음소 음향학적 변화 패턴을 이용한 한국어 음성신호의 연속 모음 분할

박창목* 왕지남*
*아주대학교 산업정보시스템공학과
(cmpark, gnwang)@madang.ajou.ac.kr

Consecutive Vowel Segmentation of Korean Speech Signal using Phonetic-Acoustic Transition Pattern

Chang Mok Park* and Gi-Nam Wang*
*Dept. of Industrial and Information Systems Engineering, Ajou University

Abstract

This article is concerned with automatic segmentation of two adjacent vowels for speech signals. All kinds of transition case of adjacent vowels can be characterized by spectrogram. Firstly the voiced-speech is extracted by the histogram analysis of vowel indicator which consists of wavelet low pass components. Secondly given phonetic transcription and transition pattern spectrogram, the voiced-speech portion which has consecutive vowels automatically segmented by the template matching. The cross-correlation function is adapted as a template matching method and the modified correlation coefficient is calculated for all frames. The largest value on the modified correlation coefficient series indicates the boundary of two consecutive vowel sounds. The experiment is performed for 154 vowel transition sets. The 154 spectrogram templates are gathered from 154 words(PRW Speech DB) and the 161 test words(PBW Speech DB) which are uttered by 5 speakers were tested. The experimental result shows the validity of the method.

1. Introduction

The automatic segmentation and labeling tools becomes one of the most important tasks in speech recognition applications, especially considering recent trends in large speech data-based and applications. Various segmentation strategies have been developed [3,4,5,6], and the vowel-consonant discrimination shows relatively good performance compared to that of consecutive vowel segmentation. Due to the abrupt change of acoustic characteristics between vowels and consonants, the segmentation boundaries between consecutive vowel and consonant are easily identified. However, the identification of segmentation boundaries between consecutive vowels is difficult. Detecting the segmentation boundaries having two consecutive vowel sounds is not easy [8]. Also the voiced-consonant and vowel has an ambiguous boundaries in some cases.

The most segmentation system employ statistical pattern-recognition approaches such as Hidden Markov models (HMMs)[9]. Training phoneme models of HMM based segmentation requires a lot of works to be robust and a certain manual correction needs afterward.

1.1 Automatic phonetic segmentation and labeling using HMM (Hidden Markov Model)[9]

Given phonetic transcription, HMM method is widely used a segmentation tool. This method can be summarized as following.

1) Train all kinds of phoneme HMMs with speech database.
2) Extract features from new speech.
3) Using the phonetic transcription of new speech and HMMs database, generate the series of states of phoneme unit.
4) Using Viterbi algorithm, perform optimum time alignment with reference and new features.
5) Phoneme boundary is extracted from labeling result.

The performance of the above algorithm depends on the training procedure and amount of training speech data. Meanwhile the proposed method just uses the phonetic-acoustic transition characteristics for segmentation.

The procedure consists of two parts, Firstly the voiced-speech is extracted by the histogram analysis of vowel indicator which consists of wavelet low pass components. Secondly given phonetic transcription and transition pattern spectrogram, the voiced-speech portion which has consecutive vowels automatically segmented by the template matching.

2. Voiced-speech Segmentation using histogram analysis

Vowels are composed of low frequency elements(formants-0,1,2,3) and consonants are composed of high frequency elements. Silence signal shows a low energy in overall frequency bands. Signal decomposition in frequency bands are implemented by DWT(Discrete Wavelet Transform) which based on Daubechies' wavelet(order 10). DWT is well localized both in time and frequency domain, so DWT is selected as a analyzing function.[10]

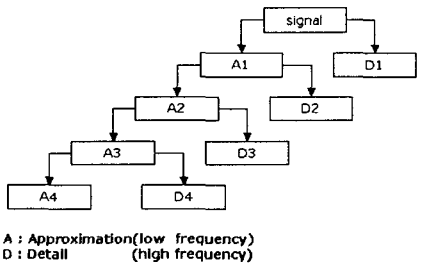


Figure 1. Wavelet decomposition

After decomposition, the RMS of each frame with a shift (3 msecs) and a hamming window(5 msecs) is extracted. The vowel sound(voiced speech) has high amplitude at A4 and D4, so we define E(=0.8*A4+0.2*D4) as a vowel indicator(low frequency component). The histograms of the amplitude of E are calculated

and the vowel threshold (λ) is determined by the peak of histogram. The voiced-signal boundaries are extracted from the threshold.

