ASRQ-VC: ASR-Guided Speech Content Quantization for High-Fidelity Voice Conversion

Songting Liu*, Deheng Ye†, Wei Yang† Haoyang Li*, Eng Siong Chng*,
*Nanyang Technological University, Singapore
Email: LIUS0114@e.ntu.edu.sg
†Tencent Inc., Singapore
Email: dericye@tencent.com

***Abstract*—Voice conversion (VC) seeks to transform the voice of a source speaker's voice to a target speaker while preserving the original linguistic content. A persistent challenge in this task is source speaker timbre leakage, where residual characteristics of the source voice remain after conversion. Existing methods face a tradeoff: using narrow bottleneck features often removes source speaker timbre but at the expense of linguistic fidelity, whereas wide bottlenecks preserve more content but risk retaining speaker traits. To address this trade-off, we propose ASRQ-VC, a voice conversion framework that integrates automatic speech recognition (ASR) loss into the VQ-VAE encoder with finite scalar quantization (FSQ) to enforce speaker-invariant linguistic content representations. These discrete representations are then utilized within a VITS-based synthesis architecture enhanced with speaker verification embeddings and a timbre encoder for fine-grained identity modeling. Experiments on LibriTTS and VCTK datasets show state-of-the-art performance, achieving 11.2% WER and 0.460 SECS, surpassing prior SOTA method by absolute value 2.6% and 2.7% respectively. Audio samples are available on https://asq-vc.github.io/**

I. Introduction

Voice conversion (VC) aims to transform source speech to match a target speaker's voice while preserving linguistic content. A key challenge lies in open-domain any-to-any scenarios, where systems must disentangle speaker-independent content (e.g., phonemes, prosody) from speaker-specific attributes (e.g., timbre, accent) without requiring paired training data.

Existing methods to generate content representation for VC frameworks face inherent trade-offs. While self-supervised learning (SSL) models like HuBERT [1] provide rich content representations, their continuous features suffer from speaker timbre leakage. To addresss leakage, quantization-based approaches (e.g., K-means clustering, VQ-VAE [2]) can be used to suppress speaker information but it could degrade content fidelity through overly restrictive bottlenecks. Recent attempts to balance these objectives (e.g. ContentVec [3], CosyVoice [4]) still struggle with residual speaker traces or linguistic inaccuracies in zero-shot settings.

To address the above, we propose ASRQ-VC, a framework which exploits automatic speech recognition (ASR) as additional cost function to focus on linguistic content, with finite scalar Quantization (FSQ [5]) as the process to suppress speaker information in the VQ-VAE encoder stage. This content representation is then applied to a VITS [6] synthesis architecture. Our key contributions include the following:

- A FSQ bottleneck feature trained via VQ-VAE with explicit ASR supervision, forcing discrete representations to retain more linguistic content while reducing speakers' timbre.
- A hierarchical synthesis model integrating transformer-based timbre encoding, speaker verification embeddings, and adaptive codebook designs for robust any-to-any conversion.

To demonstrate the effectiveness of our approach, evaluations on LibriTTS and VCTK show that our method outperforms existing approaches by 2.6% absolute WER reduction (11.2% vs. 13.8% in Vec2Wav2 [7]) and 2.7% absolute SECS improvement (0.460 vs. 0.433 in Vec2Wav2), with metrics computed using HuBERT-large ASR and WavLM-large TDNN speaker verification models. Ablation studies further validate the necessity of ASR supervision. We also note that careful tuning of codebook sizing is necessary. Our proposed ASRQ-VC advances the practical deployment of voice conversion systems by mitigating the content-speaker fidelity trade-off.

II. Related Works

A core challenge in voice conversion lies in disentangling speaker-independent linguistic content from speaker-specific characteristics. Early approaches like AutoVC [8] introduced information bottlenecks to achieve unsupervised decomposition, but suffered from limited content fidelity. Recent methods leverage self-supervised learning (SSL) models (e.g., HuBERT, WavLM [9]) for rich content representations, yet their continuous features inherently retain source speaker residual timbre information. To mitigate this, quantization-based techniques such as VQ-VAE in Speech Resynthesis [10] map SSL features to discrete tokens, trading off between content preservation (narrow bottlenecks) and speaker leakage suppression (wider bottlenecks).

