Silence Segment Detection of English Speech Signals

Z.J.CHU
School of Humanities, Nanjing University of the Arts, Nanjing Jiangsu, China

S.Q.PAN
School of Transportation, Southeast University, Nanjing Jiangsu, China

J.G.LIU
School of Foreign Languages, Southeast University, Nanjing Jiangsu, China

L.ZHAO
School of Radio Engineering, Southeast University, Nanjing Jiangsu, China

ABSTRCT: To differentiate between silence and voice segment of English speech signal is very important in English IBT systems. In order to improve the accuracy of the end detection in traditional detection method of speech end, a silence/voice detection method of speech signals using wavelet transform parameters is proposed in this paper. Based on high performance of frequency segmentation and energy focusing of the wavelet, the statistical parameters of the speech signals on different sub bands are extracted and the effectiveness of this method is validated by simulation experiment results under different noise conditions.
KEYWORD: English IBT; bark wavelet transform; silence segment; voice segment

1 INTRODUCTION

At present, oral tests are included in well-known English language tests at home and abroad such as TOEFL, IELTS, CET (College English Test) and so on. However, oral tests evaluation is influenced to a great extent by subjective factors such as the language proficiency, moods and feelings of the examiners, so that its objectivity and accuracy can hardly be guaranteed.

The project of the Internet-based college English test (iB-CET) launched by the Chinese Ministry of Education in the recent years, as a revolution, will overcome the limitations of the paper-based test system.

Statistics have shown that about half and more of the time is in the silent condition in man's common speech [5]. As a result, how to differentiate between silence and voice segment of English speech signal is a key problem to face in English IBT system. Segment with no speech signal but background noise is known as silence segment. As redundant information, it alternates with the voice segment.

In recent years, the detection method of speech signal end is usually utilized to locate the start and end of speech and then to remove the silent segment. [1][2][4] However, due to the limited validity of the detection parameters in such method, the accuracy of end detection is relatively low and the effect is worse especially when the signal-to-noise ratio is lower.

With a view to improve the accuracy, present researches mainly go to two directions: 1) pattern classification method based on fast K Nearest Feature Line is applied in silent speech identification; 2) New statistical parameters with better performance such as autocorrelation function variance are used in the detection of voice and silence segment, so as to improve the accuracy and reliability.

Based on the second method, a silence/voice detection method for speech signals using Bark wavelet transform parameters is proposed in this paper.

2 DETERMINATION OF FEATURE PARAMETERS FOR SILENCE/VOICE SEGMENT DETECTION

Silence/voice segment could be identified with several statistical parameters of speech signal, such as short-term average amplitude, short-term energy, quasi-periodicity, zero crossing rate, frequency-domain characteristics and so on. The effect of the detection method depends on whether correct statistical parameters are selected. In this paper, statistical parameters including short-term average amplitude *M*, variance of the amplitude *V*, short-term average zero crossing rate *Z*, variance of the zero crossing rate *Uk* are selected as discriminate parameters, which are defined as:

$$M = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

$$V = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - M\right)^2 \quad (2)$$

$$Z = \frac{1}{2N}\sum_{i=1}^{N}\left|\mathrm{sgn}\left[x_i\right] - \mathrm{sgn}\left[x_{i-1}\right]\right| \quad (3)$$

$$U_k = \sum_{n=-\infty}^{\infty}\left(Z_n - m_z\right)^2 w\left(n-k\right) \quad (4)$$

Where xi (i =1, 2, … , N) is the sample data of speech signal and N is the sample number in a frame. Frame length is generally chosen to be around 20ms. To validate the effectiveness of the chosen feature parameters, the above parameters are testified by the detection method based on fuzzy entropy, and the experimental results demonstrate that none of the four characteristic parameters can identify silence/voice segment very well. It is necessary that all of the four parameters be used altogether to detect the silence/voice segment.

