Autoregressive Zero-Shot Voice Conversion

Luca A. Lanzendörfer [1] **Frédéric Berdoz** [1] **Antonis Asonitis** [1] **Roger Wattenhofer** [1]

Abstract

We present LaVoco, an autoregressive model for zero-shot voice conversion built on a pretrained autoregressive text-to-speech backbone. LaVoco leverages features extracted from Whisper and XCodec2 discrete tokens to autoregressively generate the target utterance using a reference timbre. We analyze different model variants of LaVoco and find that our dual representation achieves the best content preservation and competitive speaker similarity while retaining the scalability of next-token prediction. Our approach also proves robust to sub-1 second reference timbre prompts, maintaining high speaker similarity. Finally, we release our code, pretrained models, and a 10k-hour dataset of synthetic speech pairs to support future research.

1. Introduction

Voice Conversion (VC) transforms the timbre of a source utterance to match a target speaker while preserving linguistic content and prosody. In the zero-shot setting the target speaker is unseen during training, requiring the model to generalize from a single short reference clip. This setting is particularly demanding because the model must disentangle speaker identity from content without parallel training data, and generalize across diverse speakers and recording conditions.

Early VC approaches such as AutoVC (Qian et al., 2019) and AdaIN-VC (Chou & Lee, 2019) disentangle speaker and content embeddings but often suffer from timbre leakage. Self-supervised representations from HuBERT (Hsu et al., 2021) and WavLM (Chen et al., 2022a) improved content extraction and enabled feature-based methods such as kNN-VC (Baas et al., 2023) and FreeVC (Li et al., 2023), though these still rely on simple speech generation backends. More recently, state-of-the-art VC has shifted to diffusion and flow-matching: DiffVC (Popov et al., 2022) applied score-based diffusion to VC, Seed-VC (Liu, 2024) introduced an external timbre shifter to mitigate leakage, and EZ-VC (Joglekar et al., 2025) and AdaptVC (Kim et al., 2025) leverage conditional flow matching. Notably, all current state-of-the-art VC systems are non-autoregressive.

In contrast, autoregressive (AR) modeling dominates text-to-speech (TTS). Neural audio codecs (Zeghidour et al., 2022; Défossez et al., 2023; Xin et al., 2024; Kumar et al., 2023b; Défossez et al., 2024) tokenize speech for next-token prediction, and large pretrained Transformer backbones provide strong language priors that improve naturalness and coherence. VALL-E (Chen et al., 2025) pioneered AR TTS with codec tokens, AudioLM (Borsos et al., 2023) demonstrated hierarchical AR generation, and Llasa (Ye et al., 2025b) showed that a single-codebook AR model initialized from Llama achieves state-of-the-art TTS that scales predictably with compute. These results suggest AR models are well suited for conditional speech generation, yet they remain largely unexplored for voice conversion.

We propose LaVoco, an AR VC framework built on Llasa (Ye et al., 2025b). Inspired by LLaVA (Liu et al., 2023), which projects continuous vision encoder features into a language model's embedding space via a learned linear layer, we feed continuous Whisper (Radford et al., 2023) encoder embeddings directly into the AR backbone. We compare three source content representations: (i) continuous Whisper embeddings projected into Llasa's token space, (ii) discrete XCodec2 codec tokens, and (iii) a dual-input variant that concatenates both. All three condition on discrete XCodec2 reference tokens for speaker identity and adopt the timbre-shifting strategy of Seed-VC (Liu, 2024) to prevent timbre leakage during training. We find that the dual LaVoco variant is the most robust overall, achieving competitive performance with state-of-the-art diffusion and flow-matching approaches.

Importantly, both source-side encoders we adopt are audio-only, drawn from the same self-supervised family used by recent flow-matching VC systems such as Seed-VC and EZ-VC. The novelty of LaVoco therefore does not lie in the source representation but in reusing a large pretrained autoregressive generator (Llasa, trained on roughly 250k hours of speech) as the decoder rather than training a flow or diffusion decoder from scratch. This single change yields

[1]ETH Zurich, Switzerland. Correspondence to: Luca A. Lanzendörfer <lanzendoerfer@ethz.ch>.

