pISSN 2005-8063
eISSN 2586-5854
2023. 9. 30.
Vol.15 No.3
pp. 69-74

말소리와 음성과학
Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2023.15.3.069

Zero-shot voice conversion with HuBERT

Hyelee Chung · Hosung Nam*

Department of English Language and Literature, Korea University, Seoul, Korea

Abstract

This study introduces an innovative model for zero-shot voice conversion that utilizes the capabilities of HuBERT. Zero-shot voice conversion models can transform the speech of one speaker to mimic that of another, even when the model has not been exposed to the target speaker's voice during the training phase. Comprising five main components (HuBERT, feature encoder, flow, speaker encoder, and vocoder), the model offers remarkable performance across a range of scenarios. Notably, it excels in the challenging unseen-to-unseen voice-conversion tasks. The effectiveness of the model was assessed based on the mean opinion scores and similarity scores, reflecting high voice quality and similarity to the target speakers. This model demonstrates considerable promise for a range of real-world applications demanding high-quality voice conversion. This study sets a precedent in the exploration of HuBERT-based models for voice conversion, and presents new directions for future research in this domain. Despite its complexities, the robust performance of this model underscores the viability of HuBERT in advancing voice conversion technology, making it a significant contributor to the field.

Keywords: voice conversion, deep learning, machine learning

1. Introduction

Voice conversion technology, which alters the speech sound of one speaker to mimic another while maintaining the integrity of the original content, has broad implications across sectors like entertainment and dubbing. The challenge with successful voice conversion lies in separating the content and speaker information from the speech signal. It often results in an unintentional alteration of content when trying to modify the speaker's information or an insufficient change of timbre in the output (Sisman et al., 2021). Despite this challenge, deep learning advancements have brought about significant improvements in the field of voice conversion.

AUTOVC is a prominent exemplar of neural voice conversion, utilizing an autoencoder approach to diminish and subsequently restore the dimensionality of the input (Qian et al., 2019). This model uses the source speech as the input and the target speech as the condition, which is the most typical form of neural voice conversion. One distinct advantage of this method is its applicability to non-parallel data, which was a limitation with previous voice conversion models that required multiple speakers to read identical sentences (Sisman et al., 2021). The autoencoder approach enables the use of data from multiple speakers reading different sentences, thus allowing for a more extensive data collection.

AUTOVC stands as a pioneering work in the field of neural voice conversion, having established a foundational framework upon which subsequent research has built. However, it is important to

* hnam@korea.ac.kr, Corresponding author

Received 5 July 2023; Revised 23 July 2023; Accepted 23 July 2023

note that its performance with respect to zero-shot voice conversion leaves room for substantial improvement.

Indeed, the overarching objective in the realm of voice conversion, and more broadly in machine learning, is to achieve what is known as zero-shot learning. In simple terms, zero-shot learning can be likened to a person's ability to correctly answer a question on a topic that they have never studied, purely based on their broader knowledge and understanding of related subjects.

For a machine learning model, zero-shot learning translates into the model's capacity to perform tasks for which it was not explicitly trained (Palatucci et al., 2009). For example, in the case of voice conversion, a zero-shot learning model can transform the speech of one speaker to mimic that of another, even if the model has never encountered the target speaker's voice during its training. This is a complex and challenging goal, as the model must infer the acoustic traits of the target speaker without any direct exposure.

On the other hand, the few-shot learning is akin to someone learning a new concept or skill with only a few examples or a minimal amount of instruction. In machine learning, this implies training a model on a small amount of data related to the task it will be asked to perform (Logan et al., 2021). In the voice conversion example, few-shot learning would involve training the model on a limited number of voice samples from the target speaker, just enough for it to understand the speaker's essential voice traits.

Zero-shot learning presents a significant challenge in the realm of voice conversion, requiring sophisticated techniques for unsupervised learning and inference. Furthermore, it must contend with the inherent variability in human speech, affected by factors ranging from the speaker's emotional state to their physical environment.