For any-to-any voice conversion, recent works focus on zero-shot generalization to unseen speakers. YourTTS [11] integrates text supervision during training but requires transcriptions, limiting scalability. FreeVC [12] achieves strong zero-shot performance through continuous SSL features and data augmentation, yet suffers from timbre leakage. Discrete token-based approaches like Vec2Wav 2.0 and CosyVoice combine

979-8-3315-7206-8/25/$31.00

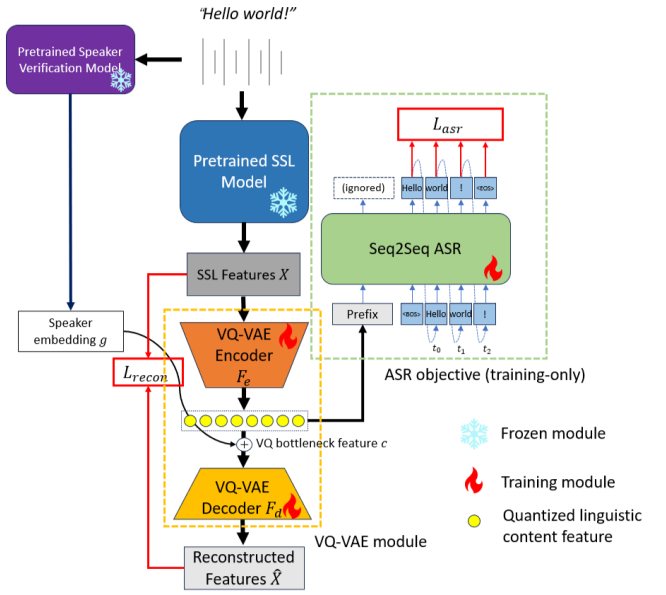

Fig. 1. This figure illustrates the ASR-guided VQ-VAE module in ASRQ-VC framework. The dash-box shows the additional ASR objective to guide the discrete bottleneck in preserving linguistic content while discarding speaker information.

quantized SSL or ASR tokens with advanced vocoders, achieving conversions with high speaker similarity, but struggles in keeping content fidelity. While these methods demonstrate progress, they either rely on suboptimal quantization strategies or lack explicit supervision for content preservation. Our proposed ASRQ-VC bridges this gap through ASR-guided finite scalar quantization and hierarchical synthesis with timbre-aware conditioning. Our approach is described below.

III. Methods

A. Problem Formulation

Voice conversion aims to transform speech from a source speaker's voice to a target speaker's voice while preserving the linguistic content. Formally, given a source speech utterance A and a reference speech utterance B, the goal is to synthesize speech C that maintains the linguistic content of A while adopting the speaker identity of B. This task requires two key components: a speaker-independent content extractor and a speaker information extractor. The content extractor removes speaker-specific characteristics from the source speech A, while the speaker information extractor captures the target speaker's identity information in reference speech B. During training, we simplify this process by using the same utterance as the source, reference, and the target (A = B = C), allowing the model to learn to map content and speaker information back to the original speech. This approach provides direct supervision for the synthesis process while maintaining the essential disentanglement between content and speaker characteristics.

B. Content Extraction with Guided Quantization

We propose to add an additional ASR objective to guide VQ-VAE training, as depicted in Figure 1. Our key insight is that traditional VQ-VAE training with L1/L2 reconstruction loss treats all information equally, lacking explicit guidance for preserving content while discarding speaker characteristics. We address this by incorporating an ASR objective that actively guides the quantization process toward maintaining linguistic content. Our model architecture consists of three primary components:

VQ-VAE Encoder: Maps SSL features $\mathbf{X}$ to a latent space through vector quantization:

$$\mathbf{c} = \mathrm{VQ}(\mathbf{F_e}(\mathbf{X})) \tag{1}$$

The output from VQ-VAE encoder are quantized linguistic content features, which will serve as the VC model input.

VQ-VAE Decoder: Reconstructs the SSL features from the quantized representations. Since SSL features inherently contain speaker-dependent characteristics. This contradicts the objective of learning speaker-invariant representations via the VQ bottleneck. To resolve this conflict, we explicitly provide the speaker embedding g from pretrained speaker verification model CAM++ [13] as conditioning inputs to the decoder:

$$\hat{\mathbf{X}} = \mathbf{F_d}(\mathbf{c}, g) \tag{2}$$

VQ-VAE decoder is only used during VQ-VAE training. It only provides gradient flow to train VQ-VAE encoder but will not be used in VC model training or inference.

