EZ-VC: Easy Zero-shot Any-to-Any Voice Conversion

Anonymous ACL submission

Abstract

Voice Conversion research in recent times has increasingly focused on improving the zero-shot capabilities of existing methods. Despite remarkable advancements, current architectures still tend to struggle in zero-shot cross-lingual settings. They are also often unable to generalize for speakers of unseen languages and accents. In this paper, we adopt a simple yet effective approach that combines discrete speech representations from self-supervised models with a non-autoregressive Diffusion-Transformer based conditional flow matching speech decoder. We show that this architecture allows us to train a voice-conversion model in a purely textless, self-supervised fashion. Our technique works without requiring multiple encoders to disentangle speech features. Our model also manages to excel in zero-shot cross-lingual settings even for unseen languages. We provide demo samples for our model here: https://ez-vc.github.io/EZ-VC-Demo/

1 Introduction

Zero-shot Voice Conversion (VC) is the task of transforming a source speaker's voice characteristics into that of a target speaker while preserving linguistic content and prosodic attributes, even for speakers unseen during training. Over the years with the advancement of modern deep learning techniques and substantial improvements in speech encoders and speech generation systems, numerous and vastly different approaches have been proposed to address this challenge.

Textless VC architectures have become the primary area of research in this domain since cascaded ASR+TTS systems are known to lose the non-verbal characteristics of the source speech such as laughs, whispers and other filler sounds. They also lead to cascaded errors. To overcome this, many textless VC systems these days employ either self-supervised speech encoders (SSL) or neural audio codecs (NAC) to extract speaker features or linguistic content before feeding them to a speech generation decoder. These speech representations are also often disentangled to obtain certain composite characteristics such as timbre or style. Sometimes quantized speech representations are used which form as the input for a speech generation or language model. Speech synthesis systems, which are a key component of VC architectures, have of late greatly benefited from the advancements in diffusion and continuous normalizing flow (CNF) based techniques. Voicebox(Le et al., 2023) and its successors that use these methods are able to produce high quality audio outputs that are almost undistinguishable from real speech. These models thus show great promise for zero-shot VC tasks and yet architectures based on these methods remain under-explored.

In this work we contribute the following,

- We propose EZ-VC, a simple self-supervised any-to-any zero-shot voice conversion architecture that generalizes for unseen speakers, accents and languages while still producing highly natural and fluent speech.
- We demonstrate that zero-shot VC is possible without requiring multiple encoders for feature disentanglement of speaker and speech attributes.
- We show that combining quantized features from a self-supervised speech encoder and a flow matching speech generation decoder is sufficient to achieve state-of-the-art results.

2 Related Work

Early research in VC focused on disentangling speaker and content information. Works like YourTTS(Casanova et al., 2023) focused on using speaker embeddings to extract speaker features from target speech but usually required reference

text to be provided as well. Recent works like SEF-VC(Li et al., 2024) now prefer textless, speaker-embedding free VC which is also able to perform better. Since the advent of SSL speech models like Hubert(Hsu et al., 2021) and WavLM(Chen et al., 2022), VC research has quickly learned to leverage them for their high correlation with both acoustic and linguistic content. kNN-VC(Baas et al., 2023) works by replacing representations of source speech with the nearest neighbour from the reference speech. Vec2wav 2.0 on the other hand, uses a combination of discrete representations from vq-wav2vec for source content and WavLM features for capturing the timbre of the target speaker. At the same time, another school of approach has emerged that utilizes neural audio encoders and combines them with language models for high quality VC. Unfortunately, these systems suffer from slow inference speeds due to their auto-regressive nature. Diffusion based techniques also have been explored by DiffVC(Popov et al., 2022) and similar works. These models are able to demonstrate natural and robust outputs. Conditional flow-matching based speech generation methods have also begun to appear in voice conversion literature. Latest works such as AdaptVC(Kim et al., 2025), StableVC(Yao et al., 2024), Seed-VC(Liu, 2024) and PFlow-VC(Zuo et al., 2025) employ this technique for their speech decoders and generally couple them with SSL encoders.

AdaptVC uses speaker and content encoder adapters on top of Hubert while StableVC includes three feature extractors for style, linguistic content, and mel-spectrograms. Seed-VC on the other hand requires a timber shifter module and speaker-embeddings besides a semantic feature extractor. PFlow-VC proposes a slightly different approach by using a timbre encoder for target speaker and semantic encoder for source speech. In contrast, with our architecture we wish to eliminate the need for multiple encoders or adapters for voice conversion while still being able to achieve state-of-the-art results for any-to-any VC.

