The University of Hong Kong
Faculty of Engineering
Department of Computer Science

Final Year Project - Voice Changer

Interim Report

Supervisor: Dr Yip Chi Lap
Contributor: Tam Man Hon (UID: 30354346856)

Submitted on 20th January 2021
Abstract

This paper is conducted with the main goal of creating a real-time voice changing computer program that can make the voice of a source speaker to be like the sound of a target speaker. To accomplish the object, the paper will look into the existing paper posted in the previous. To be specific, the paper will evaluate the voice changing approach with vector quantization in terms of speaker similarity, conversion speed and more. This paper has a side objective of implanting the program on a mobile device, which more algorithmic criteria will be put into concern, like energy consumption and resource need. In this progress report, it introduces a computer vector quantization program that transforms raw audio data into a quantized signal with a small mean squared quantization error of 0.00257. The coming step of this project is to utilize the program to try creating a voice conversion program by mapping the vector quantized sample data to the vector quantized target data.
# Table of Contents

1. List of Figures .................................................................................................................. P.4
2. List of Tables .................................................................................................................... P.4
3. Acknowledgements ......................................................................................................... P.5
4. Introduction ..................................................................................................................... P.5
   4.1 Background .................................................................................................................. P.5
   4.2 Problem ....................................................................................................................... P.5
   4.3 Objectives .................................................................................................................... P.6
   4.4 Project Outline ........................................................................................................... P.6
5. Methodology .................................................................................................................... P.6
   5.1 Training Phase ............................................................................................................ P.7
   5.2 Conversion Phase ......................................................................................................... P.8
   5.3 Comparing Method ................................................................................................. P.8
6. Current Status – Vector Quantizer ................................................................................ P.8
   6.1 Encoding ....................................................................................................................... P.9
   6.2 Decoding ..................................................................................................................... P.10
   6.3 Reason for choosing the parameters ..................................................................... P.11
7. Difficulties ....................................................................................................................... P.11
   7.1 Testifying Wrong Approaches ............................................................................. P.11
   7.2 Program Compatibility .............................................................................................. P.11
8. The Next Step and Project Timeline ............................................................................ P.13
   8.1 The Next Step ........................................................................................................... P.13
   8.2 Project Timeline ......................................................................................................... P.13
9. Conclusion ....................................................................................................................... P.14
10. Reference List ............................................................................................................... P.15
1. List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Training and conversion phases of a typical Voice Conversion system</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>The original speech signal without nominalization</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>The two audio signals after transforming from 1-Dimensional audio</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>The quantized result of applying the encoding part of the Vector Quantizer program</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>The audio signal after decoding the data</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>Recording recorded with Laptop built-in microphone</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>Sample recording in Zero Speech 2020 Challenge</td>
<td>12</td>
</tr>
</tbody>
</table>

2. List of Table

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The significant parameters used when writing Vector Quantizer</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>The timeline of the current project</td>
<td>12</td>
</tr>
</tbody>
</table>
3. Acknowledgements

I would like to express my gratitude to my report supervisor Dr Yip Chi Lap for his guidance and useful critiques of this report work.

4. Introduction

4.1 Background

Since a very long time ago, voice changing, or voice disguising, has been involved in our lives. From parents using baby talks, which is raising their pitch, to try to catch their baby’s attention, to a vigilante trying to hide his identity from the villains in a movie, or even just a mimic making a funny voice to the audience, these are all examples of voice changing. In the digital world, with the aid of modern computers, voice changing is not anymore limited to physically tuning our way of speaking to make ourselves unrecognizable. Now, we can enhance the changes by adding special features to the voice or trying to make one’s voice sounds like somebody else’s (will be referred as voice conversion throughout this proposal) - --- by using Voice Changer.

Voice Changer is a program or device that changes the voice of a speech speaker while keeping the linguistic content of the original audio.

