A C/V SEGMENTATION ALGORITHM FOR MANDARIN SPEECH SIGNAL BASED ON WAVELET TRANSFORMS

Jhing-Fa Wang[*+], *Shi-Huang Chen*[*]

[*]Department of Electrical Engineering, [+]Department of Information Engineering
National Cheng Kung University, Tainan, Taiwan 701, R.O.C.
wangjf@server2.iie.ncku.edu.tw & shchen@cad.ee.ncku.edu.tw

ABSTRACT

This paper proposes a new consonant/vowel (C/V) segmentation algorithm for Mandarin speech signal. Since the Mandarin phoneme structure is a combination of a consonant (may be null) followed by a vowel, the C/V segmentation is an important part in the Mandarin speech recognition system. Based on the wavelet transform, the proposed method can directly search for the C/V segmentation point by using a product function and energy profile. The product function is generated from the appropriate wavelet and scaling coefficients of input speech signal, and it can be applied to indicate the C/V segmentation point. With this product function and the additional verification of energy profile, the C/V segmentation point can be accurately pointed out with a low computation complexity. Experiments are provided that demonstrate the superior performance of the proposed algorithm. An overall accuracy rate of 97.2% is achieved. This algorithm is suitable for Mandarin speech recognition task.

1. INTRODUCTION

Mandarin speech is a monosyllabic language and its structures are different from the other languages, such as English and etc.. The phoneme structure of Mandarin speech can be basically expressed as a consonant (C) followed by a vowel (V). The consonant part may be null. Hence, the description of the Mandarin speech can be represented as follows: [Consonant,]Vowel, where brackets is optional [1]. It is essential that there is a transient region in between consonant and vowel part. Classically, there are 37 vowels and 21 consonants for the constitution of 408 phonologically allowed tone-independent syllables. For the large vocabulary Mandarin speech recognition system, the C/V segmentation is one of the important steps in the preliminary processing. The consonant and vowel part must be identified in order to decode isolated syllable to character. Low consonant recognition rate degrades the possibility for extending the system to large vocabulary size. The main reason is that the coarticulation effect between consonant and vowel parts is an important factor for Mandarin speech recognition [1].

Although various methodologies like Hidden Markov Modeling and Neural Network have been applied to large vocabulary Mandarin speech recognition, the C/V segmentation is used in recognition or training depending on the techniques embedded [1-3]. In [3], syllables are not segmented into consonant and vowel part before recognition. Consonant and vowel parts are segmented and trained using separate consonant and vowel template models. Whole syllable template models are constructed by concatenation of consonant and vowel templates. Syllable recognition is done by whole syllable template matching. For the techniques proposed in [1], the consonant and vowel parts are segmented during recognition and training. Then, the consonant and vowel parts are recognized separately in recognition process. Therefore, C/V segmentation is indispensable both in template training and syllable recognition.

There are many techniques for detecting C/V segmentation points that include counting zero crossing, energy profile, and pitch information [1, 2]. The difficulty in using zero crossing rate and energy profile for detecting C/V segmentation points is in setting the appropriate thresholds. Whether choosing higher threshold or whether setting lower threshold, it will cause detecting error. The other C/V segmentation algorithms presented in [1, 2] begin by estimating pitch information from the peak amplitudes in the vowel part and searching backward across the syllable. Since a peak amplitude will appear in each pitch period, the C/V segmentation point is detected when there is a sharp energy drop within a pitch period. However, the computation complexities of these algorithms are not economical due to estimate pitch positions and backward tracing process. In the meantime, false C/V segmentation point may be caused when the false pitch position appears in the transient region.

To overcome these problems cited above, this paper develops a new algorithm based on the wavelet transform for accurately detecting C/V segmentation point of Mandarin speech signal. The new algorithm defines a product function which is generated from the appropriate wavelet and scaling coefficients of input speech signal. In contrast to the consonant parts, the values of this product function during vowel parts are much larger than those during consonant parts. Therefore, the product function can be applied to indicate the C/V segmentation point. A C/V segmentation point is considered to be correct only when both product function and energy profile are changed sharply during 10ms. From the experimental results, the performance of the proposed algorithm is satisfactory and an overall accuracy rate of 97.2% is achieved.

2. IMPLEMENT OF WAVELET TRANSFORMS

The wavelet transforms considered in this paper are meant the decomposition of a signal with a family of real orthonormal bases $\psi_{m,n}(t)$ obtained through dilations and translations of a kernel function $\psi(t)$ (the "mother wavelet") such as

$$\psi_{m,n}(t) = 2^{m/2}\psi(2^m t - n), \tag{1}$$

where *m* and *n* are integers.