Machine Learning for Audio Workshop at the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. 2026.

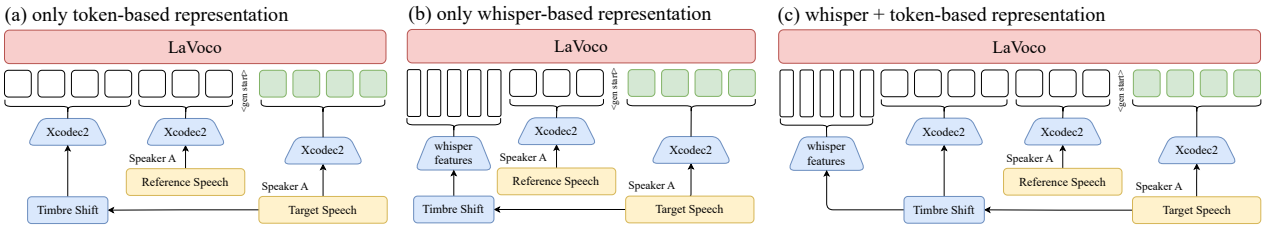

Figure 1. Comparison of the three LaVoco approaches. During training, paired samples from the same speaker are encoded with XCodec2 and feature extracted using Whisper and/or XCodec2 tokens after timbre shifting. Timbre shifting is necessary to avoid timbre leakage. The cross-entropy loss is only calculated on the green output tokens.

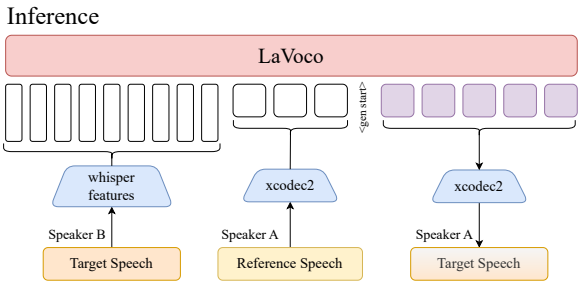

Figure 2. Inference layout for the Whisper-only LaVoco model. During inference, the timbre of speaker A is transferred to the target speech segment of speaker B (purple tokens).

the highest UTMOS and naturalness Elo in our evaluation while preserving content quality on par with the strongest non-autoregressive baselines, and it produces the most robust speaker similarity for sub-1 s reference clips. To the best of our knowledge, LaVoco is the first autoregressive zero-shot voice conversion model competitive with leading non-autoregressive approaches.

We summarize our contributions as follows:

- We introduce LaVoco, the first autoregressive voice conversion approach competitive with state-of-the-art diffusion and flow-matching methods. We systematically compare LaVoco on three source representations, showing that a dual Whisper-XCodec2 variant achieves the best overall performance.
- We show robustness to very short (as low as 0.2 seconds) reference prompts, obtaining the highest speaker similarity among tested models.
- We release our code, pretrained models, and a 10k-hour dataset of synthetic speech pairs to support future VC research.[1]

[1] https://github.com/ETH-DISCO/lavoco

2. Methodology

2.1. Model Architecture

LaVoco builds on Llasa-1B (Ye et al., 2025b), a Llama 3.2-based (Grattafiori et al., 2024) autoregressive TTS model that predicts XCodec2 (Ye et al., 2025b;a) tokens at 50 Hz. For voice conversion, we replace Llasa's text input with speech-derived representations of the source content (see Figures 1 and 2). We explore three variants that differ in how source content is represented.

LaVoco$_{\text{Whisper}}$. Given timbre-shifted source audio, the frozen Whisper-small (Radford et al., 2023) encoder produces hidden states $H \in \mathbb{R}^{T\times 768}$ at 50 Hz. Whisper pads to 30 s internally, so we keep only the first T_v frames corresponding to the actual utterance length. A trainable linear projection maps these to Llasa's 2048 hidden dimension:

$$E_{\text{sem}} = H_{1:T_v}\, W + b, \quad W \in \mathbb{R}^{768\times 2048} \tag{1}$$

The sequence is then constructed as $[\, E_{\text{sem}} \,\|\, E_{\text{ref}} \,\|\, E_{\text{tgt}} \,]$.