Table 2. Frequency range of Wavelet decomposition (sampling rate=16 kHz)[10]

Level	Range of Frequency (Hz)
D1	4780 ~ 8000
D2	2914 ~ 4737
D3	1457 ~ 2369
D4	729 ~ 1184
A4	1 ~ 716

λ can be determined by Eq. (1)

$$\lambda = \max(e_{I+\alpha}) \qquad (1)$$

where

$$I = \arg\max_i [h(e_i)], 1 < i < N$$

where e_i is a i^{th} interval of amplitude of E and $h(e_i)$ is frequency of e_i. N and α was selected as 50 and 8 respectively in this experiment. Figure 2 shows a speech signal[g-i-o-k](지옥), wavelet decomposition, E and histogram for e_i

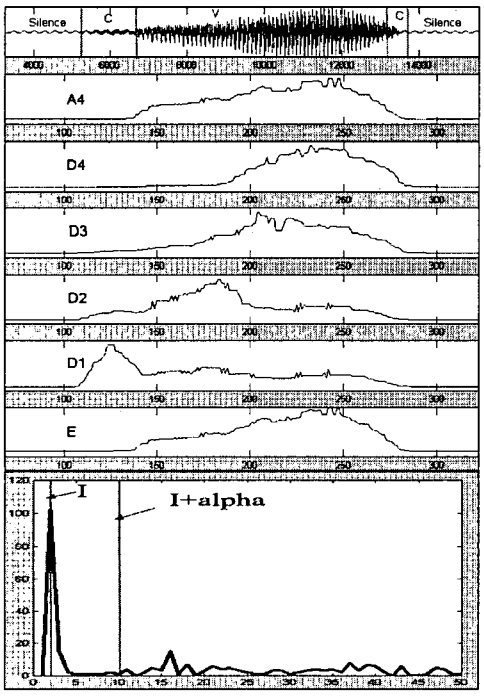


Figure 2. Speech Signal, Decomposition, E and Histogram

3. The proposed template matching method for V-V segmentation

Our research is highly motivated by the belief that the visual representation of spectrogram in vowel transition shows identical characteristics in each various two adjacent vowels. Even though a single speaker is arbitrarily selected for making template sets, the method fits well in multiple speaker environment and various words. The proposed template matching method shows a potential possibility to be a robust boundary detector by considering well-designed template sets.

3.1 Spectrogram

Spectrogram of speech signal contains significant meanings [7]. The temporal formant pattern is well described by spectrogram. One can plot the short-term powers in different frequencies as a function of time. The intensities of spectrogram are also characterized by the degrees of the powers of each frequency. The short-term powers could be estimated by among the discrete Fourier transform, Filter Bank analysis, Linear Prediction analysis, and Wavelet transform. The Linear Prediction analysis is widely used to estimate the smoothed powers in different frequencies and frequencies and shows good capabilities in vowel analysis.

3.2 Vowel transition

It is difficult to segment the adjacent vowel sounds accurately due to co-articulations between neighboring phonemes. In addition, diphthongs make it more difficult to segment consecutive vowel sounds.

The linguistic vowels in Korean are 21 and the phonetic vowel sounds can be abbreviated to 16. So 240 (16×15) vowel transitions can be constituted considering all possible vowel transitions. The vowel sounds can be characterized by first-second formant frequencies and each bandwidth [1]. Those formants range from 200 Hz to 4000 Hz [1]. Also the vowel transition portions can be characterized by the time varying pattern in those frequency bands. Transition portions are based on 100 msec duration as a template portions. Spectrogram image of transition portion is useful as a reference of vowel transition area.

3.3 Template matching using cross correlation

If $f(x,y)$ and $g(x,y)$ are functions of discrete variables, their cross correlation is defined as Eq. (2)

$$f(x,y) \circ g(x,y) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m,n) g(x+m, y+n) \qquad (2)$$

By correlation theorem, Eq. (2) is defined as

$$f(x,y) \circ g(x,y) = \Im^{-1}\{F^*(u,v) G(u,v)\} \qquad (3)$$

where $F^*(u,v)$ is conjugate of Fourier transform of $f(x,y)$, $G(u,v)$ is Fourier transform of $g(x,y)$ and $\Im^{-1}\{\cdot\}$ is inverse Fourier transform. Fast cross correlation can be implemented by Eq. (3). To apply above theorem in this application $f(x,y)$ is defined as a template spectrogram, $g_i(x,y)$, $1 < i < K$, is defined as a i^{th} portion(100 msec duration) of spectrogram of speech signal, $x, 0 < x < M-1$ is defined as a frame index and $y, 0 < y < N-1$ is defined as a frequency index

The closest match can then be found by selecting the spectrogram portion that yields the correlation function with the largest value. Correlation function is usually rearranged so that the peak is located at center like figure 3.

