

# A Speech Endpoint Detection Method Based on Wavelet Coefficient Variance

Xueying Zhang, Jing Bai, Gaofeng Zhao  
College of Information Engineering  
Taiyuan University of Technology  
Taiyuan, Shanxi, China

**Abstract** - *Speech endpoint detection is a key technology for speech recognition. A new speech segmentation algorithm based on the variance of the wavelet coefficients was proposed. Speech signal with noise was decomposed by wavelet into some layers to research the statistic characteristics of every layer wavelet coefficient. The variances of the wavelet coefficients are calculated to form detecting features for endpoint detection. Simulations were made under different signal-to-noise ratios and the results show that this method is efficient to segment noisy speech even at a low signal-to-noise ratio.*

**Keywords:** endpoint detection, wavelet coefficient variance, Zero-Crossing Rate

## 1 Introduction

The method of speech endpoint detection is aimed at correctly discerning the speech signal from the background noise condition. This is a basic issue in speech signal processing. No matter in the field of military use or civil use, the speech endpoint detection has a wide variety of applications. The difficulty in endpoint detection is that the noises from respiration and the environmental interference make speech endpoint fuzzy. Meanwhile, the circumstance existing weak spirant or explodent sound at the endpoints, as well as the circumstance existing the rhinolalia, also add great difficulties in segment discerning. The thesis is focused on detecting speech endpoint by using the wavelet's multi-resolution property to decompose the speech signal into some layers and calculate the variance of every layer wavelet coefficients. And then the results are compared with the experience knowledge to find speech endpoint. It can effectively overcome the traditional methods' flaws and increase the detect accuracy.

## 2 The wavelet transformation and its application

The wavelet transformation is an effective tool in analyzing and handling the non-stationary signals. It is a time-scale analyzing method with multi-resolution property in signal processing; therefore it can effectively extract the message from the original signal. The wavelet transformation represents the signal  $f(t)$  as weighting sum of a series of functions. The series of functions are formed

by companding and shift of the base function  $\psi(t)$ . If the scale is  $a$ , time shift is  $\tau$ , the wavelet function is shown as equation (1):

$$WT_f(a, \tau) = \frac{1}{\sqrt{a}} \int f(t) \psi\left(\frac{t-\tau}{a}\right) dt \quad (1)$$

The wavelet analysis's main feature is to be able to analyze the local features of signal. By using the wavelet transformation, we can easily find the time when the signal distorts itself, and discover many features that other methods fail to detect.

## 3 The endpoint detection based on wavelet coefficient variance (WCV)

### 3.1 The features of wavelet coefficient variance

The speech signals are statistically self-similar random processes. Its statistical features in the time domain do not change along with the waveform's compression and extension, so it has the features of  $1/f$  process [1]. According to this property, we know that after the signal is decomposed by wavelet transformation, the wavelet coefficients of every sub-band are of the same statistical property [2]. Therefore, we can make endpoint detection by using the variances of the wavelet coefficients.

Suppose there is a discrete speech signal  $f[n]$ , after using the wavelet transformation, its wavelet coefficient is  $f_k$ , and variance is  $(\sigma_f)^2$ , as shown as follows in equation (2):

$$(\sigma_f)^2 = \frac{1}{N} \sum_{k \in N} (f_k - E(f_k))^2 \quad (2)$$

Where,  $N$  stands for the total number of the wavelet coefficients.

According to the property of the  $1/f$  process, after using the wavelet transformation to the original signal, the wavelet coefficients can be viewed as a random variable with the zero mean value, so the equation (2) are changed to the following equation:

$$(\sigma_r)^2 = \frac{1}{N} \sum_{k \in N} (f_k)^2 \quad (3)$$

Extracting the noise, unvoiced and clean speech signals' wavelet coefficients as the known knowledge, as shown in equation (4) as follows:

$$\begin{aligned} (\sigma_n)^2 &= \frac{1}{N} \sum_{k \in N} (n_k)^2 \\ (\sigma_q)^2 &= \frac{1}{N} \sum_{k \in N} (q_k)^2 \\ (\sigma_c)^2 &= \frac{1}{N} \sum_{k \in N} (c_k)^2 \end{aligned} \quad (4)$$

Where,  $(\sigma_n)^2$ ,  $(\sigma_q)^2$  and  $(\sigma_c)^2$  stand for the wavelet coefficients variances of noise, unvoiced, clean speeches' respectively.  $N$  stands for the total number of the wavelet coefficients.

