ROBUST SPEECH END-POINT DETECTION AND F0 EXTRACTION BY APPLYING IMAGE PROCESSING TECHNIQUES[1]

ZHANG Bo, PENG Gang

J3, Department of Electronic Engineering

City University of HongKong
Tat Chee Avenue
Kowloon, HongKong
Email: 96420180@plink.cityu.edu.hk

ABSTRACT

This paper addresses two problems of robust speech recognition in noisy environment, i.e., robust speech end-point detection and F0 extraction. Conventional speech end-point detectors base on the features such as signal energy, zero crossing rate, energy of linear prediction error, etc. By observing more than 130 kinds of real-world environmental noises, we may believe that the essential features of speech signal may be its harmonic structure and its formant structure. An end-point detector based on detecting valid speech harmonic structure was designed and studied in this research. It operates on speech spectrogram by making use of some image processing techniques. The detector first extracts speech harmonic structure from the noisy speech signal. Then the F0 contour is calculated from the harmonic structure. Finally, the speech end-point can be deduced from the boundary of the F0 contour. The algorithm shows good robustness when tested on various noisy speech signals.

1. INTRODUCTION

There are many unsolved problems for robust speech recognition in noisy environment. One of them is robust speech end-point detection. False, missed or inaccurate speech end-point detection may cause the speech recognizer to produce word insertion, deletion or substitution errors. In a recent real-world evaluation of an isolated word recognizer, more than half of the recognition errors was due to the speech end-point detector [2]. Another problem is robust F0 extraction. This problem is especially important for tonal languages such as Chinese. Although there are many F0 extraction algorithms in literature, few of them take the real-world environmental noises into consideration explicitly.

For the speech end-point detection problem, most algorithms are based on the features such as signal energy, zero-crossings, duration, energy of linear prediction error or band-limited (e.g. 250--3500 Hz) signal energy ([4], [5]).

This research has analyzed and studied more than 130 kinds of environmental noises, such as sound of rain, thunder, wind, bird, telephone, hammering, motorcycle, door bell, various musical instruments, dog, etc. Based on the observation, it may be believed that features mentioned above may not be the essential difference between the speech signal and the environmental noises. The essential features of the speech signal may be the features that associated with the unique production mechanism of human speech, that is, the *speech harmonic structure* (along with its temporal movement) that corresponds to the human phonation part and the *speech formant structure* (along with its temporal movement) that corresponds to the human articulation part.

Although other noise sounds may also have harmonic structure, the speech harmonic structure has some sole features due to its unique production mechanism. Fig. 1 shows the spectrogram of a Chinese sentence /Hui Fu Tu Pian Si/ (in Pinyin), in which all of the harmonic lines of one voiced syllable form a speech harmonic structure. We can observe that the structures have the following properties: 1) Strict harmony, i.e., all of the lines within a structure move upward and downward harmonically. 2) Valid F0 range (50—500Hz). 3) Valid F0 jittering pattern, i.e., the lines within a structure can neither keep strict constant nor move very sharply. This is because human vocal organ is made up of muscles. It can not move very regularly and sharply 4). Valid prosody, e.g., duration of the structure can not be too long. 5). Normally, a speech harmonic structure has more than 3 harmonic lines.

Contrarily, many noises can not possess all of the properties. For example, noise of thunder does not have the speech harmonic structure, as depicted in Fig.2. Noise of telephone does not have valid F0 range, although it has harmonic structure, as shown in Fig. 3. Sounds of some music instruments have constant (thus invalid) harmonic lines. Fig. 4 shows such an example, in which the harmonic lines of the doorbell keep constant along time axis.

So, an end-point detector based on detecting valid speech harmonic structure was designed and studied in this research. It runs as follows.

[1] This research was supported by Grant 7000574-570 awarded to Prof. Willam S-Y Wang.