However, recent advancements in machine learning and artificial intelligence have opened new possibilities in this field. Powerful deep learning architectures and novel learning strategies have brought us closer than ever to achieving reliable zero-shot voice conversion.

Moreover, recent research trends include the exploration of multi-speaker text-to-speech (TTS) systems alongside voice conversion. Zero-shot multi-speaker TTS and voice conversion share similarities, though they differ primarily in their inputs: text for the former and speech for the latter. Models like VITS have leveraged normalizing flows (Rezende & Mohamed, 2015) in their architecture, conditioning the flow on speaker information and implementing it within their duration predictors (Kim et al., 2021). Furthermore, NaturalSpeech 2 from Microsoft Inc., which incorporates a diffusion model, is regarded as one of the top-performing zero-shot speech synthesizers, particularly for its ability to produce high-quality speech (Shen et al., 2023; Sohl-Dickstein et al., 2015).

The field of voice conversion has recently seen an emerging interest in the utilization of HuBERT (Hsu et al., 2021), as large-scale acoustic model. A pivotal study by van Niekerk et al. (2022) demonstrated that the employment of HuBERT as a content encoder in voice conversion outperforms models based on traditional acoustic features such as Mel-frequency cepstral coefficient (MFCC).

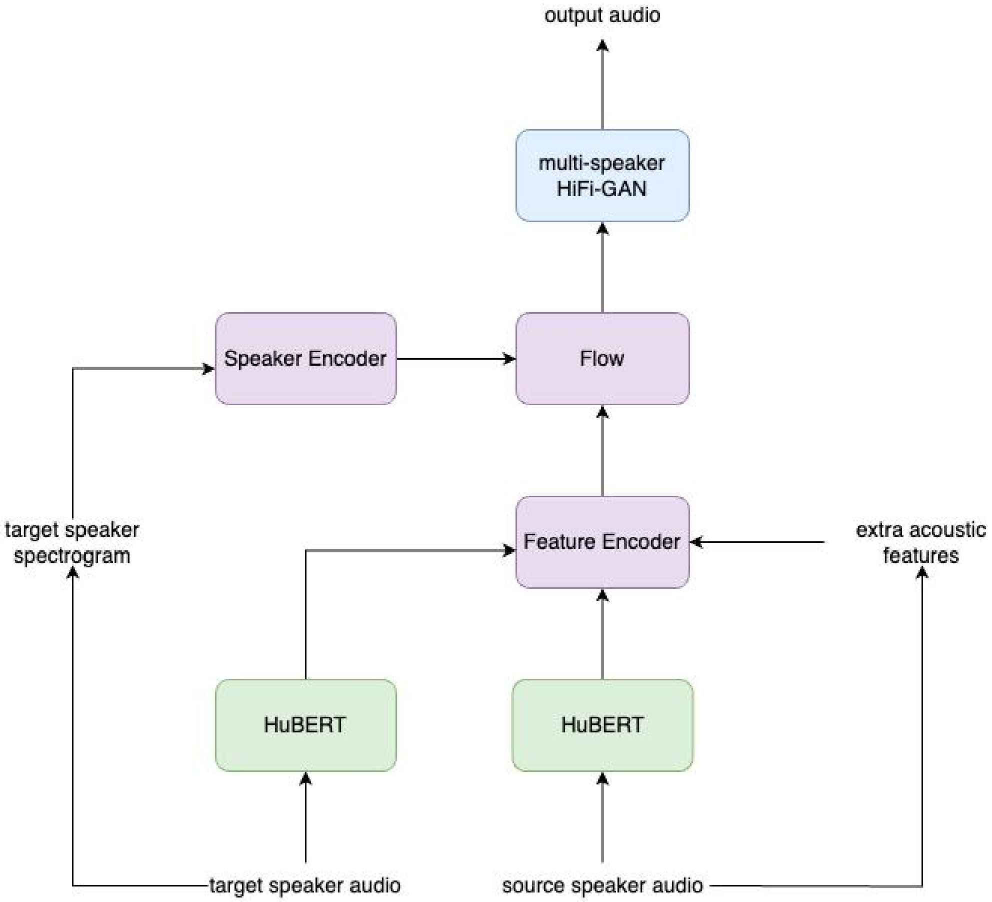


Figure 1. Proposed model

While open-source voice conversion projects such as so-vits-svc and RVC are incorporating HuBERT into their systems (RVC-Project, 2023; svc-develop-team, 2023), the application of HuBERT in voice conversion remains largely unexplored. Therefore, this paper introduces a model for zero-shot voice conversion using HuBERT, discussing its performance and viability.

2. Model

The structure of the model proposed in this paper is shown in Figure 1. The model consists of five main modules: HuBERT, Feature Encoder, Flow, Speaker Encoder, and Vocoder.

2.1. HuBERT

2.1.1. Transformer

Transformer is a model architecture which employs a unique self-attention mechanism that allows it to capture long-range dependencies and effectively model sequential data (Vaswani et al., 2017). The transformer architecture consists of an encoder and a decoder, both consisting of multiple identical layers. Each layer incorporates self-attention and feed-forward neural networks to capture dependencies and generate higher-level representations. Residual connections and layer normalization are used to retain important information and stabilize training. The encoder processes the input sequence, while the decoder generates the output sequence, attending to relevant parts of the input.

2.1.2. BERT

