Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 12, No. 3, 2024

Research on the Recognition and Application of Montreal Forced Aligner for Singing Audio

Jinyu Liu*

Guangdong Haige ICreate Technology, Guangzhou 510627, China

* **Corresponding author**

Abstract: This paper discusses the feasibility of obtaining phoneme-aligned time segments for singing audio using the Montreal Forced Aligner (MFA) tool. Initially, the recognition effectiveness of singing audio data is tested using the open-source MFA model. Subsequently, samples with high recognition accuracy are manually annotated and used to train the MFA model. Finally, the recognition effectiveness of the open-source MFA model and the MFA model trained on singing audio data is observed. It is found that the performance of the trained MFA model is significantly improved. This also confirms the feasibility of MFA in recognizing singing audio.

Keywords: MFA; Singing audio; Phoneme alignment; Model training.

1. Introduction

The Montreal Forced Aligner (MFA) is an open-source tool for aligning speech phonemes with text. Its operation involves providing a piece of speech and its corresponding text, after which MFA can identify the time segments corresponding to each word or phoneme. In MFA, we generally assume that the alignment between text and speech segments is one-to-one. This means that if a phoneme appears after another in the text, the same relationship holds true in the speech, and vice versa [1-4]. Currently, MFA is primarily used for everyday conversation audio. However, for singing audio, aligning lyric phonemes with text data is beneficial for evaluating the singer's pitch, rhythm, and other standards [5-6]. Due to differences in pronunciation between singing audio and everyday conversation, as well as differences in frequency ranges during singing compared to normal speech, the open-source models of MFA or those trained specifically for everyday conversation struggle to effectively recognize singing audio. This paper explores the feasibility of obtaining phoneme-aligned time segments for singing audio using MFA. It tests the recognition effectiveness of singing audio data using the MFA open-source model and then trains the MFA model with samples obtained through manual refinement. Finally, it compares the model's effectiveness to study the application of MFA in recognizing singing audio.

2. Methodology

2.1. The Acquisition of Singing Audio Data

A relatively straightforward method to obtain singing audio data is to directly collect audio recordings from students using the solfa singing method during music teaching sessions. Solfa singing is a type of solmization, and syllable names are used to represent the degrees of the scale for ease of singing melody [6]. These singing audio recordings typically use Western traditional solmization, which consists of only seven types, as shown in Table 1. Since the pronunciation of syllable names is derived from Italian pronunciation, the MFA open-source model 'italian_cv' can be used for testing.

Table. 1 Comparison of Pitch Name, Syllable Names, and Scale Degree

Pitch Name	C	D	E	F	G	A	B
Syllable Names	do	re	mi	fa	sol	la	si
Scale Degree	I	II	III	IV	V	VI	VII

2.2. Testing the Open-Source MFA Model

Preprocessing is conducted on singing audio data using the solfa singing method, including extracting lyrics from the staff notation, removing any piano sounds or noise at the beginning and end of the singing audio, and manually filtering out high-quality singing audio. The preprocessed singing audio is then tested using the MFA open-source model. As shown in the Figure 1, the recognition accuracy of samples with high accuracy is relatively low, only 10.08%. Therefore, the recognition effectiveness of the MFA open-source model for singing audio is not ideal.

2.3. Training of MFA Model Based on Singing Audio and Comparative Analysis of Effects

The samples with recognition accuracy greater than or equal to 80% are manually calibrated, with data type being TextGrid. Using Praat, one of the TextGrid samples is visualized as shown in Figure 2. The calibrated samples are then fed into the MFA tool for training. Samples with recognition results falling between 60% and 80% are selected for comparison. Figure 3 illustrates the comparison of power spectrogram with recognition results from different MFA

models. The green line represents the recognition results of the trained model, while the blue line represents those of the open-source model. It can be observed through auditory inspection of audio and comparison of power spectrograms that the trained model achieves basic accuracy in aligning audio segments, indicating a significant improvement in model performance before and after training.

