On Detecting the Steady State Segments of Phonemes by Using the Magnitude Distribution of Speech Waveforms.

음성파형의 진폭분포를 이용한 음소의 정상상태 구간 검출

(Deok Cho Chung, Myung Jin Bae, Sou Guil Ann)

정 덕 조*, 배 명 진*, 안 수 길**

ABSTRACT

To recognize continued speech, it is necessary to segment the connected acoustic signal into phonetic units. In this paper, as a parameter to detect the steady state segments in continued speech, we propose a new magnitude distribution. The suggested parameter represents a change rate of the magnitude of speech signals. As comparing this distribution with the other in adjacent frame, the state of the frame can be distinguished between steady state and transient state.

요 약

연속음 인식을 위하여 연결된 음향 신호를 음소단위로 분할하는 것이 필요하다. 본 논문에서는 연속 음성에서의 정상상태 구간 검출을 위한 파라미터로서 진폭분포를 이용하는 방법을 제안하였다. 제안된 진폭분포는 음성신호의 변화특성을 정확히 나타내며 이러한 프레임사이의 진폭 분포 차이값을 비교하여 프레임의 안정구간과 천이구간을 구분할 수 있었다.

I. Introduction

It is difficult to segment automatically the variation of phonological structure, due to the coarticulation between phonemes uttered continuously. If phonemes combined complicatedly can be segmented in preprocessing, the recognition technique for the isolated words can be extended up to the connected and continuous words. Also the computation complexity in analysis for recognition can be reduced.

*Hoseo Univ.
**Seoul National Univ.

Segmentation methods after Zue[1] have been developed actively. Zue's method is to perform the classification of phonemes through the complicated decision rule using several feature parameters. The parameters include the linear predictive coefficients (LPC), the prediction errors, the spectrum energy, the pitch, and the formant informations.

The study on the detection of the transient segments can be classified into three groups : the time domain methods, the spectrum domain methods, and the hybrid methods. The time domain methods have the advantages of simplicity in calculation, and several methods using the continuity of VOT(Voice Onset Time) and the energy

contour of speech signal in time domain have been proposed[4-5, 7]. In the spectrum domain methods, several methods using the transient characteristics of the formants have been proposed[2-4]. The hybrid methods also use feature parameters[3-6] in the conversion domains, such as variation features of the LPC coefficients and errors.

The detection of parameters in the time domain is easy, but the decision logic is difficult compared to other domains. The spectrum domain or the hybrid methods are accurate but they are easy to be influenced by orders of transformation[8]. And the complexity of the calculation is greater than that of the time domain, in the aspect of the preprocessing.

In this paper, the problems about the detection of the steady state segments by the energy contour mainly used in the time domain methods are reviewed and a new parameter to solve these problems for the segmentation is proposed.

Ⅱ. Detection of the Steady States Segment Using the Magnitude Contour

Phonemes of continuous or connected words are varying in time. The acoustic waveform for utterance /ouyukou/ is depicted in Fig 1(a) and the average magnitude contour in Fig 1(b), in which frame length is 20msec and overlapped by 10msec. Fig 1(b) shows well the overall variations of phonemes. There are three high peaks and two deep valleys in Fig. 1. We can see each peak corresponds to syllable /ou/, /yuk/, /ou/ and each valley is the combination frame of articulation /ou/ and /yuk/, /yuk/ and /ou/. Because these frames are transition segments varying between peaks and valleys in the average magnitude contour, the analysis result for these frames are complicated by the mixture of two phonemes.

Henec if we can find the frames with magnitude peaks, consequently the analysis can be reduced. However, because the averabe magnitude countour may have local peaks and with slow slope extended over serval frames can be appeared, the detection of frames in the steady states is not so easy.

For the detection of steady states segment of phonemes using the average magnitude contour, the calculation of the average magnitude is influenced by the window applied to frame. Influences by the window are by the length and by the type of the window. A good choice of window length for the voiced segments would be a windowing having a multiple duration of the pitch period. Though we adjust the window length as being multiples of the pitch period, speech components which are not multiples of pitch period can be existent. Typical examples of window type are rectangular window, Hamming window, and Blackman window according to the ratio of the pass bandwidth and the stop bandwidth.

