# An Automatic Diphone Segmentation for Korean Speech Synthesis-by-Rule

# 한국어 규칙 합성을 위한 다이폰의 자동 추출

InJong Jeong*, YounJeong Kyung**, HanWoo Kim*, YangHee Lee**

정인종*, 경연정**, 김한우*, 이양희**

## ABSTRACT

In this paper, a method is proposed for automatically segmenting diphones from two-syllabic natural speeches for speech synthesis.

The natural speeches are analyzed by Improved Cepstral parameters from which diphone extraction parameters are derived. These parameters express the dynamic variation of energy level(zero-th cepstrum coefficient) and spectral envelopes(cepstral coefficient), also express zero crossing rate and cepstral Euclidean distance. As to detect ambiguous phoneme boundaries in vowel-vowel pairs, we split spectral envelopes into fine structure and smooth structure, and use two parameters for the dynamic variation of them.

About 120 words consisted of VV (vowel-vowel), VCV(vowel-consonant or semivowel-vowel) and VCCV (vowel-consonant-consonant-vowel) are tested for diphone segmentation.

The result of this test verify that the phoneme boundaries can be detected at almost 85% accuracy in spite of including many VV types in test words. The listening test proved that the speech synthesized by the diphones is very intelligible.

## 요 약

본 논문에서는 무제한 음성 생성을 위한 단위음성으로서의 다이폰을 2음절 자연음성으로부터 자동 추출하는 알고리즘을 제안한다. 입력음성을 개량켑스트럼 파라미터로 분석하여 이로부터 다이폰 추출 파라미터들을 도출한다. 제안된 파라미터로는 에너지 레벨을 나타내는 0차 켑스트럼의 동적변화량, 스펙트럼의 시간 변화량, 영교차율, 켑스트럼의 유클리디안 거리이다. 스펙트럼 포락의 변화가 완만한 모음연쇄등의 음소 경계를 보다 효율적으로 검출하기 위해 스펙트럼의 시간 변화를 미세 부분과 개형부분으로 나누어 각각을 파라미터로 사용한다. VV(모음연쇄), VCV(C : 반모음, 자음), VCCV형들로 이루어진 2음절 단어들에 대해 실험한 결과, 모음연쇄 동이 포함되어 있음에도 약 85% 정확도의 음소경계검출을 얻었다. 본 논문에 의한 다이폰을 이용한 합성음의 청취실험 결과 명료도가 높음을 확인하였다.

*Dept. of Computer Science & Engineering, Han-Yang University

**Dept. of Computer Science & Dong-Duck Women's University

접수일자 : 1993. 4. 27

## Ⅰ. INTRODUCTION

The speech synthesis units should satisfy naturality and intelligibility on concatenation. A diphone is generally defined as a speech unit which starts in the stable state center of a phone, and ends in the stable state center of next phone and includes the transitions between the two phones. Since the boundaries of diphone are steady state center of phones, spectral distortions by concatenation are reduced and it is possible to generate intelligible speech[1].

In the case of systematic extraction of such diphones, there are several advantages over manual extraction method. The systematic approaches make it possible to prepare consistent diphones of speech data base and to reduce the amount of time to extract diphone elements. As the methods to extract diphones, there are studies using HMM [1] or centroid[2], and they show good result but it is difficult to detect phoneme boundaries in semi-vowel to vowel or vowel to vowel pairs. The study of phoneme detection for speech recogniton[3], [4] shows good result on phoneme segnentation. This method used mel-cepstrum from 1st order to 7th order to obtain the dynamic variation of spectral envelopes. But in this case vowel pairs are not included in tested words.

In this paper, two-syllable nonsense words are used as source words on diphone segmentation which consist of vowel to vowel pairs, vowel to semi vowel to vowel pairs, VCV units and VCCV units.

We introduce parameters into automatic diphone segmentation which represent fine structures and smooth structures of spectral envelope expressed by low and high order of improved cepstral coefficients. We propose an algorithm to detect phoneme boundaries and to extract diphone units automatically.

In section II, speech analysis system is described, in section III, the parameters to extract diphones are described. And in section IV, diphone extraction algorithm is proposed and described and synthesis-by-rule by using diphone units is described in section V, and in section VI, VII, experimental results and conclusion is described repectively.