The correlation coefficient is defined as Eq. (4)

$$R_i = \frac{E(f(x,y)g_i(x,y)) - E(f(x,y))E(g_i(x,y))}{\sqrt{Var(f(x,y))}\sqrt{Var(g_i(x,y))}} \qquad (4)$$

Substituting $E(f(x,y)g_i(x,y))$ to $\frac{L}{MN}$ (L: the largest value of cross correlation function), we can estimate the correlation coefficients of two variables of what we don't know the shift information. When the peak of cross correlation locates in center, it represents almost perfect matching. To evaluate a perfect matching, a modified correlation coefficient can be

defined as Eq. (5).

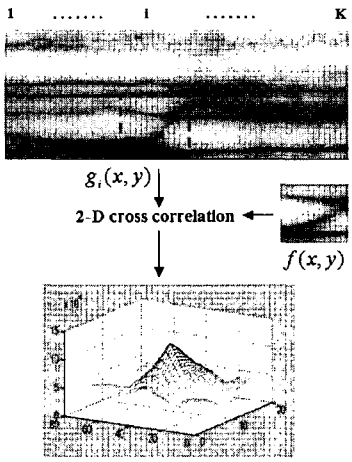


Figure 3. 2-Dimensional cross correlation

$$R'_i = R_i + \delta$$
$$\delta = \sum_{n=N/2-5}^{N/2+5} C_i(M/2,n) - \sum_{n=N/2-5}^{N/2+5} C_i(M-1,n) \quad (5)$$

where $C_i(x,y)$ is the $i-th$ cross correlation function that the peak is located at center. δ measures the relative peakedness or flatness of the middle area of a cross correlation.

Using a transition template spectrogram, we can observe R'_1, $\cdots$, R'_K at K sequential portions of spectrogram of speech signal. We can easily notify that the largest R'_i indicates the vowel transition point.

3.4 Summary of the spectrogram template matching method

1) Establish template sets.
2) For test signal, Compute spectrogram.
3) Using the phonetic transcription of test signal, select a vowel transition template.
4) Compute modified correlation coefficients R'_i, $i = 1, 2, \cdots, K$.
5) Find largest R'_i.
6) Determine vowel transition point.

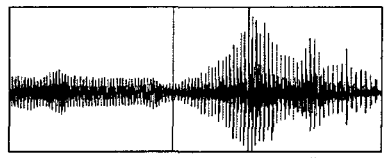
4. Examples

Figure 4 shows the segmentation of speech signal that has two adjacent single vowels $[o-ae]$. It shows (a)speech signal, (b)spectrogram, (c)the sequential R'_i and (d)transition template. The vertical line is the manual segmentation of $[o-ae]$ vowels. Figure 5 shows the segmentation of speech signal that has a single vowel and a diphthong $[eo-ya]$.

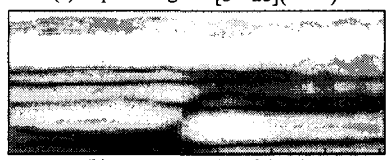
5. Experiments

The speech data used in this experiment was recorded at ordinary in-door environment with pc and headset. The speech signal was sampled at a rate of 16 kHz and 16 bits resolution. Spectrogram is computed with 25 msec window duration and 10 msec frame period. Short time power spectrum is computed by Linear Predictive analysis.

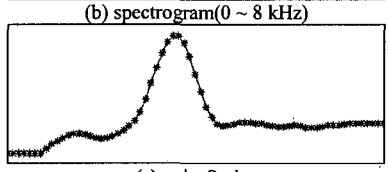
We used a template set consisting of 154 vowel transition spectrograms which are extracted from male speaker A (20 years old) and 154 words(PRW Speech DB). The test speech data consist of 161 test words(PBW Speech DB) that each one includes a vowel transition. The test words uttered by male speaker B (10 years old), male speaker C(20 years old), female speaker D(10 years old), female speaker E(20 years old) and female speaker F(40 years old).



