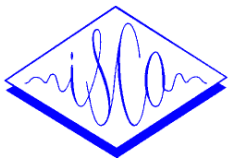

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Phoneme Segment Concatenation and Excitation Control Based on Spectral Distortion Criterion for Speech Synthesis

Kenzo Itoh, Hideyuki Mizuno, Tetsuya Nomura and Hirokazu Sato

Speech and Acoustics Laboratory, NTT Human Interface Laboratories
Yokosuka-shi, Kanagawa 238-03 Japan

ABSTRACT

This paper proposes two new methods based on spectral distortion criteria that produce high quality speech synthesis. One is a phoneme segment selection method using an objective continuity measure, and the other is an excitation signal extraction method for pitch and duration control. The continuity measure is expressed using continuity of the LPC spectrum envelopes. When this measure is used for optimum selection, natural sounding synthetic speech is produced without any smoothing technique. For pitch and duration control, an automatic excitation signal extraction method is proposed that also uses the spectral distortion criteria between original and synthetic speech based on residual excited LPC vocoder. When this new pitch and duration control method is used, the average LPC cepstrum distortion (CD) is decreased from 1.90 dB to 1.01 dB.

1. INTRODUCTION

When using phoneme-like segments, the main problems are selecting the most appropriate kind of synthesis unit and achieving effective concatenation. Many types of synthesis units have been proposed in previous text-to-speech systems, such as CV (Consonant-Vowel)[1], VCV or CVC[2] and demisyllable[3]. Generally, synthesis units can easily be connected at voiceless consonants, but, it is difficult to connect them at voiced consonants and vowels unless a smoothing technique is used.

The context oriented clustering procedure has been proposed for the generation of context sensitive synthesis units[4]. In this procedure, synthesis units can be generated automatically without any prior phonological knowledge. In addition, variable length phoneme synthesis units have been proposed[5]. In both proposals, the synthesis units are concatenated without any consideration of the continuity of acoustical feature parameters. However, to reproduce higher quality synthetic speech, it is necessary to optimize the unit selection by considering the continuity between units. This paper proposes an optimum phoneme segment selection method[6]. The proposed method uses segments with triphone labels and connects them taking account of the continuity of the LPC (Linear Predictive Coding) spectral envelope at each segment junction.

In text-to-speech systems, the LPC type vocoder is widely known. However, the synthetic speech quality of a vocoder that uses a pulse and white noise generator as a simple excitation source signal is insufficient. The residual excited LPC vocoder is one of the most effective waveform synthesis techniques for producing high quality synthetic speech[7]. However, it is difficult to control the pitch frequency and duration of each segment. In previous methods, the LPC residual waveforms that were extracted synchronously with the pitch were used to control pitch frequency. However, the extraction position and window length are very critical to synthetic speech quality. This paper proposes an optimum residual waveform extraction method that uses the spectral envelope distortion criteria between original and synthetic speech[8].

2. SYNTHESIS SYSTEM OUTLINE

Figure 1 shows the outline of the speech synthesis system proposed in this paper. The input phoneme string is decomposed into a sequence of triphone labels. Several segment candidates, which carry triphone labels corresponding to each decomposed triphone

label, are selected from synthesis units in the pre-selection stage.

In the next stage, optimum segment units are determined by an objective continuity measure. Each segment duration and power are calculated using the center phoneme information in the pre-selected triphone speech data. Finally, the selected segment sequence, pitch, duration and power are used to drive a pitch synchronous synthesizer as will be described later.

3. OPTIMUM SYNTHESIS UNIT SELECTION

The optimum selection process for phoneme segment units is shown in Fig. 2. An example input phoneme string, *bakuoNga* is decomposed into a sequence of triphone labels such as /*ba/, /bak/, /aku/,... and so on. In this figure, the asterisk /*/ means silence. Phoneme /b/ must be synthesized by taking care of the triphone context of /*ba/ which induces preceding phoneme /*/ and succeeding phoneme /a/. First, words or passages that contain triphones which must be synthesized are selected from the phoneme information table. Next, the continuous objective measure Dab is calculated for optimum selection. Figure 3 is a schematic locus of the LPC cepstrum coefficient $Cep(i)$ of segment A and B for time sequence from $t1$ to $t6$. Dab is defined by