ASR Head: To include ASR cost function into VQ-VA, an auto-regressive transformer [14] is included during training to predict the N-token transcript $t_0, t_1, ..., t_N$ given the reconstructed features:

$$P(t_n|\hat{\mathbf{X}}, t_0, t_1, ..., t_{n-1}), n <= N \tag{3}$$

The ASR head takes the reconstructed features from decoder output as prefix conditions and performs next-token prediction of the corresponding transcript. Same as the VQ-VAE decoder, ASR head is only used in VQ-VAE training.

The overall training objective combines reconstruction loss with teacher-forced cross-entropy loss from the ASR prediction:

$$\mathcal{L}_{tot} = \lambda_1 \mathcal{L}_{recon} + \lambda_2 \mathcal{L}_{asr} \tag{4}$$

where $\mathcal{L}_{recon}$ is the L1 loss between original and reconstructed features, and $\mathcal{L}_{asr}$ is the cross-entropy loss for transcript prediction. For simplicity, we set both λ_1 and λ_2 to be 1.0 in the following experiments.

C. Finite Scalar Quantization

We adopt finite scalar quantization [5] to discretize SSL features into speaker-agnostic content representations. Unlike standard vector quantization (VQ), FSQ applies per-dimension scalar quantization with hierarchical codebooks, enabling efficient codebook utilization without auxiliary losses. Concretely, given SSL features $\mathbf{X} \in \mathbb{R}^{\mathbf{T}\times\mathbf{D}}$, the content encoder projects

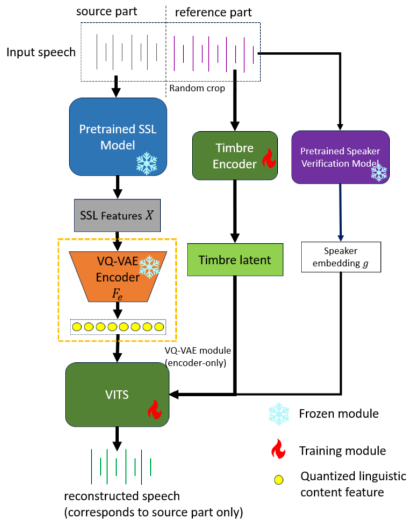

Fig. 2. Training architecture of the VITS-based voice conversion system. During training, the VITS model takes in linguistic content feauture, speaker embedding and timbre latent from input speech as optimized to reconstruct the input speech waveform. The timbre encoder is introduced to provide more fine-grained timbre details besides speaker embedding. To avoid inference time out-of-distribution, input speech is randomly segmented into source part and reference part, where the source part serves as VITS reconstruction target, and the reference part is used as input of timbre encoder and SV model to capture speaker identity features.

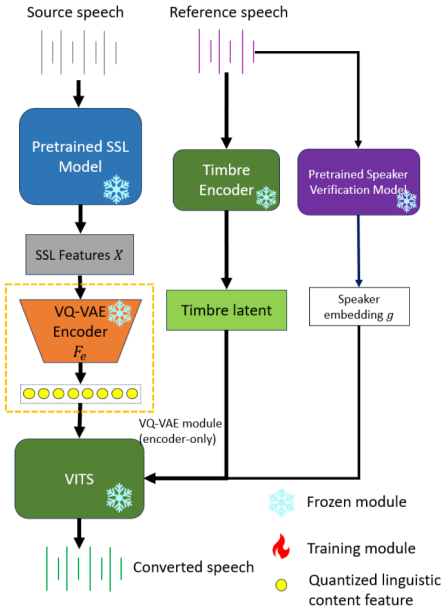

Fig. 3. Inference pipeline of ARSQ-VC. the VITS-based voice conversion system. The difference between training is that timbre latent and speaker embedding are now from reference speech rather than source speech. The model performs voice conversion by performing non-parallel reconstruction with linguistic content feature from source speech and speaker information from reference speech.

them to a latent space $\mathbf{c} \in \mathbb{R}^{\mathbf{T}\times\mathbf{d}}$, which is then quantized into discrete codes through scalar thresholding across multiple resolution levels (e.g., [8,5,5] for 200-vocab). The key innovation lies in joint training with ASR supervision: the cross-entropy loss from the ASR head explicitly penalizes quantization-induced content distortion, forcing the bottleneck to preserve phonemic information while discarding speaker traits.

D. Voice Conversion Framework

The discrete content representations obtained from the VQ-VAE encoder are applied into a modified VITS [6] architecture for high-quality speech synthesis. VITS (Variational Inference Text-to-Speech) originally is a state-of-the-art, end-to-end neural network architecture designed for text-to-speech (TTS) synthesis. VITS takes text as input along with desired target speaker's identify to generate speech in target speaker's timbre. Its internal modules are a text encoder, conditional VAE framework, adversarial learning (GANs), normalizing flows, and vocoder.