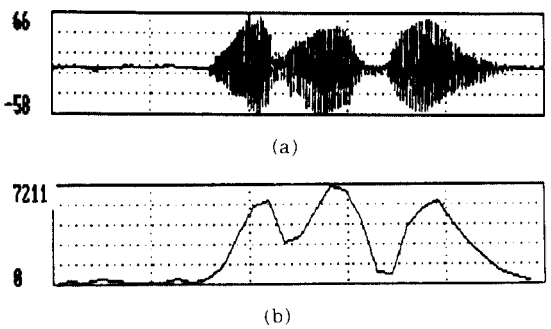


Fig.1 Energy contour for connected word speech /OUYUKOU/ :
(a) Speech waveform for /OUYUKOU/ :
(b) Energy contour.

The average magnitude contour is dependent on the length of window. If the length of window is longer than the fundamental period, the average magnitude contour can't show well the rapid varying features of phonemes. On the other hand, if the length of window is too short, it is difficult

to know the variation features of phonemes due to local peaks in the average magnitude contour.

In addition to the good choice of the window to obtain the energy contour for continuous speech, decision logic to show the phoneme variation is necessary. Two difficulties are anticipated in the application of the decision logic. First, the local peaks unnecessary to decision logic are appeared in the average magnitude contour due to the complexity and the variety of speech components, in spite of the good choice of the window. It is difficult to isolate these local peaks and the true peaks of phonemes. Second, there are various types of phonemic peaks. For example, it is difficult to find a standard type of peak in the connections of voiced sounds and nasals, unvoiced sounds and nasals or unvoiced sounds and voiced sound.

Ⅲ. Magnitude Distribution of Speech Waveforms

It is usual to analyze the speech signals by a frame unit because it is changing slowly compared to the waveform fluctuation. One convension techniques is to decide whether the present frame is in the transient segment or in the steady state is that compares the ratio of the average magnitude of the first half frame to the latter half one :

$$MR(fr)=\frac{\sum_{k=N/2}^{N-1}|S(n-k)|}{\sum_{k=0}^{k=N/2-1}|S(n-k)|} \qquad (1)$$

Where n is the starting point of the analysis frame, and N is the length of the frame. This ratio shows the average magnitude ratio of the adjacent frames and it is also influenced by the window as we discussed at the chapter Ⅱ.

We suggest the magnitude distribution diagram (MDD) as a new parameter to detect the state of the present frame :

$$MDD(\ fr,\ s(n)+\beta\ ) = MDD(\ fr,\ s(n)+\beta\ )+1, \quad n=0,\ 1,\ \cdots\cdots,\ N-1 \qquad (2)$$

where fr is the number of the present frame, s(n) is the speech, and β is the bias value to make the speech sample into a positive value. For example, if the speech waveforms are sampled by 8-bits, the level of the magnitude is ranging over -128 to +127, thus β is 128.

The MDD, as defined in the equation(2), shows the normalized distribution with N levels. Because the influence by edges of window to magnitude distribution is lessened by effect of 1/N, the influene by the window type is lower compared to the magnitude contour. And if the window length become twice, the average magnitude contour is smoothed but the smoothing effect on the distribution is reduced comparing to using the energy contour, because the distributions are added by the isolated 1/N level. The Fig. 2 shows the

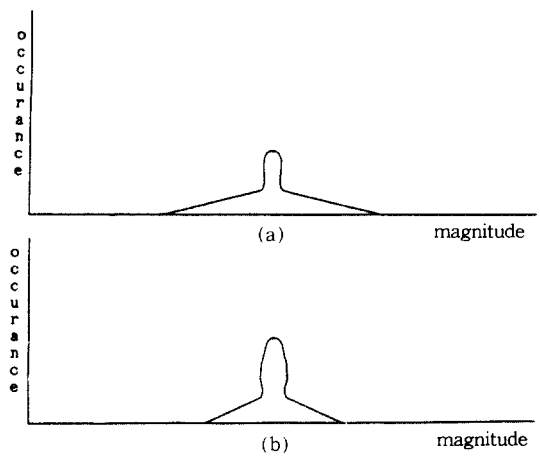


Fig.2 Example of the magnitude distribution diagram for a voiced and a unvoiced segment :
(a) The magnitude distribution diagram for a voiced segment :
(b) The magnitude distribution diagram for an unvoiced segment.

magnitude distribution for the voiced sound /AH/ and the unvoiced sound /S/. The horizontal axis indicates the speech level and the vertical axis indicates the number of samples. We can see in the Fig. 2 that magnitude distributions for the unvoiced sound are concentrated near the zero level and that the magnitude distribution for the voiced sound shows the feature with slow slope.