## Ⅱ. SPEECH ANALYSIS/SYNTHESIS SYSTEM

In order to derive parameters to extract diphone elements, we use analysis system specified below:

– A/D CONVERT: 5 kHz LPF, 10 kHz Sampling,<br>
12bit Quantization<br>
– ANALYSIS PARAMETER: Improved Cepstrum<br>
– ANALYSIS(VOCAL TRACT APPROXIMATION)<br>
Frame period - 10ms for male<br>
5ms for female<br>
Window function W = 25.6ms Blackman window<br>
Wp = 40ms Blackman window<br>
Cepstrum order - 30 for male<br>
25 for female<br>
Spectrum Envelope by Improved cepstral method<br>
Voiced/Unvoiced determination -<br>
by Spectrum Envelope parameter<br>
Pitch Determination - by Cepstrum peak picking<br>
– SYNTHESIS FILTER: LMA(Log Magnitude<br>
Approximation) Filter[8]

In the first place, input speech signal is analyzed into improved cepstral parameters[5], and from those parameter we derives segmentation parameters.

## Ⅲ. PROPOSED PARAMETERS TO EXTRACT DIPHONE ELEMENTS

Because diphones include transitional interval [6], phoneme boundaries should be detected acculately prior to the diphone extracting process. The parameters used to detect phoneme boundaries are voiced/unvoiced detection parameter, dynamic variation parameter of zero-th cepstral coefficient[3][4], two dynamic variation parameters of log spectral envelopes represented by high and low order cepstral, zero crossing rate and cepstral

Euclidean distance.

Figure 1 shows waveform and log spectral envelope of /ede/ of male speaker and phoneme detection parameteres B(i), a(i), e₁(i), e₂(i) and zero crossing rate.

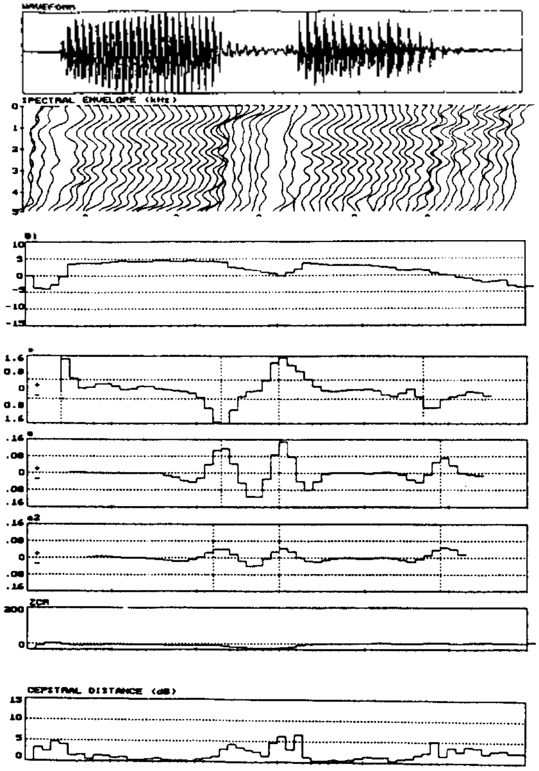


Figure 1. Waveform of utterance /ede/ and segmentation parameters

Parameter B(i) defined by (1) is determined by the averaged value of the spectral envelope between 120Hz and 400Hz of frequency band for male, between 200Hz to 480Hz for female, since in the selected band voiced speech has high energy owing to first formant and unvoiced speech has low energy. Parameter B(i) is used to classify the voiced/unvoiced interval.

For male, parameter B(i) is

$$B(i) = \frac{1}{8} \sum_{l=3}^{10} V_i(\Omega_l) \quad (\Omega_l = \frac{2\Pi l}{N}, N = 256) \qquad (1)$$

and for female,

$$B(i) = \frac{1}{8} \sum_{l=5}^{12} V_i(\Omega_l) \quad (\Omega_l = \frac{2\Pi l}{N}, N = 256)$$

where Vi is log spectral envelope.

Parameter a(i) is obtained by quasi-derivative filtering of zero order improved cepstral coefficient. Zero order cepstral coefficient represents the energy level of waveform. Parameter a(i) is defined by (2) where window function w(n) is Blackman window and 2M+1=9. The window size M is got empirically.

$$a(i) = K_M \sum_{n=-M}^{M} w(n) n c_{i+n}$$

$$K_M = (\sum_{n=-M}^{M} w(n) n^2)^{-1} \qquad (2)$$

In voiced interval, a phoneme is detected between local maximum value and local minimum value of parameter a(i). In VCV units, most of phoneme boundaries in voiced interval are detected.