In recent years, VITS has been adapted for voice conversion (VC) to enabling it to transform source speaker's voice into another while preserving the source content and prosody For voice conversion, the text input of VITS is replaced with linguistic content and prosody features from the source audio.

Figures 2 shows the training and pipeline. During training, each source utterance is randomly split into two segments: source part and reference part. The reference part serves as a timbre reference, while the source segment is used as the reconstruction target.

During training, the VITS model learns to resynthesize speech using the quantized linguistic content features alongside speaker identity features. Specifically, the linguistic content is extracted using the pretrained VQ-VAE encoder (described in Figure 1), while speaker identity is represented through embeddings from a pretrained speaker verification (SV) model and an additional timbre encoder. The timbre encoder, jointly trained with the VITS model, captures more nuanced speaker characteristics beyond the global speaker embedding, and passed the timbre latent information to VITS through cross-attention mechanism, thus enhancing voice similarity in the converted speech. This dual-path speaker conditioning allows the model to perform high-fidelity any-to-any voice conversion with improved control over fine-grained timbre attributes.

Finally, Figure 3 illustrates the inference pipeline. Here the reference speech is a segment of the desired target speaker's voice, while the source speech comes from a different person.

IV. Experiments

A. Implementation Details

1) VQ-VAE Training: The VQ-VAE was trained on a combined English-only dataset of Gigaspeech [15] (10k hours) and MLS English subset [16] (40k hours) for 300k steps with equivalent batch size of 64. The ASR vocabulary consist of 32

tokens, including 26 English letters, space, and special characters. The VQ-VAE architecture employs 12-layer ConvNeXt v2 [17] blocks for both encoder and decoder. We experiment with two codebook configurations: a smaller version with 40 vocab size and a larger version with 200 vocab size. The SSL features are extracted from the 18th layer of HuBERT-large model. The ASR head is a decoder-only transformer incorporating RoPE [18] embedding, RMS Norm [19], and SwiGLU [20] activation, taking VQ-VAE output as prefix conditions.

2) Voice Conversion Model Training: We modify the VITS architecture by replacing its phoneme encoder with a content encoder of similar design to our ASR head: 8 transformer layers with 512 hidden dimension and 1536 intermediate dimension. Each layer includes an additional cross-attention module to attend to timbre encoder outputs. The timbre encoder processes downsampled (2×) reference audio mel-spectrograms through an identical transformer architecture. We utilize the pretrained CAM++ [13] model to extract speaker embeddings. The vocoder operates at 22.05kHz sampling rate with window size 1024, hop size 256, and 80 mel bins. The model was trained on LibriTTS (train-clean-100, train-clean-360, and train-other-500) for 300k with equivalent batch size of 32. We employ AdamW optimizer with initial learning rate 0.0001 and exponential decay of 0.999875 per step.

B. Evaluation Setup

For evaluation, we randomly selected 100 utterances from VCTK as source speech and 10 speakers as targets, with one reference utterance per target speaker. We assess the conversion quality through multiple metrics:

- Word Error Rate (WER) using HuBERT-large-ls960-ft[1] for speech intelligibility and source content preservation;
- Speaker Embedding Cosine Similarity (SECS) using WavLM-large TDNN[2] to measure the speaker similarity of the converted speech to both source speech and reference speech;
- DNSMOS [21] for speech quality (SIG, BAK, and OVRL scores) assessment.

C. Results and Analysis

The experimental results, as summarized in Table I, demonstrate that our proposed model achieves state-of-the-art performance in both source speaker information removal and target speaker similarity.

Effective Removal of Source Speaker Information. Our ASRQ-VC achieves the lowest similarity to the source speech (**SECS=0.103** with the 40-vocab configuration), indicating its effectiveness in suppressing residual timbre leakage. This suggests that our ASR-guided quantization successfully eliminates speaker-specific characteristics while preserving linguistic content.

Superior Target Speaker Mimicry. Simultaneously, our approach achieves the highest similarity to the reference speaker (**SECS=0.460**), outperforming previous methods such as Vec2Wav2. This demonstrates the effectiveness of our timbre encoder and speaker verification embeddings in capturing and reconstructing speaker identity.

