2024 년 봄 POSTECH 컴퓨터공학과 과제연구 최종 보고서

[One-shot Voice Conversion Using VQVC Model]

학 번: 20200437
이 름: 김채현
연구 지도교수: 서영주
학 과: 컴퓨터공학과

연구 목적 (Problem statement)

This study is part of the research on Voice Conversion (VC), which aims to transform the voice of the source speaker into that of the target speaker by modifying only the voice, tone, accent, and pronunciation while keeping the linguistic information intact. Among the various approaches, this study focuses on One-shot Voice Conversion, a recently prominent area in the field of speech processing. One-shot VC performs the task of converting one speaker's voice into another's using only a single example. The significance and necessity of One-shot VC can be outlined as follows:

Firstly, One-shot VC models can be developed with limited training data. Typically, models used in one-to-one or many-to-one VC, such as GMM-based or regression-based models, are supervised learning models that require a substantial amount of parallel data. However, collecting sufficient voice data for specific speakers can be challenging. Unlike the aforementioned models, One-shot VC leverages minimal training data to provide a practical and scalable solution. This approach offers flexibility and adaptability, making it applicable in scenarios where data collection is restricted or impractical.

Secondly, One-shot VC models demonstrate the advancement of speech processing technology by effectively separating content and speaker information. This separation allows for more precise control over the VC process and facilitates the development of more sophisticated speech synthesis systems.

Lastly, for the reasons mentioned above, One-shot VC models can generalize to new speakers. Since One-shot VC models can effectively separate content and speaker information using only a

single example of speaker data, they are capable of generalization, thereby offering practical applicability in domains with high diversity.

In summary, the study highlights the importance of One-shot Voice Conversion in creating efficient, adaptable, and advanced voice conversion systems with minimal data requirements, contributing to the broader field of speech processing technology.

연구 배경 (Motivation and background)

In the field of voice conversion technology, the application of the Vector Quantization (VQ) method in performing One-shot VC tasks has led to the development of the VQVC model, which successfully separates content and speaker information. However, the speech generated through this model exhibits limitations in terms of naturalness and interpretability. Specifically, VQ-based voice conversion models often produce speech with unnatural tone and pronunciation. Due to these shortcomings, voice conversion technology using the VQ method still requires improvements to achieve better speech quality.

Furthermore, one of the significant challenges of traditional supervised learning methods is the difficulty in collecting large amounts of parallel data. However, One-shot VC models show potential in overcoming this limitation by learning to separate content and speaker information from limited training data. This suggests a new research direction in voice conversion technology, aiming to develop more generalized and practical models by addressing the difficulties in data collection.

This advancement points towards the development of voice conversion models that can better generalize and adapt to realistic scenarios, thus overcoming the existing limitations and paving the way for more refined and effective speech processing technologies.

연구 방법(Design and implementation)

Overview and Approach

The objective of this study is to achieve high-quality voice conversion using limited data. To this end, I use Waveform Language Model (WavLM), for Self-Supervised Learning (SSL) feature

extraction and the VQ technique to remove speaker identity, thereby separating content and speaker information.

1. **Generation of Self-Supervised Learning (SSL) Features Using WavLM and Removal of Speaker Identity with VQ Method**

   By utilizing the WavLM, pre-trained through self-supervised learning, I can extract richer embeddings from voice data. Subsequently, I apply the VQ method of the VQVC model to the SSL features obtained from WavLM, removing speaker identity and retaining only content information.

2. **Extraction of Speaker Information Using Content Information**

   By effectively removing the quantized content information of the target speaker from the raw vectors, I can objectively extract the speaker information.

3. **Synthesis of Converted Speech**

   The content information and speaker information extracted from the above methods are combined to synthesize speech. This approach results in synthesized speech with higher quality and similarity, enabling more natural voice conversion.

Detailed 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. This rich dataset provides an ideal environment for training the voice conversion model, as it is expected to significantly aid the model in synthesizing speech that closely resembles the original speaker's voice.

2. **Generation of SSL Features Using WavLM**

   The WavLM model employed in this study performs self-supervised pre-training using raw audio data (waveform) as input. The input data, in the form of PCM (pulse-code modulation) time-series data, is directly processed by WavLM through complex convolutional layers and a transformer encoder to extract 1024-dimensional feature vectors.

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.

- **Encoder**

The encoder transforms the 1024-dimensional SSL feature extracted by WavLM into a low-dimensional latent vector. This process involves 1D convolutional layers and residual blocks, which extract and emphasize the temporal and spatial characteristics of the audio, ultimately outputting a 32-dimensional feature vector.

- **Vector Quantization (VQ)**

The continuous 32-dimensional feature vectors obtained from the encoder are mapped to discrete codebook vectors. This process involves finding the closest vector using the VQVC model's codebook and quantizing the features to that vector. The codebook size is set to 256. The quantized vector purely represents content information independent of speaker identity, enabling consistent voice conversion across various speakers. Additionally, the residual (the result of subtracting the quantized vector from the raw vector) contains the speaker identity information. This residual information is used as the speaker embedding, which is applied in the decoder to re-synthesize the Mel spectrogram with the desired speaker characteristics.

- **Decoder**

The decoder combines the quantized content vector and the separately extracted speaker identity vector to re-synthesize an audio signal similar to the original Mel spectrogram. The decoder takes these two vectors as inputs and generates the final audio signal through complex convolutional and residual layers. The aim of this process is to produce converted speech that retains the original speaker's content characteristics while adopting the voice characteristics of another speaker.

4. Waveform Conversion Using Vocoder

The generated Mel spectrogram is converted into the final audio waveform using a high-performance vocoder called VocGAN. VocGAN is specifically designed to preserve naturalness, resulting in converted waveforms that closely resemble the original audio quality.

