



# WESPAC IX 2006

The 9th Western Pacific Acoustics Conference  
Seoul, Korea, June 26-28, 2006

## STATUS-CHANGE BASED SYLLABLE DETECTION IN CHINESE CONTINUOUS SPEECH

Xiaoliang FENG, Shuiyuan YU

*Communication Acoustic Lab, Communication University of China*

*Chaoyang Dist. Communication University of China 126 mailbox Beijing, 100024, P.R.C*

*E-mail: banban220@gmail.com, yusy@cuc.edu.cn*

### ABSTRACT

In this paper, a syllable detection method of Chinese continuous speech is studied and implemented. Different from other detection methods, this method detects at more detailed level. In this method the detection does not stop at syllable level, but goes deep into vowel and consonant level. The method uses a variety of features including the frame energy, the zero crossing rates to separate into different segments which can also be called statuses. Seven kinds status are defined in continual speech (No speech, consonant start, consonant persistence, consonant end, vowel start, vowel persistence, vowel end). The continuous speech can be passed into the detection to get the syllable detection results. In addition, the detection gives the ranges for each status defined above.

**KEYWORDS:** Syllable detection, speech signal process

### INTRODUCTION

Syllable detection is one of key problems in continual speech recognition. The task of detection is judging the start point and end point of a syllable in a continual speech. Especially in the noise environment, the detection is more useful for the speech recognition system<sup>[1-3]</sup>. An accurate detection can make the recognition processing focus on the speech segment to improve the accuracy.

Generally, Chinese speech voice is classified into three kinds: No voice segment; consonant segment; vowel segment, each part should be detected before recognition. Energy and zero cross rate and their product are three basic features used to syllable detection. These features are effect in word-segment endpoint detection but the static threshold method can not meet the variability of continual speech. This paper introduces a status-change based method to overcome that weakness.<sup>[4]</sup>

## CLASSIC DETECTION METHOD

In syllable detection, the classic features<sup>[5]</sup> include energy (E), zero-cross rate (Z), and their product (EZ). These features are defined as follow:

$$E_n = \sum_{l=1}^L S_l^2 \quad (1)$$

$$Z_n = \sum_{l=2}^L |\text{sgn } S_l - \text{sgn } S_{l-1}| \quad (2)$$

$$EZ_n = E_n \times Z_n \quad (3)$$

In the equations the  $E_n$  means one frame energy,  $Z_n$  means the zero-cross rate and the  $EZ_n$  means their product.  $n$  is the number of frame,  $L$  is the total sample point in one frame and  $S_l$  is the  $l^{\text{th}}$  sample point in one frame.

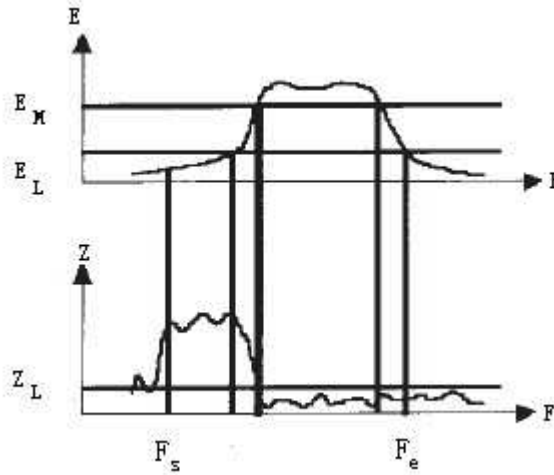


Figure 1 Classic Detection

Figure 1 showed the classic syllable detection method. The classic detection uses many thresholds to detect the start frame ( $F_s$ ) and end frame ( $F_e$ ). There are many ways to define the threshold. Whatever how the threshold is defined, it can not avoid some weaknesses:

1. The threshold is fixed and can not adapt to the variety in the speech. It may miss detect some low energy voice to noise.
2. The detection only detect the word-segment the detail information (consonant; vowel) can not be used.

## STATUS-CHANGE BASED DETECTION

The cause of the weaknesses is the variety of speech. Some methods were introduced for improvement. In the paper<sup>[6]</sup>, the author found the energy distribution obeys the Gauss distribution. So if use the maximum probability energy as the threshold, the result could be

improved. In the paper<sup>[7]</sup>, the author use an enhanced zero-cross definition:

$$Z_n = \sum_{l=2}^L \frac{1}{2} \{ |\text{sgn}(S_l - T_0) - \text{sgn}(S_{l-1} - T_0)| + |\text{sgn}(S_l + T_0) - \text{sgn}(S_{l-1} + T_0)| \} \quad (4)$$

$T_0$  is the noise threshold, only two adjacent points having absolute value which exceeds  $T_0$  and different signs can be counted a zero-cross. But in these methods, the object of detection is finding some quantitative thresholds to divide the speech, so the speech is seen as some segments combination which is a static model. The status-change based method, however, sees the speech as a different statuses change process. The goal of detection is to describe how the statuses change, then find where the interesting statuses appear. This method defines seven statuses: No voice; consonant start; consonant persistence; consonant end; vowel start; vowel persistence; vowel end. Each status can change to other status under some conditions which is an open set of rules. Table 1 shows the change relationships in Chinese.