BERT consists of a stack of transformer encoders that capture complex representations of input sequences (Devlin et al., 2018). Each transformer encoder layer incorporates self-attention and feed-forward neural network sub-layers to model dependencies between words and generate higher-level contextualized representations. BERT follows a two-step process, starting with pre-training on unlabeled text data using masked language modeling and next sentence prediction tasks (Devlin et al., 2018). This pre-training enables BERT to learn contextual word representations. In the fine-tuning phase, BERT is trained on specific labeled datasets for downstream tasks.

2.1.3. HuBERT

HuBERT, a speech embedding model, employs the Transformer's encoder to extract hidden units from speech data (Hsu et al., 2021). Its self-supervised learning methodology allows it to learn effectively from unlabeled data. The architectural design of HuBERT closely mirrors that of the wav2vec 2.0, comprising several primary components, but they differ in training processes (Baevski et al., 2020; Schneider et al., 2019).

The initial step of the training process involves generating hidden units. This begins by extracting MFCC from the waveform, which serve as fundamental acoustic features for speech representation. Each audio segment is subsequently subjected to the K-means clustering algorithm and allotted to one of the K clusters. Each audio frame is labeled according to its respective cluster. Following this, these hidden units are transformed into embedding vectors for usage in the subsequent training phase.

The second phase aligns closely with the training process of BERT, employing masked language modeling (Devlin et al., 2018). In this phase, the Convolutional Neural Network is tasked with generating features from the raw audio. These features are randomly masked and introduced into the Transformer encoder. The encoder then produces a feature sequence, filling in the gaps left by the masked tokens. The resulting output undergoes a projection into a lower dimension to align with the labels. Finally, the cosine similarity between the output and hidden unit embedding created during the first phase is computed.

With fine-tuning through supervised learning, HuBERT can be employed for tasks such as speech recognition and synthesis. Meta AI has released pre-trained HuBERT models that were trained using Librispeech data or Libri-Light data. For the purpose of this study, the HuBERT base model was used, which boasts of 95 million parameters.

2.2. Feature Encoder

The Feature Encoder is a novel module designed to encode target speaker features, source speaker features, as well as auxiliary acoustic features. Comprised of several convolutional blocks, each block is built upon a CNN layer, followed by batch normalization, an activation function, and dropout. This configuration of the convolutional blocks demonstrates that the activation function should be fed with a well-calibrated distribution, as proposed by Ioffe & Szegedy (2015).