The wavelet $\psi(t)$ can be generated from a companion $\varphi(t)$, which is known as the scaling function and satisfies the two-scale difference equation [4, 5]

$$\varphi(t) = \sqrt{2}\sum_{k} h(k)\varphi(2t-k). \tag{2}$$

Then, the wavelet kernel $\psi(t)$ is related to the scaling function via

$$\psi(t) = \sqrt{2}\sum_{k} g(k)\varphi(2t-k). \tag{3}$$

The $h(k)$ and $g(k)$ in (2) and (3), respectively, are a pair of (quadrature-mirror) lowpass and highpass filters that are related through

$$g(k) = (-1)^{k} h(N-k-1) \tag{4}$$

where N is the number of filter coefficients.

The filter coefficients $h(k)$ and $g(k)$ play a very crucial role in a given discrete wavelet transform and have to satisfy orthonormal and a certain degree of regularity. To perform the wavelet transform dose not require the explicit forms of $\psi(t)$ and $\varphi(t)$ but only depends on $h(k)$ and $g(k)$. In other words, the wavelet transforms can be implemented through the two-channel filter banks which as filtering a signal by a pair of lowpass filter $h(k)$ and highpass filter $g(k)$. Consider a Jth order wavelet decomposition of a function $f(x)$ which can be written as

$$\begin{aligned} f(x) &= \sum_{n} s_0(n)\varphi_{0,n}(t) \\ &= \sum_{n} s_{J+1}(n)\varphi_{J+1,n}(t) + \sum_{n}\sum_{j=0}^{J} w_{j+1}(n)\psi_{j+1,n}(t) \end{aligned} \tag{5}$$

where coefficients $s_0(n)$ are given and the scaling coefficients $s_j(n)$ and the wavelet coefficients $w_j(n)$ at scale j are related to the coefficients $s_{j-1}(m)$ at scale j via

$$\begin{cases} s_j(n) = \sum_{m} h(m-2n)s_{j-1}(m) \\ w_j(n) = \sum_{m} g(m-2n)s_{j-1}(m) \end{cases} \tag{6}$$

where $0 \le j \le J$. Thus, equation (6) provides a recursive algorithm for wavelet decomposition through $h(k)$ and $g(k)$. The coefficients $h(k)$ of Haar and the 4-tap Daubechies wavelet transforms are listed in Table I. Other different sets of the orthogonal filter coefficients $h(k)$ and $g(k)$ can be found in [4, 5].

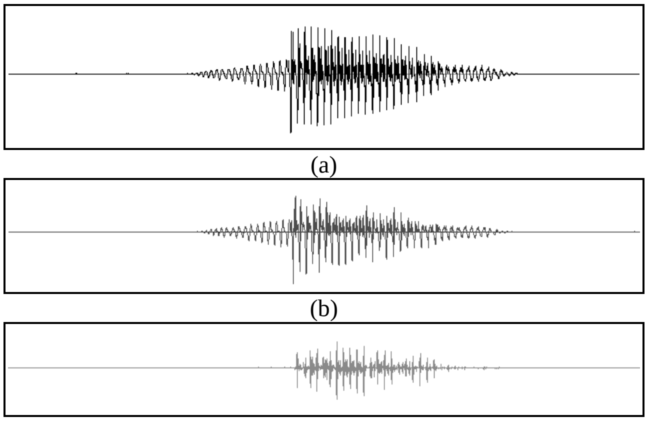
TABLE I
Wavelet Transform Filter Coefficients

	Haar	4-tap Daubechies
$h(0)$	0.7071067811865	0.48296291314453
$h(1)$	0.7071067811865	0.83651630373708
$h(2)$	—	0.22414386804201
$h(3)$	—	-0.12940952255126

3. THE C/V SEGMENTATION ALGORITHM

For Mandarin speech, the consonant parts can be divided into voiced (such as /m/, labeled /ㄇ/, and etc.) and unvoiced (such as /chi/, labeled /ㄑ/, and etc.) classes. By the reason of the differences between the voiced and unvoiced consonants are great, it is very difficult to separate them. Most of the classic algorithms can not directly search for the C/V segmentation point, so they must begin by estimating pitch information from the peak amplitudes in the vowel part and searching backward across the syllable. Above processes embedded in the C/V segmentation algorithms will lead to copious computation complexities.