$$Dab=\sum_{i=1}^{n}\frac{Da(i)+Db(i)+2dab(i)}{2} \qquad (1)$$

where n is the number of LPC cepstrum parameters. $Da(i)$ and $Db(i)$ are slope coefficients of i-th cepstrum coefficient of segment A and B. For example, $Da(i)$ is defined by

$$Da(i)=\frac{dat_1(i)+dat_2(i)+dat_3(i)}{3} \qquad (2)$$

where $dat1$, $dat2$, $dat3$ are the differential coefficients around time points $t1$, $t2$, $t3$ respectively. $dab(i)$ in the expression (1) is cepstrum distance between time $t3$ and $t4$. $dab(i)$ is defined by

$$dab(i)=\frac{|Ct_3(i)+Ct_4(i)|}{2} \qquad (3)$$

where $Ct3$ and $Ct4$ are cepstrum coefficients of segment A at time $t3$ and of segment B at time

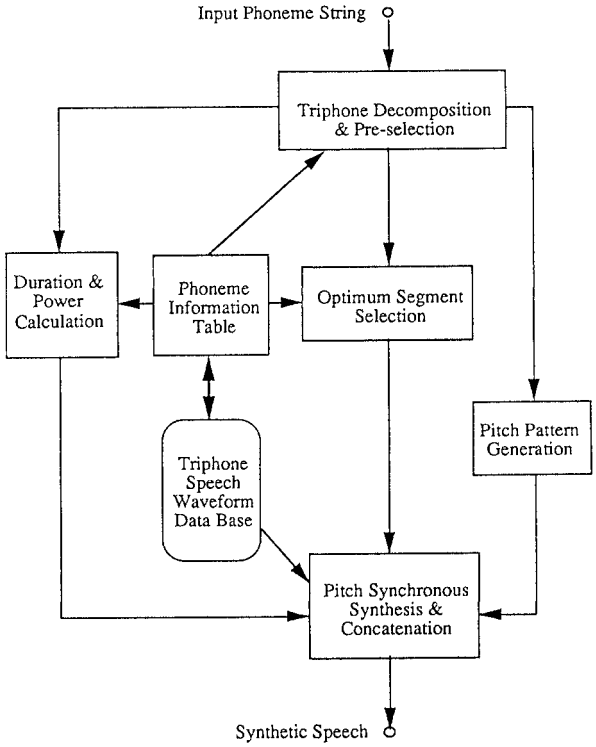

Fig.1 Outline of Speech Synthesis System

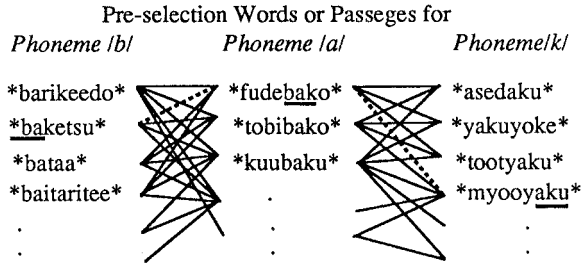

Fig.2 Optimum Selection for Synthesis Phoneme Segments (Input Phoneme String: *bakuoNga*)

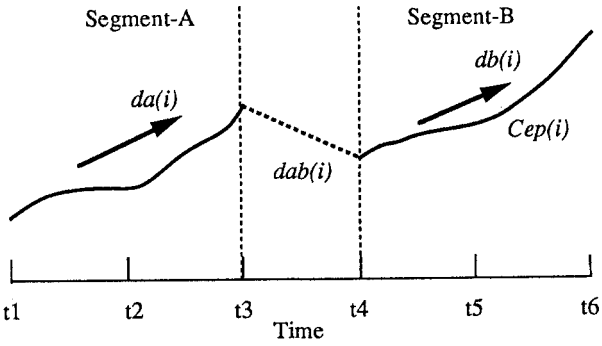

Fig.3 Cepstrum Coefficients Locus of Segment-A and B

t4, respectively. *Dab* measure is calculated at all segment junction points. However, to speed up the calculation time, a pruning technique is used for optimum pass selection.

4. SYNTHESIZER

A residual excited LPC synthesizer is used to produce natural sounding speech. For pitch and duration control, it is necessary to achieve pitch synchronous extraction for the residual waveforms. However, the extraction position (Cp) and window length (Cw) are very critical to synthetic speech quality. Moreover, sufficient study has not been made on the residual extraction problem.