In the case that source words have smooth dynamic variation of spectral envelope like vowel pair, the phoneme boundary detection rate by parameter a(i) is low, so we use two dynamic variation parameters of spectral envelopes. In order to obtain the dynamic variation of spectral envelopes, we define dynamic variation quantity of spectal envelope, $d_i$, by (3).

$$d_i = (\sum_{m=a}^{b} (d_i^m)^2)^{\frac{1}{2}}$$

$$d_i^m = K_M \sum_{n=a}^{b} w(n) n c_{i+n}[m] \qquad (3)$$

Where a=10, b=25 in high order part of cepstral coefficients, and a=1, b=10 in low order part. For high order part 2M+1=9 and for low order 2M+1=15, window function w(n) is Blackman window of size M. The smoother the variation is, the longer the window size is.

The quantity, $e_1(i)$ and $e_2(i)$ spectral envelopes dynamic variation of are defined by (4) and (5), respectively. The waveform of utterance /ada/ of female speaker, and its dynamic variation of spectral envelope obtained from whole cepstral coefficients, low order part and high order part of cepstral coefficients are shown in figure 2. In figure 2, spectral envelope obtained by high order part cepstral shows fine structure and spectral envelope obtained by low order part cepstral shows smooth structure.

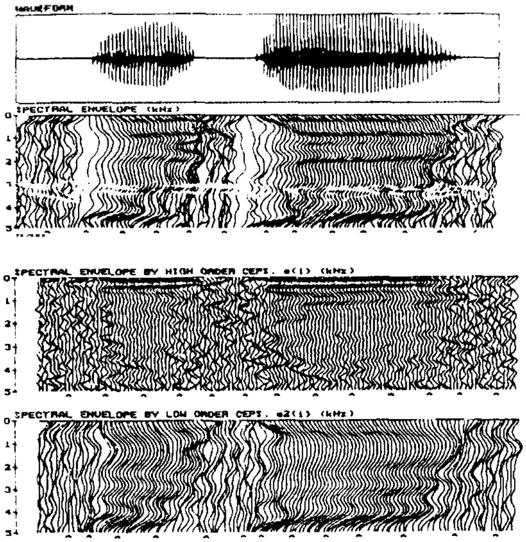


Figure 2. Spectral envelopes from high order and low order cepstral coefficients, utterance of /ada/ of female speaker

The parameter $e_1(i)$ represents the dynamic variation of the fine structure of spectral envelopes obtained by high order part of cepstral coefficients, and $e_2(i)$ represents it of the smooth structure of spectral envelopes obtained by low order part of cepstral coefficients.

$$e_1(i) = -L_4 \sum_{n=-4}^{4} w(n)(n^2-\overline{n_w}^2)(d_{i+n}-\overline{d_w})$$

$$(L_4 = (\sum_{n=-4}^{4} w(n)(n^2-\overline{n_w}^2)^2)^{-1}),$$

$$\overline{n_w}^2 = \frac{1}{\overline{w}} \sum_{n=-4}^{4} w(n)n^2,$$

$$\overline{d_w} = \frac{1}{\overline{w}} \sum_{n=-4}^{4} w(n)d_{i+n}, \quad \overline{w} = \sum_{n=-4}^{4} w(n)) \qquad (4)$$

Where $|n|>4$ for male spealcer, $|n|>7$ for female spealcer, Blackman window w(n) is 0.

$$e_2(i) = -L_7 \sum_{n=-7}^{7} w(n)(n^2-\overline{n_w}^2)(d_{i+n}-\overline{d_w})$$

$$(L_7 = (\sum_{n=-7}^{7} w(n)(n^2-\overline{n_w}^2)^2)^{-1}),$$

$$\overline{n_w}^2 = \frac{1}{\overline{w}} \sum_{n=-7}^{7} w(n)n^2,$$

$$\overline{d_w} = \frac{1}{\overline{w}} \sum_{n=-7}^{7} w(n)d_{i+n}, \quad \overline{w} = \sum_{n=-7}^{7} w(n)) \qquad (5)$$

Where $|n|>7$ for male spealcer, $|n|>10$ for female spealcer, Blackman window w(n) is 0.

So the parameter $e_1(i)$ and $e_2(i)$ are independent parameters and used to detect phoneme boundaries in voiced interval.

Conceptually, the peak values of $e_1$ and $e_2$ in voiced intervals represent the phoneme boundary candidates. Parameter $e_2(i)$ is specially useful to detect phoneme boundaries in vowel pair. Figure 3 shows dynamic variation of spectral envelopes obtained by whole cepstral coefficients. In this case, incorrect detections of phoneme boundaries occur, i.e., insertion error occurs in parameter e (1 25).

