Proceedings of the 2017 IEEE
International Conference on Information and Automation (ICIA)
Macau SAR, China, July 2017

Audio Similarity Comparison System of English Dubbing on Android Platform

Yuqing Ye, and Shaoran Zhang

School of Automation & School of Humanities

Beijing University of Posts and Telecommunications

Haidian, Beijing, China

Sunnyyyq@bupt.edu.cn & vans364566665@126.com

Weijun Hong, Xiaolu Wang, Danting Huang and Bocheng Chen

School of Computer & ScienceSchool of Automation

Beijing University of Posts and Telecommunications

Haidian, Beijing, China

18810560683@163.com

Abstract - This paper studies an audio similarity comparison system on Android platform to judge the similarity of dubbing audio in some contexts. The system mainly uses the image similarity comparison method and the MFCC algorithm as its algorithm basis. Considering the limitation of mobile platform, the current mobile audio comparison methods are studied and improved, to reduce operation time and improve accuracy of the comparison.

Index Terms - Android, Audio Similarity Comparison, Situational English, Evaluation Mechanism

I. INTRODUCTION

With the increasing demand for smart mobile applications, more mobile developers start doing multimedia research and have developed a lot of innovative features. Android platform is very popular now and easy to develop. Thousands of applications based on the Android platform not only facilitate people's lives, but also brings more new technologies to mobile terminal development. Therefore, how to make audio comparison more convenient, quick and correct on mobile terminals has become one of the research directions for some Android developers. Researching and upgrading audio contrast technology based on Android platform, and applying it on fields such as language learning, artificial intelligence identification, will have some development assistance. This paper, combined with audio comparison technology, mainly studies an audio similarity comparison system on Android platform.

At present, current audio comparison technology on Android platform has been able to provide a certain degree of audio similarity comparison functions. It is by calculating the audio signal’s short-term energy, and calculating the cosine distances of original audio between every period and compare audio to get similar conclusions. So there will appear some problems like low accuracy, out of memory, greatly varied score with the volume level or not intuitive score and so on.

Take "English fun voice" APP, which uses the most mature domestic audio contrast technique, as an example. The project scores audio by clauses. It mainly compares different short-time energy of audio, and obtains the average score of these clauses. In this way, the disadvantages of other existing APP projects based on pitch scoring are overcame, and the accuracy of scores is improved. But because “English fun voice” only considers the short-term energy of the audio, it is affected by the external noise and different personnel's own conditions, and the score evaluation is often in an unstable state. This problem is common in the current mobile audio contrast technology.

In order to overcome the above defects, adapt to the development environment of Android, make the score stable, and reduce interference of the noise or different vocal conditions, we have improved the audio based on the original technology and completed an audio similarity comparison system on Android platform.

Our audio similarity comparison system mainly includes three parts, which are audio pretreatment, audio similarity comparison and similarity scoring. The audio pretreatment part is used as the information acquisition part of the whole system to ensure the accuracy of the result. The audio similarity comparison part obtains five characteristic parameters of audio and gets equalized aspects of the

978-1-5386-3154-6/17/$31.00 ©2017 IEEE

frequency domain audio similarity. By using the method of time domain waveform comparison, the time domain waveform is adjusted to reduce the influence from the whole volume change on the similarity degree score. And considering different users' different focuses on oral practice, we have added similarity scoring part, which performs weighting on three aspects of audio similarity, then converts difference to an intuitive score. The Fig.1 shows system architecture.

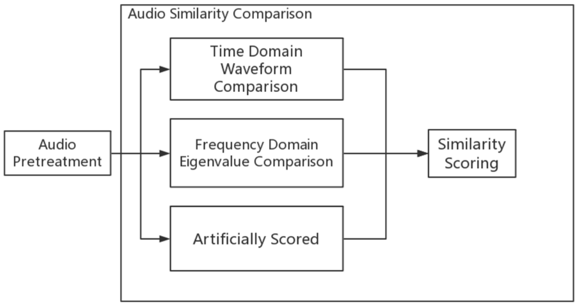

Audio Similarity Comparison
Time Domain Waveform Comparison
Frequency Domain Eigenvalue Comparison
Artificially Scored
Audio Pretreatment
Similarity Scoring

Fig. 1 System architecture

II. Basic Parameters of The Audio Feature

In a certain context, the audio features of English conversations can be broadly divided into two aspects: different tone in different contexts and different pronunciation features due to different English levels of dubbing. As different conditions can cause different emotions, users will show different volume, pitch and pause intervals. And different pronunciation features will lead to slur words, non-standard pronunciation and other issues [1]. The volume indicates amplitude of sound, the tone indicates the change in the frequency of sound vibration, and the pause interval indicates the time of very low volume. Slured sound will cause incomplete waveform, while non-standard pronunciation will lead to changes in the overall pronunciation content. The Fig.2 is the audio feature chart:

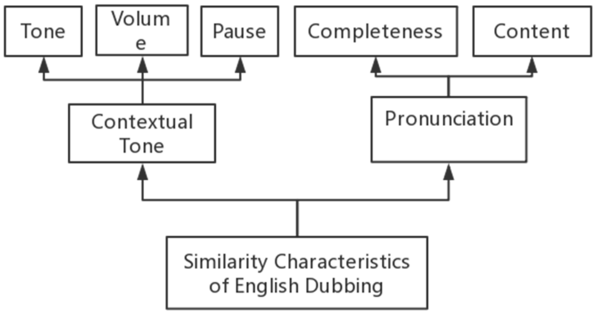


Fig. 2 Audio feature

III. Module Model

A. Audio Pretreatment Module

The function of this module is to achieve the audio acquisition and pretreatment. Fig.3 represents the flow of the module.

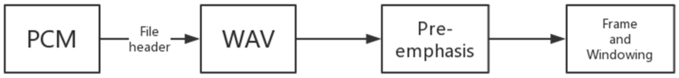


Fig. 3 Module's flow diagram

The voice of a person is emitted to the channel due to vocal cords. In this process, lips and other parts cause a radiation effect, and the circular head causes a diffraction effect. Studies have shown that lips radiation in the high frequency band is more obvious, less affected by the low frequency band. Therefore, in order to remove the effect of lips radiation and increase the high frequency resolution of speech, we need to pre-emphasis the recording file. That is, adding a high-pass filter, so that we cab reduce the impact of glottal pulse to the minimum, leaving only the channel part, making it easy for subsequent analysis [10].

FIR high pass filter can guarantee arbitrary amplitude frequency characteristics while having strict linear phase frequency characteristic.

So our module uses the first-order FIR high-pass filter to achieve pre-emphasis, the transfer function is:

$$H(z)=1-0.98z^{-1}$$

Due to Gibbs effect, the signal needs to be windowed. From Fig.4 we can see rectangular window will produce frequency leakage, so this module uses Hamming window to make the original audio show some characteristics of the periodic signal. The hamming window's expression is

$$\mathrm{w}(n)=0.54-0.46\cos(2\pi\frac{n}{N}),0\le n\le N$$

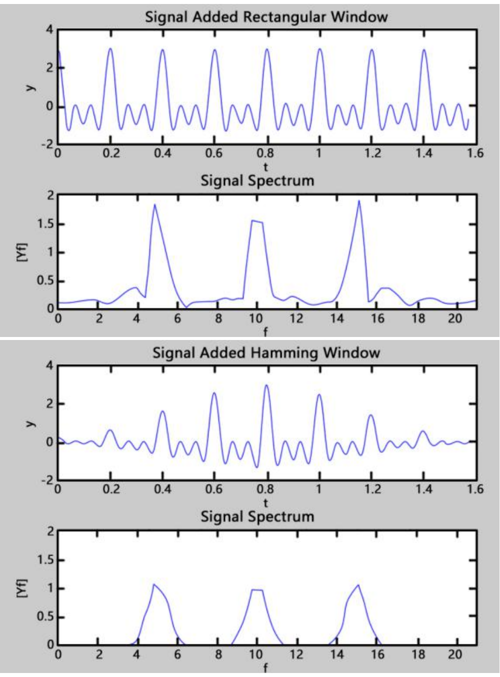


Fig. 4 Spectral changes of rectangular window and Hamming window

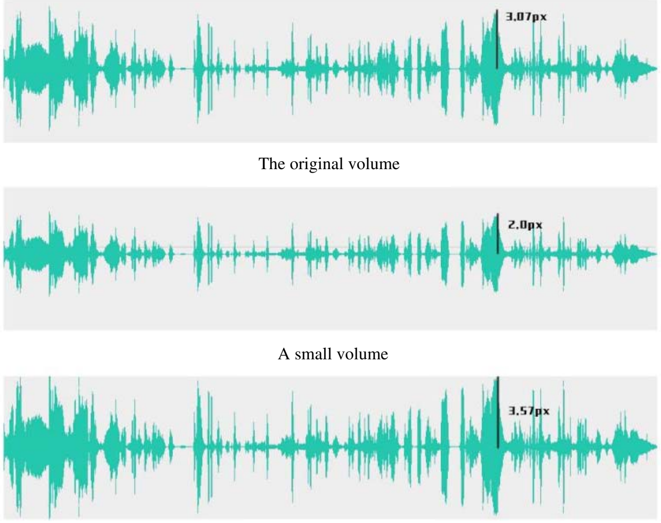
B. *Time Domain Waveform Comparison Module*

In order to avoid the influences from different distances between the user and the microphone, or different microphone radio performances which lead to different volume and thus affecting similarity of the results, we refer to the image similarity comparison method and extend it to the audio comparison direction. This module can reduce the influence of the whole volume on the similarity results by drawing time domain waveform.