Figure 4 shows Cp and Cw on the residual waveform for pitch and duration control. This paper proposes a new effective and simple method of residual waveform extraction which automatically determines Cp and Cw. It uses the spectral envelope distortion criteria between speech with labels and synthetic speech.

Figure 5 gives a block diagram of the proposed pitch and duration control method which is based on the residual excitation PARCOR vocoder technique. The input speech signal is analyzed pitch synchronously using the LPC method. The analyzed residual signal *E(f)* is cut by Cp and Cw. The original pitch period To is manipulated by the synthesis pitch period Tm, and the new residual signal sequence *E(f)'* is produced. The spectral envelope distortion between original and synthetic speech is calculated at all Cp and Cw instances. The synthetic speech that has the minimum spectral envelop distortion is selected as the output speech. Synthetic speech is generated for each pitch period as shown in Fig.6. The pitch synchronous synthesis and the overlap-add technics are used.

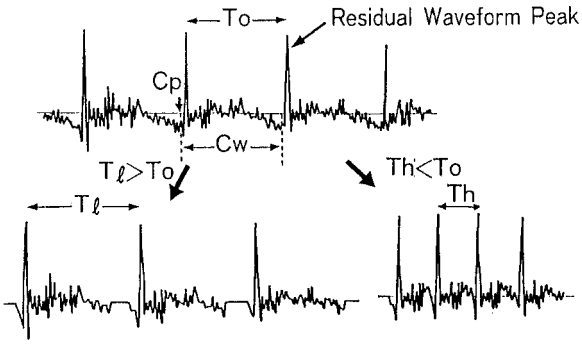

Fig.4 Cutting Window Length (Cw) and Start Point (Cp) for Pitch or Duration Control

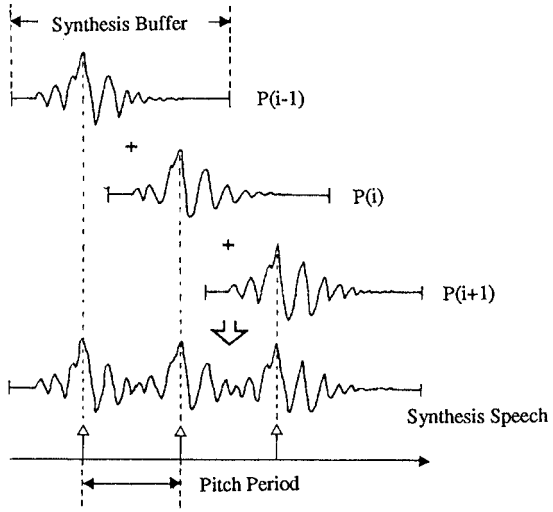

Fig.6 Pitch Synchronous Overlap Add Process

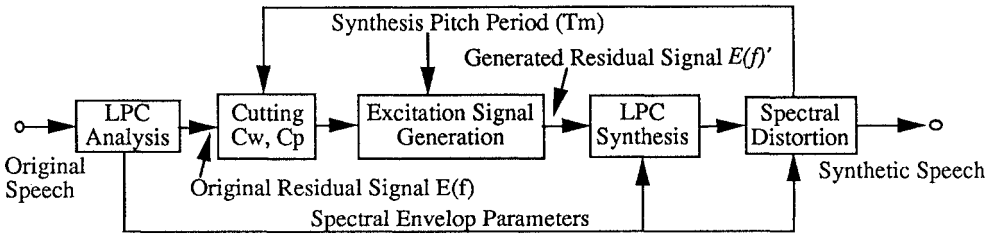

Fig.5 Block Diagram of Cutting Window (Cw) and Point (Cp) Determination Process for Residual Waveform Based on Residual Excited PARCOR Vocoder

Duration and power of the segments are controlled by mean values of the segments which have the pre-selected triphone labels from the speech data base. The speech signal analysis conditions are as follows, sampling frequency : 12 KHz, low-pass filter : 6.0 KHz, LPC analysis window : 20 ms and order of LPC analysis : 12. The order of LPC cepstrum coefficients is set to 16 to calculate the spectral envelope distortion.

Figure 7 shows an example of synthetic speech waveform and sound spectrogram of a short sentence. It is seen from the figure that the spectral changes are very smooth at the segment concatenation points. The averaged LPC cepstrum envelope distortion (CD)[9] of the example synthetic speech is decreased from 1.90 dB to 1.01 dB by introducing the new method.

