A ROBUST AND FAST ENDPOINT DETECTION ALGORITHM FOR ISOLATED WORD RECOGNITION

Yiying Zhang Xiaoyan Zhu Yu Hao
State Key Laboratory of Intelligent Technology and Systems , dept. of CS
Tsinghua University, Beijing, 100084
Email: zxy-dcs@mail. tsinghua. edu. cn

Yupin Luo
Dept. of Automation
Tsinghua University, Beijing, 100084
Email: luo@iris.au.tsinghua.edu.cn

Abstract — **The problem of automatic word boundary detection in quiet environment and in the presence of noise is addressed in this paper. A fast and robust algorithm for accurately locating the endpoints of isolated words is described in detail. This algorithm utilizes energy and zero-crossing parameters to acquire the reference endpoints, and then the principle of variable frame rate(VFR) is adopted and cepstrum is used to accurately define the boundaries of isolated words. Experimental results show that the accuracy of the algorithm is quite acceptable. Moreover, the computation overload of this algorithm is low since the cepstrum parameters will be used in later recognition procedure.**

I . INTRODUCTION

A major cause of errors in isolated word automatic speech recognition systems is the inaccurate detection of the beginning and ending boundaries of test and reference patterns [1]. It is essential for automatic speech recognition algorithms that speech segments be reliably separated from nonspeech. Recently, a real-world evaluation of a discourse system using an isolated word recognizer showed that more than half of the recognition errors were due to the word boundary detector [2]. According to Savoji [10], the required characteristics of an ideal word boundary detector are: reliability, robustness, accuracy, adaptation, simplicity, real-time processing and no a priori knowledge of the noise.

For the detection of the endpoints of speech signals in speech recognition systems several algorithms have been proposed during the last years. Most of these algorithms are based on simple parameters and have a low accuracy which decreases significantly the overall accuracy of the speech recognition system [3][4]. On the contrary, algorithms with high accuracy are based on acoustic parameters, the calculation of which is rather time consuming making the response time of the system slow [5].

The followings are typical word boundary detection algorithms[1].

1) Energy-based algorithms with automatic threshold adjustment. they are intuitive approaches based on energy levels and durations of silence and speech. Sometimes several pairs of boundaries are yielded in order of their rank of being correct.

2) Use of pitch information. This kind of algorithms relies on pitch extraction and energy variations.

3) Noise adaptive algorithms: use the log of the rms signal energy, the zero-crossing rate, duration information, and a set of heuristics. The thresholds used for the energy and the zero-crossings are adapted automatically from a few frames provided by the signal environment.

4) Voice activation algorithms, which are based on energy and zero-crossing parameters and a set of decision rules and threshold settings.

5) Algorithms using frequency-based features. Two typical methods of this kind of algorithms are [1] and [6]. [1] applies time and frequency-based features, performs FFT transformations and computes the energy in the frequency-band 250-3500Hz, and logarithm of the rms energy. [6] is a two-dimensional cepstrum (TDC) approach for the recognition of Mandarin syllable initials. A TDC matrix is calculated to decide the boundary of Mandarin initials.

The algorithm for word boundary detection proposed in this paper is based on time and frequency features. Time features are energy and zero-crossing. Frequency feature is cepstrum. Compared with [6], although they use the same feature, the proposed algorithm decides the word boundaries according to the change of spectrograph of signals, and it has the advantage of low computation overhead.

II . DESCRIPTION OF THE ALGORITHM

The proposed algorithm consists of three steps. At first, the speech signal corresponding to a single word is preprocessed and the background noise is estimated which is used to decide the threshold values of the following steps. In the second step, the starting-point of the first and the ending-point of the last voiced sound are located to be used as reference endpoints based on time features energy and zero-crossing. And then, the accurate endpoints of the utterance are located according to frequency parameter cepstrum of the sequence of speech signals between the reference endpoints,

A. Background Noise Estimation

The input signal is pre-emphasized to eliminate the d-c component and to emphasize the higher frequency component. Let the sequence of pre-emphasized speech signal be:

$S'_1, S'_2, \cdots, S'_N$

in which N is the number of samples included in the current utterance.

From samples taken at the beginning and the ending of the input signal, the background noise (acoustic

0-7803-4253-4/97/$10.00 © 1997 IEEE

environment) is estimated. Energy level is computed as (1).

$$E_k = \sum_{n=(k-1)F/2+1}^{(k+1)F/2} (S'_n)^2 \qquad (1)$$

where, $k=1, 2, \cdots, K-1, K$, F is the length of a frame (we use F=256), K is the number of total frames.