LaVoco$_{\text{XCodec2}}$. Instead of continuous embeddings, the timbre-shifted source audio is encoded directly into discrete XCodec2 tokens, which are embedded through Llasa's token embedding table to obtain E_{src}. This avoids the Whisper encoding and projection entirely, giving the input sequence $[\, E_{\text{src}} \,\|\, E_{\text{ref}} \,\|\, E_{\text{tgt}} \,]$.

LaVoco. The dual variant combines both representations, concatenating the continuous Whisper embeddings with the discrete XCodec2 source tokens:

$$[\, E_{\text{sem}} \,\|\, E_{\text{src}} \,\|\, E_{\text{ref}} \,\|\, E_{\text{tgt}} \,] \tag{2}$$

This provides complementary information: Whisper embeddings capture high-level semantic and prosodic content, while XCodec2 tokens retain fine-grained acoustic detail.

Common design. In all variants, a reference utterance from the target speaker is encoded by the frozen XCodec2 encoder into discrete tokens and embedded through Llasa's token embedding table, yielding E_{ref}. During training, E_{tgt}

Table 1. Zero-shot voice conversion performance on 100 random pairs of source and reference samples from Expresso (Nguyen et al., 2023). Elo ratings are derived from pairwise subjective evaluations using the Bradley-Terry model (Bradley & Terry, 1952) for speaker similarity (Sim) and naturalness (Nat). LaVoco using both Whisper and XCodec2 source representations is more robust compared to the variants using only one.

Model	WER ↓	CER ↓	UTMOS ↑	WavLM ↑	HuBERT ↑	SQ-SDR ↑	ELO_{Sim} ↑	ELO_{Nat} ↑
KNN-VC (Baas et al., 2023)	0.491	0.341	2.025	0.778	0.896	19.389	1216	1170
Diff-HierVC (Choi et al., 2023)	0.037	0.025	3.330	0.825	0.937	22.514	1385	1518
DDDM-VC (Choi et al., 2024)	0.060	0.036	2.963	0.839	0.939	12.757	1462	1500
Seed-VC (Liu, 2024)	0.012	0.005	3.471	0.881	0.939	15.827	**1748**	1588
EZ-VC (Joglekar et al., 2025)	0.158	0.101	3.579	**0.882**	0.934	19.779	1527	1400
FreeVC (Li et al., 2023)	0.055	0.031	3.611	0.842	0.939	**24.054**	1469	1552
$\text{LaVoco}_{\text{Whisper}}$	0.091	0.039	3.702	0.858	0.939	21.193	1445	**1630**
$\text{LaVoco}_{\text{XCodec2}}$	0.073	0.043	**3.712**	0.859	0.938	22.068	1529	1532
LaVoco	**0.007**	**0.004**	3.610	0.847	**0.942**	21.878	1720	1612

are ground-truth XCodec2 tokens of the target speech. During inference they are generated autoregressively. The cross-entropy loss is computed only over E_{tgt}, with all context positions masked. We fine-tune the full Llasa-1B backbone jointly with any trainable projection layers, applying a three times higher learning rate to the projection in variants that use it. Whisper and XCodec2 remain frozen throughout. We use AdamW with a base learning rate of 1×10^{-5} (three times higher for the projection layer), cosine scheduling with 5% warmup, weight decay of 0.01, and gradient clipping at 1.0. Training is done for one epoch over the 10k-hour dataset in bf16 mixed precision with gradient checkpointing on 4 NVIDIA A100 GPUs.