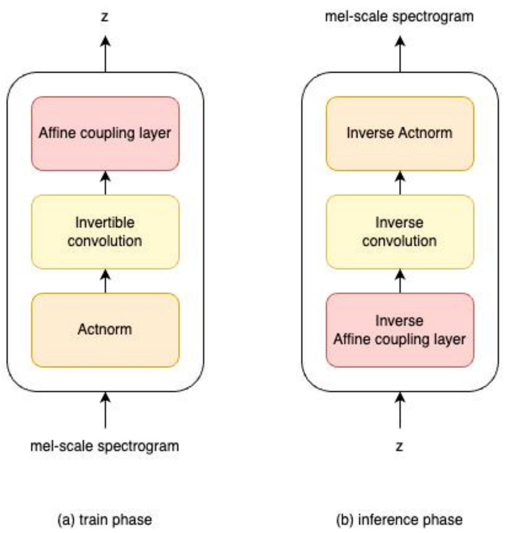
The encoder ingests both the target speaker's voice and the source voice, which have been processed through HuBERT. Features of the source speaker and the target speaker are amalgamated and go through multiple convolutional blocks. The output from the feature encoder is 'z', which is subsequently used as the input to the inverse flow.

Supplementary acoustic features, such as pitch, energy, and voicing, can also be incorporated at this stage. They were obtained using a signal processing package Librosa (McFee et al., 2015). The modulation of pitch information based on the gender difference between the target and source speaker has proven beneficial during inference. For instance, when the target speaker is male and the source speaker is female, the pitch derived from the source speaker is reduced approximately by an octave. This adjusted pitch is then supplied to the Feature Encoder.

2.3. Flow

Normalizing flow is a type of generative model adept at learning a probability distribution from given data (Rezende & Mohamed, 2015). This model works by transforming a straightforward, easily sampled distribution into a more intricate distribution that mirrors the data.

The primary concept behind normalizing flows involves the use of a series of invertible transformations, which map a simple distribution to a more complicated one. These transformations are typically executed as neural networks, with the parameters of these networks being learned from data.



TTS, text-to-speech.

Figure 2. Spectrogram decoder in Glow-TTS.

The flow component for the proposed model is inspired by the spectrogram decoder of Glow-TTS (Kim et al., 2020). Each flow block is composed of an activation normalization layer, an affine coupling layer, and an invertible convolution layer (Kingma & Dhariwal, 2018). During the training phase, the system learns to convert the mel-scale spectrogram into a latent representation 'z', which follows a Gaussian distribution. Conversely, during the inference phase, 'z' is transformed back into the mel-scale spectrogram. This reciprocal process is feasible due to the invertible nature of all operations in the flow block.

The output derived from the Speaker Encoder, details of which will be elaborated subsequently, is utilized as a condition in the Affine Coupling Layer. Rather than being incorporated as a complex expression, the condition is imply added to the input. This straightforward integration facilitates its usage both during the training phase and the inference phase.

2.4. Speaker Encoder

The Speaker Encoder, a key component of the voice conversion model, is designed to accept the mel-scale spectrogram of the target speaker as its input. This module is elegantly simple, composed of a bidirectional LSTM layer and a linear layer.

An LSTM layer is known for its ability to capture long-term dependencies (Hochreiter & Schmidhuber, 1997). A bidirectional LSTM layer is a modified version of a traditional LSTM Layer. It processes the input data in both forward and backward directions, allowing it to capture temporal dependencies from both past and future contexts (Graves, 2013; Schuster & Paliwal, 1997).

The bidirectional LSTM layer processes the mel-scale spectrogram to generate a sequence of hidden states, ensuring a more comprehensive understanding of the speaker characteristics. These hidden states, carrying intricate temporal information, are then processed by the linear layer.

Unlike typical conventional embedding layers that rely on speaker IDs to generate speaker vector, this encoder has the capacity to infer characteristics of speakers unseen during the training phase. This design of the speaker encoder caters to the challenge of zero-shot voice conversion: capturing and emulating speaker characteristics without having prior exposure to the target speaker during training.