5. CONCLUSION

Two new effective methods were proposed that produce high quality synthetic speech based on the residual excited LPC vocoder. First, an optimum synthesis unit selection method using the objective continuity measure was described. Second, for pitch and duration control, an automatical excitation signal extraction method was proposed that also uses spectral envelope distortion criteria between original and synthetic speech. With these new methods, natural sounding synthetic speech was produced.

ACKNOWLEDGMENT

The authors would like to thank Dr. Sadaoki Furui for his continuous support of this research.

REFERENCES

[1] Y. Tohkura and Y. Sagisaka, "Speech synthesis using CV-syllable assembly," Proc. of Meeting of Acoust. Soc. of Japan, 3-4-3, March, 1980

[2] H. Sato, "Speech synthesis using PARCOR-VCV synthesis units," J. of IEICJ, Vol.J61-D, No.11, p.858, 1978

[3] J. B. Lovins M. J. Macchi and O. Fujimura, "A demisyllable inventory for speech synthesis," 97th Meeting of the Acoust. Soc. Amer., YY4, 1979

[4] S. Nakajima and H. Hamada, "Automatic generation of synthesis units based on context oriented clustering," Proc. of ICASSP, p.659, April, 1988

[5] Y. Sagisaka, "On the design of a speech synthesis unit set using entropy measure," Proc. of meeting of Acoust. Soc. of Japan, 1-7-20, March, 1989

[6] H. Mizuno and T. Nomura, "A speech synthesis using phoneme segments with multi-phonetic environments," Proc. of Meeting of Acoust. Soc. of Japan, 1-4-10, March, 1990

[7] H. Sato, " Speech synthesis using phoneme sequence and LPC residual waveform signal," Proc. of Meeting of Acoust. Soc. of Japan, 1-2-6, Oct. 1978

[8] K. Itoh and H. Sato, "Excitation waveform extraction for pitch control in residual excited LPC speech synthesis," 118th Meeting of the Acoust. Soc. of America, Vol.86, 1989

[9] K. Itoh, N. Kitawaki and K. Kakehi, " Objective quality measures for speech waveform coding systems," Review of the E.C.L., Vol.32, No.2, p.220, 1984

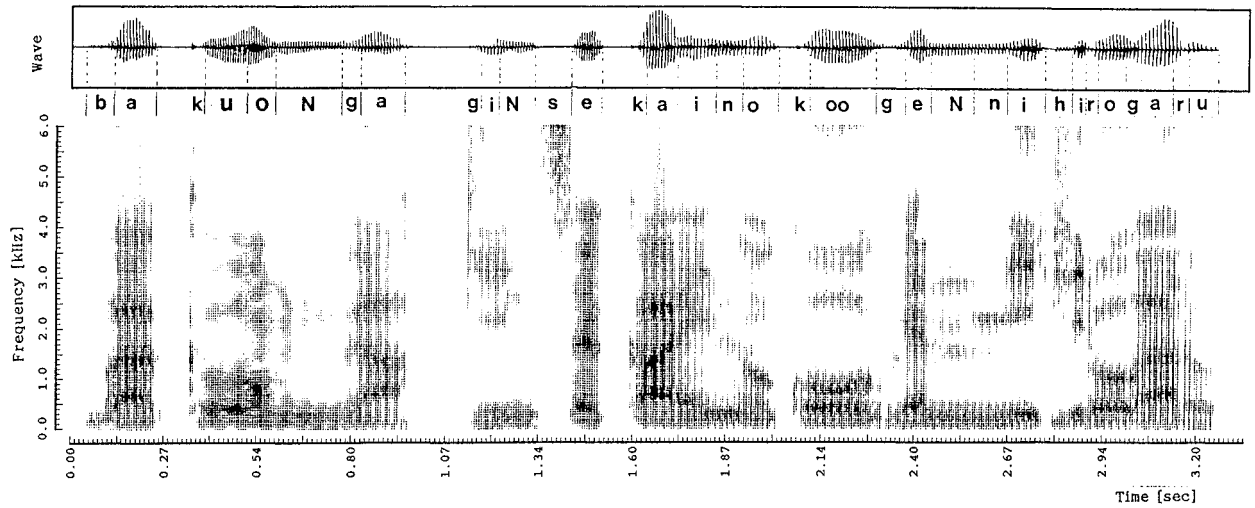

Fig.7 An Example of Synthetic Speech Waveform and Sound Spectrogram Using the Proposed New Methods