According to the research on statistical parameters, it is found that under high signal-to-noise ratio, short-term average amplitude of the voice segment which is mainly determined by speech signal, is greatly different from that of the silence segment and is effective in differentiating the silence and voice segment. However, as the signal-to-noise ratio decreases, the influence of short-term average amplitude of the noise increases, discriminable ability of this parameter decreases.
Short-term average energy of the voice segment is composed of signal energy and noise energy. Under high signal-to-noise ratio, short-term average energy of the voice segment which is mainly determined by signal energy, is greatly different from that of the silence segment and is effective in differentiating the silence and voice segment. However, as the signal-to-noise ratio decreases, so do the difference and the discriminable ability of this parameter.

Short-term zero crossing rate (ZCR) of the voice segment is composed of noise and clean speech signal. Under high signal-to-noise ratio, ZCR of the voice segment is higher than that of the silence segment. However, as the signal-to-noise ratio decreases, the influence of the noise increases and the discriminable ability of this parameter weakened accordingly.

Based on the analysis above, in order to reduce the effect of noise on silence/voice segment identification, it is necessary that frequency distribution characteristics of their signal should be considered. The noise signal mainly ranges in high frequencies while the speech signal mainly ranges in relatively low frequencies. As a result, the frequency sub-band statistical parameters can be used to project speech and noise signals to different frequency levels to get statistical features of the mapping coefficient. The silence segment and voice segment are identified by different frequency characteristics of speech and noise signals, which reduces the effects of noise on identification.
As wavelet transform is widely used in coding and has high performance of frequency segmentation and energy focusing, the signal features of silence/voice segment in different frequencies levels and time slots are “projected” in the wavelet coefficient of different levels. Features of silence and voice segment will be enhanced in “subspace” after wavelet transform and such kind of parameters will have better performance in detecting silence and voice segment.

Considering the features of English speech signals and the requirements to simplify computation, Bark wavelet transform parameters are adopted in this article as the method to detect silence/voice segments.

3 THE DETECTION METHOD OF THE SILENCE/VOICE SEGMENT

The specific procedure for the method is as follows: Firstly wavelet transform is carried out for the speech signals to get its wavelet transform coefficient and the statistical feature of the wavelet transform coefficient is calculated. Then statistical parameters of wavelet coefficient in different frequencies levels are regarded as discriminating parameters of the frame of signal to detect whether it is silence or voice segment.

Since wavelet transform is successfully applied in audio and video coding and de-noising, the process of extracting wavelet coefficient could be combined with that of speech coding and de-noising, in order to reduce the computation. Besides, the fact that wavelet coefficients are sensitive to the opening and closing of the “voice door” can not only achieve frequency segmentation but also improve the accuracy of detection. The flow chart for the whole algorithm is shown in Fig.1:

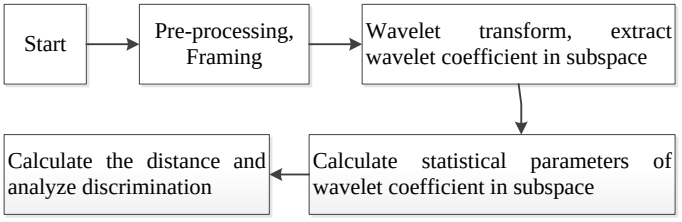

Fig.1 The flow chart of silence/voice segment discrimination

Pre-processing includes signal filtering and such process as signal regulating, which is the same with general speech processing. Processed speech signal needs to be framed, which is about 33-100 frames per second. To guarantee a smooth transition, generally there are overlaps between frames and frames, with the general frame shift to fame length ratio chosen to be 0 to 0.5.

The function of signal framing is to guarantee a stationary signal. Frame length is extremely important for traditional speech signal processing method. As the wavelet transform adopted in this paper is non-stationary signal analysis tool in itself, the effect of frame length choice is found to be negligible. It is discussed in detail by Shubha Kadambe & G. Faye Boudreaux-Bartels(1992). [3]

Speech signals are carried through wavelet transform frame-by frame after framing to derive wavelet coefficient of the subspace. Considering the requirement of calculation amount, Mallat algorithm is adopted and "db3"wavelet is chosen to decompose wavelet into 3 layers. An approximate coefficient ca3 and three detail coefficients cd3, cd2, cd1 in subspace are acquired. The specific algorithm is discussed in the previous section and not to be repeated here.