In unvoiced interval, zero crossing rate determined by (6), is used to segment phoneme boundary between unvoiced consonants and silence.

$$Z(n) = \sum 0.5\,|\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]|\,w(n-m)$$

$$\text{where, } \mathrm{sgn}[x(n)] = 1, \quad x(n) \geq 0$$
$$= -1, \quad x(n) < 0 \qquad (6)$$

The Euclidean distances between adjacent frames defined by (7) are used to extract diphone boundaries when phoneme boundaries are to be found accurately.

$$Ed(n) = (\mathrm{dist}(n)^2 - \mathrm{dist}(n-1)^2)^{\frac{1}{2}}$$

$$\mathrm{dist}(n) = \sum_{n=0}^{30} c_i(n) \qquad (7)$$

where n is frame number.

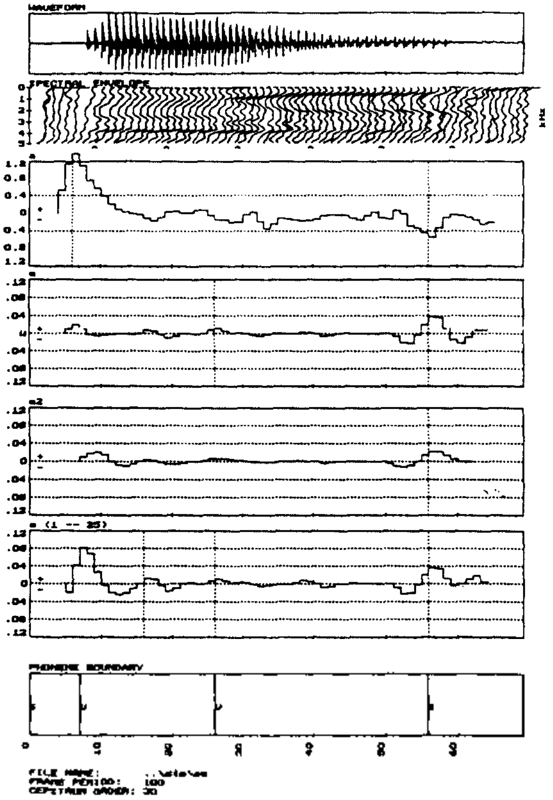


Figure 3. The dynamic variation parameters of spectral envelopes obtained by entire cepstral coefficients show phoneme insertion error, utterance of /ae/ of male spearker

## Ⅳ. DIPHONE EXTRACTION ALGORITHM

In order to extract diphone elements, phoneme boundaries in source speech should be detected correctly. After the cepstral analysis on source speech and calculations of parameters, phoneme boundaries are detected and diphones are extracted from source speeches by extraction algorithm. For the detection of phoneme boundaries, first, voiced and unviced intervals for by the parameter B(i). In voiced interval, paramter a(i), $e_1(i)$ and $e_2(i)$ are applied sequentially, and in unvoiced interval parameter a(i) and zero crossing rate are used to detect phoneme boundary. After the phoneme boundaries are detected, diphone units are extracted from source. Figure 4 shows the flow of proposed diphone extraction algorithm.

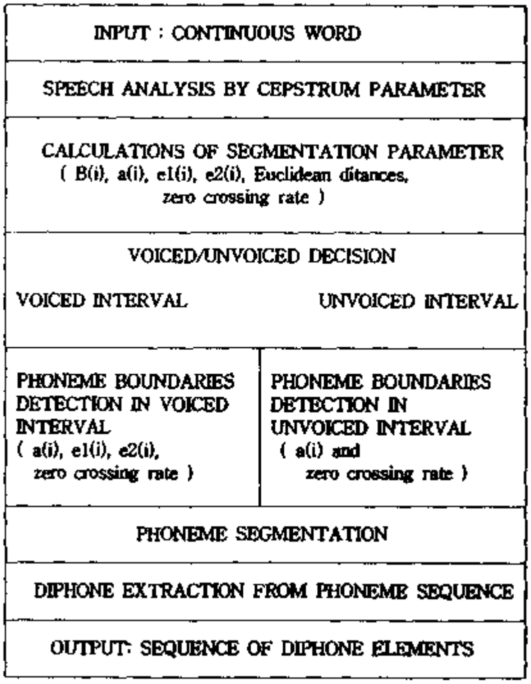

INPUT : CONTINUOUS WORD
SPEECH ANALYSIS BY CEPSTRUM PARAMETER
CALCULATIONS OF SEGMENTATION PARAMETER ( B(i), a(i), e1(i), e2(i), Euclidean ditances, zero crossing rate )
VOICED/UNVOICED DECISION
VOICED INTERVAL
UNVOICED INTERVAL
PHONEME BOUNDARIES DETECTION IN VOICED INTERVAL ( a(i), e1(i), e2(i), zero crossing rate )
PHONEME BOUNDARIES DETECTION IN UNVOICED INTERVAL ( a(i) and zero crossing rate )
PHONEME SEGMENTATION
DIPHONE EXTRACTION FROM PHONEME SEQUENCE
OUTPUT: SEQUENCE OF DIPHONE ELEMENTS