3 EZ-VC

EZ-VC is a simple architecture that only requires one pre-trained speech encoder and a trainable speech decoder. Unlike most other works, we do not need multiple encoders for disentanglement of speech features. Our architecture also benefits from using an off-the-shelf encoder. Other than training a simple k-means model, we do not train our speech encoding module. This helps reduce the compute and training time requirements compared to existing methods that usually ask for training both the encoder and decoder modules.

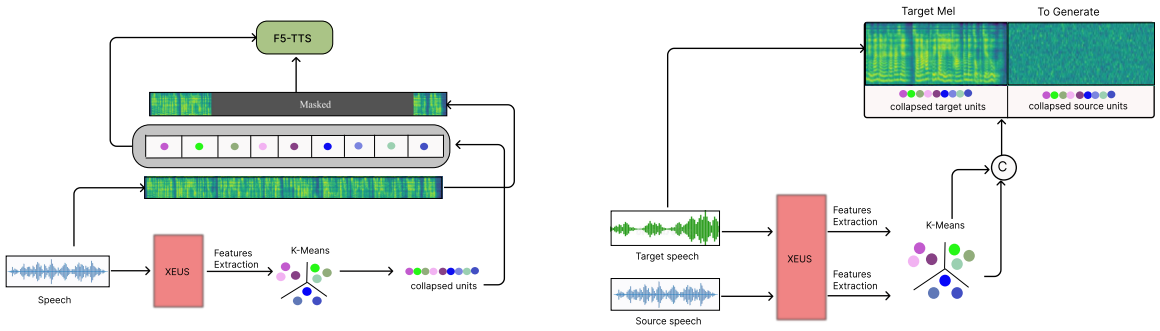
Figure 1 provides a description of our model's architecture for both training and inference. At the time of training, our model does not require any supervised or labeled data. To prepare our training set, we extract the mel-spectogram for every speech sample. These are then passed through the speech encoder first and then the resultant speech features from the 14th layer are taken and quantized using a k-means clustering model. The features are extracted at 75% of the model depth consistent with previous works(Maiti et al., 2024; Communication et al., 2023). We also de-duplicate adjacent discrete units for all samples. The mel and the corresponding discrete units become the input for our speech decoder training. With this, the model is able to learn to produce mel-spectogram from these given discrete representations and is also able to condition them based on the provided speech prompt. During inference, we pass both the source and target speech through our speech encoder system. The mel-spectogram of the target speech and its discrete units form the reference for our CFM model and the source discrete units form the prompt to generate the corresponding mel. The target and source units are concatenated and given as input to the model. The target mel is then discarded upon inference. This generated mel inherits the speaker attributes from the reference target mel while the content and style is obtained from the source units.

3.1 Speech-to-Units

To extract high-quality speech representations, we employ Xeus (Chen et al., 2024a), a self-supervised learning (SSL) encoder trained on an extensive multilingual dataset encompassing 4,000 languages. Given its exposure to such linguistic diversity, we expect Xeus to provide robust, language-agnostic representations, enabling our model to generalize effectively to unseen languages.

Similar to WavLM, Xeus processes speech by generating frame-level embeddings. Each output embedding corresponds to a 25ms window size with a 20ms stride, effectively producing 50 embeddings per second of speech.

For the purpose of enabling speech reconstruction, we apply a quantization step using k-means clustering. Specifically, we train a 500-cluster k-



(a) Training (b) Inference

Figure 1: An overview of EZ-VC

means model using embeddings extracted from the 14th layer of Xeus. This clustering process provides us discrete speech units that can be used to train a units-to-speech model for resynthesis. Our k-means training dataset comprises 100 hours of English speech and 50 hours each from five Indian languages, ensuring a balanced and representative distribution of phonetic variations. This dataset is a subset of the one used for training EZ-VC.

3.2 Units-to-Speech

We choose the F5-TTS(Chen et al., 2024b) architecture for our speech generation system. Building upon the work of E2-TTS(Eskimez et al., 2024) and Voicebox, F5-TTS manages to alleviate several of their shortcomings such as duration modelling, phoneme alignment and slow convergence. We train our model for speech generation with discrete units as input. The model learns to reconstruct speech from these condensed speech representations via an infilling task. The speaker attributes are derived from the unmasked mel-spectogram and the speech content comes from the input units. This disentangles the speaker and speech, allowing us to achieve zero-shot voice conversion.

4 Experiment

4.1 Datasets

We select a wide variety of publicly available datasets for English and 5 Indian languages comprising of a total 12840 hours of speech. We hope that using a diverse set of languages and accents will help the model to generalize in unseen settings.