Maintaining a Relatively Low WER. Despite its strong speaker information suppression, ASRQ-VC maintains a competitive word error rate (**WER=11.2%** with 200-vocab), showing that it effectively balances linguistic content preservation and speaker conversion performance.

Trade-off Between WER and Speaker Similarity. Comparing the results for the 40-vocab and 200-vocab configurations reveals a trade-off between WER and speaker similarity. Increasing the vocabulary size from 40 to 200 leads to an

TABLE I
PERFORMANCE COMPARISON WITH BASELINE METHODS ON VCTK TEST SET. WER EVALUATE CONTENT PRESERVATION, SECS MEASURES SPEAKER SIMILARITY TO BOTH SOURCE AND REFERENCE SPEECH, AND DNSMOS (OVRL) ASSESSES SPEECH QUALITY. BEST RESULTS ARE MARKED IN **BOLD**. * INDICATES DATA-INTENSIVE BASELINES.

Model	Training Data	WER↓	SECS↑ (to *ref*)	SECS↓ (to *src*)	DNSMOS (OVRL)↑
Ground Truth	-	6.1	0.645	0.062	3.19
YourTTS	LibriTTS (585h)	12.1	0.331	0.191	3.27
FreeVC		11.3	0.373	0.182	3.21
Diff-Hier VC [22]		16.3	0.298	0.197	3.13
DDDM-VC [23]		18.5	0.276	0.182	3.06
Vec2Wav2		13.8	0.433	0.164	3.27
FAcodec*	LibriLight (60kh)	**10.9**	0.349	0.183	3.23
CosyVoice*	Proprietary (170kh)	19.0	**0.464**	0.175	3.28
ASRQ-VC-40 (Ours)	LibriTTS (585h)	13.0	**0.460**	**0.103**	3.28
ASRQ-VC-200 (Ours)		**11.2**	0.452	0.121	**3.30**

TABLE II
ABLATION STUDY RESULTS COMPARING DIFFERENT CODEBOOK SIZES, THE IMPACT OF ASR LOSS AND TIMBRE ENCODER. INTEGER LISTS IN BRACKETS INDICATES FSQ CODEBOOK LEVEL CONFIGURATIONS

Vocab Size	ASR Loss	Timbre Encoder	WER↓	SECS↑ (to *ref*)	SECS↓ (to *src*)
GT	-	-	6.1	0.645	0.062
8 ([8])	✗	✗	25.3	0.4211	0.106
	✓	✗	19.8	0.43	0.085
	✗	✓	27.6	0.461	0.105
	✓	✓	**20.0**	**0.467**	**0.088**
40 ([8, 5])	✗	✗	17.2	0.415	0.122
	✓	✗	12.9	0.424	0.107
	✗	✓	16.5	0.448	0.129
	✓	✓	**13.0**	**0.46**	**0.103**
200 ([8, 5, 5])	✗	✗	11.7	0.407	0.136
	✓	✗	11.1	0.414	0.120
	✗	✓	12.4	0.435	0.132
	✓	✓	**11.2**	**0.452**	**0.121**
1000 ([8, 5, 5, 5])	✗	✗	11.2	0.397	0.164
	✓	✗	10.1	0.404	0.139
	✗	✓	11.3	0.429	0.159
	✓	✓	**10.6**	**0.439**	**0.137**

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

[2]https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification

improvement in WER (**13.0%** $\rightarrow$ **11.2%**) but slightly reduces SECS to the reference speaker (**0.460** $\rightarrow$ **0.452**). This suggests that while larger codebooks retain more linguistic details, they also preserve minor speaker-specific features.

Comparison with Large-Scale Baselines. Even when compared with baselines trained on significantly larger datasets, such as FAcodec (LibriLight 60k hours) and CosyVoice (proprietary dataset 170k hours), ASRQ-VC demonstrates remarkable efficiency. Specifically, it significantly outperforms CosyVoice in WER (**11.2%** vs. **19.0%**) and FAcodec in SECS (**0.460** vs. **0.349**). This highlights the effectiveness of our method in achieving a strong balance between content preservation and speaker adaptation, even with limited training data.

These results validate our ASRQ-VC's ability to perform high-quality voice conversion with a strong trade-off between linguistic fidelity and speaker similarity, making it a viable solution for practical applications.

