Selection of Optimal Solution for Example and Model of Retrieval Based Voice Conversion

Zhongxi Ren

Faculty of Data Science, City University of Macau, Macau, 999078, China
D21090102776@cityu.mo

Abstract. Since 2010, the computer has been developing continuously in the field of speech conversion, and now the speech-to-text technology has become mature, but the development of timbre conversion and imitation is not perfect. Recently a new tone imitation program has become a focus, but this program model training options are still lacking. This paper hopes to train the model through the in-depth practical operation of this program and the custom value in the model training step of this program. Multiple training processes of Retrieval Based Voice Conversion (RVC) model will be practiced, and the timbour produced by the model with different number of rounds will be compared with the sound source. After the model training, two evaluation methods were used to check the similarity of the evaluation model. One is the objective evaluation method based on Mel cepstral distortion principle, which is realized by software. The other is a subjective evaluation method based on the principle of directly collecting human sensory data. The similarity statistics are obtained respectively, the selection criteria of the general optimal solution model are obtained, and the relative standard training reference values are provided for users.

Keywords: Timbre conversion, Mel cepstral distortion,Model training, Objective evaluation, Subjective evaluation.

1 Introduction

Retrieval based Voice Conversion (hereinafter referred to as RVC) is a new type of voice conversion program developed in June 2023. Compared with traditional training timbre imitation program, Retrieval based Voice Conversion eliminates timbre leakage by replacing input source features with training set features using top1 retrieval.At the same time, RVC has the following characteristics: even on relatively poor graphics cards can be fast training; Training with a small amount of data can also get better results (it is recommended to collect at least 10 minutes of low-noise speech data); The timbre can be changed by model fusion (with the help of CKpt-merge in the ckpt processing TAB); Easy to use web interface; The Ultimate Vocal Remover 5 (UVR5) model can be invoked to quickly separate vocals from accompaniment [1]. However, RVC also cannot intelligently select the number of rounds with the highest similarity to the training sound source in the process of model training, so it needs to

B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_48

manually customize the number of training rounds and provide multiple model results for manual screening. In this paper, multiple training processes of RVC model will be practiced, and the timbour produced by the model with different number of rounds will be compared with the sound source, so as to find out the most similar training rounds between the model and the sound source.

2 Model training practice based on RVC

The next audio source to be trained is the voice from the cartoon character. The reason for this choice is that the voice of the cartoon character has the characteristics of stable voice line and clear audio source. Because the training steps of each model are exactly the same, and the values set during training are exactly the same, so in this part, we will take out the training process of one of the models, according to audio acquisition, format conversion, set program parameters, limit the number of training rounds, show the training process of the model, and show the process of output training results in detail.

2.1 Audio collection

In terms of audio source material selection, RVC requires the file format of wav, and the total audio source material duration is preferably about ten minutes. At the same time, it is best to choose materials with no background music or small background music. If background music is unavoidable, third-party software or RVC's own Ultimate Vocal Remover 5 routine can be used to separate and batch process vocal accompaniment [2]. It is worth mentioning here that Ultimate Vocal Remover 5 is a deep neural network-based instrument separation software that trains the model to accurately separate drums, bass, vocals and other vocal parts. Compared to RX10, RipX, and SpectraLayers, UVR5 shows significant advantages in terms of model generation quality and selectivity.

When the audio file is imported into the program (all audio formats can be automatically recognized), the UVR5 model drives the program to generate two files: the voice file and the background file (Fig.1).

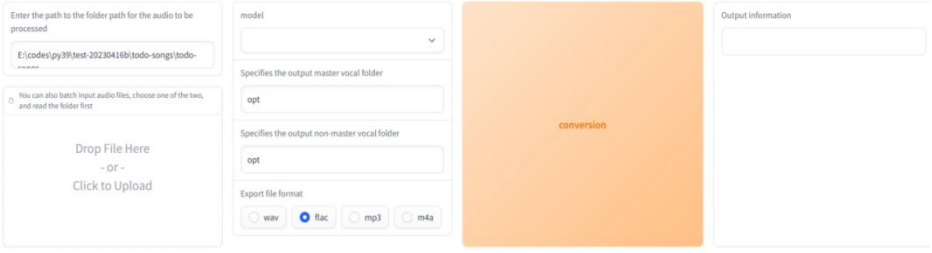

Fig. 1. Voice separation program based on UVR5 (Original)

