MATEC Web of Conferences **22**, 02017 (2015)
DOI: 10.1051/matecconf/20152202017

A Study on English Audio Segmentation Methods Based on Threshold Value and Energy Sequence

Kun Li, Yan Zhang, Jing Li & Wei Dong
Qinhuangdao Institute of Technology, Qinhuangdao, Hebei, China

ABSTRACT: As a new method of learning, the mobile platform is playing an increasingly important role in English learning. In order to fit in with the construction requirements of the English mobile learning platform, it is obliged to carry out methods of positive studies on English audio segmentation. In this study, the author makes a brief analysis on English audio segmentation based on threshold value and energy sequence, provides the basic process of English audio segmentation, and carries out a discussion on key technical points of the scheme, which including windowing process, audio feature extraction and dual threshold. In the end, the author makes a research on the operation of English audio segmentation, aiming at drawing more attentions to this study.

Keywords: English audio; segmentation; threshold value; energy sequence; method

1 INTRODUCTION

With the increasing development of wireless communication technology and the gradually mature mobile information platform, the mobile learning, as a new method of learning, has made a long-term progress in various industries and fields. Nowadays, mobile learning has become one of popular topics of research about distance education and digital learning system. It is now playing an indispensable role with values in the reform of traditional education system and learning mode due to its interactivity. Making discussions on the organic combination of a discipline and mobile-learning mode is one of the problems that should be attached great importance to, especially for English, a discipline with a large scale of learning resource and high pertinence. As one of the key parts of English mobile learning, the English audio is of great significance. Meanwhile, in order to meet the construction requirements of the multimedia base of mobile platform, the accurate segmentations of English audio should be also attached great importance to it. This paper carries out an analysis on this point.

2 BASIC IDEAS

It can be found from previous statistics analysis on a large number of samples that phonetic characteristics have obvious changes in segmentation positions in terms of different semantic units. For example, audio energy characteristics can present a significant downtrend at the boundary of two sentences; therefore, boundary detection can be conducted through applications of the energy characteristics so as to realize reasonable division of the boundary of two sentences. For instance, the MPEG-1/Audio Layer3 corresponding frame header formats are shown in the following table (See Table 1).

With this feature, the paper adopts the following technical approach in the study on English audio segmentation methods: First, predict the boundary of semantic unit through the analysis on the variation trend of audio energy sequence features; second, introduce characteristics of time delay and calculate the corresponding threshold value of the audio sample based on the analysis of a specified audio (which including mute-energy threshold value and mute-delay threshold value); and at last, carry out a final detection on the boundary of the audio sample through the secondary judgment on this basis. It is undoubted that the audio data should be segmented into frames by windowing in the process of audio segmentation based on threshold value and energy sequence. Then, a series of time domain features are extracted according to frame unit so as to construct a complete energy sequence. The operation flow chart of English audio segmentation based on threshold value and energy sequence is shown in the following figure (See Figure 1).

It can be seen from Figure 1 that, as for an energy sequence, the counter will carry out "+1" processing if the corresponding energy of each frame satisfies the mute-energy threshold value. And the counter will carry out "0" processing if the corresponding energy of each frame satisfies the mute-delay threshold value. In the process of the above-mentioned loop operation, the frame unit processed with "+1" can be incorporated into the sequence of boundary detection point when the counter satisfies the requirement of mute-delay threshold value so as to provide a necessary support and reference for audio segmentation.