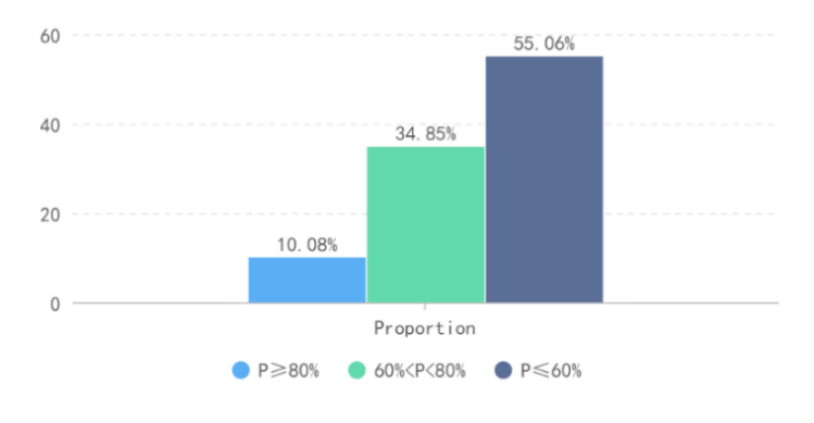

Figure. 1 Bar chart of recognition accuracy for the open-source MFA model

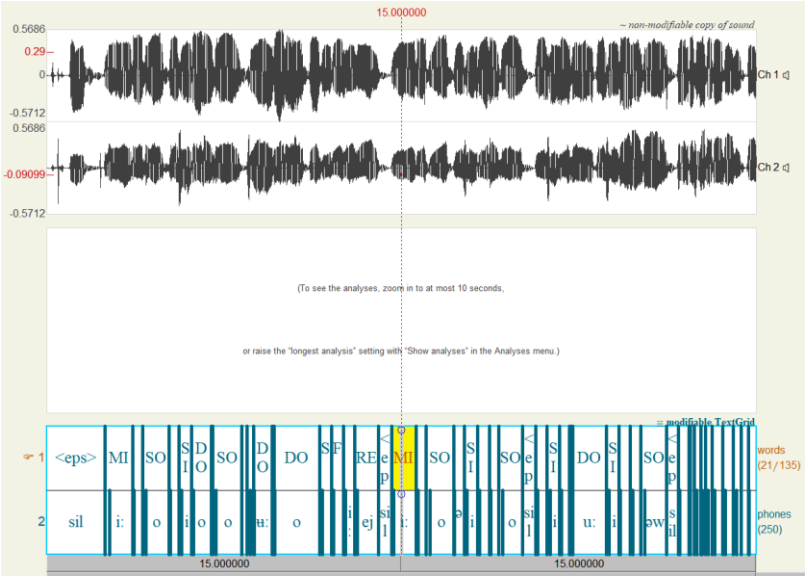

Figure. 2 Visualizing the singing audio and its corresponding TextGrid using Praat

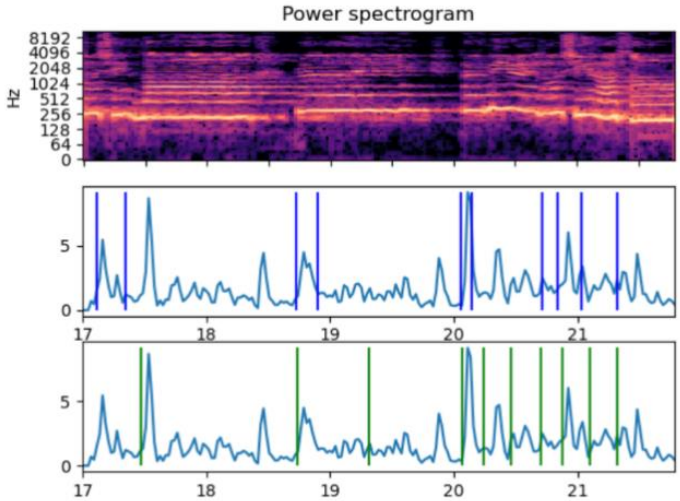

Figure. 3 The comparison of power spectrogram with the recognition results of the MFA pre-trained model (middle graph) and the MFA trained model (bottom graph).

3. Results and discussion

Based on the experiments described above, it can be observed that continuously training the model through this iterative process theoretically improves the model's

effectiveness over time. In fact, the main issue addressed by this cyclic process is the problem of sample labeling. Labeling and generating TextGrid files for music samples are inherently challenging tasks. After completing this iterative process, we only need to label samples with P≥80%, theoretically reducing the workload by nearly 80%.