The noise level at the front-end of the signal (E_F) is estimated using the first two energy frames, i.e.

$$E_F = \begin{cases} \dfrac{E_1 + E_2}{2} & if\ 0.5 \le E_1 / E_2 \le 2 \\ min(E_1, E_2) & otherwise \end{cases} \qquad (2)$$

The noise level at the back-end of the signal (E_B) is estimated in the same way, using the last two frames, i. e.

$$E_B = \begin{cases} \dfrac{E_{K-1} + E_K}{2} & if\ 0.5 \le E_{K-1} / E_K \le 2 \\ min(E_{K-1}, E_K) & otherwise \end{cases} \qquad (3)$$

Finally, the background noise level of the input signal (E_N) is estimated using the noise levels at the front and back ends as the following:

$$E_N = \begin{cases} \dfrac{E_F + E_B}{2} & if\ 0.5 \le E_F / E_B \le 2 \\ rejected & otherwise \end{cases} \qquad (4)$$

However, the energy of background noise obtained should lie within two limits E_L and E_H, i.e. $E_L<E_N<E_H$, otherwise the speech signal is not acceptable as being either too noisy or underamplified.

Another parameter zero-crossing of background noise is also been estimated in the similar way as that of the parameter energy. The following formulae (5)-(8) are the estimation of noise zero-crossing:

$$Z_k = \frac{1}{2}\{ \sum_{n=(k-1)F/2+1}^{(k+1)F/2-1} |sgn(S'_{n+1}) - sgn(S'_n)| \} \qquad (5)$$

where $sgn[x] = \begin{cases} 1, & if\ x \ge 0 \\ -1, & if\ x < 0 \end{cases}$

$$Z_F = \begin{cases} \dfrac{Z_1 + Z_2}{2} & if\ 0.5 \le Z_1 / Z_2 \le 2 \\ min(Z_1, Z_2) & otherwise \end{cases} \qquad (6)$$

$$Z_B = \begin{cases} \dfrac{Z_{K-1} + Z_K}{2} & if\ 0.5 \le Z_{K-1} / Z_K \le 2 \\ min(Z_{K-1}, Z_K) & otherwise \end{cases} \qquad (7)$$

$$Z_N = \begin{cases} \dfrac{Z_F + Z_B}{2} & if\ 0.5 \le Z_F / Z_B \le 2 \\ rejected & otherwise \end{cases} \qquad (8)$$

However, the zero-crossing of background noise obtained should lie within two limits Z_L and Z_H, i.e. $Z_L<Z_N<Z_H$, otherwise the speech signal is not acceptable as being either too noisy or underamplified.

B. Location of Reference Endpoints

The starting-point of the first voiced sound of the input utterance and the ending-point of the last one are located to be used as reference points for the location of the actual endpoints of the speech signal.

Searching the energy function from left to right with a frame, in a frame shift step. The first frame whose energy is above an energy threshold T_E is assumed to lie at the beginning of the first voiced sound of the utterance. Thus, the starting-point (P_{F3}) of the front voiced sound is obtained by

$$P_{F3} = \arg\min_k \{E_k > T_E,\ k = 1, 2, \cdots, K\} \qquad (9)$$

where E_k is defined by (1).

The energy threshold is experimentally derived from the background noise E_N, using the relation

$$T_E = C_E \cdot E_N \qquad (10)$$

where C_E is an experimentally derived constant.

In the same way, searching the energy function backwards from right to left, the ending-point (P_{B3}) of the last voiced sound is obtained by

$$P_{B3} = \arg\max_k \{E_k > T_E,\ k = K, K-1, \cdots, 1\} \qquad (11)$$

If the relations (9) and (11) can not be satisfied or if the distance between the points P_{F3} and P_{B3} is below a certain threshold, i.e. $P_{B3} - P_{F3}<t_{min}$, the algorithm recognizes absence of speech in the input signal and the procedure terminates. Otherwise, the speech signal between P_{F3} and P_{B3} is assumed to be voiced speech segment.

Then, utilizing the parameter zero-crossing to relax the endpoints. Searching the zero-crossing function from point P_{F3} backwards, and the reference starting point (P_{F2}) is obtained by

$$P_{F2} = \arg\max_k \{Z_k > T_{ZF},\ k = P_{F3}, P_{F3}-1, \cdots, 1\} \qquad (12)$$

where Z_k is defined by (5), T_{ZF} is the zero-crossing threshold defined by

$$T_{ZF} = C_{ZF} \cdot Z_N \qquad (13)$$

where C_{ZF} is obtained by experiments.