4.2 Problems

Adding distortions to a speech signal to make it unrecognizable is not a challenging task. For instance, to change the pitch of a voice recording, only the playback speed of the signal needs to be changed¹. Decelerating the signal will lower the pitch and increasing the speed will higher the pitch. Performing these changes, even to current mobile applications, is by no means heavy task. In fact, there are many mobile applications that can accomplish the job. Voice conversion, on the other hand, is a much more challenging procedure.

To perform voice conversion, there are multiple speech processing techniques involved, including, speech analysis, spectral conversion, speaker identification, vocoding, etc.² And many of the techniques cannot be applied easily. For instance, Speech analysis is a process of identifying and extracting relevant information from the speech signal. To perform it, one needs to know what to identify and how to extract without losing crucial information. As a result, comprehensive research needs to be performed in different phases during the project.

4.3 Objectives

The objective of this project is to build a standalone voice conversion program that uses parallel modelling method and runs in real-time.

To achieve the goal of the project, the main objective can be further divided into the following:

i. To find out the acoustic features that are essential for humans to recognised other’s voice
ii. To develop a program function that extracts the corresponding acoustic features
iii. To implement the algorithm to a standalone program that can perform voice conversion in real-time
iv. To investigate if the possibility to implement the program into a mobile application. If time allows, build an IOS application with the function of changing a recorded

4.4 Report Outline

The remainder of this progress report is structured as follows. Chapter 5 includes the methodology of how a voice changer will be implemented and compared. In chapter 6, it contains the current status and the reasons for selecting some of the parameters adopted. Chapter 7 contains the difficulties encountered in the progress. Chapter 8 includes the future plan and timeline of the project. A conclusion of the whole process is included at the end in Chapter 9.

5. Methodology

The project will investigate a parallel voice conversion model, Vector Quantization, based on the resource required, speaker similarity, speech quality and conversion speed. Then select the appropriate method for the product program.

The method Vector Quantization is referenced from Nakamura’s “Voice conversion through vector quantization”\(^3\). This technique belongs to parallel voice conversion model. The term, parallel, implies that the source speaker and the target speaker of the training recording are required to being uttering a same sentence or same set of words. This method helps make a mapping codebook, that shows the binding between the codebook of the source speaker and the target speaker. The method is expected to control voice individuality accurately.

---

Figure 1. Training and conversion phases of a typical Voice Conversion system.

The pipeline of this project will follow the procedure shown in Figure 1. As seen in Figure 1, to perform a voice conversion, there are in principle two main phases, one of them is the Training Phase, and the other one is the Conversion Phase.

5.1 Training Phase

In the training phase, it first takes a set of parallel sentences (sentences including the same linguistic contents) from a source speaker and a target speaker respectively.

Then, a speech analysis is performed to extract the acoustic properties, including but not limited to the average spectrum, formants, and the average pitch level (as these three factors are considered as the most relevant data\(^6\)), of the two speakers. After acquiring the data sets, feature mapping computation will be performed to increase the relevance of the feature sets.

---


The next step in the training phase is time alignment. Since the sample sentences are acquired from two different people, the recording time and the phoneme durations within the utterance can then also be different. This step indicates the differences.

5.2 Conversion Phase

In the conversion phase, it takes an input utterance and also first performs the speech analysis and feature mapping steps like in the training phase.

After deciding the modelling method, the next step is to transfer the data gotten from the processed sources, target recordings and input utterance into the converted features and output as the utterance result.

5.3 Comparing Method

As mentioned in the report earlier, the voice conversion model will be based on the resource required, speaker similarity, speech quality and conversion speed (more criteria will be added judging from the actual situation). And when doing the compartment, 12 testers will be asked to listen to the two converted speech and rate the speaker similarity and speech quality from one to five, which five indicating the best result.

Since that it is the beginning part of the project, the training and conversion phases of the voice conversion model are what is focused on currently. As a result, more comparison criteria will be added after further research is completed.