5. **Application of Loss Functions and Model Evaluation**

Two types of loss functions are applied during the model training process.

- **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. Lower MSE values indicate higher reconstruction accuracy.

- **Latent Loss**

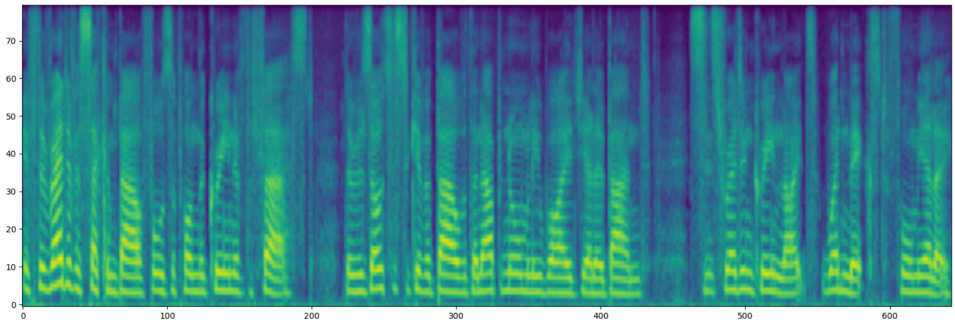
This loss minimizes the distance between the encoder output and the codebook vectors, ensuring quantization efficiency and model stability in the VQ-VAE model. It is used to minimize information loss that occurs during the mapping of encoder outputs to codebook vectors.

연구 결과 및 평가 (Methodology and evaluation)

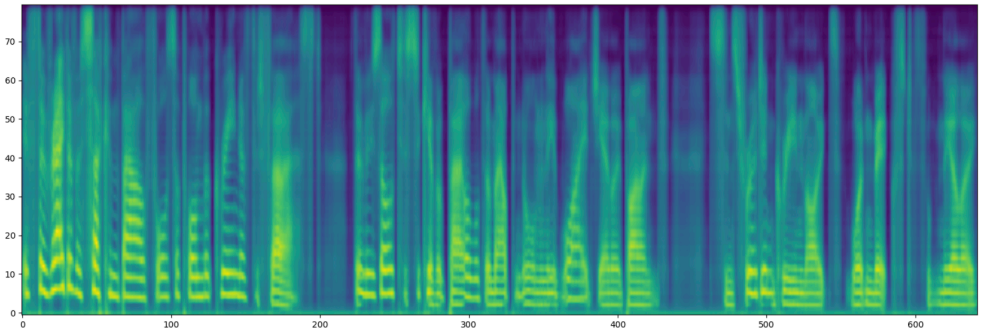
Analysis of Mel-spectrograms for the Proposed Model

To evaluate the performance of the proposed VQVC model, a comparative analysis of the mel-spectrograms was conducted. This allows for a visual assessment of the similarity between the original and converted speech. The figure below compares the mel-spectrograms of the original speech (Ground-truth), reconstructed speech (Reconstructed).

- **Ground-truth Mel-spectrogram**



- **Reconstructed Mel-spectrogram**



When comparing the Ground-truth mel and Reconstructed mel, I observe a high degree of similarity in the temporal frequency distribution. Specifically, the following aspects show notable similarity:

- **Frequency Bands**

  Both high and low-frequency regions display similar patterns, indicating that the model effectively reproduces the full frequency spectrum.

- **Energy Distribution**

  The bright areas in both mel-spectrograms represent high energy, and this energy distribution is very similar. This demonstrates that the model maintains the original speech's intensity and tone.

- **Temporal Dynamics**

  The temporal changes in the speech signal, such as the variations at the beginning and end of the speech, are consistently represented in both the original and reconstructed mel-spectrograms.

Performance Comparison of MOS/SMOS between Baseline and Proposed Models

Experimental Setup

- **Evaluation Method**: Subjective listening tests to measure MOS (Mean Opinion Score) and SMOS (Speaker Mean Opinion Score).
- **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).

Result

The table below compares the MOS and SMOS scores of the two models.

	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

Analysis

The Baseline VQVC model performed poorly in both scenarios, with a slight improvement in the unseen-to-unseen scenario but still overall low performance.

The Proposed VQVC model received high ratings in both scenarios, with a slight decrease in performance in the unseen-to-unseen scenario compared to the seen-to-seen scenario.

These results clearly demonstrate that the Proposed VQVC model outperforms the Baseline VQVC model in both quality and speaker similarity. The significant improvements in MOS and SMOS scores highlight the effectiveness of the modifications introduced in the proposed model, particularly its superior generalization performance to new speakers.

토론 및 전망 (Discussion and future work)

This study has demonstrated the efficacy of the proposed VQVC model in enhancing the performance of One-shot Voice Conversion (VC) models. 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.

Future research can extend this work in several key directions. Firstly, it is imperative to validate the generalizability of the proposed model using datasets that encompass various languages and dialects. Secondly, further research should focus on refining the model architecture and training methods to achieve even better audio quality and speaker similarity. Thirdly, exploring the expansion of voice conversion technology to real-time applications is essential.

Moreover, the potential applications of voice conversion technology are extensive. In the entertainment industry, it can be employed to naturally convert the voices of various characters in animations, games, and movies. In the education sector, it can facilitate multilingual learning, while in the medical field, it can enhance communication with patients through voice-based interfaces.

In conclusion, this study lays a crucial foundation for improving the performance of One-shot VC models. Future research will contribute to achieving better performance and exploring diverse applications, thereby advancing the field of voice conversion technology.

참고 문헌(References)

Van Den Oord, A., & Vinyals, O. (2017). Neural discrete representation learning. Advances in neural information processing systems, 30.

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.

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.