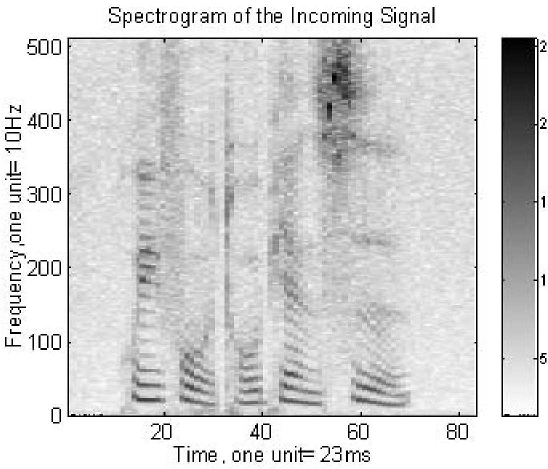

Fig. 1 Speech harmonic structure.

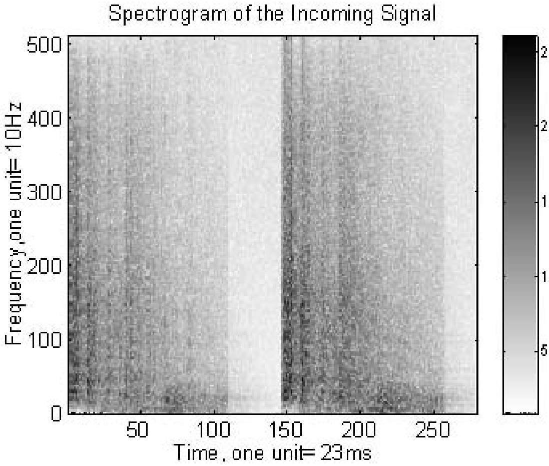

Fig.2 Spectrogram of thunder does not have harmonic structure.

Firstly, 1024-dots short-time Fourier analysis is applied on every 23ms frame of the incoming signal and the conventional narrow band spectrogram can be obtained. Every 25 frames are grouped to form a *sub-image*. Two consecutive sub-images have an overlapping region of 6 frames.

Secondly, a filter is designed and applied on the sub-image in order to enhance the harmonic lines.

Thirdly, for every sub-image, a line detection algorithm [1] is used to track and extract the harmonic lines of the speech harmonic structure. The resultant lines may contain some noise lines.

Fourthly, also for every sub-image, the sole features of the speech harmonic structure, such as strict harmony property, valid pitch range (from 50Hz to 500Hz), valid F0 jittering pattern and valid temporal variation rate are used to group the harmonic lines to form a valid speech harmonic structure.

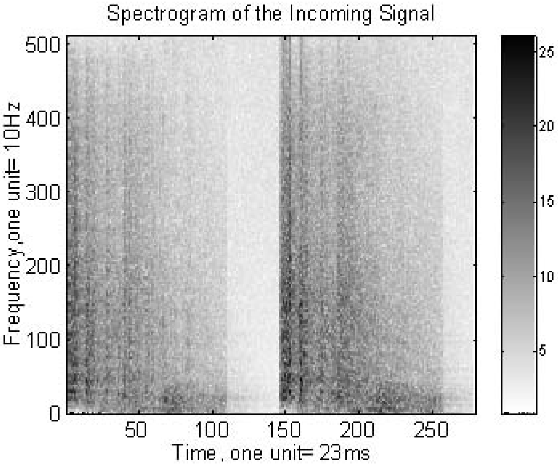

Fig.3 Harmonic structure of telephone does not have valid F0 range.

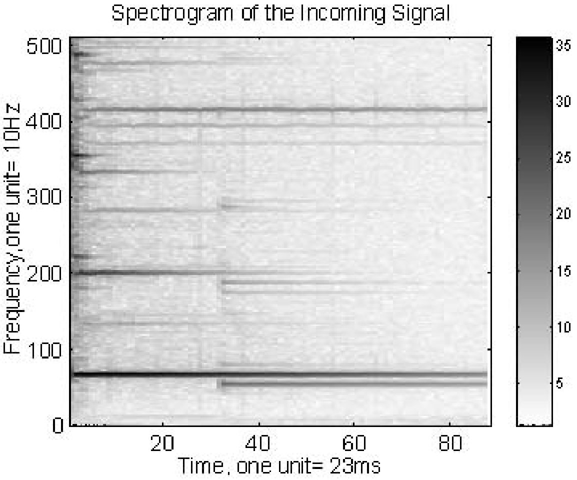

Fig. 4 Harmonic structure of doorbell has constant (invalid) harmonic lines.