For English, we use 3060 hours of speech which includes a range of American, European and Indian accents. American accents come from Librispeech while European accents appear in Vox Populi(Wang et al., 2021) dataset. For Indian English accent we use 1100 of speech from NPTEL[1] lectures.

We also select 5 Indian languages, namely Bengali, Hindi, Tamil, Telugu and Kannada to introduce diversity to our training set. We obtain in total 9780 hours of data from these languages. We procure unlabeled speech from several sources including Vaani(Bhogale et al., 2022), Commonvoice(Ardila et al., 2020) and datasets from IIIT-H and IIT-M. Table 5 contains a full breakdown.

We downsample all data, wherever neccessary to 16KHz. We further pass this data through our speech decoder combination of Xeus and k-means model to obtain discrete speech representations of each audio sample.

4.2 Training setup

We adopt the original implementation of F5-TTS for training our model. We use the base model configuration(300M params) which consists of 22 layers, 16 attention heads. For the audio samples we set sampling rate to 16KHz and use 80-dimensional log mel-filterbank features with hop length of 160. We also train a base BigVGAN(gil Lee et al., 2023) model on Libri-TTS(Zen et al., 2019) with the same configuration for a million steps. For our tokenizer, we use character level tokens with a vocabulary which includes all the 500 different discrete units.

We train this F5-TTS model from scratch with a batch size of 64 samples for 1.35 million updates on 4 NVIDIA RTX 6000 ADA GPUs. We use a peak learning rate of 5e-5 with 100k warmup steps. The rest remains the same as the original F5-TTS configuration.

[1]https://nptel.ac.in/

	SSIM ↑	NMOS ↑	SMOS ↑	UTMOS ↑
Seed-VC	0.69	3.55	3.78	3.02
kNN-VC	0.59	1.94	2.05	2.42
Vec2Wav2.0	0.61	3.67	3.55	3.55
Diff-HierVC	0.44	3.30	3.33	3.16
EZ-VC (Ours)	**0.71**	**3.91**	**3.90**	**3.56**

Table 1: Performance metrics comparison of different VC baselines

5 Evaluation

Subjective and objective measures are equally important for evaluating voice conversion systems. In our test we use Naturalness Mean Opinion Score (NMOS) and Similarity Mean Opinion Score (SMOS) as our subjective evaluations. For objectivity, we utilize Speaker Similarity (SSim) and UTMOS(Saeki et al., 2022) scores for comparing our models. We measure speaker similarity by using cosine similarity scores between our target speech and that of our output speech by using embeddings from a speaker verification model called ECAPA-TDNN(Desplanques et al., 2020).

For our baselines, we select few of the most recent and best performing open-source voice conversion models. This makes sure that we evaluate our model against the current state-of-the-art architectures available. We select SeedVC, vec2wav 2.0, Diff-HierVC(Choi et al., 2023) and kNN-VC as our baselines. Vec2wav and kNN-VC use primarily units-to-speech vocoders, while Diff-HierVC employs diffusion based methods. SeedVC and our work meanwhile uses CFM based speech models.

We choose 10 samples for our evaluations. These samples are selected from various languages and accents. We prepare a variety of source and target speech combinations based on gender, interlingual and cross-lingual speech. We also include combinations of seen and unseen languages to test the robustness and generalization capabilities of these models. All audios are resasmpled to 16KHz to ensure fair comparison.

For our subjective evaluation, we provided these 10 samples to 20 student volunteers for comparison. Each volunteer was asked to evaluate each sample based on it's naturalness which evaluates for mainly intelligibility, style preservence, and sound quality of the output speech in comparison to the source speech. In contrast, the similarity mean opinion score judges the similarity of the speaker in the output speech to that of the target speaker. We take the average of all the samples from all the volunteers which becomes the results of our NMOS and SMOS scores.

We further objectively compare our model with Seed-VC on a seen language(English) and 2 unseen languages(German and Spanish). The results, as shown in Table 4, demonstrate that EZ-VC provides better naturalness according to UTMOS, while having comparable or better speaker similarity scores.

Analyzing the naturalness and similarity MOS scores from Table 1, we see that EZ-VC convincingly beats the latest state-of-the-art approaches for voice conversion. We find that Vec2wav 2.0, which uses discrete units coupled with a vocoder competes very well for naturalness but lags behind when it comes to imitating the target speaker. This shows that having a CFM based speech decoder is a major benefit for voice conversion systems as they are better able to capture speech styles. They also seem to generalize very well for unseen languages and accents.