2.5. Vocoder

HiFi-GAN is employed as the vocoder for this study. HiFi-GAN is a model that processes a mel-scale spectrogram as input to produce an audio signal as output. Its generator utilizes a transposed convolution to upscale the mel-scale spectrogram, which is then synthesized into a waveform via multi-receptive field fusion. HiFi-GAN employs two discriminators to ensure balance between the intensive work of the generator and the relatively simpler tasks of the discriminators. These two discriminators, namely the multi-period discriminator (MPD) and the multi-scale discriminator (MSD), scrutinize the input and discern its authenticity (Kong et al., 2020).

While traditional applications of HiFi-GAN focus on single speaker usage, it can be adeptly reconfigured for multi-speaker functionality by training it on data from diverse speakers. There are two prominent approaches: one entails providing distinct speaker information to the vocoder, and the other involves generating audio from spectrograms devoid of speaker information. Given the robust and stable nature of HiFi-GAN, we opted for feeding it with multi-speaker data without separate speaker information.

3. Training

The training data utilized was the VCTK dataset, comprised of 44 hours of English speech recorded by 109 distinct speakers (Yamagishi et al., 2019). This dataset provides a well-rounded and comprehensive learning environments for the model to understand and mimic various voice characteristics.

During the training process, the model was fed with same speakers as the source and the target. To ascertain that the speaker vector, generated by the speaker encoder, adequately extracts speaker information, a layer was incorporated that undertakes the speaker classification task. The classification layer was trained together with the voice conversion model and was subsequently discarded after the training process.

For the optimization strategy, the Adam optimizer was utilized. Adam, an acronym for Adaptive Moment Estimation, is a widely used optimization algorithm for training deep learning models, acting as an extension of the Momentum method and the RMSProp optimizer (Buduma et al., 2022; Kingma & Ba, 2014). Known for its efficiency, it swiftly trains deep learning models across diverse tasks by using first- and second-order moment estimates to update parameters. Despite being sensitive to hyperparameters, it stands as a potent and versatile optimization algorithm.

Hyperparameters were carefully tuned to optimize the model's performance. The learning rate was initially set as 0.0001, with a decay rate implemented after every ten epochs to ensure the model could converge smoothly. The model was trained over a total of 1,000 epochs, providing a balance between achieving satisfactory performance and preventing overfitting. Batch size was set at 32, given the trade-off between computational efficiency and the stability of the model's learning process (Keskar et al., 2016).

The model was trained on an NVIDIA Tesla V100 GPU, due to its high computational power and its large memory capacity, which is beneficial for handling complex models. The capability to process large blocks of data in parallel makes it an ideal choice for deep learning tasks.

4. Evaluation

The models deployed in this study were assessed using both the mean opinion score (MOS) and a similarity test. The MOS is determined by averaging the scores given by multiple evaluators on a scale ranging from 1 to 5, where 1 corresponds to 'Bad' and 5 denotes 'Excellent'.

Mirroring the structure of the MOS, the similarity test is also scored on a 1 to 5 scale. However, as suggested by Wester, in the context of the similarity test, a score of 5 denotes that the evaluator is highly confident that the two voices belong to the same speaker (Wester et al., 2016). Conversely, a score of 1 suggests the evaluator is less confident that the two voices represent different speakers.

A total of 20 participants, ten males and ten females, were engaged to evaluate the voice conversion outcomes. The voices were stratified into four distinct settings: seen-to-seen, seen-to-unseen, unseen-to-seen, and unseen-to-unseen, contingent on whether the source and target speakers were included in the training dataset.

The seen speakers comprise a randomized selection of 10 individuals from the training dataset, VCTK. The unseen speakers are drawn from the Hifi Multi-Speaker English TTS Dataset, which is a curated collection of voice samples from 10 distinct speakers sourced from LibriVox (Bakhturina et al., 2021). Ten output audios were synthesized for each setting and subsequently assessed by human evaluators.