Ⅳ. Detection of the Steady State Segment Using Magnitude Distribution

The normalized average difference between the adjacent distributions is defined as follows :

$$NAD(fr) = \frac{\sum_{i=0}^{2^b-1} (MDD(fr,i) - MDD(fr-1,i))}{2N} \quad (3)$$

Where b is the number of bits to quantize the speech waveforms. If the average difference value is closer to 1, it shows the transient state of phonemes ; otherwise if it is closer to 0, the frame is steady state segment.

The example of computation by Eq.(3) for speech signal /SAM/ is depicted in Fig.3. Fig. 3(a), (b), and (c) shows the speech waveforms, the average magnitude contour, and the difference contour, respectively. This simulation result is obtained by overlapping each half segment of adjacent frames.

It can be seen the magnitude distribution is larger at the beginning of phoneme. And comparing to the energy contour, the average difference value is larger and shows peaks at the transient frame. On the otherhand, if the difference value constitutes a valley, we can judge the frame is in the steady state.

Also, it can be seen that the difference value

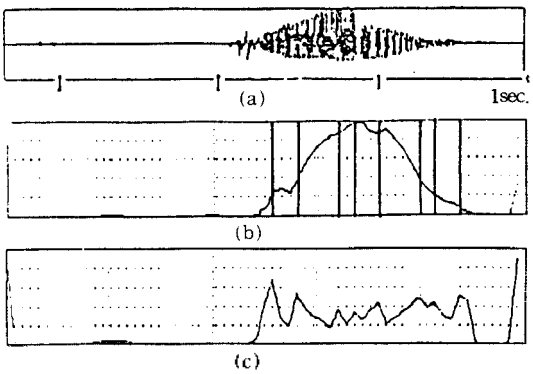


Fig.3 Average difference between adjacent magnitude distributions for speech utterance /SAM/ :
(a) Speech waveform for /SAM/ :
(b) Energy contour ;
(c) Difference contour between the adjacent distributions.

at sound /AH/, which has more complicated waveform, is larger than that at sound /M/. The slope of peak at the connected segment of consonants or at the end of phonemes, of which variations of phonemes are fast, is steeper than at the voiced segment.

Thus we can get the decision logic using the average difference value of the distribution. We can say the present frame is steady state if the difference contour shows valley.

Ⅴ. Experimental Verification and Discussion

We used an IBM PC/AT with a 12-bit A/D converter for the simulation. Two men and one woman speakers uttered continuous voice as follows:

utterance 1) 24 years old male speaker :
/INSOONAE KOMAGA CHUNJAE SONYUNWL JOAHANDA/

utterance 2) 28 years old male speaker :
/HOSEODAE JUNJAKONGHAKWA WMSUNGSINHOCHURI YUNGUTIM/

utterance 3) 25 years old female speaker :
/KAMSAHAMNIDA/

The samples with 8KHz sampling rate ware quan - tized in 12bits.

In analysis, the length of one frame is 256 samples and each adjacent frames are overlapped by 128 samples. We used the buffer with the size of 2^{12}=4096 words for the magnitude distributions per each frame. The waveform, the average en- eragy magnitude, the difference contour between the adjacent frames, the steady states for utterence 1 are depicted in Fig. 4(a), (b), (c), and (d), respectively.

We can see that the variation feature can be classified by the difference contour and the valleys in the difference contour show the steady state of phonemes well. Though the beginning of the speech waveforms in the third part of the Fig.4 is the connected word without variation, the dif- ference contour of the magnitude distributions classifies it correctly.