2.2 Training parameter setting

At this stage, in order to save hard disk space, the storage frequency of this experiment was set to once 50 rounds (Fig.2). At the same time, in order to obtain relatively comprehensive data to detect the similarity of the trained model and expand the result data set, the total number of training rounds was set to a larger 500.The Graphics Processing Unit (GPU) is divided into 10G memory for the program, it is worth mentioning that RVC currently only supports Nvidia series graphics cards. Because RVC relies on NVIDIA's CUDA technology. CUDA (Compute Unified Device Architecture) is a programming model developed by NVIDIA to take full advantage of their graphics processing units (Gpus) for large-scale computing. Many deep learning and machine learning libraries, such as TensorFlow and PyTorch, can directly leverage CUDA for efficient numerical computation. AMD graphics cards do not support CUDA because it is a proprietary NVIDIA technology. AMD has its own similar technology called ROCm, but ROCm is not as widely supported and adopted as CUDA.

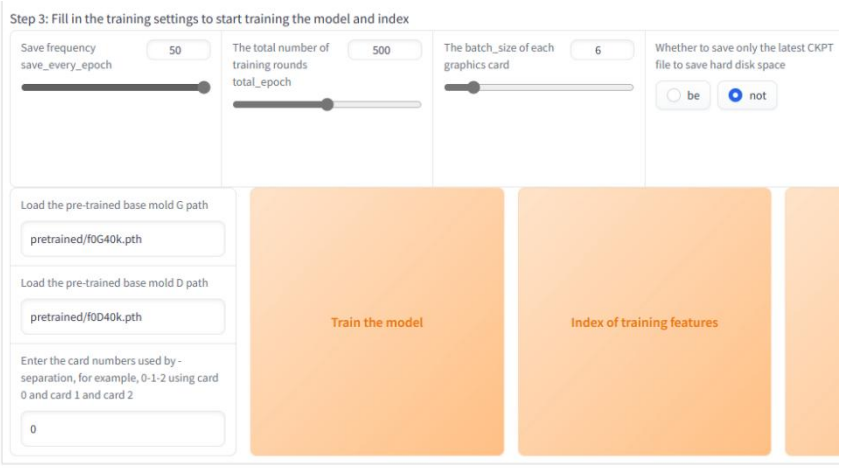

Fig. 2. Parameter setting (Original)

2.3 Program initiation

- Audio automatic segmentation

At the beginning of the training session, RVC automatically divides a single 10-minute audio file into one-sentence wav files. The entire audio segmentation process can be completed in as little as a minute or two. This capability is made possible by RVC's ability to leverage the Kaldi and Mozilla DeepSpeech programs, a powerful open-source speech recognition toolkit that provides a range of tools and libraries for speech recognition. It contains various components for training and testing speech recognition models and supports a variety of speech recognition tasks, including automatic segmentation of audio. DeepSpeech is Mozilla's open-source speech

recognition engine, based on deep learning technology. It provides tools for training and using speech recognition models with high accuracy and performance. This process allows for automatic segmentation of audio using DeepSpeech [3]. In addition, RVC can also select the appropriate audio segmentation program based on the input audio.

- Training

The RVC program starts training on a scale of 1 to 500, saves every 50 rounds as previously set, and stores the results of each 50 rounds in the logs folder.Training time will vary depending on the clarity and duration of the audio file. In addition, the RVC does not strictly require 10 minutes for the audio source file, but too long the audio source file may make the training time too long and lead to the final model timbre distortion [4].

After the model training is complete, both the model file and the base model file are output to the logs folder.

3 Result

There is no accepted standard measurement data set for speech conversion. The Speech Conversion Challenge was held once each in 2016 and 2018, and the data sets they used are expected to become the standard [5]. The evaluation methods of speech conversion are divided into objective evaluation and subjective evaluation [6]. We will use both methods for this experiment.