Table 1. The MOS with 95% confidence interval

Setting	MOS
Seen to seen	4.02±0.08
Sseen to unseen	3.96±0.12
Unseen to seen	4.01±0.07
Unseen to unseen	3.85±0.10

MOS, mean opinion score.

Table 2. The similarity scores with 95% confidence interval

Setting	Similarity score
Seen to seen	4.21±0.10
Seen to unseen	3.98±0.09
Unseen to seen	4.10±0.06
Unseen to unseen	3.92±0.05

Upon analysis of the results, it was observed that the MOS uniformly hovered around 4 across all settings. Additionally, the similarity scores, which reflect the resemblance of the converted voices to the target voices, demonstrated noteworthy performance, with a particularly impressive similarity score of 3.92 being achieved in the unseen-to-unseen setting. These findings corroborate that the model proposed in this paper exhibits efficacy, even in the challenging context of zero-shot voice conversion.

5. Conclusion

This study introduces an advanced model for zero-shot voice conversion, employing the potent capabilities of HuBERT. The proposed system is an integration of five crucial components—HuBERT, Feature Encoder, Flow, Speaker Encoder, and Vocoder. Together, they culminate in a model that delivers superior performance, outshining expectations even under challenging conditions such as unseen-to-unseen scenarios.

The strength of this model is particularly evident in its remarkable voice quality and speaker similarity metrics. It consistently generates high-quality output that maintains a high degree of similarity to the target speaker's voice, attesting to its precision and reliability. Even when faced with unfamiliar speakers, the model exhibits a robust capacity for voice conversion, showcasing its adaptability and resilience.

Given its robust performance, the model presents wide-ranging applications in real-world settings where high-quality voice conversion is demanded. Its potential impact extends across sectors from telecommunication and entertainment to accessibility services, setting new benchmarks in the field of voice conversion. The success of this research further underscores the relevance and promise of deploying advanced technologies like HuBERT in the ongoing evolution of voice conversion systems.

However, despite the significant strides made in performance and quality, one of the current limitations of the model is its substantial computational weight, largely attributed to the inclusion of the HuBERT module. While this aspect contributes significantly to the model's performance, it also necessitates extensive computational resources, thereby limiting its immediate deployment on devices with modest computational capabilities.

Future work will focus on refining and optimizing the model, particularly the HuBERT component, to reduce its computational demands without compromising the quality and effectiveness of voice conversion. In this regard, attention mechanisms could be explored within HuBERT for qualitative changes in the outcome. Introducing residual forward paths may potentially enhance the numeric flow of the model. Further, for model optimization, especially concerning weight adjustments and calculation speed, the convolution mechanics could be modified to utilize separable, depth-wise convolution. Continued efforts in this vein would lead to a more accessible and widely deployable voice conversion model, bringing the benefits of advanced voice conversion technology to a larger audience.

References

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020, December). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proceedings of the Advances in Neural Information Processing Systems*. Online Conference.

Bakhturina, E., Lavrukhin, V., Ginsburg, B., & Zhang, Y. (2021). Hi-fi multi-speaker English TTS dataset. Retrieved from https://doi.org/10.48550/arXiv.2104.01497

Buduma, N., Buduma, N., & Papa, J. (2022). *Fundamentals of deep learning*. Sebastopol, CA: O'Reilly Media.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. Retrieved from https://doi.org/10.48550/arXiv.1810.04805

Graves, A. (2013). Generating sequences with recurrent neural networks. Retrieved from https://doi.org/10.48550/arXiv.1308.0850

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451-3460.

Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the International Conference on Mmachine Learning* (pp. 448-456). Lille, France.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. Retrieved from https://doi.org/10.48550/arXiv.1609.04836

Kim, J., Kim, S., Kong, J., & Yoon, S. (2020, December). Glow-TTS: A generative flow for text-to-speech via monotonic alignment search. *Poceedings of the Advances in Neural Information Processing Systems*. Online Conference.