3 KEY TECHNOLOGIES

3.1 *Windowing technology*

Table 1. Frame header formats of MPEG-1/Audio Layer3

Marks	Meanings	Length (unit: bit)
Synchronization word FFF	Marking the beginning of one frame of MP3	
Markers	1: MPEG-1 code stream 0: MPEG-2 extended code stream of low sampling rate	1
Layer Information	11: layer I 10: layer II 01: layer III 00: reservation	2
Error detection bit	Indicating whether error detection information is added in audio bit stream 1: without CRC 0: with CRC	1
Bit rate index	Specifying different bit rates Different indexes in different layers	4
Sampling frequency	00: 44.1kHz 01: 48kHz 10: 32kHz 11: reservation	2
Padding bit	1: an additional slot to adjust the central bit rate to the sampling rate 0: no containing	1
Private bit	Bit for personal use	1
Modes	00: stereophonic sound 01: joint stereophonic sound 10: dual track 11: single track	2
Mode expansion	Marking which kind of joint stereophonic sound mode is used	2
Copyright	0: without copyright 1: copyright protection	1
Original/copy	0: copy 1: original stream	1
Accentuating mode	00: accentuation 10: reservation 01: 50/15ms accentuation 11: CCIT J.17 accentuation	2

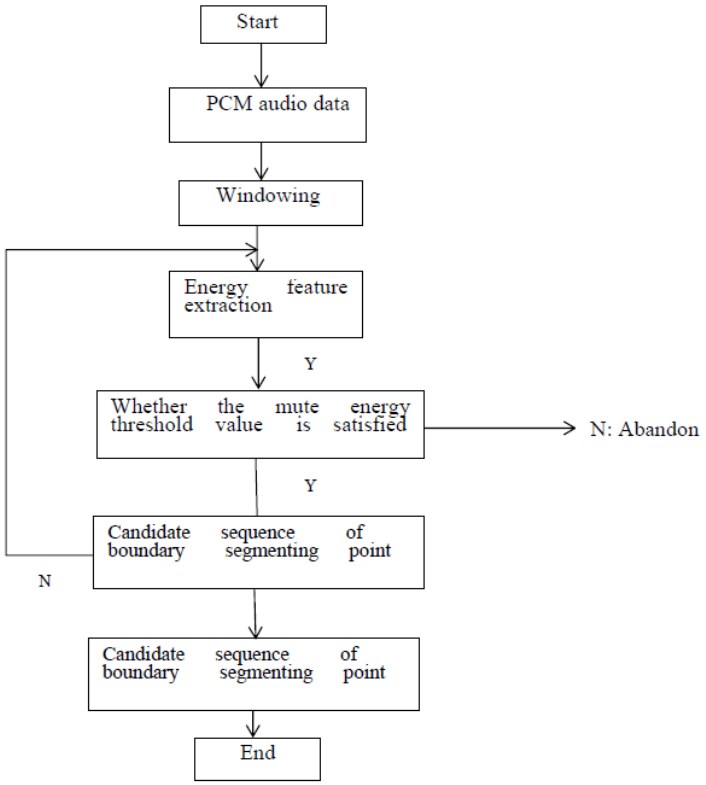

Figure 1. Operation flow chart of English audio segmentation based on threshold value and energy sequence

The sound of generation mechanism is quite special, which determines the non-stationary features of voice signals. However, the speed of voice heard by people is slower than that of the sound of vibration velocity itself. So, it is necessary to take "the short-term stationary feature of voice signals" as the assumed fundamental condition in the process of voice signals. Based on this assumed condition, features of voice signals will not change obviously with time. It can be basically seen as the voice signals have features of being stationary in a time frame unit. Therefore, the analysis on voice signals is actually a process about stationary signals. The process of studying English audio segmentation with threshold value and energy sequence including indicators such as average range, short-term average energy and average zero-crossing rate, those are generated on the premise of the short-term stationary features and separated from the time domain in the short-term stationary state.

As for a window sequence with limited length [w (m)], the window must be maintained in a sliding condition in the process of analyzing intercepted signals, thus, it can achieve the analysis on nearby signals at any time. The basic principle of short-term analysis is that the state can be expressed with the following formula:

$$Q_n = \sum_{m=-\infty}^{\infty} T\left[x(m)\right] \bullet w(n-m) \tag{1}$$

Where, x (m) can be defined as an input signal sequence, and $\bullet$ can be defined as dot product calculation.