Lastly, pitch contour is calculated from the harmonic structure by a multi-curves fitting algorithm. Furthermore, speech end-point can be deduced from the time-axis boundaries of the pitch contour by padding appropriate preceding and succeeding segment.

The remainder of this paper is organized as follows. In section 2, we present the filtering procedure. The line detection algorithm is explained in section 3. And the speech harmonic structure extraction procedure is given in section 4. Then we provide the details of the multi-curves fitting algorithm in section 5. Finally, in section 6, we present the implementation and experimental result of the whole algorithm.

2. ENHANCING HARMONIC LINES

An enhancing filter is used to remove noises and enhance the harmonic lines in the sub-image.

Let us denote one sub-image as S(n1,n2) and its two-dimensional FFT amplitude spectrum as W(w1, w2). For some noises, such as the thunder noise (in Fig.2) and hammering noise, their strong energy spread out along the whole frequency (n2) axis in the n1-n2 space, forming a vertical bar. Since the line detection algorithm in section 3 will make use of energy information to track lines, this high-energy vertical bar should be removed. When viewed in the w1-w2 space, energy of the vertical bar concentrates along the w1 axis. So, the enhancing filter should have zero response along the w1 axis.

For some other kinds of noises, their energy distributes in the rectangular region above a certain value of w2. After experimental trial, a bandpass filter is designed by using the window sampling method. It’s two-dimensional FFT amplitude response is shown in Fig. 6. In order to save computational load, the order of the filter is set to 21. This leads to the big ripple in the high frequency region along the w2 axis. But this ripple does not affect its enhancing effect.

As an example, an original sub-image of a segment noisy speech signal is shown in Fig. 5 while the enhanced sub-image is shown in Fig. 7. The noisy speech signal is a

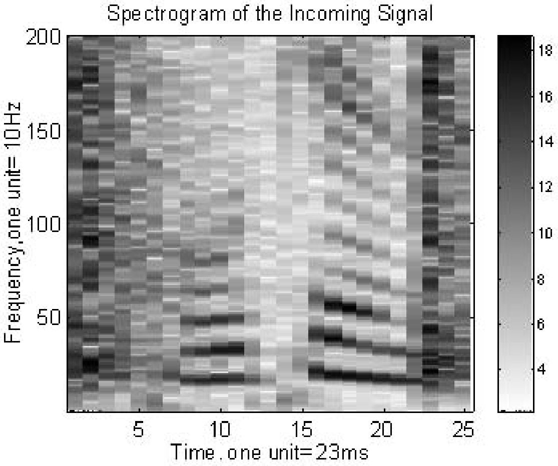

Fig. 5 Spectrogram of a segment of noisy speech signal. The speech signal is corrupted by hammering noise.

3.DETECTING HARMONIC LINES

This section presents how to detect harmonic lines from the enhanced sub-image. Although the line detection problem seems to be very simple for human eyes, it requires some careful consideration during computer processing. Many methods have been proposed for this task [1]. Compared to the general sense lines, speech harmonic lines have their own features. Their shapes are quite smooth while a general sense line can have a saw-shape. Furthermore, they are normally thicker than their surrounding noises. Also, energy of a harmonic line is stronger than other dots within a local area where the harmonic line is located in. It may be believed that we can develop a more robust and efficient algorithm if we integrate these knowledge into our algorithm.

