder, K-means quantization, a Disentangler, and a HiFi-GAN [24] decoder with a discriminator. Further details are provided in the following section.

A. Encoder

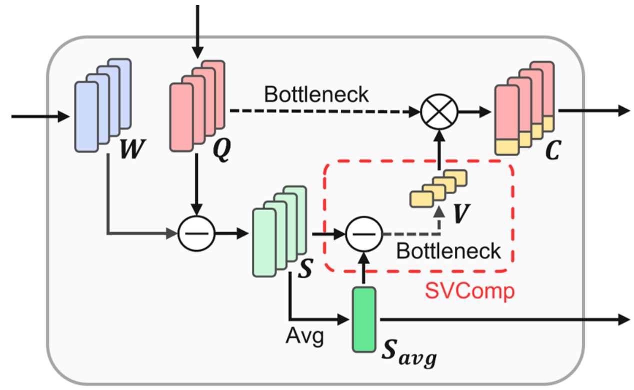
Instead of traditional representations such as spectrograms or mel-spectrograms, the proposed model utilizes SSL features obtained from the 6^{th} layer of the WavLM model. WavLM is pre-trained on large datasets and designed for various speech processing tasks, demonstrating superior performance in both ASR and non-ASR tasks. Specifically, features from the 6^{th} layer of WavLM have been shown to contain richer acoustic information, including pitch, prosody, and speaker identity, compared to later layers [17, 25]. The input waveform $X = \{x_1, x_2, \ldots, X_L\} \in \mathbb{R}^{1\times L}$, where L is the number of the waveform samples, is fed into the WavLM encoder, and continuous SSL features $W = \{w_1, w_2, \ldots, w_T\} \in \mathbb{R}^{1024\times T}$, where T is the number of frames, are extracted.

B. K-means Quantization

Discrete content embedding $Q = \{q_1, q_2, ..., q_T\} \in \mathbb{R}^{1024\times T}$ is obtained from SSL features by quantization, as



(a) SKQVC



(b) Disentangler

Fig. 1: The SKQVC model. (a) The overall architecture of SKQVC. (b) The Disentangler module.

shown in Eq. 1. The codebook used for quantization is initialized through applying K-means clustering to the SSL features of the training dataset. Each code vector in the codebook corresponds to the centroid vector of a cluster, and the codebook is fixed during the model's training process [21]. Nearby SSL features share similar phonetic features [17, 18], meaning that variations in these feature values are primarily driven by changes in phonetic content rather than other speech attributes, such as pitch, stress, or speaker identity. Therefore, the codebook initialized through K-means clustering is strongly content-related, with information about other attributes being removed. Furthermore, as the codebook size decreases, it loses the ability to capture the continuous variations in phonetic features, which reduces content accuracy.

$$KQ\left(w_t\right) = q_t; \quad q_t = \arg\min_{q\in\text{Codebook}} \left(\|w_t - q\|_2^2\right). \tag{1}$$

C. Disentangler

The residual of quantization, which is the difference between W and Q, denoted as $S = \{s_1, s_2, ..., s_T\} \in \mathbb{R}^{1024\times T}$, contains a mix of speaking variation and speaker information. The Disentangler, shown in Fig. 1(b), separates these attributes through an information bottleneck that utilizes their unique characteristics. First, inspired by VQVC [7], the speaker embedding $S_{avg} \in \mathbb{R}^{1024\times 1}$, which contains speaker information, is extracted averaging across the time dimension. Second, in the Speaking Variation Compensation layer (SVComp), the difference between S and S_{avg} is passed through a bottleneck layer, enforcing dimensionality reduction on the

TABLE I: Subjective and objective evaluation results for MOS, SMOS, WER, CER, and EER across seen-to-seen and unseen-to-unseen scenarios. Ground truth (GT) using source utterances is provided for reference.