3.2 *Audio feature extraction technology*

An audio can be segmented in line with lengths from 10.0ms to 30.0ms through windowing. Each independent segment can be called as an audio frame. Partial overlapping is permitted between two adjacent audio frames. Relevant features including time domain, auditory sense, frequency domain and cepstrum can be extracted when the influence of the short-term stationary principle on audio feature extraction is taken into account. The audio feature concept utilized in this paper mainly includes two aspects, namely short-term average energy and short-term average zero-crossing rate. Details are presented as follows:

First, from the perspective of short-term average energy, an energy sequence function is used to describe the variation of audio energy range so as to accurately classify voiced consonants and silent (non-silent) voices. As for voices of the i^{th} frame, the corresponding short-term average energy can be calculated in the following method:

$$e(i) = \frac{1}{N} \log x_i^2(n) \tag{2}$$

Where i is defined as the number of the corresponding ratio of an audio frame, N is defined as the number of samples in an audio frame, and $X_i^2(n)$ is defined as the sample value of the n^{th} point in the i^{th} frame.

Second, from the angle of short-term average zero-crossing rate, there would be a "zero crossing" when two adjacent sample values of a discrete signal have different signs. Zero-crossing rate refers to the time of a signal passing through a zero level in short term, which is used to measure the frequency of zero crossing. It is easy to distinguish the voiced sound from the unvoiced sound and voice from voiceless by taking advantage of zero-crossing rate. The calculation of short-term average zero-crossing rate can be realized in the following method:

$$Z_I(i) = \frac{1}{N} \sum_{N=0}^{N-1} \frac{1}{2} \left[1 - Sgn\left(S_i(n) - S_i(n-1)\right)\right] \tag{3}$$

Where Sgn is defined as a sign function. The other values are the same as those in the previous formula.

But the zero-crossing rate defined in this way can be easily interfered by low frequency signals. Thus, the time of modifying signal waveforms crossing the zero level into signal waveforms crossing positive and negative thresholds is generally defined as the zero-crossing rate. Zero-crossing rate defined in this way has certain anti-interference ability.

3.3 *Dual threshold technology*

We analyze the mute segment between different semantic units in each kind of audio manually, especially between two sentences, provide the statistics of average segment length and the shortest segment length, and then obtain the mute-delay threshold value through certain strategies. For example, we multiply the average segment length by a coefficient which is smaller than 1 or directly utilize the shortest segment length. If the added window is not overlapped, it only needs to select segment length by the window length, which is namely the mute-delay threshold value.

4 CASE STUDY

In order to further verify the practice effect of English audio segmentation methods based on threshold value and energy sequence, the VOA Special English is taken as an example in this study. The 28 audio samples of VOA Special English during a certain period are selected as analytic targets. The duration in each audio is about 4.0 minutes. The content covers education, news, agriculture, science and technology, economy, development and so on. There are altogether 5 announcers, including 3 male announcers and 2 female announcers.

In the process of verifying the practice effect of English audio segmentation methods based on threshold value and energy sequence, the results of text segmentation are used as contrastive methods to carry out evaluation and comparison on recall rate and precision ratio of English audio segmentation methods

based on threshold value and energy sequence.

Table 2. Comparative results of the two different methods

Methods of sentence segmentation	The number of sentences after segmentation	The practical number of sentences	Recall rate (%)	Precision rate (%)
Text segmentation	830	829	100.00	99.88
Segmentation based on threshold value and energy sequence	840	829	96.74	95.48

The computing method of recall rate is to detect the correct number of segmenting points/the practical number of segmenting points;

The computing method of precision ratio is to detect the correct number of segmenting points/the number of all segmenting points;

The comparative results of two methods are provided in the following table (Table 2). It can be concluded from data in Table 2 that the effect of text segmentation is satisfactory. It guarantees that audios can be provided with accurately contrastive characters and provide the correction based on this with reliable basis. Meanwhile, the comparative data also verifies that this method has an ideal application value in English audio segmentation, which deserves further studies.

