# Otsu 방법을 이용한 음성 종결점 탐색 알고리즘

## Otsu's method for speech endpoint detection

고 유, 장 한, 정길도
Yu Gao, Xian Zang, Kil To Chong
전북대학교 전자정보공학부

**Abstract** – This paper presents an algorithm, which is based on Otsu's method, for accurate and robust endpoint detection for speech recognition under noisy environments. The features are extracted in time domain, and then an optimal threshold is selected by minimizing the discriminant criterion, so as to maximize the separability of the speech part and environment part. The simulation results show that the method play a good performance in detection accuracy.

**Key Words** : Endpoint detection, Otsu's method, Lyapunov exponent

### 1. Introduction

The detection of the endpoints in a speech utterance for discriminating speech from background noise is a crucial tache in speech processing system. Accurate endpoint locating could ensure good speech recognition accuracy. In particular, a major source of error in automatic recognition system of isolated words is the inaccurate detection of beginning and ending boundaries of test and reference templates, thus it is essential to locate the regions of a speech signal that correspond to each word. Furthermore, an appropriate scheme for locating the beginning and end of a speech signal can be used to eliminate significant computation by making it possible to process only the parts of the input that correspond to speech.

This paper considers the time-domain features by computing the Lyapunov exponents of the time-domain speech signal, and then, the Otsu's method [1] is proposed to solve the problem of selecting threshold, which is a optimal value can discriminate the speech from the background noise. Experiments results indicated that the speech segments can be very accurately extracted in utterances consisting of background noise.

### 2. Endpoint detection

The conventional endpoint detection methods [1,2] are mainly based on the simple energy detector, which could performs adequately for clean speech. For example, short-time energy and zero-crossing rate representations can be combined to serve as the basis of a useful algorithm for locating the beginning and ending point if in the high signal-to-noise environments, but will degrade in noisy circumstance.

### 2.1 Feature extraction by Lyapunov exponent

As we know, in mathematics, the Lyapunov exponent of a dynamic system is a quantity that characterizes the rate of separation of infinitesimally close trajectories [4]. Quantitatively, two trajectories in phase space with initial separation $|\delta Z_0|$ diverge

$$|\delta Z(t)| \approx e^{\lambda t}|\delta Z_0| \tag{1}$$

where $\lambda$ is the Lyapunov exponent, which could tell how chaotic a system is. The lower, the system is less chaotic.

We adopt the rationale of Lyapunov exponent and make some conversion to serve for the speech recognition system. In our experiments, each acoustic signal was sampled at 8 kHz in 1sec, thus we got 8000 samplings distributed in time-domain. After observing, it could be found that the noise segments part during the speech interval are corresponding to high values of Lyapunov exponents , while the voiced segments are opposite. In this way, we can set a threshold to filter the high Lyapunov exponents in order to discriminate the speech from the

저자 소개
* 高　瑜 : 全北大學 電子情報工學科 博士課程
* 媛　嫻 : 全北大學 電子情報工學科 碩士課程
* 丁吉道 : 全北大學 電子情報工學科 教授 · 工博

background noise. Hence, this algorithm provides a convenient way to realize the endpoint detection.

### 2.2 Formulation of Otsu's method

Let the Lyapunov exponents of a given speech utterance be represented in $L$ levels $[1,2,...,L]$. The number of exponents at level $i$ is denoted by $n_i$ and the total number of exponents by $N = n_1 + n_2 + ... + n_L$. In order to simplify the discussion, the level histogram is normalized and regarded as a probability distribution:

$$p_i = n_i/N, \quad p_i \geq 0, \ \sum_{i=1}^{L} p_i = 1. \tag{2}$$

Now suppose that we divide the exponents into two classes $C_0$ and $C_1$ (background noise and speech) by a threshold at level $k$; $C_0$ denotes exponents with levels $[1,...,k]$, and $C_1$ denotes exponents with levels $[k+1,...,L]$. Then the probabilities of class occurrence and the class mean levels, respectively, are given by

$$\omega_0 = \sum_{i=1}^{k} p_i = \omega_k \tag{3}$$

$$\omega_1 = \sum_{i=k+1}^{L} p_i = 1 - \omega_k \tag{4}$$

and

$$\mu_0 = (\sum_{i=1}^{k} ip_i)/\omega_0 = \mu(k)/\omega(k) \tag{5}$$

$$\mu_1 = (\sum_{i=k+1}^{L} ip_i)/\omega_1 = \frac{\mu_T - \mu(k)}{1 - \omega(k)} \tag{6}$$

where

$$\omega(k) = \sum_{i=1}^{k} p_i \tag{7}$$

and

$$\mu(k) = \sum_{i=1}^{k} ip_i \tag{8}$$

are the zeroth- and the first-order cumulative moments of the histogram up to the $k$th level, respectively, and

$$\mu_T = \mu(L) = \sum_{i=1}^{L} ip_i \tag{9}$$

is the total mean level of the original speech utterance. We can easily verify the following relation for any choice of $k$:

$$\omega_0\mu_0 + \omega_1\mu_1 = \mu_T, \quad \omega_0 + \omega_1 = 1. \tag{10}$$