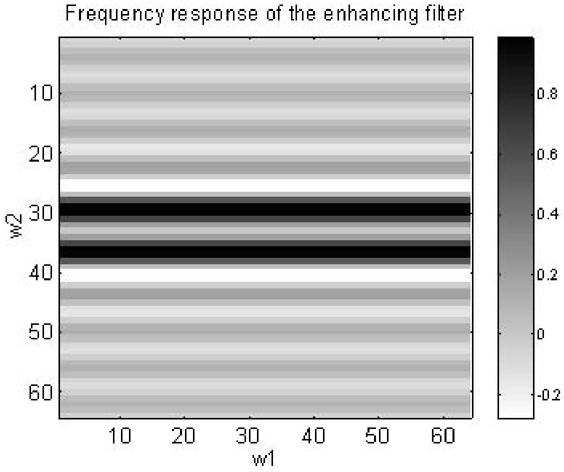

Fig.6. Frequency response of the enhancing filter.

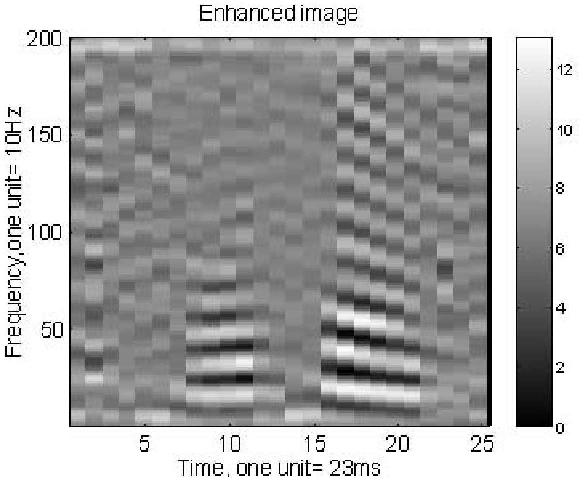

Fig. 7. Enhanced image. For clear viewing, the white color represents stronger energy while the dark color represents poorer energy.

The algorithm runs as follows.

1. Let us denotes the enhanced sub-image as Partmap (x, y) where x present the time axis and y present the frequency axis of the spectrogram.

   Copy Partmap(x, y) to OrigPartmap(x, y). The OrigPartmap will be used in step 6 – c) below.

2. Based on the energy of Parmap(x, y), set threshold GlobalThrd.

3. Find the dot (maxx, maxy) whose energy Maxvalue is the strongest in Parmap(x, y).

4. Starting from this most prominent dot (maxx, maxy), backward/forward search is conducted in the following steps. First of all, initiate (maxx, maxy) to be *current dot*.

5. Assume the currnet dot is (Nowx, Nowy). For each step of the backward search,

   5.1 As shown in Fig.8, nine left neighbors of the current dot, (Nowx –1, Nowy-4), (Nowx –1, Nowy-3), (Nowx –1, Nowy-2),…, (Nowx –1, Nowy+4), are examined to find a maximum energy dot M.

   5.2 If energy of the dot M is smaller than the threshold GlobalThrd, break the backward search.

   5.3 Starting from the dot M and limited within the nine neighbors, find M's neighbors whose energy is larger than the GlobalThrd. Set UpBound to point to the highest one and DownBound to point to the lowest one. Also, set Thickness at (Nowx-1) to be | UpBound – DownBound |.

   5.4 The dot M is then appended as a new dot of the current line if it does not lead to saw-shape.

   5.5 Set energies of the dots between UpBound and DownBound to be zero. Set the dot M to be the current dot and go to 5.1 to continue the backward search.

   The forward search is similar to the backward search except that the right neighbors are examined.

6. After the backward/forward search, a line has been detected. If

   a) the line is too short to be a harmonic line or,
   b) the Thickness values of the dots on the line is too small or,
   c) energy of the line is not so strong within its neighboring local area,

   the line will be judged as a noise line and thus discarded. Otherwise, it will be recorded as a harmonic line.

7. Go to step3, until energy of the most prominent dot is smaller than GlobalThrd.

Fig. 9 shows the result of the algorithm when it is applied on the sub-image in Fig. 7.