2.2. Training Data Construction

Feeding continuous embeddings of the source speech from models such as Whisper's encoder can lead to a problem known as *timbre leakage* (Liu, 2024), caused by a data mismatch between training and inference. During inference, when the reference speaker (Speaker A) and the target speaker (Speaker B) differ, the output retains the target speaker's voice instead of transferring the reference speaker's timbre onto the target utterance (see Figures 1 and 2). To avoid timbre leakage, we follow the timbre-shifting strategy of Seed-VC (Liu, 2024). Given a target utterance, we convert the utterance to a different speaker's voice using a pretrained VC model (Seed-VC or EZ-VC), producing a timbre-shifted copy with the same content but a different voice. Because this synthetic audio acts only as the source condition rather than the ground-truth target, our model is not bottlenecked by the capabilities or imperfections of the shifting model. In the Whisper variant, Whisper embeddings extracted from this copy become E_{sem}, in the XCodec2 variant, the timbre-shifted copy is encoded into discrete XCodec2 tokens to obtain E_{src}, the dual variant uses both. A separate utterance from the original target speaker serves as E_{ref}, and the unmodified target is E_{tgt}. In this way, the semantic input never shares the same timbre with the ground truth, forcing the model to rely only on the reference speech utterance for speaker identity and timbre transfer.

To avoid biasing the model toward any single shifter's artifacts, we split the training pool evenly between two timbre shifters: half of the pairs are produced with Seed-VC and the other half with EZ-VC. Crucially, the cross-entropy loss is computed over the clean ground-truth target rather than over the shifted source. The shifter only needs to provide a content-preserving voice swap so that the source no longer leaks the target speaker's identity. The model then learns to map source content to target timbre using the unaltered reference and ground-truth target, so the shifter's quality does not bottleneck training.

We use the English subset (10k hours) of the Multilingual Audio Alignments dataset from Hugging Face, whose word-level alignments allow us to split an utterance into reference and target segments at correct word boundaries, randomly selecting either the left or right segment. For timbre shifting, we sample source speakers from the 921 speakers in the LibriSpeech clean-360 train split (Panayotov et al., 2015).

2.3. LaVoco-10k: A Parallel Voice Conversion Dataset

We release a 10k-hour parallel voice conversion dataset generated by our trained LaVoco model, with source and target speakers drawn from GigaSpeech (Chen et al., 2021), LibriSpeech (Panayotov et al., 2015), Hi-Fi TTS (Bakhturina et al., 2021), and Expresso (Nguyen et al., 2023). Each sample is annotated with a WavLM-TDNN similarity score between output and reference, allowing downstream filtering by quality.

3. Experiments and Results

3.1. Evaluation Setup

We compare LaVoco and its variants against Seed-VC (Liu, 2024), EZ-VC (Joglekar et al., 2025), FreeVC (Li et al., 2023), Diff-HierVC (Choi et al., 2023), KNN-VC (Baas et al., 2023) and DDDM-VC (Choi et al., 2024). For ob-

jective evaluation we report Word Error Rate (WER) and Character Error Rate (CER) computed with Whisper-large-v3 (Radford et al., 2023), speaker similarity as the cosine similarity between embeddings of the generated and reference audio extracted by WavLM-TDNN (Chen et al., 2022a) and HuBERT (Hsu et al., 2021). Furthermore, we use UTMOS (Saeki et al., 2022) to predict human-perceived naturalness, and signal quality estimates from SQUIM (Kumar et al., 2023a) (SQ-PESQ for perceptual quality, SQ-STOI for objective intelligibility, and SQ-SDR for signal-to-distortion ratio) to measure acoustic clarity and artifact reduction. For the reference-length ablation we measure speaker similarity with UniSpeech-SAT (Chen et al., 2022b). For subjective evaluation we conduct pairwise head-to-head listening tests and report Elo ratings derived from win rates.

For speaker similarity we deliberately use three complementary verification systems: WavLM-TDNN and HuBERT on the general evaluation, and UniSpeech-SAT, the current SUPERB speaker-verification leader, on the short-reference ablation. Reporting multiple speaker verification systems prevents over-fitting evaluation to any single embedding's biases.

General evaluation. We use 100 randomly selected utterances from Expresso (Nguyen et al., 2023), each paired with a reference prompt of at most 5 seconds from a different speaker. A subset of 10 pairs is reserved for head-to-head subjective evaluation, split into two tasks: a naturalness test (participants pick which output more closely resembles the reference utterance) and a speaker similarity test (participants pick which output more closely resembles the source speaker). A total of 40 participants (20 per task) were recruited using an online tool.[2]

Reference speech length ablation. We run an ablation on reference-length over 100 samples from TIMIT (Garofolo et al., 1993). Each reference clip is cut to between 0.2 and 1 second, and speaker similarity (using UniSpeech-SAT (Chen et al., 2022b)) is measured between the generated audio and the full uncut source clip (at least 3 seconds duration), simulating scenarios with short reference speech.