All the steps and methods involved are summarized into the process shown in Figure 4, which enables the acquisition of an MFA model with higher recognition accuracy. Here, p_0 and p_1 can be threshold values set according to actual conditions.

Although the efficiency of labeling and training has significantly improved, there are still some labeling difficulties in this process. This is because even though the model can achieve high labeling accuracy for individual samples, manual labeling is still required, and this labeling accuracy needs to be accurate to the millisecond level. Therefore, despite the use of certain tools, there is still some margin for error. However, it is fortunate that the recognition results allow for errors greater than the millisecond level in evaluating singing standards such as pitch and rhythm.

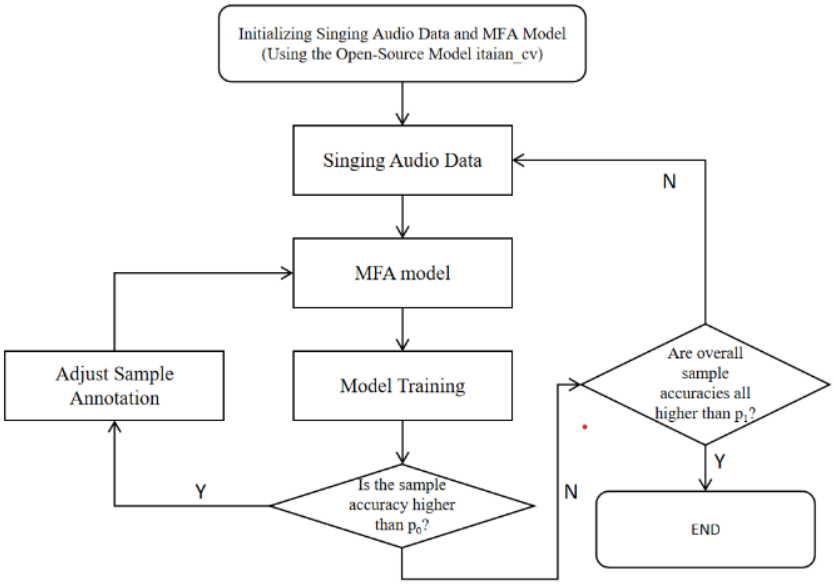

Figure. 4 Flowchart of MFA Model Training Process

4. Conclusion

Although the technology in music education evaluation is relatively niche, with the continuous development and popularization of technology, as well as the increasing emphasis on the quality and effectiveness of music education, more and more research is beginning to focus on how to improve the efficiency and accuracy of music education evaluation using advanced technology. The use of MFA to study phoneme recognition in singing audio aims to introduce mature speech phoneme recognition technology into music phoneme recognition. From the conclusion, the effect is significant. In the future, there will undoubtedly be more mature technologies introduced, which will bring better development and improvement to music education evaluation.

References

[1] McAuliffe, Michael, et al. "Montreal forced aligner: Trainable text-speech alignment using kaldi." Interspeech. Vol. 2017. 2017.\

[2] Kelley, Matthew C., and Benjamin V. Tucker. "A comparison of input types to a deep neural network-based forced aligner." (2018).

[3] Tan, Xu, et al. "A survey on neural speech synthesis." arXiv preprint arXiv:2106.15561 (2021)

[4] Rahmatullah, Griffani Megiyanto, and Shanq-Jang Ruan. "Performance Evaluation of Indonesian Language Forced Alignment Using Montreal Forced Aligner." 2023 Sixth International Symposium on Computer, Consumer and Control (IS3C). IEEE, 2023.

[5] Gupta, Chitralekha, Haizhou Li, and Ye Wang. "Perceptual evaluation of singing quality." 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2017.

[6] Sisman, Berrak, et al. "An overview of voice conversion and its challenges: From statistical modeling to deep learning." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2020): 132-157.

[7] Majaro-Majesty, Henry. "TEACHING ADULTS TO READ: INTRODUCING THE TONIC SOL-FA METHOD." Edukacja Dorosłych 2 (2017): 241-253.

Styler, Will. "Using Praat for linguistic research." University of Colorado at Boulder Phonetics Lab (2013).