Figure 4. Flow of proposed diphone extraction algorithm

### 4.1 VOICED/UNVOICED DECISION

By the parameter B(i) using averaged energy in selected frequency band, voiced and unvoiced intervals are determined. If the value of B(i) exceeds a predefined threshold, i-th frame is voiced, otherwise is unvoiced.

### 4.2 PHONEME SEGMENTATIONS IN VOICED INTERVALS

(I)Maximum/Mimimum value function of parameter a(i), A(i)

A funcion of A(i) defined by (8) represents the maximum or minimum values of the parameter a(i).

$\text{if}((a(i) \text{ and } a(i-1) < a(i) \geq a(i+1))$
$\quad \text{or } (a(i) < 0 \text{ and } a(i-1) > a(i) \leq a(i+1))$
$\text{then } A(i) = a(i)$
$\text{else } A(i) = 0 \qquad (8)$

By (9), A(i) is modified,

$\text{if}((i-5 \leq 1 < i \text{ and } 0 < A(i) < TA_1$

$$\begin{aligned}&\text{and } A(i) <> 0 \text{ and } A(i) - A(1) < TA_2)\\ &\text{and } (i < m \le i+7 \text{ and } -TA_1 < A(i) < 0\\ &\text{and } A(m) <> 0 \text{ and } A(m) - A(i) < TA_2))\\ &\text{the } A(i) = 0 \qquad (9)\end{aligned}$$

where the thresholds $TA_1$ and $TA_2$ are got empirically.

(II)Maximum value function of parameter $e_1(i)$ and $e_2(i)$, $E_1(i)$ and $E_2(i)$

Two functions of $E_1(i)$ and $E_2(i)$ to represent the maximum value of $e_1(i)$ and $e_2(i)$ respectively, are defined by (10) and (11), and are modified by (12) and (13).

$$\begin{aligned}&\text{if}(E_1(i) > 0 \text{ and } E_1(i-1) < E_1(i) \ge E_1(i+1))\\ &\text{then } E_1(i) = e_1(i)\\ &\text{else } E_1(i) = 0 \qquad (10)\end{aligned}$$

$$\begin{aligned}&\text{if}(E_2(i) > 0 \text{ and } E_2(i-1) < E_2(i) \ge E_2(i+1))\\ &\text{then } E_2(i) = e_2(i)\\ &\text{else } E_2(i) = 0 \qquad (11)\end{aligned}$$

$$\begin{aligned}&\text{if}(0 < E_1(i) \le TE_1)\\ &\text{then } E_1(i) = 0 \qquad (12)\end{aligned}$$

$$\begin{aligned}&\text{if}(0 < E_2(i) \le TE_2)\\ &\text{then } E_2(i) = 0 \qquad (13)\end{aligned}$$

Where the thresholds $TE_1$ and $TE_2$ are calculated by the equations below,

$$TE_1 = \left(\frac{\sum_{i=0}^{n-1} E_1(i)}{n}\right) * 1.2$$

$$TE_2 = \left(\frac{\sum_{i=0}^{n-1} E_2(i)}{n}\right) * 1.2$$

where n is the number of analysis frame.

In voiced interval, by applying modified $A(i)$, $E_1(i)$ and $E_2(i)$ in sequence, phoneme boundaries are determined.

## 4.3 PHONEME SEGMENTATIONS IN UNVOICED INTERVALS

Unvoiced intervals are classified into silence and unvoiced consonant by using zero crossing rate and parameter a(i). Phonemes in unvoiced interval are detected by (14).

$$\begin{aligned}&\text{if}(ZCR > T_z \text{ and } a(i) > Ta)\\ &\text{then UNVOICED CONSONANT REGION}\\ &\text{else SILENCE REGION} \qquad (14)\end{aligned}$$

Where Tz is the thereshold value of zero crossing rate defined by following equation,

$$T_z = \frac{\sum_{i=1}^{n} ZCR(i)}{n}$$

$$Ta = 0.6$$

where n is the number of analysis frame, and Ta is got empirically.