4. EXTRACTING SPEECH HARMONIC STRUCTURES

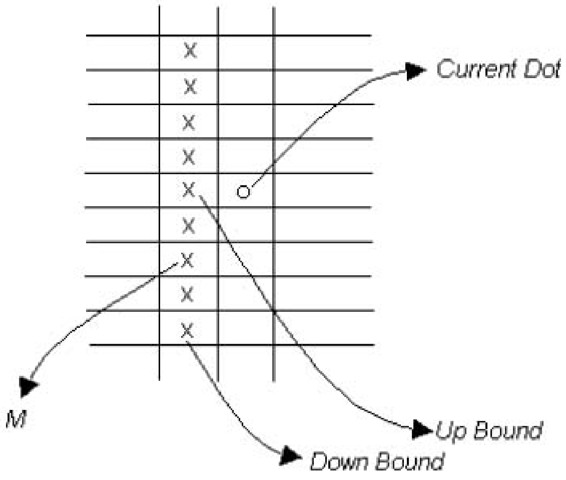

Fig. 8. One step of backward search procedure.

This section presents an algorithm to group harmonic lines that belong to one speech harmonic structure together and calculate harmonic order for each line. The facts below should be taken into consideration when design the algorithm.

a) Since the incoming signal may be a mixture of several different sound sources, several harmonic structures may co-exist simultaneously within some portion of the time-axis. Furthermore, part of the different structures may move harmonically. The method should be able to separate the different structures apart.

b) Within one harmonic structure, some of the harmonic lines may appear only at part of the whole time duration of the harmonic structure. The method should be able to group these lines into their corresponding harmonic structure. Otherwise, they will possibly form another false harmonic structure.

c) Due to noises, small portion of the harmonic lines may not move harmonically along with its harmonic structure. The method should fuse this harmonic line into its harmonic structure.

Outline of the algorithm is shown as follows.

1. For the sub-image that contains the detected lines, find the time instance xmax so that the vertical line x=xmax penetrates maximum detected lines.

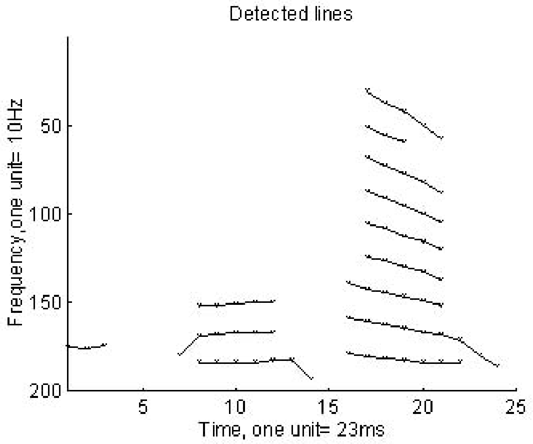

Fig.9. Detected lines from the enhanced sub-image.

2. From xmax, backward/forward search more than three lines that move harmonically within three frames. If not found, break the algorithm. Otherwise, these lines form an initial kernel.

3. Backward/forward extend the initial kernel. For each step of the extending, do

   3.1 Judge whether there are some lines that have been following the kernel moving harmonically within three frames. If so, append them into the old smaller kernel to form a larger kernel.

   3.2 For backward extending, suppose the updated kernel occupies a time segment of [x, xmax], we count a number Rnum for the kernel lines that go upward from time instance x to x-1 and another number Fnum for lines going downward from time instance x to x-1. Similarly, we can calculate Rnum and Fnum for the forward extending.

   3.3 If Rnum < 2 and Fnum < 2, the kernel contains too little lines. The backward/forward extending will be terminated.

   3.4 If Rnum > Fnum, the kernel lines that go downward will be discarded during the remainder backward/forward extending procedure. If Rnum < Fnum, the kernel lines going upward will be discarded.

   The steps from 3.1 to 3.4 will be repeated until the condition in 3.3 is satisfied or the extending reaches at the left/right side of the sub-image.