Model	seen-to-seen					unseen-to-unseen				
	MOS↑	SMOS↑	WER(%)↓	CER(%)↓	EER(%)↓	MOS↑	SMOS↑	WER(%)↓	CER(%)↓	EER(%)↓
VQMIVC [9]	3.17±0.17	2.66±0.16	49.85	29.90	30.1	1.74±0.13	1.58±0.12	62.71	39.13	39.0
YourTTS [11]	3.03±0.19	3.41±0.15	31.22	16.27	5.0	1.98±0.13	2.09±0.15	41.04	23.65	15.9
FreeVC [12]	3.88±0.14	4.10±0.13	10.36	4.02	3.4	3.79±0.13	2.56±0.16	10.64	5.70	16.4
SKQVC	**3.91±0.13**	**4.28±0.12**	**8.42**	**3.32**	**3.1**	**3.84±0.14**	**3.95±0.16**	**10.05**	**5.37**	**10.8**
GT	4.24±0.15	-	5.53	1.79	-	4.31±0.13	-	8.09	4.31	-

feature space to obtain the speaking variation embedding $V = \{v_1, v_2, ..., v_T\} \in \mathbb{R}^{8\times T}$. Speaking variation, which includes prosodic detail and phonetic variation, is a time-variant attribute, whereas speaker information remains time-invariant. Therefore, these attributes can be disentangled through information bottleneck, without additional constraint losses. Finally, the discrete content embedding Q is passed through the bottleneck layer to reduce the dimension from 1024 to 1016. It is then concatenated with V to form the continuous content embedding $C = \{c_1, c_2, ..., c_T\} \in \mathbb{R}^{1024\times T}$, thereby compensating for speaking variation.

D. Decoder

The decoder and discriminator are based on HiFi-GAN. The training losses for reconstruction include the adversarial loss $L_{adv}(G)$ and $L_{adv}(D)$ for the generator G and the discriminator D, feature matching loss L_{fm}, and mel-spectrogram loss L_{mel}, as shown in Eq. 2. Here, L_{mel} is measured as the L_1 distance between the mel-spectrograms of real and reconstructed waveforms, and L_{fm} is the L_1 feature matching loss of the intermediate output of the discriminator.

During training, the speaker embedding S_{avg} and the content embedding C, which are the outputs of the Disentangler module, are added together and fed into the decoder. The decoder then reconstructs the original waveform that was input into the encoder, and the training losses are calculated. During inference, source speech and target speech are input into the model and separated by the Disentangler module. The content embedding from the source speech and the speaker embedding from the target speech are added together and fed into the decoder to generate the converted waveform.

$$\begin{aligned} L_G &= L_{adv}(G) + \lambda_{fm} L_{fm} + \lambda_{mel} L_{mel}, \\ L_D &= L_{dav}(D). \end{aligned} \tag{2}$$

III. EXPERIMENTS

A. Experimental Setups

Experiments are conducted using the VCTK [26] and LibriTTS [27] datasets. For the VCTK dataset, the validation set consists of 214 utterances (2 per speaker), while the test set contains 1,070 utterances (10 per speaker) from 107 speakers, with the remaining utterances allocated to the training set. Additionally, the test-clean subset of LibriTTS is used as another test set. The VCTK dataset is recorded in a controlled studio for consistent quality, while the LibriTTS dataset, derived from audiobooks, has more variable recording conditions. All audio is downsampled to 16kHz. The FFT, window, and hop sizes are set to 1280, 1280, and 320, respectively. The codebook of KQ is initialized using the Minibatch K-means algorithm [28], with a batch size of 1024 and 256 clusters, implemented on the training set. The bottleneck layer is implemented using a 1D convolution layer with a kernel size of 1. The values of λ_{fm} and λ_{mel} in the training loss are set to 2 and 45, respectively. Three baseline models, VQMIVC [9], YourTTS [11], and FreeVC [12], are selected for comparison with the proposed method.