Firstly, draw the dubbing audio's waveform pattern. Then analyse the waveform and obtain high probability waveform value. Compare it with the the original audio's high probability waveform value to get the adjustment ratio. Use this ratio to adjust the dubbing waveform and redraw the waveform.

Next, calculate the distance between the sampling points in the recorded audio waveform and the original sound waveform. Then we derive the mean square of the distance error. When put it into a fractionalization, we get Score1.

In the same environment, play a same audio source with different volume and draw the waveform. The three waveform samples are essentially consistent (excluding noise) apart from the overall volume change. Fig.5 shows waveform of a large volume, a small volume, and the original volume. Process them by the method above, and compare them with the original volume waveform similarity comparison.



The original volume

A small volume

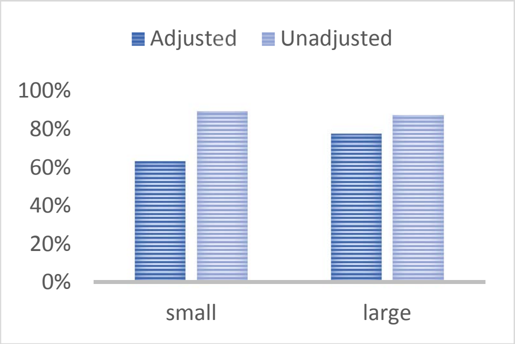
A large volume

Fig. 5 Waveform of different volumes

The waveform similarity changes are obtained as Table I:

TABLE II

PARAMETERS OF MODES



It can be seen if we only consider the volume level, the score obtained by this module is basically unchanged, while the other one has an obvious change. So the results are accurate. This part will compensate for the error caused by the audio eigenvalues comparison module, and reduce errors due to overall volume changes.

C. *Audio Eigenvalues Comparison Module*

In order to compare the similarity of two audios, it is necessary to get some characteristic values of the audio signal. In the dubbing environment, audio is susceptible to external factors and the voice of the factors of their own.

The human cochlea is actually equivalent to a filter bank, for which below 1000Hz is a linear scale, and more than 1000Hz is a logarithmic scale, which makes the human ears more sensitive to low-frequency signal than high-frequency signal. Therefore, for the listener, the audio listened and the actual audio are different.

This system uses the method of short-term energy and Mel cepstral to obtain the audio eigenvalues needed to compare the similarities. In order to adapt to the Android platform environment and limit phone memory, we use Euclidean distances to calculate similarities. The specific process is as Fig.6:

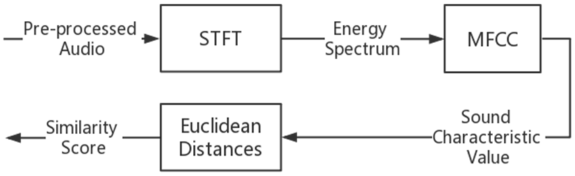


Fig. 6 Specific process

STFT can get the frequency domain information on each short period, so we can obtain the spectrum of the audio. The STFT formula is:

$$X_{\mathrm{n}}(e^{jw}) = \sum_{m=-\infty}^{\infty} x(m) \bullet w(n-m) \bullet e^{-jw}$$

Mel-Frequency Cepstral is a voice feature commonly used in speech recognition. The main principle is to calculate the MFCC parameters by combining the auditory perception characteristics of the human ears with the generation mechanism of the speech. So Mel-Frequency Cepstral can be used to simulate different sensitivity of human ears to different frequencies of audio signal, which weaken the impact of external factors. The Mel frequency's formula is as follows:

$$\mathrm{Mel\text{-}frequency} = 2595 \times \log(1 + \frac{\mathrm{F}}{700})$$

(F stands for common frequency)

Here, we use the following eigenvalues to match the audio characteristics that are needed to be compared, the calculation method is stated as Table □:

TABLE □

THE CALCULATION METHOD OF CHARACTERISTICS

Characteristic	Eigenvalues	Calculation
Pitch	Frequency Domain Energy	$E = \log(\int_0^{w_0} \lvert X(w)^2 \rvert dw)$
Volume	Short - term Energy's Square Error	$RMS = \sqrt{\frac{1}{N} \sum_{\mathrm{n}=0}^{N-1} \lvert X_i(n) \rvert^2}$
Pause and Pronunciation	MFCC	$C_n = \sum_{K=1}^{M} \log x(k) \cos[\pi(k+0.5)\frac{n}{M}],$ $n = 1,2,\cdots,L$
Pronunciation Completeness	The Number of Completed Statements	Using the pause value to get statements' number ,compare with the original one

After obtaining the above eigenvalues, the audio comparison is started. The similarity score is obtained by using the Euclidean distance between the recorded audio and the original audio. N-dimensional parameter of the standardized Euclidean distance formula is [2]:

$$Dis = \sqrt{\sum_{k=1}^{n} \left(\frac{x_{1k} - x_{2k}}{s_k} \right)}$$

There are four Euclidean distance groups [dist[1], dist[2], dist[3], dist[4]]. Use these four values to get the final score:

$$Score2 = W1 \times Dis[1] + W2 \times Dis[2] + W3 \times Dis[3]$$

W1, W2, W3, W4 are the weights of each Euclidean distance. The proportion is derived by analytic hierarchy process(AHP) [3]. According to specific situation of oral English judgment, when the degree of importance is pause and pronunciation content > pronunciation complete > pitch > volume, we can get the weight: W1 = 0.136, W2 = 0.063, W3 = 0.509, W4 = 0.291.

To test the function above, we select a very similar dubbing audio, adjust its female voice to a male voice, and score only by pitch, volume, pause, pronunciation content, and pronunciation complete respectively, then compare the unilateral score with the full score related to all these five aspects.

The score is shown as Table □:

TABLE □

THE SCORES IN THIS TEST

D. *Artificial Score*

Different learners have different degrees of inclination on pronunciation, pause or some other aspects. Therefore, in order to achieve the subjective scope of the judge, we added artificial scoring module. With this function, learners can score according to their own ideas, and the system will offer a comprehensive score:

$$Score3 = (part[1] + part[2] + part[3])$$

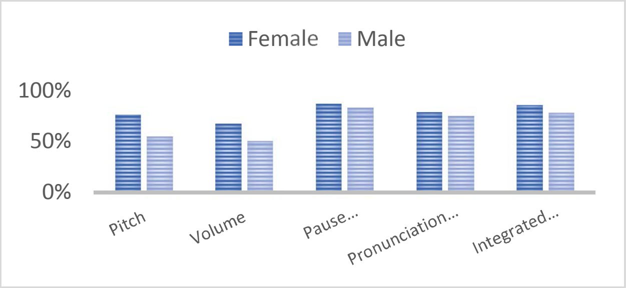
(part[1] denotes the score of emotion, part[2] denotes the score of pronunciation, part[3] denotes the score of pause.)

E. *Final Score*

From the above scores [score [1], score [2], score [3]], we can calculate a comprehensive score:

$$Score = m1 \times Score1 + m2 \times Score2 + m3 \times Score3$$

m1, m2, m3 are: 0.260, 0.633, 0.106, respectively. It is derived by analytic hierarchy process (AHP), and the degree of importance is: Score 1> Score 2> Score 3.



IV. System Test

Test environment is as Table IV:

TABLE IV

Test Environment And Configuration

Phone Model	Huawei mate S
Operating System	Android 6.0
CPU	Haisi Kirin 935 (8 cores) 2.2GHz

After testing, system can run normally.

Select English video clips with different characteristics to dub, and use this system to get similarity score. (Among them, the "The Croods" fragment has obvious changes in pitch, the "Zootopia" fragment has obvious volume changes, the "Obama speech" fragment has more pauses and more changes in pronunciation content, and the "Frozen" fragment has more change in pronunciation completeness.) At the same time, we randomly selected 10 test subjects for artificial scores. Finally, by comparing its average value with system score, the results are as Table V:

TABLE V

Results In This Test

Resources	System Score	Artificial Score Average Value	10 Scores Given by the Tester
The Croods	45	37.6	[23,56,43,32,60,34,36,27,31,34]
Zootopia	57	60.1	[74,58,71,65,34,68,70,52,49,60]
Obama's Speech	77	72.2	[79,90,56,78,66,69,78,90,68,59]
Frozen	60	56.6	[67,58,67,77,34,45,56,58,71,33]

Thus, the standard deviation of this result is 1.699 points, and the average value is 4.675 points, which reveals the error is small. Fig.7 shows the results.

