ONE-SHOT VOICE CONVERSION USING VQVC MODEL

This study leverages One-shot Voice Conversion technology using the VQVC model and WavLM to achieve high-quality voice conversion with limited data.
By separating content and speaker information, it overcomes data collection challenges and contributes to efficient, adaptable voice conversion systems.

20200437 김채현

MOTIVATION

This study focuses on One-shot Voice Conversion (VC), transforming a source speaker's voice into a target speaker's using only a single example. It requires minimal data, making it practical and scalable. By separating content and speaker information, it advances speech processing technology, allowing precise control and generalization to new speakers. One-shot VC is essential for developing efficient, adaptable voice conversion systems, significantly contributing to the field.

BACKGROUND

The original VQVC model, which successfully separates content and speaker information by applying Vector Quantization (VQ) methods, shows limitations in naturalness and interpretability. One-shot VC models learn from limited data, overcoming the challenge of collecting large amounts of parallel data and suggesting a new research direction. This advancement aims to develop more practical and generalized models to enhance speech processing technology.

IMPLEMENTATION

1 Dataset Selection and Preparation

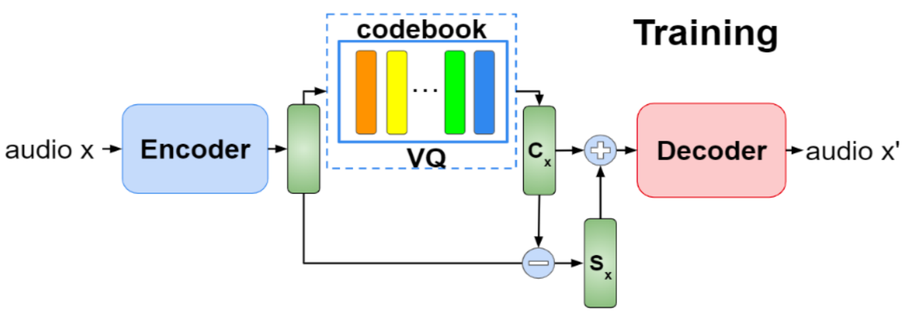
The dataset utilized in this study is the CSTR VCTK Corpus, which comprises audio data from 110 English speakers with various accents.

2 Generation of SSL Features Using WavLM

This method of generating SSL features using WavLM enables the extraction of embeddings that encompass a variety of characteristics directly from the waveform.

3 Application of the VQVC Model

The VQVC model effectively separates the content information and speaker identity from the features extracted by the WavLM. The model consists of three main components: Encoder, Vector Quantization (VQ), and Decoder.



4 Waveform Conversion Using Vocoder

The generated Mel-spectrogram is converted into the final audio waveform using a high-performance vocoder called VocGAN.

5 Application of Loss Functions and Model Evaluation

- Reconstruction Loss

  The Mean Squared Error (MSE) between the input Mel spectrogram and the output Mel-spectrogram is calculated to evaluate how well the model reconstructs the original speech.

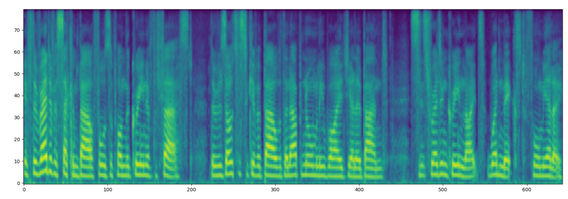
- Latent Loss

  This loss minimizes the distance between the encoder output and the codebook vectors, ensuring quantization efficiency and model stability in the VQ-VAE model.

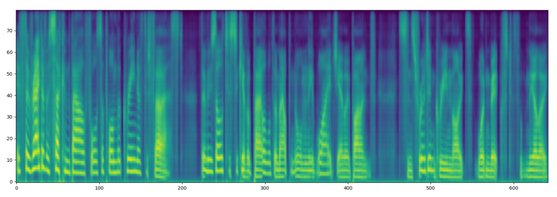
RESULT

Analysis of Mel-spectrograms for the Proposed Model

To evaluate the performance of the proposed VQVC model, a comparative analysis of the Mel-spectrograms was conducted.



Ground-truth Mel-spectrogram



Reconstructed Mel-spectrogram

When comparing the Ground-truth and Reconstructed Mel-spectrograms, we observe high similarity in frequency and energy distribution. This indicates that the model effectively reproduces the original speech's intensity, tone, and temporal changes.

Performance Comparison of MOS(Mean Opinion Score) /SMOS (Speaker Mean Opinion Score) between Baseline and Proposed Models

- Evaluation Method: Subjective listening tests to measure MOS/SMOS
- Evaluation Subjects: Baseline VQVC model and Proposed VQVC model.
- Evaluators: A total of 15 listeners (each listener rated each sample on a scale from 1 to 5).

	seen-to-seen		unseen-to-unseen	
	MOS	SMOS	MOS	SMOS
Baseline VQVC	1.14 ± 0.06	1.48 ± 0.08	1.21 ± 0.13	1.43 ± 0.13
Proposed VQVC	4.32 ± 0.09	3.67 ± 0.1	3.71 ± 0.27	3.57 ± 0.13

These results clearly demonstrate that the Proposed VQVC model outperforms the Baseline VQVC model in both quality and speaker similarity.

DISCUSSION

This study has demonstrated the efficacy of the proposed VQVC model in enhancing the performance of One-shot Voice Conversion (VC). The experimental results indicated that the proposed model significantly surpassed the Baseline VQVC model in terms of audio quality and speaker similarity, as evidenced by substantial improvements in MOS and SMOS scores. Notably, the proposed model maintained high performance with minimal data and exhibited superior generalization capabilities to new speakers.

However, this study is not without limitations. Firstly, the dataset utilized for the experiments was restricted to English speakers, necessitating further validation to assess the model's generalizability to other languages. Secondly, the evaluation method primarily relied on subjective listening tests, highlighting the need for additional validation using objective metrics.

REFERENCE

1. Van Den Oord, A., & Vinyals, O. (2017). Neural discrete representation learning. Advances in neural information processing systems, 30.
2. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., ... & Wei, F. (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.
3. Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2018, April). Generalized end-to-end loss for speaker verification. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4879-4883). IEEE.