B. Evaluation Metrics

Both subjective and objective evaluations are conducted. For the subjective evaluation, 50 participants are asked to rate naturalness and similarity on a 5-point scale using the mean opinion score (MOS) and the similarity mean opinion score (SMOS), both of which are calculated with 95% confidence intervals. Experiments are performed in both seen-to-seen and unseen-to-unseen scenarios, with 16 seen speakers from the VCTK test set and 16 unseen speakers from the LibriTTS test set. For the objective evaluation, four metrics are used. The word error rate (WER) and character error rate (CER) between the source and converted speech are obtained using an ASR model [1], assessing the intelligibility of the converted speech. The equal error rate (EER) is calculated using a pre-trained SV model [29] to evaluate speaker similarity. The F0-PCC, the Pearson correlation coefficient [30], measures the correlation between the F0 of the source and converted speech, with a score range of -1 to 1, where higher values indicate better preservation of the prosody [12]. For evaluation, 2000 utterances are randomly selected from the VCTK test set and the LibriTTS test set, respectively, and 1000 seen-to-seen and unseen-to-unseen converted utterances are generated.

C. Voice Conversion Results

As shown in Table I, the proposed method outperforms in both seen-to-seen and unseen-to-unseen voice conversion scenarios, achieving high naturalness, intelligibility, and similarity

[1] https://huggingface.co/facebook/hubert-large-ls960-ft

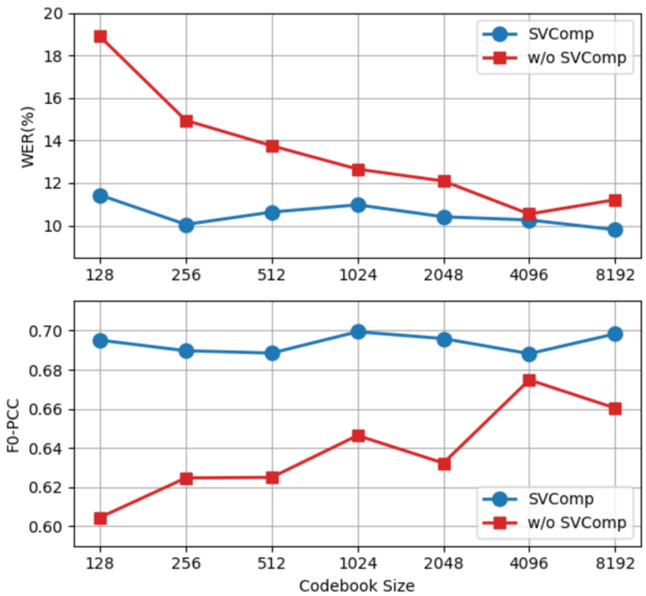


Fig. 2: Comparison of SKQVC models with and without speaking variation compensation across different codebook sizes in unseen-to-unseen scenarios.

based on both subjective and objective evaluations. Specifically, in terms of WER and CER scores, the model exhibits the highest intelligibility, indicating that content accuracy is well preserved through SVComp, even with a small codebook size of 256. SMOS and EER scores further reveal the model's ability to accurately extract speaker information from the residual without requiring external speaker embeddings. Additionally, the increased difference in SMOS and EER scores between unseen-to-unseen and seen-to-seen scenarios suggests that the model maintains robustness in similarity, even when handling unseen datasets with notable quality fluctuations.

D. Speaking Variation Compensation

Fig. 2 demonstrates the advantage of the proposed SVComp method. 6 different codebook sizes, ranging from 128 to 8192, are evaluated. The model without SVComp shows degradation in both WER and F0-PCC scores as the codebook size decreases. In contrast, the model with SVComp sustains robust performance in both WER and F0-PCC scores even as the codebook size decreases. This indicates that the proposed compensation method effectively addresses the issue of losing speaking variation by preserving the content accuracy and prosody of the source speech.

E. Ablation Studies

To investigate the importance of the proposed methods in SKQVC, experiments are conducted using three ablation systems, as shown in Table II: (1) removing the bottleneck layer in SVComp (-bottleneck), which extracts crucial time-variant information; (2) replacing the input SSL features from the WavLM 6^{th} layer with those from the last (24^{th}) layer (-WavLM 24^{th} layer); and (3) using an external speaker embedding instead of one extracted from the residual of KQ (-external speaker embedding). In (3), the external speaker embedding is not used with SVComp due to distributional differences between the embeddings.