(a) speech signal $[o-ae]$(오애)



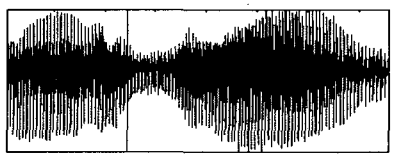
(b) spectrogram(0 ~ 8 kHz)



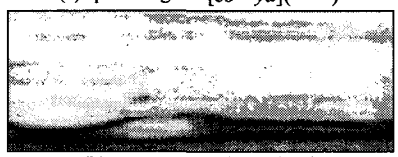
(c) R'_i Series

(d) $[o-ae]$ transition template(0 ~ 4 kHz)

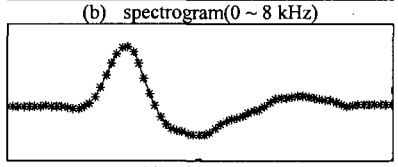
Figure 4. Segmentation of $[o-ae]$(오애) vowels



(a) speech signal $[eo-ya]$(어야)



(b) spectrogram(0 ~ 8 kHz)



(c) R'_i Series

(d) $[eo-ya]$ transition template(0 ~ 4 kHz)

Figure 5. Segmentation of $[eo-ya]$(어야)

In order to evaluate the segmentation accuracy, we compared the template matching method with the manual segmentation. In the 10, 20 and 40 msec error range, 652 utterances(81%), 684 utterances(85%) and 708 utterances(88%) are correctly segmented.

6. Discussion

The experiments for 154 vowel transitions show the validity of the methodology. The HMM based phonetic segmentation has reported 80% agreement within 20 ms for the Korean PBW Speech DB. The HMM approach should tested only considering consecutive adjacent vowels in further comparison. Also further works are under consideration for segmentation of all cases(silence, voiced-consonants, unvoiced-consonants) using phonetic-acoustic transition information.

Detecting segment boundary between adjacent vowels is considered as difficult problem. We proposed a consistent method for determining boundary of consecutive vowel sounds with template matching. The spectrogram gives us significant information of vowel transition, and we use the spectrogram pattern as a template. A modified cross correlation is used for pattern comparison techniques. A promising results are obtained: detection accuracy of segmentation boundaries between the consecutive vowels is higher than 85% agreement within 20 ms. There is high possibility for increasing the accuracy by designing robust template sets.

7. Reference

[1] L.R. Rabiner, B.H. Juang, Fundamentals of speech Recognition, Prentice Hall, 1993

[2] R.C.Gonzalez, R.E. Woods, Digital Image Processing, Addison-Wesley, 1992

[3] R. Andre-Obrecht, A new statistical approach for the automatic segmentation of continuous speech signals, IEEE Trans. ASSP 36(1988), 29-40.

[4] P.Cosi, SLAM: A PC-based multi-level segmentation tool, Speech Recognition and Coding: New Advances and Trends (A.J.R. Ayuso and J.M.L. Soler, eds.), Computer and System Sciences, vol. 147, Springer-Verlag, New York, 1995, pp. 124-127.

[5] T.Fukada, S. Aveline, M. Schuster, Y. Sagisaka, Segment Boundary Estimation Using Recurrent Neural Networks, Proc. Eurospeech '97,1997

[6] L. Hu, S. Imai, C. Furuichi, Phonetic Segmentation for Continuous Madndarin Speech Recognition, J. Acoust. Soc. Jpn. (E)18.1, 1997.

[7] D. X. Sun. Robust estimation of spectral center-ofgravity trajectories using mixture spline models. In Proceedings of the 4th European Conference on Speech Communication and Technology, Madrid, Spain, pages 749-752,1995

[8] J.H. Lee,S.B Rhee, A study on consoant/vowel phonetic segmentation of Korean isolated words based on a rule-based system for the phenomenon of Korean Vocalization, 1999 IEEE TENCON, 1999

[9] F. Brugnara, D. Falavigna and M. Omologo, Automatic segmentation and labeling of speech based on Hidden Markov Models, Speech Communication, vol. 12, no. 4, 357-3

[10] Beng T Tan, Robert Lang, Heiko Schroder, Andrew Spray and Phillip Dermody, Applying wavelet anaysis to speech segmentation and classification, Wavelet Applications, vol. Proc. SPIE 2242. pp 750-761, 1994