3.1 Objective evaluation

The objective evaluation standard of speech conversion is Mel cepstral distortion (MCD), which represents the difference between the MFCC features of the converted speech and the MFCC features of the standard output speech [7]. To measure MCD, it is necessary to align the MFCC feature sequence of the converted speech with the standard output speech. Suppose the standard output feature of a frame is y, and the feature of the converted speech is ŷ, then the MCD of this frame is defined as:

$$MCD(y, \hat{y}) = \frac{10\sqrt{2}}{ln\,10} \left\| y - \hat{y} \right\|_2 \quad (1)$$

The MCD can be averaged over all the test data, so that the comparison can produce similarity [5]. The MCD unit is the decibels (dB) (1) in front of the coefficient, is to convert the unit into decibels, which is divided by ln10 in order to convert the MFCC itself is a natural value into a common logarithm, MCD can be averaged on all test data. Next, we use Sound Similar Free based on MCD algorithm for similarity comparison to find out the training rounds model that is most similar to the sound source (Fig.3).

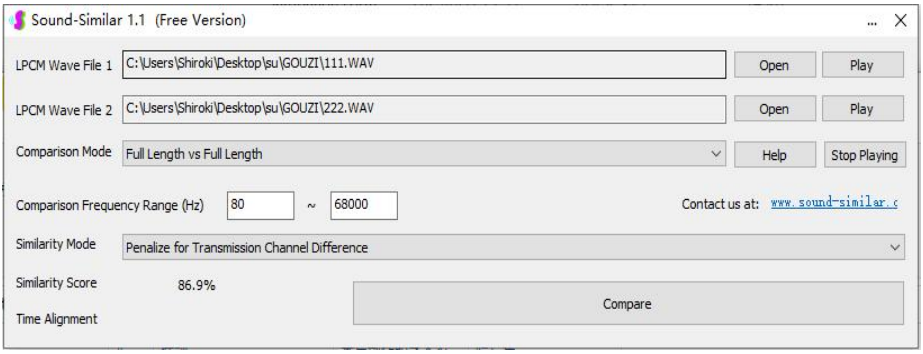

Fig. 3. Single model similarity comparison results (Original)

Table 1. The timbre similarity

Rounds	Similarity of Model 1	Similarity of Model 2	Similarity of Model 2
5000	9.1%	3.4%	6.4%
10000	33.3%	26.3%	29.3%
15000	69%	59%	60%
20000	86.9%	77.9%	88%
25000	81%	88%	87%
30000	76.7%	71.7%	61.7%
35000	68.6%	64.7%	64.8%
40000	58%	51%	50%

Through statistical data, it is found that the results obtained by objective evaluation method show that all models with 20000 and 25000 training rounds are most similar to the original sound source (Table 1). After the number of training rounds reaches 25,000, the audio model training will gradually become distorted. But here is only the software test results, for the voice timbre, the subjective hearing of people is more straightforward than the software test data, so next we will conduct subjective assessment.

3.2 Subjective evaluation

There are two main criteria for subjective evaluation: the sound quality of the converted speech and the similarity to the target speaker. When evaluating a single system, the mean opinion score (MOS) is generally used [8]. For sound quality, a 5-point system is generally used, with 1 being the worst and 5 being the best. For similarity, subjects are often asked to listen to the source speaker's speech, the target speaker's speech (in varying order), and the converted speech, and choose among the following four levels.

- The converted voice is more like the owner of the sound source, and it is very positive;

- The converted speech is more like the owner of the sound source, but it is uncertain;Completely unsure of which speaker the converted speech is more like (there may not be such a rating);
- The converted speech is more like the owner of the sound source, but it is uncertain;
- The converted voice is more like the source owner, and very sure.

When comparing two systems, we can evaluate them separately and then compare the scores; preference tests can also be performed as follows [9]. For sound quality, the subjects are generally asked to listen to the output of two systems and choose which one is better; For similarity, the subject is usually asked to listen to the output of the two systems (in varying order) and the voice of the target speaker, and choose which system's output is more like the target speaker. This latter test is often referred to as ABX test or XAB test, where A and B refer to the output of the two systems, and X refers to the speech of the target speaker [10].