5 CONCLUSION

This paper proposes an English audio segmentation method based on threshold value and energy sequence. The basic idea of the method is to predict the boundary of semantic unit through the analysis on the variation trend of audio energy sequence features, introduce characteristics of time-delay and calculate the corresponding threshold value of the audio sample based on the analysis of a specified audio (including mute-energy threshold value and mute-delay threshold value), and carry out a final detection on the boundary of the audio sample through the secondary judgment on this basis. It is undoubted that the audio data should be segmented into frames by windowing in the process of audio segmentation based on threshold value and energy sequence. A series of time domain features are then extracted according to frame unit so as to construct a complete energy sequence. Through the case study and the comparison of text segmentation results, it verifies that this method has an ideal application value in English audio segmentation and deserves further studies.

REFERENCES

[1] Liu, P. & Wang, Z.Y. 2005. Multi-mode voice endpoint detection, *Journal of Tsinghua University (Science Edition)*, 45(7): 896-899.

[2] Chen, J.Y., Li, Y.H., Wu, L.D., et al. 2004. Automatic audio classification and segmentation assisting segmentation of soccer video, *Journal of National University of Defense Technology*, 26(6): 49-53.

[3] Lu, G.Y., Jiang, D.M., Fan, Y.Y., et al. 2009. Audio/video speech recognition and phoneme segmentation based on multi-streaming three phonemes DBN, *Journal of Electronics and Information*, 31(2): 297-301.

[4] Lu, G.Y., Jiang, D.M., Zhang, Y.N., et al. 2008. A study on continuous speech recognition and phoneme segmentation of large vocabulary based on the dynamic Bayesian network, *Journal of Northwestern Polytechnical University*, 26(2): 173-178.

[5] Lu, G.Y., Jiang, D.M., Jiang, X.Y., et al. 2007. Continuous speech recognition and phoneme segmentation of audio and video based on the dynamic Bayesian network, *Computer Application*, 27(7): 1670-1673.

[6] Izadinia, H., Saleemi, I., Shah, M., et al. 2003. Multimodal analysis for identification and segmentation of moving-sounding objects, *IEEE Transactions on Multimedia*, 15(2): 378-390.

[7] Mohammad A. Haque, Jong-Myon Kim. 2013. An enhanced fuzzy c-means algorithm for audio segmentation and classification, *Multimedia Tools and Applications*, 63(2): 485-500.

[8] Chung-Hsien Wu, Chia-Hsin Hsieh. 2006. Multiple change-point audio segmentation and classification using an MDL-based Gaussian model, *IEEE Transactions on Audio, Speech, and Language Processing: A Publication of the IEEE Signal Processing Society*, 14(2): 647-657.

[9] Makoto Yamamoto, Miki Haseyama. 2009. An Accurate Scene Segmentation Method Based on Graph Analysis Using Object Matching and Audio Feature, IEICE Transactions on Fundamentals of Electronics, *Communications & Computer Sciences*, E92/A(8): 1913-1919.

[10] Kiranyaz S., Ahmad Farooq Qureshi, Gabbouj M., et al. 2006. A generic audio classification and segmentation approach for multimedia indexing and retrieval, *IEEE Transactions on Audio, Speech, and Language Processing: A Publication of the IEEE Signal Processing Society*, 14(3): 1062-1081.

[11] Rongqing Huang, Hansen J.H.L. 2006. Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora, *IEEE Transactions on Audio, Speech, and Language Processing: A Publication of the IEEE Signal Processing Society*, 14(3): 907-919.

[12] Kotsakis, R., Kalliris, G., Dimoulas, C., et al. 2012. Investigation of broadcast-audio semantic analysis scenarios employing radio-program-adaptive pattern classification, *Speech Communication: An International Journal*, 54(6): 743-762.