Figure 5 shows the phoneme boundaries detected by using described parameters from the utterance /ae/ of male speaker. Sequentially applying the parameters a(i), e1(i) and e2(i), the boundary of /a/ and /e/ are detected.

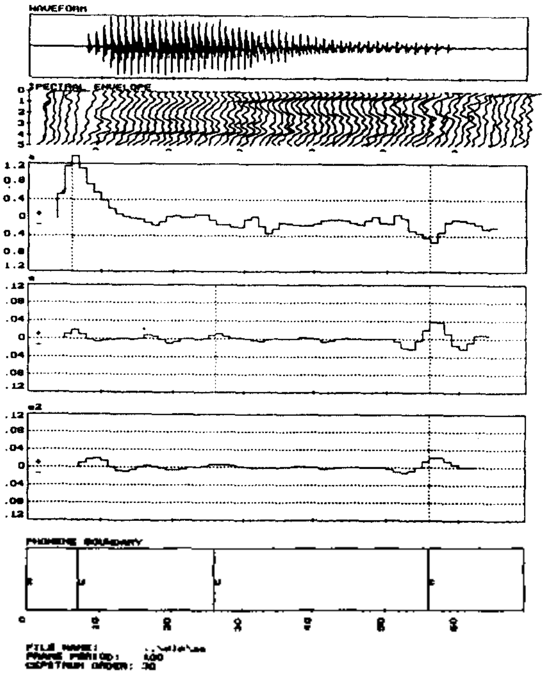


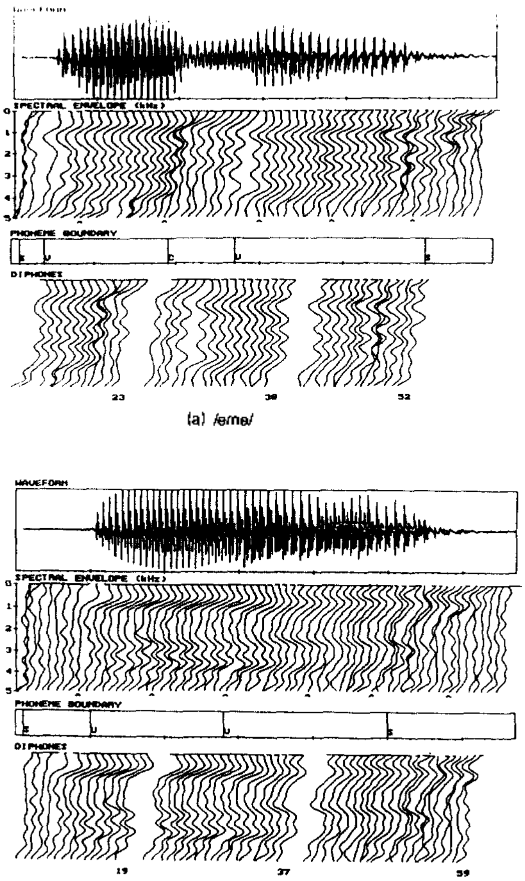
Figure 5. Automatically detected phoneme boundaries of /ae/ of male speaker

### 4.4 DIPHONE EXTRACTION

Diphone units are segmented by the following algorithm. The length of a phone and cepstral Euclidian distances to find steady-state center of phones are used in the algorithm.

```
if (male speaker) {
  LOOKAHEAD=2 ;
  LENGTH =5 ;
}
else (female speaker) {
  LOOKAHEAD=4 ;
  LENGTH=10 ;
}
while (No_of_Phone) {
  p=FRAME_NUMBER_OF_PHONE_START ;
  q=FRAME_NUMBER_OF_PHONE_END ;
  r=p+(q-p)/2 ;
  if((q-p)<LENGTH) DONE=1 ;
for (r-LOOKAHEAD ; r <= r+LOOKAHEAD ; r++)
  {
  FIND_MINIMAL_DISTANCE_FRAME_NUMBER( ) ;
  }
  No_of_Phone--;
}
```

When the boundaries of phonemes are detected accurately, proposed algorithm extracts a diphone from a phoneme pair for input phoneme sequence. The steady state center of a phoneme is selected first by choosing the middle position of a phoneme then by picking a frame in the neighborhood of this middle position which has minimal Euclidean distances. Since the phoneme which has turbulent variation of spectral envelope or has short length, for example, the phoneme /r/, is regarded as a transitional interval, proposed algorithm finds steady-state center in the next input phoneme. Automatically extracted diphones from phoneme sequences are shown in figure 6.