4 EXPERIMENTAL RESULTS

In this experiment, random noises generated by computer are attached to the speech and to both ends of speech signals. The speech is recorded under common laboratory conditions and the above method is used to discriminate silent segment from the voice segment. Wavelet statistical parameters of the signals are calculated with the respective SNR of 15dB, 10dB, 5dB and 0dB. The noise attached to both ends of the sentence lasts 1 second.

In the experiment, 3000 English sentences pronounced by 30 speakers (15 male and 15 female) are obtained, with 100 sentences pronounced by each person. 1500 of the 3000 sentences are used for the purpose of studying while 150 of them are used for identification. The average reading speed is 7.7 words/s. Statistics sampling frequency is 12 kHz, window length is 21.33ms and window move is 10ms. Experimental results are shown in table 1:

Table.1 The simulation result of silence/voice segment discrimination

SNR	15dB	10dB	5dB	0dB
stating point	100	100	97.8	96.5
ending point	100	100	97.2	95.3
average	100	100	97.5	95.9

In the above simulation result, the silence/voice segment discrimination is carried out frame-by-frame and endpoint detection error within 10 frames is considered to be correct. As it is revealed in table 2, the silence/voice segment detection proves to be effective.

The following is an experiment eliminating silent segment from a piece of actual speech signals based on the algorithm discussed above. The speech signals last about 5 seconds, the sampling frequency is 8 kHz, 8 bit quantization and actual signal to noise ratio is estimated to be about 17 dB. The results are shown in Figure 2.

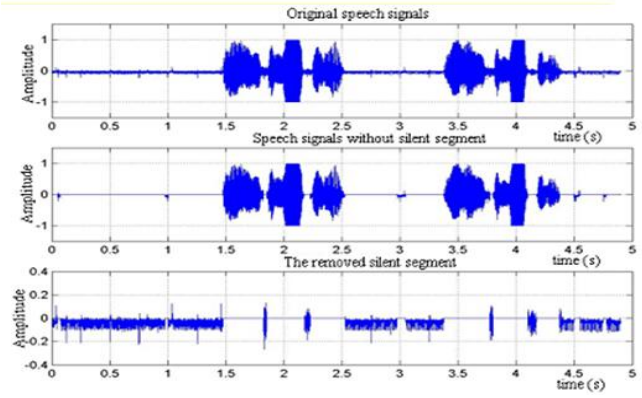

Fig.2 The instance of English silence/voice elimination

5 CONCLUSION

In the internet-based English test system, whether it is for speech recognition or for speech evaluation, detection of voice segment is very important. The voice/silence identification method for speech signals using Bark wavelet transform parameters proposed in this article is validated by experiments and proves to be simple in computation with low false drop rate and to be effective and worthy promoting in practice.

ACKNOWLEDGEMENT

This paper is supported by Philosophy and Social Sciences Research Funding for Jiangsu Higher Learning Institutions (2014SJD152) and Foreign Language Excellent Project for Jiangsu Social Sciences Application Research (14jsyw-54).

REFERENCES

[1] Baraniuk R G. Compressive sensing. *IEEE Signal*
[2] *Processing Magazine*, 2007, 24(4): 118- 121.
[3] Giacobello D. et al. 2010. Retrieving sparse
[4] Patterns using a compressed sensing framework: applications to speech coding based on sparse linear prediction. *IEEE Signal Pro-cessing Letters* 17(1): 103-106.
[5] Kadambe S. & Boudreaux-Bartels G. F. 1992. Application of the wavelet transform for pitch detection of speech signals. *IEEE Trans Inform Theory* 38 (2):917-924.
[6] Xu W. et al. 2003. A speech endpoint detector based on eigenspace-energy-entropy. *Journal of China Institute of Communications* 24(11): 125-132.
[7] Zhao L. 2003. *Speech signal processing*. Beijing: China machine press.