4. Record the detected harmonic structure and delete all of the lines of the structure from the harmonic line set. Go to step 1 if the size of the new line set is bigger than 3, otherwise stop the algorithm.

After grouping the harmonic lines together, we can calculate harmonic order for each line based on their frequency value. This step is relatively simple.

As an example, Fig.10 shows two speech harmonic structures extracted from the harmonic lines in Fig. 9. The numbers associated with the harmonic lines indicate the harmonic orders of the lines.

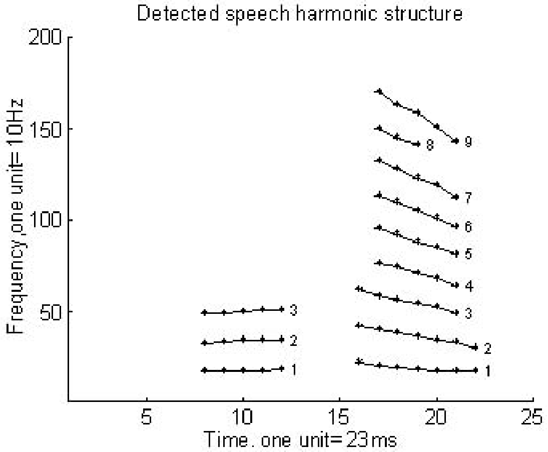

Fig.10.Detected speech harmonic structure by grouping and examining the detected lines.

5. MULTI-CURVES FITTING TO CALCULATE F0 CONTOUR

The polynomial curve fitting idea is widely used to estimate a smooth curve that is expressed as a polynomial of degree P when given a set of experimental data {(xi, yi)}. This idea is extended in this paper to calculate a polynomial to express the F0 contour. The problem is formulated below.

Assume there are ***m*** harmonic lines within a speech harmonic structure, the length of the ***i***th line is ***in*** (in frames) and the harmonic order of the ***i***th line is ***Ni.*** So the ***i***th line can be expressed as

$$X_i = \{x_{i1}, x_{i2}, x_{i3}, ..., x_{in}\},$$
$$Y_i = \{y_{i1}, y_{i2}, y_{i3}, ..., y_{in}\}. \qquad (1)$$

The objective of the multi-curves fitting is to find a polynomial of degree ***Q***

$$y = f(x) = \sum_{q=0}^{Q} a_q x^q \qquad (2)$$

which can minimize the fitting error

$$E = \sum_{i=1}^{m} \sum_{j=1}^{in} \left[Ni \cdot f(x_{ij}) - y_{ij} \right]^2. \qquad (3)$$

We first substitute (2) to (3):

$$E=\sum_{i=1}^{m}\sum_{j=1}^{in}\left[Ni\cdot\sum_{q=0}^{Q}a_q x_{ij}^{q}-y_{ij}\right]^2. \qquad (4)$$

Then we differentiate the right side of (4) with respect to a_k and solve its zeros:

$$\begin{aligned}\frac{\partial E}{\partial a_k}&=2\cdot\sum_{i=1}^{m}\sum_{j=1}^{in}\left[Ni\cdot\sum_{q=0}^{Q}a_q x_{ij}^{q}-y_{ij}\right]\cdot Ni\cdot x_{ij}^{k}\\&=2\cdot\sum_{i=1}^{m}Ni^2\sum_{j=1}^{in}x_{ij}^{k}\left[\sum_{q=0}^{Q}a_q x_{ij}^{q}-y_{ij}\right]=0 \qquad (5)\end{aligned}$$

By algebraic manipulation, we have

$$\sum_{q=0}^{Q}a_q\sum_{i=1}^{m}\sum_{j=1}^{in}Ni^2 x_{ij}^{k+q}=\sum_{i=1}^{m}\sum_{j=1}^{in}Ni\cdot x_{ij}^{k}\cdot y_{ij} \qquad (6)$$

Thus, the polynomial coefficients in (2),

$$a_k, k=0...Q$$

can be solved by the linear equations in (7).