(a) /eme/

(b) /oa/

Figure 6. The Examples of extracted Diphones

## V. SYNTHESIS-BY-RULE USING DIPHONES

Some sentence speeches are synthesized by using diphones extracted automatically.

figure 7 shows the Korean speech synthesis-by-rule system. It is largely divided into three parts. In text analysis part, input text is analyzed into phonemic symbols, in rule generation part, by phonetic and prosody rule, control parameter is generated and in synthesis part, speech is synthesized by LMA (Log Magnitude Approximation) filter.

We have implemented sub-systems represented in double lined boxes in figure 7. In rule generation part, in the concatenation of diphones, 4 frames are interpolated linearly[7]. In synthesis part, LMA filter[8] is used, which is featured by cepstrum parameters derived from vocal tract cepstral parameters. The log maginitude response of LMA filter is finite Fourier series and represent the transition function of the vocal tract.

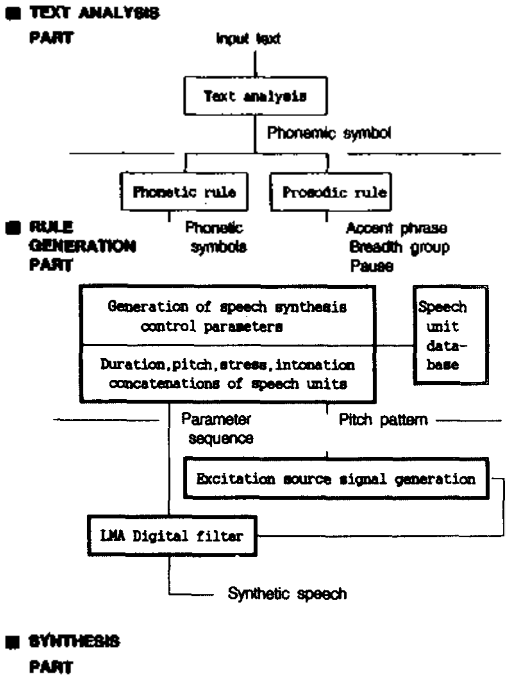


Figure 7. The block diagram of korean speech synthesis-by-rule system

Pitch contour of synthetic speech is consist of concatenation of the original pitches of speech unit because original pitches decrease distortions of synthetic sounds. It is verified by listening test that the synthetic speech are very intelligible. Figure 8 shows an example of synthetic speech. The waveform of original utterance /dongdʌkjəzatæhakyo/ of female speaker and waveform of synthetic speech by diphones automatically extracted, are shown in figure 8.

## VI. EXPERIMENTAL RESULTS AND ESTIMATION

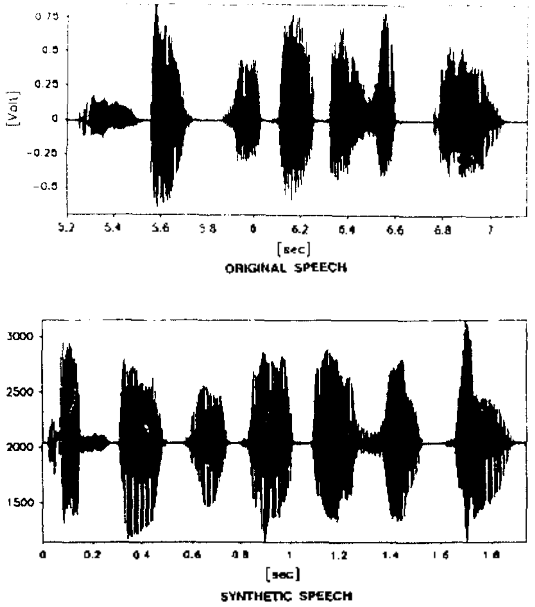


Figure 8. Original and synthetic speech /dongdʌkjəjatæhakyo/ of female speaker

The test words are total 116 two-syllable nonsense words which consist of 49 vowel to vowel pair, 16 vowel to semi-vowel to vowel, 40 VCV units and 11 VCCV units spoken by 1 male and 2 female speakers.

The proposed algorithm is implemented in C language on 486 IBM PC compatible system. Results of phoneme detection rates are shown in table 1.