6 Conclusion

EZ-VC hopes to make a substantial advancement in the field of zero-shot voice conversion, demonstrating that high-quality voice transformation can be achieved with a minimal architecture. By leveraging discrete speech representations from self-supervised models and a non-autoregressive speech decoder, EZ-VC balances both naturalness and speaker similarity without the need for complex feature disentanglement or multiple encoders.

The model's ability to generalize across diverse linguistic settings highlights its robustness in cross-lingual contexts. Our findings may also suggest that discrete representations capture deeper, more universal representations of speech.

Our comprehensive evaluations show that EZ-VC achieves significantly improved capabilities for zero-shot voice conversion. We hope that our work inspires further efforts to simplify voice conversion techniques.

7 Potential Risks

Given the highly realistic quality of voice synthesis and the ability to achieve cross-lingual voice conversion for even unseen languages, our model carries the risk of enabling dangerous deepfakes.

Limitations

Despite the benifits of our approach, it has a few limitations,

- The EZ-VC architecture is reliant on the quality of the pretrained speech encoder. It is likely that using an encoder trained on only one language may not achieve the level of generalization that our model does.
- Although our approach introduces a much simpler architecture than previous works, the computational requirements are still comparable or higher.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. *Preprint*, arXiv:1912.06670.

Matthew Baas, Benjamin van Niekerk, and Herman Kamper. 2023. Voice conversion with just nearest neighbors. *Preprint*, arXiv:2305.18975.

Kaushal Santosh Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages. *Preprint*, arXiv:2208.12666.

Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir Antonelli Ponti. 2023. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. *Preprint*, arXiv:2112.02418.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024a. Towards robust speech representation learning for thousands of languages. *Preprint*, arXiv:2407.00837.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024b. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *Preprint*, arXiv:2410.06885.

Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. 2023. Diff-hiervc: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation. *Preprint*, arXiv:2311.04693.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *Preprint*, arXiv:2308.11596.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. *Preprint*, arXiv:2406.18009.

Sang gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2023. Bigvgan: A universal neural vocoder with large-scale training. *Preprint*, arXiv:2206.04658.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Jaehun Kim, Ji-Hoon Kim, Yeunju Choi, Tan Dat Nguyen, Seongkyu Mun, and Joon Son Chung. 2025. Adaptvc: High quality voice conversion with adaptive learning. *Preprint*, arXiv:2501.01347.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *Preprint*, arXiv:2306.15687.

Junjie Li, Yiwei Guo, Xie Chen, and Kai Yu. 2024. Sef-vc: Speaker embedding free zero-shot voice conversion with cross attention. *Preprint*, arXiv:2312.08676.

Songting Liu. 2024. Zero-shot voice conversion with diffusion transformers. *Preprint*, arXiv:2411.09943.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. 2024. Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13326–13330. IEEE.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Kudinov, and Jiansheng Wei. 2022. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. *Preprint*, arXiv:2109.13821.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *Preprint*, arXiv:2204.02152.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *Preprint*, arXiv:2101.00390.

Jixun Yao, Yuguang Yang, Yu Pan, Ziqian Ning, Jiaohao Ye, Hongbin Zhou, and Lei Xie. 2024. Stablevc: Style controllable zero-shot voice conversion with conditional flow matching. *Preprint*, arXiv:2412.04724.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *Preprint*, arXiv:1904.02882.

Jialong Zuo, Shengpeng Ji, Minghui Fang, Ziyue Jiang, Xize Cheng, Qian Yang, Wenrui Liu, Guangyan Zhang, Zehai Tu, Yiwen Guo, and Zhou Zhao. 2025. Enhancing expressive voice conversion with discrete pitch-conditioned flow matching model. *Preprint*, arXiv:2502.05471.

Appendix

Dataset	Hours
Librispeech	960
Vox Populi	1000
NPTEL	1100
Total	3060

Table 2: English Datasets

	SSIM	UTMOS
English(EZ-VC)	87.3	3.76
English(Seed-VC)	83.9	3.51

Table 3: EZ-VC Vs Seed-VC on seen languages

	SSIM	UTMOS
German(EZ-VC)	91.4	3.71
German(Seed-VC)	90.8	2.83
Spanish(EZ-VC)	84.2	3.49
Spanish(Seed-VC)	84.2	3.24

Table 4: EZ-VC Vs Seed-VC on unseen languages

	Bengali	Hindi	Tamil	Telugu	Kannada	**Total**
Vaani	1420	-	-	980	1390	3790
Common Voice	-	-	420	-	-	420
Shrutilipi	620	-	950	-	-	1570
IIIT-H	-	-	-	2600	-	2600
IITM	-	1400	-	-	-	1400
Total	**2040**	**1400**	**1370**	**3580**	**1390**	**9780**

Table 5: Indian Language Datasets