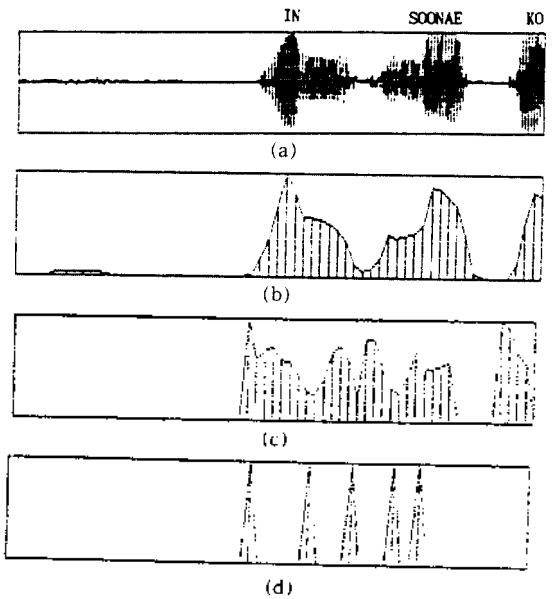


Fig.4-1 Simulation results for speech / INSOONAE KOMAKA CHUNJAE SONYUNWL JOAHANDA / :
(a) Speech waveform for / INSOONAE KOMAKA CHUNJAE SONYUNWL JOAHANDA / :
(b) Energy contour : (c) Difference contour between the adjacent distributions : (d) Steady segment.

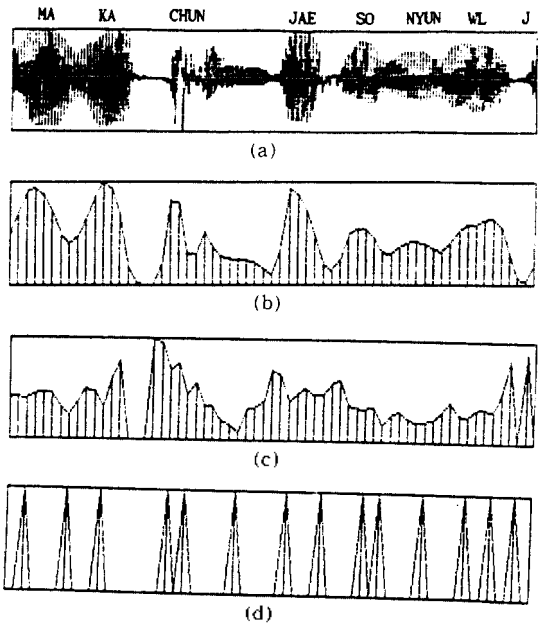


Fig.4-2 Simulation results for speech / INSOONAE KOMAKA CHUNJAE SONYUNWL JOAHANDA / :
(a) Speech waveform for / INSOONAE KOMAKA CHUNJAE SONYUNWL JOAHANDA / :
(b) En ergy contour : (c) Difference contour be- tween the adjacent distributions : (d) Steady seg- ment.

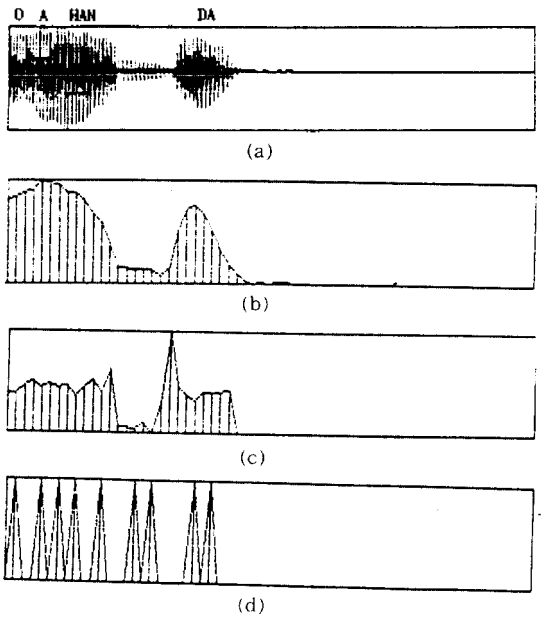


Fig.4-3 Simulation rsesults for speech / INSOONAE KOMAKA CHUNJAE SONYUNWL JOAHANDA / :
(a) Speech waveform for / INSOONAE KOMAKA CHUNJAE SONYUNWL JOAHANDA / :
(b) Energy contour : (c) Difference contour between the adjacent distributions : (d) Steady segment.