Kim, J., Kong, J., & Son, J. (2021, July). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *Proceedings of the International Conference on Machine Learning* (pp. 5530-5540). Online Conference.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. Retrieved from https://doi.org/10.48550/arXiv. 1412.6980

Kingma, D. P., & Dhariwal, P. (2018, December). Glow: Generative flow with invertible 1x1 convolutions. *Proceedings of the Advances in Neural Information Processing Systems*. Montreal, QU, Canada.

Kong, J., Kim, J., & Bae, J. (2020, December). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Proceedings of the Advances in Neural Information Processing Systems*. Online Conference.

Logan IV, R. L., Balažević, I., Wallace, E., Petroni, F., Singh, S., & Riedel, S. (2021). Cutting down on prompts and parameters: Simple few-shot learning with language models. Retrieved from https://doi.org/10.48550/arXiv.2106.13353

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015, July). librosa: Audio and music signal analysis in python. *Proceedings of the 14th Python in Science Conference*. Austin, TX.

Palatucci, M., Pomerleau, D., Hinton, G. E., & Mitchell, T. M. (2009, December). Zero-shot learning with semantic output codes. *Proceedings of the Advances in Neural Information Processing Systems*. Vancouver, BC.

Qian, K., Zhang, Y., Chang, S., Yang, X., & Hasegawa-Johnson, M. (2019, June). AutoVC: Zero-shot voice style transfer with only autoencoder loss. *Proceedings of the International Conference on Machine Learning* (pp. 5210-5219). Long Beach, CA.

Rezende, D., & Mohamed, S. (2015, July). Variational inference with normalizing flows. *Proceedings of the International Conference on Machine Learning* (pp. 1530-1538). Lille, France.

RVC-Project. (2023). RVC-Project/Retrieval-based-Voice- Conversion-WebUI. GitHub. Retrieved from https://github.com/RVC-Project/ Retrieval- based-Voice-Conversion-WebUI

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. Retrieved from https://doi.org/10.48550/arXiv.1904.05862

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673-2681.

Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., Qin, T., ... Bian, J. (2023). NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. Retrieved from https://doi.org/10.48550/arXiv.2304

Sisman, B., Yamagishi, J., King, S., & Li, H. (2021). An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 132-157.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015, July). Deep unsupervised learning using nonequilibrium thermodynamics. *Proceedings of the International Conference on Machine Learning* (pp. 2256-2265). Lille, France.

svc-develop-team. (2023). svc-develop-team/so-vits-svc. GitHub. Retrieved from https://github.com/svc-develop-team/so-vits-svc

van Niekerk, B., Carbonneau, M. A., Zaïdi, J., Baas, M., Seuté, H., & Kamper, H. (2022, May). A comparison of discrete and soft speech units for improved voice conversion. *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6562-6566). Singapore, Singapore.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., ... Polosukhin, I. (2017, December). Attention is all you need. *Proceedings of the Advances in Neural Information Processing Systems*. Long Beach, CA.

Wester, M., Wu, Z., & Yamagishi, J. (2016, September). Analysis of the voice conversion challenge 2016 evaluation results. *Proceedings of the Interspeech* (pp. 1637-1641). San Francisco, CA.

Yamagishi, J., Veaux, C., & MacDonald, K. (2019). CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. University of Edinburgh. *The Centre for Speech Technology Research (CSTR)*, *6,* 15.

• **Hyelee Chung**
Ph. D. Student, Dept. of English Language and Literature
Korea University
145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea
Tel: +82-2-3290-1991
Email: hyeleech@korea.ac.kr
Fields of interest: Phonetics, Phonology, Language engineering

• **Hosung Nam,** Corresponding author
Professor, Dept. of English Language and Literature
Korea University
145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea
Tel: +82-2-3290-1991
Email: hnam@korea.ac.kr
Fields of interest: Phonetics, Phonology, Language engineering