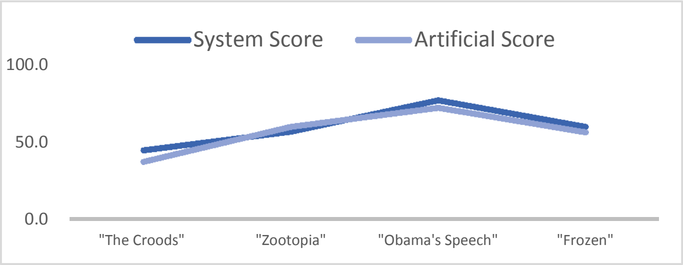


Fig. 7 Comparison Between System scores and Artificial Scores

V. Summary

A. *Background Music and Projection in Scene modes*

This paper analyzes the method of audio similarity comparison on mobile terminal, discusses and studies how to design and implement a more accurate and efficient system, improves it on the basis of existing technology, proposes a comprehensive scoring model based on time domain waveform comparison, frequency domain eigenvalue comparison and artificial score, then designs an audio similarity comparison system of English dubbing on Android platform and encodes to achieve an available application.

Compared with the existing Android based audio similarity comparison system, our system's score is more close to manual scoring, and more practical. It is easy to be disturbed by external noise, and can automatically screen the errors caused by different users' voice characteristics. Our system only focuses on whether the voice content is correct, and whether the pronunciation and intonation of the dubbing is standard. It can basically meet the learning needs of English teaching.

B. Issues and Prospects

As shown in the test results, similarity scores given by system are slightly different from those given by the testers, but the error is not significant. The reason for this problem on the one hand is that there are certain subjective factors when the testers score, because of the insufficient number of testers, the fractional float is large; on the other hand, the weight of the system is still open to question.But overall, the results of this study are basically in line with expectations. How to make improvements in practical application, make the situation more in line with the subjective score, is the focus of the next step.

ACKNOWLEDGE

Here, sincerely thanks for Research Innovation Fund for College Students of Beijing University of Posts and Telecommunications. Our university supports us in technology and fund.

And we would like to thanks for our guidance counsellor, Mr. Shaoran Zhang and Mr. Weijun Hong, who has provided valuable guidance in constructing our system and writing this paper.

REFERENCES

[1] TENG Haikun, LIU Xinsheng, WANG Lihong,“The Research of English Pronunciation Evaluation System,” *Journal of Yancheng Institute Of Technology,* vol. 29, No.1, pp 18-19, Mar.216.

[2] Tu Huiyan, Chen Yining, "English Learning System of Oral Phonation Based on ASR and Smart Phone Platform,” *Computer Applications and Software*, vol. 28, No. 09, pp. 65-66, Jun. 2007.

[3] YANG Da-Li, XU Ming-Xing, WU Wenhu, “Study of Feature Selection for Speech Recognition,” *Journal of Computer Research and Development*,” vol. 40, No.07, pp. 968, Jul. 2003.

[4] ZHANG Wan-ii, LIU Qiao, “Mel-freguency Cepstral Coefficients Extraction and Its Application on Voiceprint Recognition,” *JournaI of Guizhou University（NaturaI Sciences）*,Vol. 22, No. 02, pp. 208-209, May. 2005.

[5] XU Guang-lie, “Comparison and Contrast of Speech Sounds in Teaching Overseas Students Chinese,” Soft Limit Switch’s Establishment Principle and Application,” *Joumal of Guangzhou University(Social Science Edition)*, Vol.5, No. 08, pp. 88-89, Aug. 2006.

[6] LIU Zhen-an, LUO Yong-zhao, “Research of Speech Assessment Based on Reature Comparison,” *Computer Appiications*, vol. 25, No. 12, pp. 2929-3005, Dec. 2005.

[7] ZHONG Chen, XIAO Nanfeng, “Design and Implementation of an Automatic Recognition Robot for Fingerprints and Voices,” *Computing Technology and Automation*, vol. 25, No. 02, pp. 114-116, Jun. 2006.

[8] Yue Dongjian, Ji Hongfei, “The Application of Speech Technology in Language Learning,” *Computer Engineering and Applications*, vol. 25, No. 02, pp. 1-3, Feb. 2000.

[9] HE Dongzhi, HUANG Zhangqin, HOU Yibin, DiNG Zhihao, “Research on Key Technologies of Front-end and Back-end for Embedded Automatic Speech Recognition,” *Computer Simulation*, vol. 27, No. 02, pp. 192-195, Feb. 2010.

[10]XU Qunfeng, “The principle and protection of vocal cord phonation,” *Biology Bulletin*, vol. 11, No. 01, pp. 6-8, Nov. 1986.