This paper presents a new algorithm that can directly search for the C/V segmentation point. As a result of unvoiced consonants are composed of high frequency elements and the peak amplitude of the consonant part is much less then that of the vowel part, its scaling coefficients and wavelet coefficients will be very small. Due to the voiced consonants are composed of low frequency elements, its wavelet coefficients will be very small under appropriate wavelet decomposed order. Figure 1 and 2 are shown the second order scaling and wavelet coefficients of the syllable /ma/ (labeled /ㄇㄚ/) and the syllable /chii/ (labeled /ㄑㄧ/), respectively. For display purposes, the data points of scaling and wavelet coefficients shown in Figure 1 and 2 are connected by cubic interpolation.

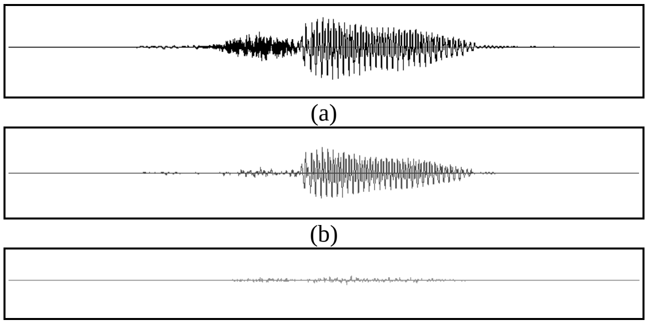


(a)

(b)

(c)

Figure 1. (a) Original waveform of syllable /ma/. (b) The waveform of the second order scaling coefficients of syllable /ma/. (c) The waveform of the second order wavelet coefficients of syllable /ma/.



(a)

(b)

(c)

Figure 2. (a) Original waveform of syllable /chii/. (b) The waveform of the second order scaling coefficients of syllable /chii/. (c) The waveform of the second order wavelet coefficients of syllable /chii/.

Based on the above properties, the consonant and vowel parts of a Mandarin speech can be effectively separated by using the product function, $p(n)$, which is obtained as

$$p(n) = \frac{s_j(n) \cdot w_j(n) \cdot 2^j}{16} \tag{7}$$

where $s_j(n)$ and $w_j(n)$ were given in (6), and j is an appropriate wavelet decomposed order which can be discussed by

$$j = \left\lfloor \log_2\left(\frac{F_S}{1000}\right)\right\rfloor . \tag{8}$$

In the equation (8), F_s (unit: Hz) is the bandwidth of the input speech signals, and 1000 (unit: Hz) is the dominant frequency bandwidth that can ensure the vowel signals are never filtered out and the consonant signals will be suppressed.

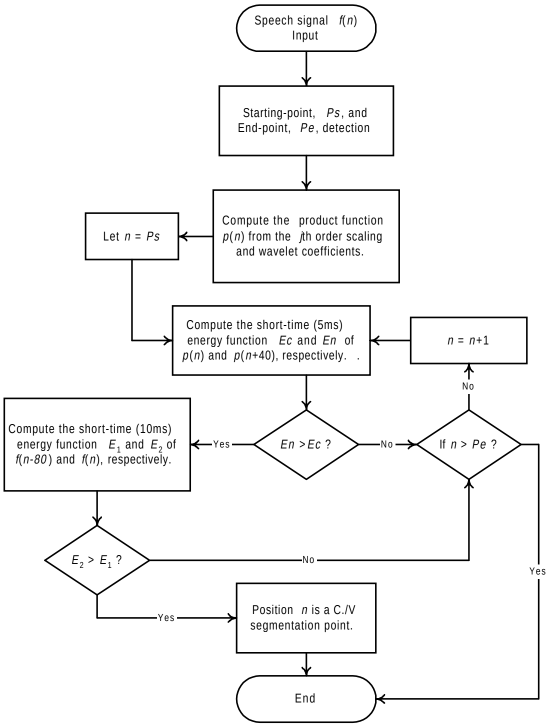

Speech signal f(n)
Input
Starting-point, Ps, and
End-point, Pe, detection
Compute the product function
p(n) from the jth order scaling
and wavelet coefficients.
Let n = Ps
Compute the short-time (5ms)
energy function Ec and En of
p(n) and p(n+40), respectively. .
n = n+1
No
Compute the short-time (10ms)
energy function E1 and E2 of
f(n-80) and f(n), respectively.
Yes
En >Ec ?
No
If n > Pe ?
E2 > E1 ?
No
Yes
Yes
Position n is a C./V
segmentation point.
End

Figure 3. The flow chart of the proposed algorithm.