Ⅵ. Conclusion

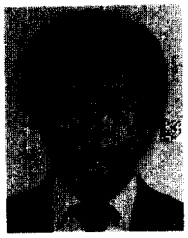
If the segmentation of the syllable unit is performed well, techniques used in the isolated word can be applied for connected or continued speech recognition. Various methods for the detection of the steady state segment so far have been proposed, and the method to detect it using the average magnitude contour is easy. But the average magnitude contour is influenced by the applied window and the decision logic is complicated.

In this paper, we propose a new method which uses the average difference value in the normalized magnitude distribution to detect the steady state segment. Using the proposed method, we can see the variation feature of the continued speech without being influenced by the length or the type of the window. This algorithm is easy to detect the steady state segment by the simple decision logic and it is possible to understand approximately the characteristics of the voiced word.

References

1. C.J. Weinstein, S.S. McCandless, L.F. Mondshenin, and V.W. Zue, "A System for Acoustic-Phonetic Analysis of Continuous Speech", IEEE Trans, on ASSP, Vol. ASSP-23, No.1, pp.54-67, Feb. 1975.
2. W.F. Ganong, and R.J.Zatorre, "Measuring Phoneme Boundaries Four Ways," J. Acoust. Soc. Am, Vol. 68, No.2, pp.431-439, Aug, 1980.
3. S.J. KIM and two man, "A Segmentation Algorithm of the Connected Word Speech by Statistical Method", KITE REVIEW, Vol. 26, No.4, pp. 151-162, Apr., 1989.
4. R. Mori, P. Laface, and E. Piccoo, "Automatic Detection and Description of Syllabic Features in Continuous Speech", IEEE Trans. On ASSP, Vol. ASSP-24, No.2, pp. 880-883, Oct., 1976.
5. L.R. Rabiner, and M.R. Sambur, "Some Preliminary Experiments in the Recognition of Connected Digits", IEEE Trans. on ASSP, Vol. ASSP-24, No.2, pp. 170-182, Apr., 1976.
6. R. Mori, and P. Laface, "Use of Fuzzy Algorithms for Phonetic and Phonemic Labeling of Continuous Speech", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAM-2, No.2, pp. 436-448, Mar., 1980.
7. M. BAE, J. R.HHEM, and S ANN, "A Study on the Energy Extraction using G-Peak from the Speech Production Model", KIEE, Vol. 24, No.3, pp.381-386, May, 1987.
8. M. BAE and S. ANN, "On Improving the Effects of Varying the Window Length on Speech Energy Computation", J., Acoust., Soc., Korea, Vol.9, No. 2, pp. 34-41, April, 1990.

▲Deok Cho Chung



was born in Incheon, Korea on May 7, 1954. He received B.S. degree in 1977 from the Civil Aviation Colleage of Korea and M.S. degree in 1991 from Hoseo University both in electronics engineering. Since 1977 he has been with the system integration division of the missile system development group, ADD(Agency for Defence Development), Daejon, Korea. His current research interests are speech signal processing and its applications.

▲Souguil Ann

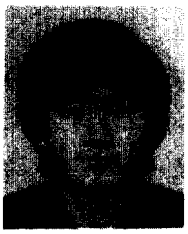
1956.5 : Department of Electronics Engineering, Seoul National University (B.S)

1957.4 : Department of Electronics Engineering, Seoul National University (M.S)

1974.3 : Department of Electronics Engineering, Seoul National University (Ph.D)

1989.1~ : the president of the Acoustical Society of KOrea

▲Myung Jin Bae



1981.2 : Department of Electronics Engineering, Soongsil University (B.S)

1983.2 : Department of Electronics Engineering, Seoul National University (M.S)

1987.8 : Ph.D course from Department of Electronics Engineering, Seoul National University

1986.3~ : Department of Electronics Engineering, Hoseo- University