V. Ablation Studies

We conduct comprehensive ablation studies to investigate two key aspects of our method: (1) the impact of codebook size on the overall performance, (2) the effectiveness of incorporating ASR loss during content extractor training, and (3) the impact of adding timbre encoder for local timbre feature. We experiment with four different bottleneck sizes: 8, 40, 200, and 1000. For each configuration, we train two variants of the content extractor: one with ASR loss and one without. Subsequently, we train separate VITS voice conversion models with/without timbre encoder using these content extractors. The results are presented in Table II.

Table II reveals three key insights: First, ASR supervision consistently improves WER across all vocab size, confirming its role in enhancing content preservation. Second, the timbre encoder significantly boosts reference similarity, proving its effectiveness in capturing fine-grained vocal characteristics. Third, increasing vocab size progressively improves both WER (25.3$\rightarrow$10.6) and source similarity (0.088$\rightarrow$0.137), while reducing reference similarity (0.467$\rightarrow$0.439), indicating a vocab size around 200 better resolves the content-speaker trade-off.

VI. Conclusion

We present ASRQ-VC, a novel voice conversion framework that effectively disentangles speech content from speaker characteristics through ASR-guided VQ-VAE training and VITS-based synthesis. Our method achieves state-of-the-art performance in both speaker similarity and content preservation, significantly outperforming existing approaches. The ablation studies reveal crucial design insights: ASR supervision strengthens content retention, timbre encoding enhances speaker mimicry, and larger codebooks better balance these objectives. Future work will explore cross-lingual conversion cases and accent conversion task.

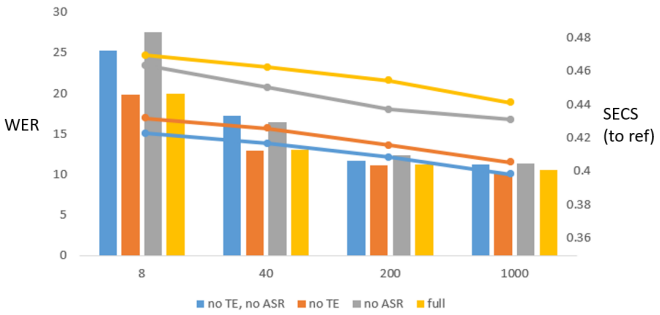

Fig. 4. Visualization of Table II. The bar chart clearly shows that WER decreases significantly with increasing vocab size, while the line chart shows that SECS (to reference) decreases slightly with increasing vocab size. Removing ASR loss or removing timbre encoder (TE) decreases WER and SECS (to reference) significantly.

Acknowledgement

This work was supported by Tencent and the Tencent-NTU Joint Research Laboratory (CENTURY), Nanyang Technological University, Singapore. The computational work for this article was fully performed on resources of the National Supercomputing Centre (NSCC), Singapore.

References

[1] 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.

[2] A. Baevski, S. Schneider, and M. Auli, "Vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[3] K. Qian, Y. Zhang, H. Gao, *et al.*, "Contentvec: An improved self-supervised speech representation by disentangling speakers," in *International Conference on Machine Learning*, PMLR, 2022, pp. 18 003–18 017.

[4] Z. Du, Q. Chen, S. Zhang, *et al.*, "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.

[5] F. Mentzer, D. Minnen, E. Agustsson, and M. Tschannen, "Finite scalar quantization: Vq-vae made simple," *arXiv preprint arXiv:2309.15505*, 2023.

[6] 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.

[7] Y. Guo, Z. Li, J. Li, *et al.*, "Vec2wav 2.0: Advancing voice conversion via discrete token vocoders," *arXiv preprint arXiv:2409.01995*, 2024.

[8] 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.

[9] S. Chen, C. Wang, Z. Chen, *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.

[10] A. Polyak, Y. Adi, J. Copet, *et al.*, "Speech resynthesis from discrete disentangled self-supervised representations," *arXiv preprint arXiv:2104.00355*, 2021.

[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] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "Cam++: A fast and efficient network for speaker verification using context-aware masking," *arXiv preprint arXiv:2303.00332*, 2023.

[14] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[15] G. Chen, S. Chai, G. Wang, *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.

[16] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[17] S. Woo, S. Debnath, R. Hu, *et al.*, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 133–16 142.

[18] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127 063, 2024.

[19] B. Zhang and R. Sennrich, "Root mean square layer normalization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] N. Shazeer, "Glu variants improve transformer," *arXiv preprint arXiv:2002.05202*, 2020.

[21] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 886–890.

[22] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, "Diff-hiervc: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation," *International Speech Communication Association*, pp. 2283–2287, 2023.

[23] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, "Dddm-vc: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 17 862–17 870.