The flow chart of the proposed C/V segmentation algorithm is shown in Figure 3. Before segmentation, the input speech signal is starting / end-point detected and its algorithm can be found in [6]. The C/V segmentation point will be searched for in between starting-point and end-point. And the short-time energy function E_f of $F(n+a)$ used in Figure 3 is given by

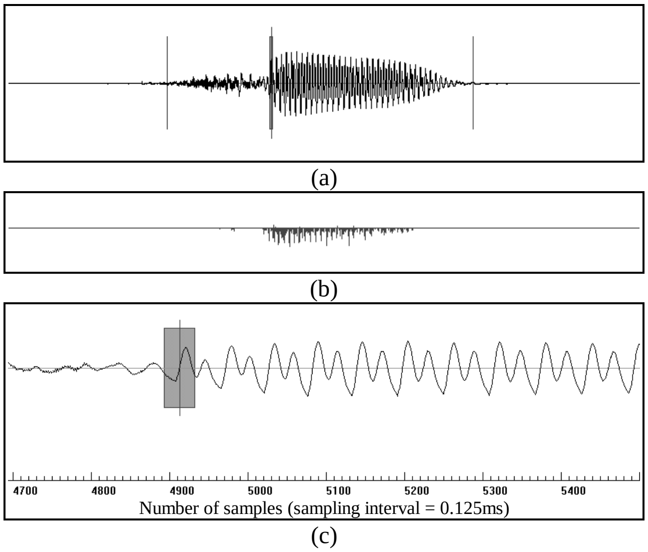
$$E_f = \sum_{x=n+a}^{M} |F(x)| \tag{9}$$

where M will be 80 or 40 depending on the short-time interval be 10ms or 5ms, respectively. The proposed algorithm exploits the characteristic that the values of the product function, $p(n)$, during the consonant parts are much smaller than those during the vowel parts. Therefore, one can directly detect the C/V segmentation point of Mandarin speech by comparing the variation of the product function with checking of energy profile in addition to confirming whether the detected point is a correct C/V segmentation point instead of a noise disturbance.

4. EXPERIMENTAL RESULTS

The performance of the proposed algorithm is tested by 82 Mandarin syllables contain (voiced/unvoiced) consonant part and 18 Mandarin syllables contain only vowel part. These Mandarin syllables are spoken by male and female Chinese and sampled at 8000 Hz with 8-bit resolution. Because of the input signals are sampled at 8000 Hz, the bandwidth of the input signals is 4000 Hz. By the equation (8), the wavelet decomposed order, called j, can be computed to equate 2. As far as different wavelet functions are concerned previously, the 4-tap Daubechies wavelet function gives the better result than the others. Therefore, the following experimental results are generated by using the 4-tap Daubechies wavelet function.

First, The performance of the proposed algorithm in the syllable contains unvoiced consonant is demonstrated. In Figure 4(a), the thin vertical lines located on the left and right terminal are the starting and end point of the syllable /chii/ (labeled /ㄑㄧ/), respectively, and the thick vertical line located on the midpoint is the C/V segmentation point of this speech signal. Figure 4(b) shows the corresponding product function, $p(n)$, of this signal and Figure 4(c) shows the strict location of the C/V segmentation point in this signal. The gray block in the Figure 4(c) represents the C/V transient region and the vertical line located on the middle of the gray block is the C/V segmentation point. From this experimental result, one can observe that the proposed algorithm has successfully separated the consonant and vowel part of the syllable /chii/.



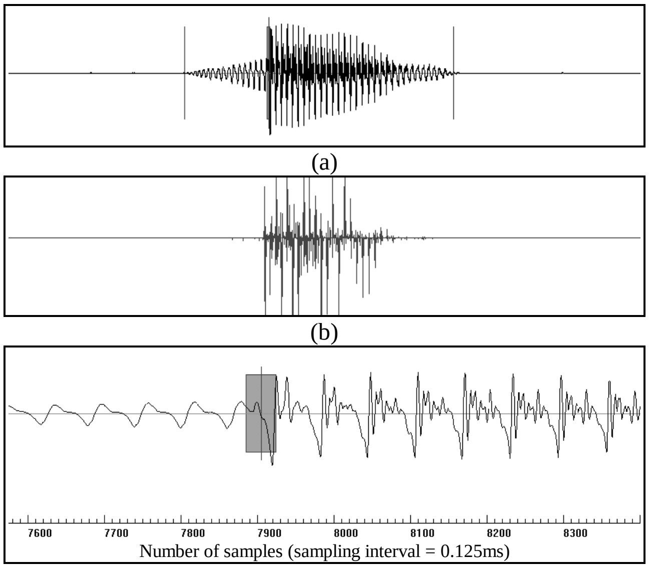
(a)

(b)

(c)