TABLE II: Objective evaluation results of ablation systems in unseen-to-unseen scenarios.

Model	WER(%)↓	CER(%)↓	EER(%)↓
SKQVC	10.05	5.37	10.8
-bottleneck layer	9.96	5.09	16.8
-WavLM 24^{th} layer	10.94	5.27	31.8
-external speaker embedding	14.27	7.6	15.3

When the bottleneck layer is removed, the WER decreases by 0.09, while the EER increases by 6.0. Although there is a slight trade-off between intelligibility and similarity, this suggests that the bottleneck layer preserves crucial speaking variation information while filtering out speaker information. When the WavLM 24^{th} layer is used instead, a significant increase in EER is observed. This indicates that the WavLM 6^{th} layer, which contains richer acoustic information, plays a critical role in extracting speaker information from the residual of quantization. When the external speaker embedding is used without SVComp, a notable increase in WER/CER is observed, demonstrating that SVComp plays an essential role in preserving content accuracy. The increase in EER further suggests that extracting speaker embedding from residual yields more accurate speaker information compared to using external speaker embedding.

IV. CONCLUSION

In this work, we propose a simple yet effective approach to one-shot voice conversion. By applying the methods of K-means quantization and an information bottleneck, the model disentangles speech attributes from SSL features without additional constraint losses. This approach addresses the common issue of losing speaking variation, while also eliminating the need for external speaker embeddings. The results demonstrate that the proposed model outperforms existing methods in both objective and subjective evaluations, proving its robustness across diverse conditions and datasets. Furthermore, the experiments underscore the benefit of the proposed compensation method across varying codebook sizes, and ablation studies validate the importance of key components in the system's performance.

V. ACKNOWLEDGMENT

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH)) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1A6A1A03052954).

REFERENCES

[1] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

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

[3] D. Li, X. Li, and X. Li, "Dvqvc: An unsupervised zero-shot voice conversion framework," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[4] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Vq-cl: Learning disentangled speech representations with contrastive learning and vector quantization," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

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

[6] J.-c. Chou, C.-c. Yeh, and H.-y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *arXiv preprint arXiv:1904.05742*, 2019.

[7] D.-Y. Wu and H.-y. Lee, "One-shot voice conversion by vector quantization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7734–7738.

[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] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," *arXiv preprint arXiv:2106.10132*, 2021.

[10] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[11] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[12] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] 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.

[14] 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.

[15] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[16] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[17] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," *arXiv preprint arXiv:2305.18975*, 2023.

[18] E. Dunbar, N. Hamilakis, and E. Dupoux, "Self-supervised language learning from raw audio: Lessons from the zero resource speech challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1211–1226, 2022.

[19] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," *arXiv preprint arXiv:2104.00355*, 2021.

[20] J. Li, Y. Guo, X. Chen, and K. Yu, "Sef-vc: Speaker embedding free zero-shot voice conversion with cross attention," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 296–12 300.

[21] L. Ma, X. Zhu, Y. Lv, Z. Wang, Z. Wang, W. He, H. Zhou, and L. Xie, "Vec-tok-vc+: Residual-enhanced robust zero-shot voice conversion with progressive constraints in a dual-mode training strategy," *arXiv preprint arXiv:2406.09844*, 2024.

[22] X. Zhu, Y. Lv, Y. Lei, T. Li, W. He, H. Zhou, H. Lu, and L. Xie, "Vec-tok speech: Speech vectorization and tokenization for neural speech generation," *arXiv preprint arXiv:2310.07246*, 2023.

[23] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[24] 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.

[25] G.-T. Lin, C.-L. Feng, W.-P. Huang, Y. Tseng, T.-H. Lin, C.-A. Li, H.-y. Lee, and N. G. Ward, "On the utility of self-supervised models for prosody-related tasks," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1104–1111.

[26] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, pp. 271–350, 2019.

[27] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[28] O. Kramer and O. Kramer, "Scikit-learn," *Machine learning for evolution strategies*, pp. 45–53, 2016.

[29] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

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