3.2. Results

General evaluation findings. We report the performance of the general evaluation on the Expresso dataset in Table 1. LaVoco achieves the lowest WER and CER, indicating strong content preservation, and the highest HuBERT-based similarity, suggesting that combining Whisper and XCodec2 source representations provides complementary information that benefits linguistic fidelity. For speaker similarity, Seed-VC remains state-of-the-art in ELO_{Sim}, with LaVoco obtaining competitive scores. On perceived naturalness, $\text{LaVoco}_{\text{Whisper}}$ achieves the highest ELO_{Nat}, followed by LaVoco, confirming that autoregressive generation produces speech which listeners consistently rate as highly natural. Comparing the three LaVoco variants, the Whisper-only and XCodec2-only models achieve marginally higher UTMOS, but LaVoco is substantially more robust on content metrics while maintaining competitive naturalness and speaker similarity. This confirms that fusing semantic and acoustic source representations reduces errors without losing perceptual quality.

[2] https://www.mabyduck.com/

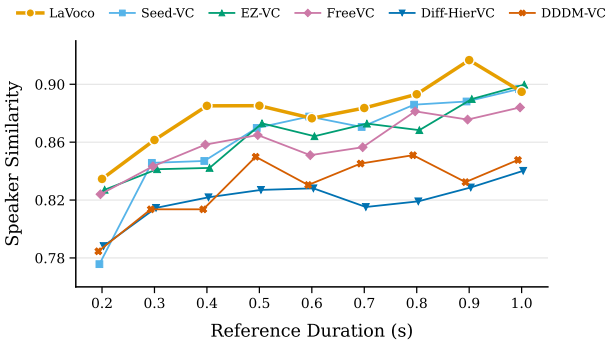

Figure 3. Zero-shot speaker similarity across varying reference prompt durations. We compare LaVoco against voice conversion baselines using 100 randomly drawn samples from TIMIT (Garofolo et al., 1993) test set. LaVoco overall outperforms models when the reference is below one second in duration.

Reference length ablation. As shown in Figure 3, we analyze speaker similarity as a function of reference length. LaVoco demonstrates superior performance over baseline models, especially when provided with shorter reference utterances. We attribute this to the autoregressive design, which allows the model to leverage pretrained priors when predicting from a small set of reference codes.

4. Conclusion

We presented LaVoco, an autoregressive approach to zero-shot voice conversion that repurposes a pretrained TTS backbone by replacing text input with speech-derived representations. Our comparison of three source content representations shows that the dual Whisper-XCodec2 variant is the most robust overall, achieving the lowest WER and CER, the highest HuBERT-based similarity, and competitive Elo scores with the strongest non-autoregressive baselines. Our experiments show that the AR formulation handles short reference prompts well, retaining high speaker similarity even when the reference clip is below one second. However, this advantage comes at the cost of decoding speed which is slower than non-autoregressive alternatives, constituting a path for future research. We release our code, pretrained models, and a 10k hour synthetic speech pair dataset to support future work on voice conversion.

Impact Statement

This work aims to advance zero-shot voice conversion by improving content preservation, speaker similarity, and robustness to short reference prompts. This can benefit accessibility, speech restoration, localization, and creative audio production. However, voice conversion also carries risks of speaker impersonation, fraud, and misleading synthetic media. We therefore encourage deployment only with safeguards against malicious or deceptive use.

References

Baas, M., van Niekerk, B., and Kamper, H. Voice Conversion With Just Nearest Neighbors. In *Proc. Interspeech*, 2023.

Bakhturina, E., Lavrukhin, V., Ginsburg, B., and Zhang, Y. Hi-Fi Multi-Speaker English TTS Dataset. In *Proc. Interspeech*, 2021.

Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., et al. AudioLM: A Language Modeling Approach to Audio Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023.