### 3.2 Bayes classification model

The Bayes Classification Model is a typical mathematical classification model based on the statistical methods. The Bayes Theorem is one of the most important equations in the Bayes Theory, and is also the fundamental theory in the Bayes learning methods. It combined artfully the known probability and the conditional probability, using the known message and the sample data message to determine the conditional probability.

Suppose  $U = \{A_1, A_2, \dots, A_n, C\}$  is a finite discrete random variable set, and  $A_1, A_2, \dots, A_n$  are attribute variables. The value range of class variable  $C$  is  $\{c_1, c_2, \dots, c_l\}$ , and  $a_i$  is the value of variable  $A_i$ . The probability of the example  $\mathbf{x}_i = \{a_1, a_2, \dots, a_n\}$  belonging to  $c_j$  can be deduced by Bayes Theorem, as shown as follows (5):

$$\begin{aligned} P(c_j | a_1, a_2, \dots, a_n) &= \frac{P(a_1, a_2, \dots, a_n | c_j) \cdot P(c_j)}{P(a_1, a_2, \dots, a_n)} \\ &= \alpha \cdot P(c_j) \cdot P(a_1, a_2, \dots, a_n | c_j) \end{aligned} \quad (5)$$

Where,  $\alpha$  is the normal factor,  $P(c_j)$  is the known probability of class  $c_j$ ,  $P(c_j | a_1, a_2, \dots, a_n)$  is the conditional probability of class  $c_j$ . The known probability is independent of the sample data, but the conditional probability reflects the influence of the sample data to

class  $c_j$ . From formula (5), we can calculate out the probability of the sample  $\mathbf{x}_i$  belonging to class  $c_j$ .

### 3.3 Property classifications

In the endpoint detection, we make statistical classification to the variance equation (4) by using the Bayes classification algorithm.  $V_0$  stands for the pre-extracted noise variance,  $V_1$  stands for pre-extracted clean speech variance and  $V_2$  stands for the unvoiced's variance. According to the Bayes classification principles, they should be as shown in equation (6)-(8) as followings:

$$\begin{aligned} P(\{s_k^m\} | V_0) &= \prod_{m \in M} \prod_{k \in N(m)} p(\{s_k^m\} | V_0) \\ &= \prod_{m \in M, k \in N(m)} \frac{1}{\sqrt{2\pi(\sigma_n^m)^2}} \text{EXP} \left\{ -\frac{(s_k^m)^2}{2(\sigma_n^m)^2} \right\} \end{aligned} \quad (6)$$

$$\begin{aligned} P(\{s_k^m\} | V_1) &= \prod_{m \in M} \prod_{k \in N(m)} p(\{s_k^m\} | V_1) \\ &= \prod_{m \in M, k \in N(m)} \frac{1}{\sqrt{2\pi(\sigma_c^m)^2}} \text{EXP} \left\{ -\frac{(s_k^m)^2}{2(\sigma_c^m)^2} \right\} \end{aligned} \quad (7)$$

$$\begin{aligned} P(\{s_k^m\} | V_2) &= \prod_{m \in M} \prod_{k \in N(m)} p(\{s_k^m\} | V_2) \\ &= \prod_{m \in M, k \in N(m)} \frac{1}{\sqrt{2\pi(\sigma_q^m)^2}} \text{EXP} \left\{ -\frac{(s_k^m)^2}{2(\sigma_q^m)^2} \right\} \end{aligned} \quad (8)$$

Where,  $M$  is the number of wavelet layers,  $m$  stands for different wavelet sub-layers.  $N(m)$  stands for the total coefficient number of the  $m$  sub-layer.  $P(\{s_k^m\} | V_0)$  is the probability of  $s(t)$  being noise,  $P(\{s_k^m\} | V_1)$  is the probability of  $s(t)$  being clean speech, while  $P(\{s_k^m\} | V_2)$  is the possibility of  $s(t)$  being unvoiced.

If  $P(\{s_k^m\} | V_0) < P(\{s_k^m\} | V_1)$ , then we deduce the signal should be speech, otherwise we compare  $P(\{s_k^m\} | V_0)$  and  $P(\{s_k^m\} | V_2)$ . If  $P(\{s_k^m\} | V_0) < P(\{s_k^m\} | V_2)$ , we decide that the signal is speech, otherwise is noise.

## 4 Endpoint detecting algorithm

(1) Sample the input speech signal, and then we make the discrete speech samples into frames. We use  $R_i$  ( $0 < i < D$ ) to denote the frames. Where,  $D$  is the frames' total number.