*Table 1 Status change relationship*

Next current	No Voice	Consonant start	Consonant persistence	Consonant end	Vowel start	Vowel persistence	Vowel end
No Voice	Default	Condition N2C-S	No	No	Condition N2S	No	No
Consonant start	No	No	Immediately	No	No	No	No
Consonant persistence	No	No	Default	Condition C-P2C-E	No	No	No
Consonant end	No	No	No	No	Immediately	No	No
Vowel start	No	No	No	No	No	Immediately	No
Vowel persistence	No	No	No	No	No	Default	Condition V-P2V-E
Vowel end	Default	Condition V-E2C-S	No	No	Condition V-E2V-S	No	No

The first column lists the current statuses and the first row lists the target statuses. The cells contain four kinds content. If the across cell is “No”, it means from current statuses can not change to the target status. “Immediately” means current status should change to the target status unconditionally. “Condition X2Y” means if current status meets the condition the status X will change to Y. X,Y is one of the statuses formatted “a-b”, “a” is N(no voice),C(consonant) or V(vowel); “b” is S(start); P(persistence) or E(end). If the current status does not meet any change requirements, the status change to “Default”

The rules are some characters in status change which described by the basic features  $E_n$ ,  $Z_n$  and  $EZ_n$ , because the rules are too many to be shown in this paper. So take the condition N2C for example. This condition describes how the consonant start.

Rule 1: Four frames energy after this frame is average 100 times than the current frame

energy and the average zero-cross rate of those four frames is above 5.

Rule 2: Or the total energy of 3 frames after this frame is above 30% of the maximum energy and the zero-cross rate of 4 frames and so on.

Those rules could be set or added by test and experience but should avoid the confliction. So this method is more flexible than the old method and can adapt to the variety of continual speech easily.

## EXPERIMENT AND RESULT ANALYSIS

In the experiment, we select sixty different long articles and fifteen short sentences for recording. The long article and short sentence contain 500 words and 20 words separately on average. Each article and sentence we record ten females' voice. The sample rate is 16 KHz and the resolution is 16 bit. The frame length is 20ms and the overlap window is 10ms. The speeches of four articles and all short sentences are labeled manually, including the consonant segment start position; the consonant segment end position; the vowel segment end position. These speeches are used for test and the others are used for training. In the test if the error between the result and manual label is under 1 frame the result is consider to be correct, otherwise the result is wrong.

*Table 2 Test result*

	Total consonant segment	Correct consonant detection	Total vowel segment	Correct vowel detection
Article 1	610	569	628	580
Article 2	482	467	497	477
Article 3	598	576	607	581
Article 4	435	397	444	402
Total Sentence	331	312	339	317

Table 2 shows the accuracy is above 90% and have good result. But the continual two different vowel detection is not as accurate as others, because using those features could not describe how this kind of changes occurs, so that new feature like time duration should be added the rules.

## CONCLUSIONS

This paper introduces a status-change based syllable detection method. This method sees the speech as a status change process and uses the description way to find the syllable endpoint. This method not only could detect the word-segment but also could divide into consonant and vowel. Although more complex, especially in the rule design which needs much experience in endpoint detection and should be checked carefully to avoid confliction, this method is flexible and extendable which make it easy to adapt the continual speech. In future, more features such as time duration will be used to improve the detection result.

## REFERENCES

- [1] Rabiner L.R, Juang B.H. *Fundamentals of Speech Recognition* (Tsinghua University Press, Beijing, 1999) 82-85
- [2] Rabiner L.R, Sambur M. R. "An Algorithm for determining the Endpoint of Isolated Utterance" *Bell System Tech J* 1975, 54(2); 297-315.
- [3] Junqua J.C, Mak B, Reaves B. "A Robust Algorithm for Word Boundary Detection in the Presence of Noise" *IEEE Transactions on Speech and Audio Processing*, 1994; 2(3); 406-412
- [4] Ma Bin,Huang Taiyi,Xu Bo, et al. "Context-dependent Acoustic Models for Chinese Speech Recognition" *IEEE International Conference on Acoustic, Speech and Signal Processing* 1996:455-458.
- [5] Tian Ye Wang Zuoying Lu Dajin "Adaptive Algorithm of Speech Detection Based on Statistical Classification in the Presence of Noise" *Computer Engineering and Application* 14-15 (2002.01) (in Chinese)
- [6] He Zhiyuan, Hu Qixiu, Xu Guangyou "Research on Speech Segmentation Algorithm in Speaker Recognition" *Computer Engineering and Application* 55-58 (2003, 06) (in Chinese)
- [7] LU Yan-Ling, HOU Yuqing. et al "An Endpoint Detection and Syllable Separation Algorithm Based on Multi-characteristic for Noise signal " *Audio Engineer* 60-62 (2005,07)(in Chinese)