6. Current Status – Vector Quantizer

Since this is the beginning phase of my project, there is not any notable work delivered previously.

In this project, one of the two approaches to create a voice conversion model is through vector quantization. To date, a vector quantization programme has been created using python. The programme will be used to generate a codebook.

The programme uses the setting listed in Table 1:

<table>
<thead>
<tr>
<th>Input audio file format</th>
<th>Waveform Audio File Format (.wav)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling Frequency</td>
<td>44100Hz</td>
</tr>
<tr>
<td>Number of the dimensional signal vector</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 1. The significant parameters used when writing Vector Quantizer.

The program can be separated into two parts, encoding and decoding.

6.1 Encoding
In the encoding part of the program, the program first takes a speech recording (Figure 2) as the input audio signal.

![Figure 2. The original speech signal without nominalization.](image)

Then, the program applies normalization to the audio signal to make sure the signal will match with other training and target signals that will be involved in the future.

After normalization, the signal is redistributed into two arrays of the same size, transforming the audio signal array from \([1, 2, 3, \ldots, 22, 23, 24]\) to \([\[1, 3, \ldots, 21, 23\], [2, 4, \ldots, 22, 24]\])

This helps transform the signal from one-dimensional into two-dimensional. The two generated signals will have a length of one half of the original signal. The signals transformed is shown in Figure 3.

![Figure 3. The two audio signals after transforming from 1-Dimensional audio. As the sampling frequency is not very high frequencies (currently at 44100Hz), the curves through the samples are more or less smooth, and the two generated audio signals have roughly the same shape.](image)

Then the data in the two signal arrays will be fitted into a k-means++ clustering model by putting them into groups having approximately the same number of points closest to them to generate a codebook, which is a collection of all the codewords.
And after fitting, the indices of the codeword are taken to represent the corresponding signal. The vector quantized result of the coded signal is shown in Figure 4.

![Coded Signal]

Figure 4. The quantized result of applying the encoding part of the Vector Quantizer program to the original audio signal (shown Figure 1). Note that the sample size of the encoded result is half of the original signal.

The final step of the encoding part is to create a dictionary that matches the centroids of the two arrays in Figure 3 to the indices. The dictionary will be used in the decoding part.

### 6.2 Decoding

The following part of the Vector Quantizer program is to decode the signal as seen in Figure 4 back to a meaningful audio signal.

To decode the signal, the program utilizes the dictionary created in the encoding part by matching the dictionary to the quantized result. The final result is shown in Figure 5.

![Decoded Signal]

Figure 5. The audio signal after decoding the data (shown in Figure 4).

When comparing with the original signal (shown in Figure 2), the result (shown in Figure 5) is not as smooth. The resulted signal, however, keeps most of the features of the original signal and holds a small mean squared quantization error of 0.00257. Comparing the result to that of mid-rise’s quantization (another quantization method), which has an error of 0.01214 on the same recording, the vector quantization method shows a significant 471% improvement. In addition to its advantage in data restoring, owing to the fact that the data points of vector quantization are represented by the index of their closest centroid, it provides a low error method for voice conversion, while using less space than processing the original signal.
6.3 Reason for choosing the parameters

Input audio file format

The programme is set to take Waveform Audio File Format (WAV) file as input. Unlike many commonly used audio formats, like MPEG-1 Audio Layer III (MP3), Advanced Audio Coding (ACC) and Ogg Vorbis, WAV is uncompressed audio format which indicates the files remain the same size from origin to destination. This helps to keep all the acoustic features that recorded file has for further processing.

Sampling Frequency

The sampling rate for the audio signals is set at 44100 Hz. At this frequency, the details of the recording can be captured while keeping the file stay in a reasonable size.

Number of the dimensional signal vector

The programme originally uses four-dimensional signal vector (transforming the signal into four 1-D arrays). Through experiments, although the decoded audio generated from four-dimensional signal vectors is significantly distorted, the fitting process is lengthened exponentially (from 2^3 to 4^3). As a result, the signal vector dimension number is now set as 2 instead of higher.