(2) As db4 wavelet shows a good orthogonal property, after using a 5 layer db4 wavelet transformation to the  $i$ -th frame  $R_i$ , the resulting wavelet coefficient is  $r_k^m$ .

(3) Using equations (6) and (7) separately, if  $P(\{s_k^m\} | V_0) < P(\{s_k^m\} | V_1)$ , then the signal is speech,

otherwise using equation (8). If  $P(\{s_k^m\} | V_0) < P(\{s_k^m\} | V_2)$ , then we decide the signal should be speech, otherwise it should be noise.

(4) If  $i > D$ , the algorithm ends, otherwise goes to step (2).

(5) After all frames are marked separately, the process will be started. We define that the minimum speech span is 6 frames and the minimum noise span is 8 frames. Thus, when the time spans are shorter than the defined time period, it will be discarded.

## 5 The experimental results and conclusions

The experiment is carried under different noise conditions. First, the speech signal is sampled at the rate of 11025 Hz and quantized into data of 16bits, and then added with different levels of white noise. In all experiments, the speech signals are divided into frames with 220 sample points each. The neighboring frames shared 50% overlapping area. Marking manually each speech file to distinguish speech endpoint from the noisy background, and then we can use these marks to obtain the accuracy of the speech endpoint detecting method.

Figure 1, Figure 2 and Figure 3 are the results of using the method mentioned above. Figure 1 is the speech signal without noise interference. Figure 2 is the waveform of speech adding specific noise interference, and Figure 3 is the output waveform of using the algorithm.

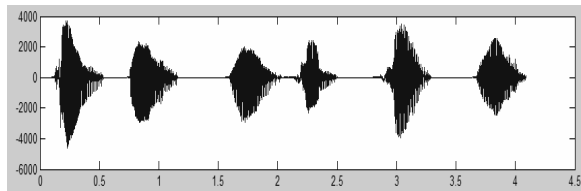


Figure 1. The original speech without noise

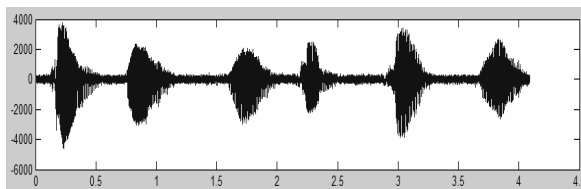


Figure 2. The speech of SNR=15dB with Gaussian white noise

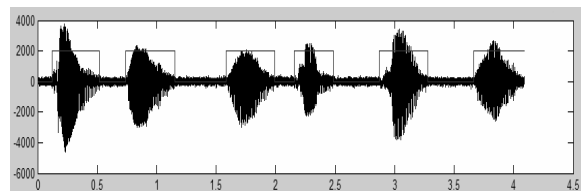


Figure 3. The result of speech endpoint detecting with noise

We can also use the calculation methods of Energy and Zero-Crossing Rate (EZCR) to detect the speech signal endpoint. The endpoint detection result of 60 speech sentences using two methods are shown in the Table 1. From the table we can see that along with the increasing in the noise interference, the result of WCV method is super to the one of EZCR method. It shows that the method based on the wavelet coefficient variance can perform better detection comparing with the ordinary energy and zero crossing rate method. It can meet the demand of endpoint detection in practical applications, such as the speech strengthening under strong noise interference and the speech recognition etc.

Table 1. The result of 60 speech sentences endpoint detection

method	Clean speech	SNR=15dB	SNR=10dB	SNR=0dB
EZCR	97.9%	96.6%	75.6%	64.0%
WCV	97.2%	96.7%	90.4%	85.6%

## 6 Acknowledgements

The project is sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry of China ([2004] No.176), Natural Science Foundation of China (No.60472094), Shanxi Province Natural Science Foundation (No.20051039), and Shanxi Province Scientific Research Foundation for University Young Scholars ([2004] No.13). The authors gratefully acknowledge them.

## 7 References

- [1] G. Wornell, "Wavelet-based representation for the 1/f family of fractal process", *Proceeding of IEEE*, 1993, vol.81, no.10, pp.1428-1450.
- [2] A.M. Nassar, N.S. Kader, A.M. Refat, "Endpoints detection for noisy speech using a wavelet based algorithm", *EUROSPEECH'99*. Budapest Kluwer Academic Publishers, pp.903-906.
- [3] J.B. Xu, C.S. Ran, "The application of adaptive wavelet transformation in speech signal processing", *Computer Engineering and Science*, vol.26, no.7, 2004.