Table 1. Results of Phoneme Boundaries Detection

| | DELETION ERROR | | INSERTION ERROR | |
|---|---|---|---|---|
| | TOTAL | % | TOTAL | % |
| VOWEL TO VOWEL | 3 (49) | 6.1 | 7 (49) | 14.2 |
| VOWEL TO SEMI VOWEL | 6 (16) | 37.5 | 1 (16) | 6.2 |
| VCV UNIT | 0 (10) | 0.0 | 0 (10) | 0.0 |
| VCCV UNIT | 0 (11) | 0.0 | 1 (11) | 9.0 |
| TOTAL | 9(116) | 7.7 | 9(116) | 7.7 |

The most difficult areas to detect phoneme boundary candidates are those containing semi-vowel bacause the transition between phonemes are very smooth. The boundaries of 90% of automati cally extracted diphones are within the 30ms of diphones boundaries of manually segmented.

## Ⅶ. CONCLUSIONS

In this paper, an algorithm is proposed to extract diphones automatically from two-syllabic natural speeches consist of vowel pairs, VCV units(C is consonants or semi-vowels) and VCCV units, which minimizes spectral discontinuities introduced by concatenations.

The proposed algorithm shows good results on not only VCV unit but also vowel pairs. Using the automatic diphone segmentation it is easier to estimate consistent speech unit database and possible to reduce the amount of time to implement diphone datebase.

LMA filter is used which is pole-zero model and sentence speeches are synthesized with intelligibility and naturalness by diphones automatically segmented.

Further studies will be continued to modify proposed algorithm to extract diphones from continuous natural speech sentence.

## ACKNOWLEDGEMENTS

This work was supported by ETRI(Electronics and Telecommunications Research Institute) Grant in 1993.

## REFERENCES

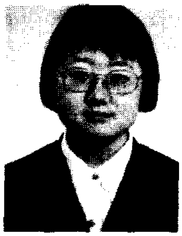
1. P.A. Tayor and S.D. Isard, "Automatic diphone segmentation," Eurospeech_90, pp.709-711, 1990.
2. H. Kaeslin, "A systematic approach to the extraction of diphone elements from natural speech," IEEE Vol.ASSP-34 No.2, pp.264-270, 1986.
3. Chieko FURUICHI, Satoshi IMAI, "Speaker-Dependent Phoneme Recognition of Unspecified Vocabulary Japanese Speech," IECE Vol.J73-D-II No.4, pp.501-511, 1990.
4. Chieko FURUICHI, Satoshi IMAI, "Segmentation of continuous speech into phoneme units," IECE Vol.J72-D-II No.1, pp.11-21, 1989.
5. S. Imai and Y.Abe, "Spectral envelope extraction by improved cepstral method," Trans IECE Vol.62-A No.4, pp.217-223, 1979.
6. Georg E. Ottesen : "An automatic diphone segmentation system," Eurospeech_90, pp.713-716, 1990.
7. Y.H. Lee, "A study on demi-syllable based Korean synthesis-by-rule by using mel-cepstrum parameter," Doctoral Disertation, Tokyo Institute of Technology, 1988.
8. S. IMAI, et al, "Log Magnitude Approximation (LMA) filter," Trans. IECE Vol.J63-A No.12, pp.886-893, 1981.

**▲InJong Jeong**

InJong Jeong was born on August 2, 1968.

He received the B.S. degree in computer science from HanYang University, in 1991. He is currently enrolled in a M.S. degree at HanYang University. His research interests include speech synthesis and speech recognition.

**▲Youn Jeong Kyung**



Youn Jeong Kyung was born on October 1, 1970.

She received the B.S. degree in computer science from DongDuck Women's University, Seoul, in 1992. She is currently enrolled in a M.S. degree at Dongduck Women's University. Her research interests include speech synthesis, speech recognition and natural language processing.

**▲HanWoo Kim**

1976 B.A : HanYang University, Dept. of Electronical Engineering
1978 M.A : HanYang University, Dept. of Electronical Engineering
1980 ~ 1981 : Research staff in Kyoto University, Japan
1981 ~ present : Associate Professor, Dept. of Computer Science & Engineering, HanYang university.
Research interest : Natural Language Processigng & Speech information Processing

**▲YangHee Lee**

YangHee Lee was born in Kyunggi-do, on Oct. 5, 1948.

He received the B.S degree in Electronics Engineering from Dong-kuk University, Seoul and the M.S and Ph.D. degree in Information Processing Engineering from Tokyo Institute of Technology at Tokyo, Japan, in 1976, 1984, and 1988, respectively.

From 1988 to 1992. 3, he was an assistant professor and since 1992. 4, he has been an associate professor of computer science at Dong-Duck women's University. His research interests include speech synthesis, speech recognition and natural language processing.