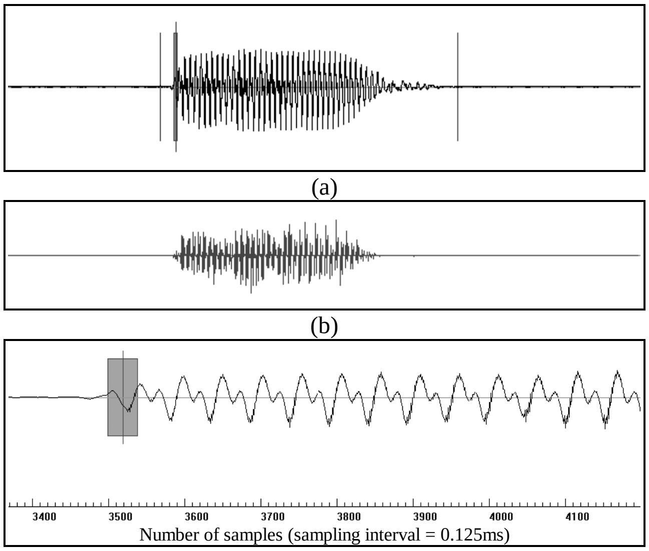
Figure 4. (a) The starting/end points and the C/V segmentation point of syllable /chii/, (b) the corresponding product function, $p(n)$, of above signal, (c) the strict location of the C/V segmentation point in above signal.

Figure 5 shows the experimental result of the syllable /ma/ (labeled /ㄇㄚ/) which contains voiced consonant. Moreover, the explanations of Figure 5(a), 5(b) and 5(c) are the same as Figure 4(a), 4(b) and 4(c), respectively. The C/V segmentation point has again been detected accurately.


(a)
(b)

(c)

Figure 5. (a) The starting/end points and the C/V segmentation point of syllable /ma/, (b) the corresponding product function, $p(n)$, of above signal, (c) the strict location of the C/V segmentation point in above signal.


(a)
(b)

(c)

Figure 6. (a) The starting/end points and the C/V segmentation point of syllable /i/, (b) the corresponding product function, $p(n)$, of above signal, (c) the strict location of the C/V segmentation point in above signal.

Besides segmentation of syllable which contains the consonant and vowel parts, a good C/V segmentation algorithm should be able to handle the syllable which contains only vowel part. Figure 6 shows the experimental result of the syllable /i/ (labeled /一/) which only contains vowel part and the explanations of Figure 6(a), 6(b) and 6(c) are the same as Figure 4(a), 4(b) and 4(c), respectively. The proposed algorithm has been shown that it was also suitable for syllable contains vowel part only. Table II gives the overall results of the proposed algorithm.

TABLE II

Syllable Structure	Accurate Rate (%)
Unvoiced consonant, Vowel	98.6
Voiced consonant, Vowel	99.2
[Null], Vowel	93.8
Average Results	97.2

5. CONCLUSIONS

In this paper, the proposed C/V segmentation algorithm for Mandarin speech is shown to have an excellent performance. Based on the wavelet transform, the proposed algorithm first applies the product function to search for the C/V segmentation point. A result of 97.2% accuracy is obtained in various real Mandarin speech experiments. It is worthwhile to point out two advantages of the proposed algorithm in comparison with the other existing C/V segmentation schemes. First, the proposed algorithm can directly and accurately search for the C/V segmentation point without backward tracing process. Second, the computational complexity of the proposed algorithm is quite simple. The future work will focus on algorithmic development and experimental justification with other languages and the applications of speech recognition.

6. ACKNOWLEDGMENT

This work was supported by the National Science Council (NSC), Taiwan, R.O.C., under Grant NSC 88-2215-E-006-022.

7. REFERENCES

[1] J. F. Wang, C. H. Wu, S. H. Chang, and J. Y. Lee, “A Hierarchical Neural Network Model Based on a C/V Segmentation Algorithm for Isolated Mandarin Speech Recognition,” *IEEE Trans. on Signal Processing*, vol. 39, No. 9, pp. 2141-2146, September 1991.

[2] Stephen W. K. Fu, C. H. Lee, O. L. Clubb, “A Robust C/V Segmentation Algorithm for Cantonese,” IEEE TENCON, pp. 42-45, 1996.

[3] L. S. Lee, C. Y. Tseng, H. Y. Gu *et al*., “ Golden Mandarin (I) - A Real-Time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary”, *IEEE Trans. on Speech and Audio Processing*, vol. 1, No. 2, pp. 158-179, April 1993.

[4] C. Sidney Burrus, Ramesh A. Gopinath and Haitao Guo, *Introduction to Wavelets and Wavelet Transforms.* Upper Saddle River, NJ: Prentice-Hall, 1998.

[5] Gilbert Strang and Truong Nguyen, *Wavelets and Filter Banks*. Wellesley, MA: Wellesley-Cambridge Press, 1996.

[6] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.