The class variances are given by

$$\sigma_0^2 = \sum_{i=1}^{k} (i - \mu_0)^2 p_i/\omega_0 \tag{11}$$

$$\sigma_1^2 = \sum_{i=k+1}^{L} (i - \mu_1)^2 p_i/\omega_1 \tag{12}$$

These requires second-order cumulative moments.

In order to evaluate the "goodness" of the threshold (at level $k$), we shall introduce the following discriminant criterion measures (or measures of class separability) :

$$\sigma_W^2 = \omega_0\sigma_0^2 + \omega_1\sigma_1^2 \tag{13}$$

Then the problem is reduced to an optimization problem to search for a threshold $k$ that minimizes the function.

### 3. Simulations

We use the software "GoldWave" to record English digitals 0-9 in the lab environments in our experiments, each acoustic signal was low-pass filtered at 4 kHz and sampled at 8 kHz, also an 8 bit quantization of the signal amplitude was used. Each frame length was set to32ms (256samples), and there was a 16ms (128 samples) overlap between two adjacent frames. Then a 250-points Hamming window was applied to each frame to select the data points to be analyzed. Silent and background noise regions were removed by endpoint detection algorithm proposed in this paper.

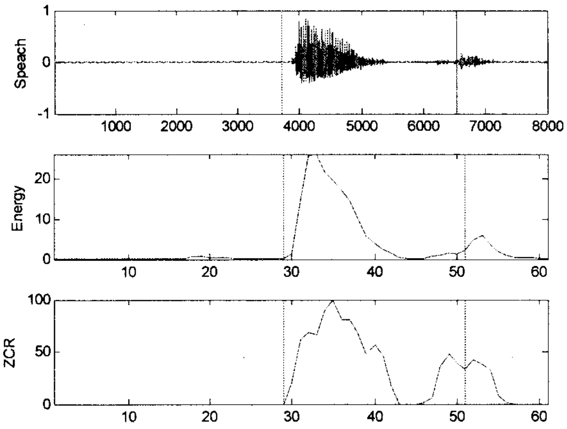

Figure. 1(a)

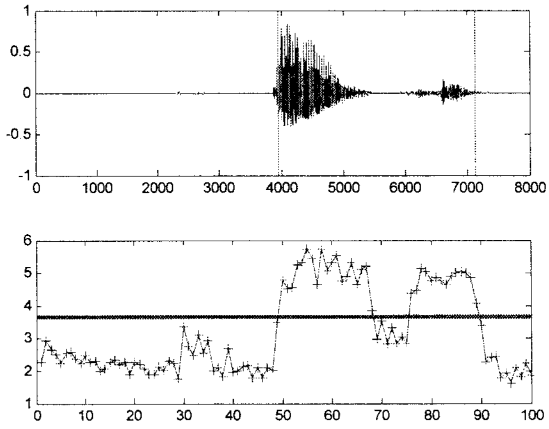

Figure. 1(b)

Let's take the simulation results of "eight" for example. Figure. 1(a) shows the results based on short-time energy and zero-crossing rate methods. We can see that the bound of the speech part is not accurate because two isolated energy wave crests. Figure. 1(b) demonstrated the detection based on our method. The threshold (the red dashed) of Lyapunov exponent is computed by Otsu's method, and the end detection performance is better.

## 4. Conclusions

An good algorithm for accurate and speech endpoint detection is proposed in this paper. The features are extracted by computing the Lyapunov exponents of the time-domain speech signal, and then the threshold is chosen by Otsu's method to get the bound of speech part. The simulation results show that the performance is better than the conventional endpoint detection methods based on the simple energy detector.

## 참고문헌

[1] N. Otsu, "A threshold selection method from gray-level histograms". IEEE Trans. Sys., Man., Cyber. vol. 9, pp: 62 - 66. 1979.

[2] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", Bell Syst. Tech. J., vol. 54, No. 2, pp. 297-315, February 1975.

[3] M. R. Sambur and L. R. Rabiner, "A Speaker Independent Digit-Recognition System", Bell Syst. Tech. J., vol. 54, No. 1, pp. 81-102, January 1975.

[4] A. Petry. D. A. C. Baronr, "Preliminary experiment in speaker verification using time-dependent largest Lyapunov exponent," Computer Speech and Lanuage, vol. 17, pp. 403-413, 2003.