7. Difficulties

Despite the long history of voice conversion and the thoroughly research all the digital libraries are filled with, there are some challenges encountered in the process.

7.1 Wrong Approaches Testifying

Since that the research field is overwhelmed with all kinds of successful approaches of conducting voice conversion, it is less easy to find a place to testify unworkable methods and thus, consumes time for testifying some wrong approaches. At the beginning of the project, the originally planned approach is to try finding a constant equation that matches the original speaker’s speech signal directly to the target speaker. And because of the popularity of the topic, these kinds of simple and not working approaches often get pushed to the very bottom of the search result. In the end, a sample program was built to check a constant equation is not the correct approach.

7.2 Program Compatibility

The original plan of this research is to conduct the result on MATLAB as the program is well designed for signal processing. Since there exists some library compatibility issue, the current process was mainly achieved on python platform. However, even though the program is currently built with python, there is a programming issue concerning one of the processes in training the model, which is the function of Dynamic Time Warping (DTW).
7.3 Hardware Limitation

The sample audio recordings are recorded using the built-in microphone of a laptop. As a result, the signal quality is highly sensitive to the surrounding noise. Figure 6 shows one of the sample recordings recorded with the built-in microphone with the audio program, Praat. Comparing it to the Figure 7, which is a sample recording of Zero Speech 2020 Challenge, the signal shows an intensive distortion. The distortion can be problematic for speech analysis. Currently, the solution to the problem is to use the sample audio in the past voice conversion challenge as the training data.

Figure 6. Recording recorded with Laptop built-in microphone.

Figure 7. Sample recording in Zero Speech 2020 Challenge.
8. The Next Step and Project Timeline

8.1 The Next Step

The immediate next step of this project is to solve the compatibility issue mentioned in the Difficulties Section. As mentioned, the current process was mainly achieved on python platform, and there is a technique coding issue in dynamic time warping. The first step after this report is to investigate a solution to the issue and generate a mapping codebook that will be used in synthesizing a sample audio.

After synthesizing a sample audio, the direction of the project will be towards improving the program runtime and achieve a real-time conversion.

8.2 Timeline

The project will follow the schedule below:

<table>
<thead>
<tr>
<th>Timeline</th>
<th>Milestones</th>
<th>Deliverables</th>
</tr>
</thead>
</table>
| September                   | Proposal                        | • Project plan  
|                              |                                 | • Project Webpage                                                           |
| October - November          | Design and Research             | • Research on voice changing theories  
|                              |                                 | • Select appropriate methods for voice changing  
|                              |                                 | • Building a vector quantizer                                               |
| Mid December to February    | Research and Development        | • Code migration  
|                              |                                 | • Implementing a voice changer based on vector quantization  
|                              |                                 | • Synthesize a sample audio  
|                              |                                 | • More research on making                                                   |
| February to March           | Development                     | • Improve the algorithm to make the voice changing process to real time  
|                              |                                 | • Try implementing the program into a mobile app                           |
| April                       | Deployment                      | • Finalise the result  
|                              |                                 | • Final Report                                                              |

Table 2. The timeline of the current project
9. Conclusion

In conclusion, this project is conducted to understand the acoustic features that are essential for humans to recognize other’s voice. It will evaluate the result of voice conversion with vector quantization.

The current deliverable of this paper is a vector quantization program that transforms raw audio into a vector quantized signal that can be used to match with the quantized signal. The error of the quantizer is relatively small with a mean squared quantization error of 0.257%. An important next step is to utilize the program to try creating a voice conversion program by mapping the vector quantized sample data to the vector quantized target data and synthesize a sample audio.

In the end, it is to conclude that voice conversion is not a simple technology that substituting the original speaker’s speech data to a constant equation can solve.
10. Reference List