Here, the experiment set up a post on the network, put the audio generated by the three models and the audio of the original sound source in different posts, and launched a vote, so that users can judge whether the two audio is the same person, a total of 103 people participated. The statistical results are as follows:

Table 2. Model 1 Results of similarity network survey for each round

Rounds	affirmative	indeterminacy	repudiate
5000	0	0	103
10000	0	8	95
15000	30	13	60
20000	77	20	6
25000	91	5	7
30000	40	60	3
35000	24	66	19
40000	3	10	90

Table 3. Model 2 Results of similarity network survey for each round

Rounds	affirmative	indeterminacy	repudiate
5000	0	0	103
10000	1	7	95
15000	28	13	62
20000	79	15	9
25000	96	4	3
30000	37	60	6
35000	28	61	14
40000	1	12	90

Table 4. Model 3 Results of similarity network survey for each round

Rounds	affirmative	indeterminacy	repudiate
5000	0	0	103
10000	2	7	94
15000	29	14	60
20000	89	10	4
25000	96	4	3
30000	29	60	14
35000	24	67	18
40000	1	13	89

Judging from the evaluation results, no matter the subjective evaluation results or the objective evaluation results (Table 2-4), the model of 20000 rounds and 25000 rounds has the highest similarity with the original sound source. However, if only from the subjective evaluation results, in addition to the 5,000-round model, all the other round number models in the investigation process have listeners think that there is a similar tone, but the 20,000 round and 25,000 round model are considered by most listeners to be the most similar. In general, the timbre similarity of the model will gradually increase from the beginning of training round 0 until 25000, and the similarity will reach the highest when the model reaches 25000. However, after the model is trained to 25000, due to the overtraining of the model, the timbre will gradually be distorted, and the higher the degree of overtraining, the higher the degree of timbre distortion will be. Therefore, the experiment judged that with the increase of training times, the timbre similarity showed a trend of first increasing and then decreasing.

4 Conclusion

This paper makes a practical application of the RVC program and records the model training process in detail, showing the best value of model training in this program. By using the objective evaluation method and the subjective evaluation method, the optimum number of model rounds is obtained, which provides a relatively objective training reference value for the new program RVC. In addition to the detailed recording of the model training process, this paper also introduces the MCD measurement method in the acoustic feature conversion, and realizes the MCD measurement through the program. This provides some help for other researchers to further study the sound transformation. By measuring MCD, we can assess the gap between the generated speech and the target speech, which can guide the improvement and optimization of the model. With further development of the technology, RVC or similar programs could make the training of AI models unnecessary to manually screen models. This means models can be trained more efficiently, saving time and resources. However, there is also a risk that such timbre imitation procedures could be abused by criminals. To ensure the legal and ethical use of technology, relevant authorities need to establish strict regulatory mechanisms and laws and regulations to prevent potential abuses.

References

1. Qian K, Zhang Y, Chang S, Yang X, Hasegawa-Johnson M. AutoVC: Zero lens voice style transmission, only autoencoder loss. Attended the 36th International Machine Learning Conference. 2019, 10.48550/arXiv.1905.05879.
2. Toda T, Black AW, Tokuda K.Spectral transformation based on maximum likelihood estimation considering the global variance of the transformation parameters. 2005.
3. Nakashika T, Takiguchi T, Minami Y. Speech conversion non-parallel training based on adaptive constrained Boltzmann machine. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2016, 24(11):1-1.
4. Hwang Hsin-Te, Tsao Yu, Wang Hsin-Min, Wang Yih-Ru, Chen Sin-Horng. Global variance analysis of speech conversion training phase based on gmm.Signal and Information Processing Association Annual Summit and Conference (APSIPA). 2013, Asia-Pacific.
5. Toda T, Chen LH, Saito D, Villavicencio F, Yamagishi J. The Voice Conversion Challenge 2016. Interspeech, 2016.
6. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. NIPS, 2014.
7. Lorenzo-Trueba J, Yamagishi J, Toda T, Saito D, Villavicencio F, Kinnunen T, Ling Z. Speech Conversion Challenge 2018: Facilitating the development of parallel and non-parallel approaches, arxiv:1804.04262.
8. Hsu CC, Hwang HT, Wu YC, Tsao Y, Wang HM. Non-parallel corpus speech conversion using variational autoencoders. Signal and Information Processing Association Annual Meeting (APSIPA). 2016.
9. Wu Z, Virtanen T, Kinnunen T, Chng ES. Paradigm-based speech conversion uses non-negative spectral deconvolution. Workshop on Speech Synthesis, 2013,5:201-206.
10. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cyclic consensus against networks. IEEE. 2017, arxiv:1703.10593.