The degree of the polynomial in (2), Q, varies according to the width of the speech harmonic structure. It is increased by 2 when the width increases 10 frames, with a minimum value 3. Fig. 11 shows two F0 contours fitted from the two speech harmonic structures in Fig. 10.

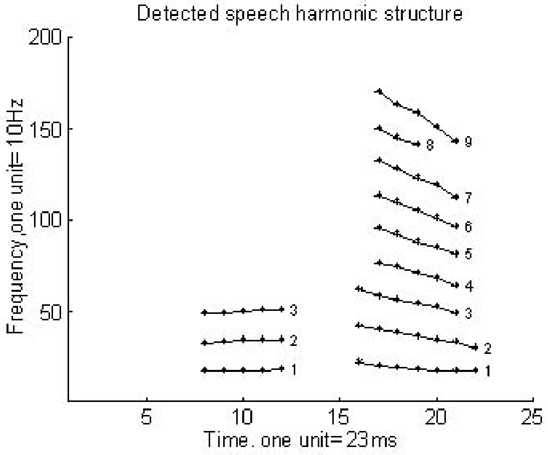

Fig.11. Calculated F0 contour by the multi-curves fitting algorithm.

$$\left|\begin{array}{cccc}\sum\sum Ni^2 & \sum\sum Ni^2x_{ij} & ... & \sum\sum Ni^2x_{ij}^{Q}\\ \sum\sum Ni^2x_{ij} & \sum\sum Ni^2x_{ij}^2 & ... & \sum\sum Ni^2x_{ij}^{Q+1}\\ \sum\sum Ni^2x_{ij}^2 & \sum\sum Ni^2x_{ij}^3 & ... & \sum\sum Ni^2x_{ij}^{Q+2}\\ ... & ... & ... & \\ \sum\sum Ni^2x_{ij}^{Q} & \sum\sum Ni^2x_{ij}^{Q+1} & ... & \sum\sum Ni^2x_{ij}^{2Q}\end{array}\right|$$

$$\bullet\begin{pmatrix}a_0\\a_1\\a_2\\...\\a_Q\end{pmatrix}=\begin{pmatrix}\sum\sum Niy_{ij}\\ \sum\sum Niy_{ij}x_{ij}\\ \sum\sum Niy_{ij}x_{ij}^2\\ ...\\ \sum\sum Niy_{ij}x_{ij}^{Q}\end{pmatrix} \qquad (7)$$

6. EXPERIMENTAL RESULT AND CONCLUSION

The algorithm is implemented on a Pentium-266 PC. It works in five times real-time. It has been tested for clean speech signal and noisy speech signal. The performance of the algorithm is good for the clean speech signal. For the noisy speech, the overall result is also good. But in some cases the algorithm will detect human's laughing, yowling of cats and some other noises as speech. This is because the harmonic structures of these noises are very similar to that of human speech. To solve this problem, the detected signal segment should be submitted to a speech recognizer which can judge whether the detected segment contains speech signal by the confidence measure technique.

REFERENCES

[1]. Dana H. Ballard, Christopher M. Brown, Computer Vision, Pentice-Hall, Inc. 1982.

[2].Jean-Claude Junqua, Brian Mak and Ben Reaves. A Robust Algorithm for Word Boundary Detection in the Presence of Noise.IEEE Transaction on Speech and Audio Processing, Vol.2, No.3, July 1994.

[3] Jean-Claude Junqua, Jean-Paul Haton. Robustness in Automatic Speech Recognition, Fundamentals and Applications. Kluwer Academic Pulishers,1996.

[4].Lamel, L., Rabiner, L., Rosenberg A., and Wilpon, J. (1981). An improved endpoint detector for isolated word recognition. IEEE Trans. ASSP, ASSP-29:777-785.

[5].Rabiner, L. and Sambur, M. (1975). An algorithm for determining the endpoints of isolated utterances. Bell Syst. Tech. J., 54(2):297-315.