Searching the zero-crossing function from point P_{B3} forwards, and the reference ending point (P_{B2}) is obtained by

$$P_{B2} = \arg\min_{k} \{Z_k > T_{ZB},\ k = P_{B3},\ P_{B3}+1,\ \cdots,\ K\} \tag{14}$$

where T_{ZB} is the zero-crossing threshold defined by

$$T_{ZB} = C_{ZB} \cdot Z_N \tag{15}$$

where C_{ZB} is obtained by experiments.

Due to the different characteristics of starting and ending phonemes of an isolated word, different zero-crossing thresholds are utilized for determining the starting-point and ending-point, respectively.

C. Accurate Endpoints Detection

Variable frame rate (VFR) is a technique used in speech processing and recognition for discarding frames that are too much alike. It is indeed possible to prune in the time domain and to discard certain frames. During rapid changes, successive frames will be very different and will be retained. During slow speech, successive frames will look alike and can be left out.

VFR method emphasizes the transient regions, which are more relevant for speech recognition. Studying the spectrograph of speech signal, radical changes can be found at the point from unvoiced segment to voiced segment or from voiced segment to unvoiced segment. Thus, the boundary between voiced and unvoiced speech signal can be determined by adopting the principle of VFR methods.

The classical approach for VFR is to compute the Euclidean distance $D(i, j)$ between the current frame i and the last retained frame j and to compare this distance to some threshold T [7][8][9]. The decision criterion then becomes the following: leave the current frame out if

$D(i, j)<T$.

Since the changes of speech signal can be better embodied in the frequency domain and cepstrum can be measured by Euclidean distance, so cepstrum is adopted to determine the accurate endpoints.

Let $D(i, j)$ be the Euclidean distance between the cepstrum vectors of frame i and j. If $D(i, j)$ is greater than some threshold T_D in the searching procedure, the transient of voiced and unvoiced speech segment is assumed to occur. In order to avoid the sudden high energy noise, three frames are detected.

Searching from frame P_{F2} forwards until frame P_{B2}-1, the accurate starting-point (P_{F1}) is determined by

$$P_{F1} = \arg\min_{k} \left\{ \begin{array}{l} D(k,\ k+1) > T_D\ \&\& \\ D(k,\ k+2) > T_D\ \&\&\ D(k,\ k+3) > T_D \end{array} \right\} + 1 \tag{16}$$

where

$$P_{F2} \le k \le P_{B2} - 1$$

Searching from frame P_{B2} backwards until frame $P_{F2}+1$, the accurate ending-point (P_{B1}) is determined by

$$P_{B1} = \arg\max_{k} \left\{ \begin{array}{l} D(k,\ k-1) > T_D\ \&\& \\ D(k,\ k-2) > T_D\ \&\&\ D(k,\ k-3) > T_D \end{array} \right\} - 1 \tag{17}$$

where $P_{B2} \ge k \ge P_{F2} + 1$.

III. EXPERIMENTS

The described algorithm has been evaluated with several tests using 100 isolated Chinese words ranging from two to five syllables. The recordings were made in three different acoustic environments:

- Anechoic chamber
- Quiet laboratory environment
- Noisy environment

The test database consists of the utterances of 10 speakers (5 males and 5 females) who have repeated the vocabulary 20 times in anechoic environment, quiet laboratory environment and noisy environment, respectively.

The algorithm was applied to the words of the test database and the results, boundary definitions or rejections of the speech signal are recorded. Table I shows the rejection of the speech signal tested, which made it impossible to detect the endpoints.

The accuracy of endpoint results obtained from the application of the algorithm were then checked both acoustically and optically by skilled personnel. Acoustic tests included gradual segmentation of the speech signal and listening to locate the endpoints. Optical tests included comparisons between automatically and manually located endpoints. Both tests showed that the accuracy achieved by the proposed algorithm is quite acceptable. Results of these tests are shown in Table II.

TABLE I

Rejections during the steps of background noise estimation and location of reference endpoints in the three acoustic environments, respectively.

Environment	Rejection/Tests
Anechoic chamber	0/2000
Quiet laboratory	8/2000
Noisy environment	40/2000

TABLE II

Results of detection for the three environments.

Environment	Errors/Tests	Accuracy(%)
Anechoic chamber	0/2000	100
Quiet laboratory	14/2000	99.3
Noisy environment	32/2000	98.4

The words in the vocabulary are classified into nine categories according to the starting or ending phoneme of each word. Table III shows the test results of each category.

Fig. 1 and fig. 2 are endpoint results of word /Kun4

nan2/. Fig. 1 shows the reference endpoints acquired by utilizing parameters energy and zero-crossing. The final endpoints by employing cepstrum are shown in Fig. 2.