Bradley, R. A. and Terry, M. E. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Chen, G., Chai, S., Wang, G., Du, J., Zhang, W., Weng, C., et al. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Proc. Interspeech*, 2021.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022a.

Chen, S., Wu, Y., Wang, C., Chen, Z., Chen, Z., Liu, S., et al. UniSpeech-SAT: Universal Speech Representation Learning with Speaker Aware Pre-Training. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022b.

Chen, S., Wang, C., Wu, Y., Zhang, Z., Zhou, L., Liu, S., et al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718, 2025.

Choi, H., Lee, S., and Lee, S. Diff-HierVC: Diffusion-based Hierarchical Voice Conversion with Robust Pitch Generation and Masked Prior for Zero-shot Speaker Adaptation. In *Proc. Interspeech*, 2023.

Choi, H., Lee, S., and Lee, S. DDDM-VC: Decoupled Denoising Diffusion Models with Disentangled Representation and Prior Mixup for Verified Robust Voice Conversion. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

Chou, J. and Lee, H. One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization. In *Proc. Interspeech*, 2019.

Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. High Fidelity Neural Audio Compression. *Transactions on Machine Learning Research*, 2023.

Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., et al. Moshi: A Speech-Text Foundation Model for Real-Time Dialogue, 2024. arXiv:2410.00037.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. Technical Report NISTIR 4930, National Institute of Standards and Technology, 1993.

Grattafiori, A. et al. The Llama 3 Herd of Models, 2024. arXiv:2407.21783.

Hsu, W., Bolte, B., Tsai, Y., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Joglekar, A., Singh, D., Bhatia, R. R., and Umesh, S. EZ-VC: Easy Zero-shot Any-to-Any Voice Conversion. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025.

Kim, J., Kim, J.-H., Choi, Y., Nguyen, T. D., Mun, S., and Chung, J. S. AdaptVC: High Quality Voice Conversion with Adaptive Learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

Kumar, A., Tan, K., Ni, Z., Manocha, P., Zhang, X., Henderson, E., et al. TorchAudio-Squim: Reference-less Speech Quality and Intelligibility Measures in TorchAudio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023a.

Kumar, R., Seetharaman, P., Luebs, A., Kumar, I., and Kumar, K. High-Fidelity Audio Compression with Improved RVQGAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Li, J., Tu, W., and Xiao, L. FreeVC: Towards High-Quality Text-Free One-Shot Voice Conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Liu, H., Li, C., Wu, Q., and Lee, Y. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Liu, S. Zero-shot Voice Conversion with Diffusion Transformers, 2024. arXiv:2411.09943.

Nguyen, T. A., Hsu, W.-N., D'Avirro, A., Shi, B., Gat, I., Fazel-Zarani, M., et al. Expresso: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis. In *Proc. Interspeech*, 2023.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: An ASR Corpus Based on Public Domain Audio Books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., Kudinov, M., and Wei, J. Diffusion-Based Voice Conversion with Fast Maximum Likelihood Sampling Scheme. In *International Conference on Learning Representations (ICLR)*, 2022.

Qian, K., Zhang, Y., Chang, S., Yang, X., and Hasegawa-Johnson, M. AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. In *International Conference on Machine Learning (ICML)*, 2019.

Radford, A., Kim, J., Xu, T., Brockman, G., Mcleavey, C., and Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning (ICML)*, 2023.

Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S., and Saruwatari, H. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Proc. Interspeech*, 2022.

Xin, D., Tan, X., Takamichi, S., and Saruwatari, H. BigCodec: Pushing the Limits of Low-Bitrate Neural Speech Codec, 2024. arXiv:2409.05377.

Ye, Z., Sun, P., Lei, J., Lin, H., Tan, X., Dai, Z., et al. Codec Does Matter: Exploring the Semantic Shortcoming of Codec for Audio Language Model. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025a.

Ye, Z., Zhu, X., Chan, C., Wang, X., Tan, X., Lei, J., et al. Llasa: Scaling Train-Time and Inference-Time Compute for Llama-based Speech Synthesis, 2025b. arXiv:2502.04128.

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. SoundStream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022.