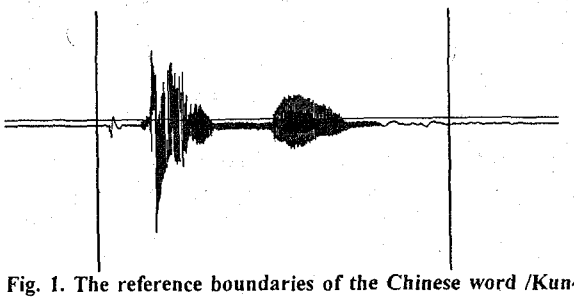

Fig. 1. The reference boundaries of the Chinese word /Kun4 nan2/ in quiet laboratory environment.

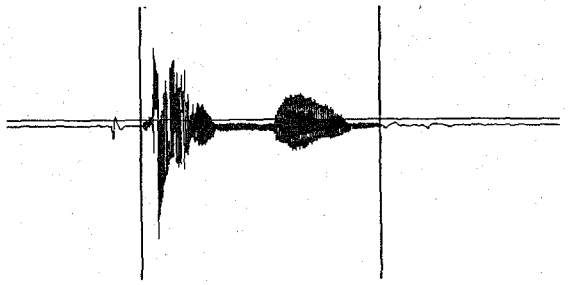

Fig. 2. The final boundaries of the Chinese word /Kun4 nan2/ in quiet laboratory environment.

In order to verify the effectiveness of the adoption of the principle of VFR and the utilization of cepstrum coefficients, we tested the endpoint results without using cepstrum described in sections 2.3, i.e. the results acquired by only utilizing energy and zero-crossing in quiet laboratory environment. The errors are listed in parentheses in the third column in table Ⅲ.

TABLE Ⅲ

Test results (errors/tests) of each category in three environments.

Words	Anechoic chamber	Quiet laboratory	Noisy environment
Beginning with *Stopped Consonant*	0/500	3/500 (70/500)	9/500
Beginning with Stop Fricative Consonant	0/740	0/740 (114/740)	6/740
Beginning with Fricative Consonant	0/460	2/460 (74/460)	2/460
Beginning with Nasal Consonant	0/140	0/140 (22/140)	7/140
Beginning with Semivowel	0/120	2/120 (14/120)	5/120
Beginning with Vowel	0/40	0/40 (2/40)	3/40
Ending with Simple Vowel	0/620	2/620 (90/620)	8/620
Ending with *Compound Vowel*	0/600	0/600 (92/600)	12/600
Ending with a Vowel followed by a Nasal Consonant	0/760	5/760 (114/760)	12/760

Ⅳ. CONCLUSION

A noise adaptive endpoint detection algorithm based on time and frequency parameters are described in this paper. The principle of VFR is adopted by this algorithm and the transient between unvoiced and voiced speech segments can be determined. The algorithm was tested on a vocabulary of 100 isolated Chinese words. Tests in three different acoustic environments show that the results of the endpoint detection algorithm are acceptable and this algorithm is effective in most environment, especially in low noisy environment.

Ⅴ. REFERENCES

[1] J-C Junqua, Brian Mak, and Ben Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3; Jul. 1994.

[2] J-C. Junqua, "Robustness and cooperative multimodel man-machine communication applications," *in Proc. Second Venaco Workshop and ESCA ETRW*; Sept. 1991.

[3] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.* , vol. 54, no. 2; , Feb. 1975, pp. 297-315.

[4] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE ASSP*, vol. 29 (4); Aug. 1981, pp. 777-785.

[5] L. R. Rabiner, C. E. Schmidt, B. S. Atal, "Evaluation of a statistical approach to voiced-unvoiced-silence analysis for telephone-quality speech," *Bell System Technical Journal*; Mar. 1977, pp. 455-487.

[6] Hsiao-Fen Pai and Hsiao-Chuan Wang, "A two-dimensional cepstrum approach for the recognition of Mandarin syllable initials", *Pattern Recognition*, Vol. 26, No. 4; 1993, pp. 569-577.

[7] Philippe Le Cerf and Dirk Van Compernolle, "A new variable frame rate analysis method for speech recognition," *IEEE Signal Processing Letters*, vol. 1, no. 12; Dec. 1994, pp. 185-187.

[8] S. M. Peeling and K. M. Ponting, "Variable frame rate analysis in the ARM continuous speech recognition system," *Speech Commun.*, vol. 10; 1991, pp. 155-162.

[9] K. M. Ponting and S. M. Peeling, "The use of variable frame rate analysis in speech recognition," *Computer Speech Language*, vol. 5; 1991, pp. 169-179.

[10] M. H. Savoji, "A robust algorithm for accurate endpointing of speech," *Speech Commun.*, vol. 8; 1989, pp. 45-60.