2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

High Quality Speech Synthesis based on Wavelet Compilation of Phoneme Segments

Tomohisa HIROKAWA, Kenzo ITOH and Hirokazu SATO
NTT Human Interface Laboratories
Yokosuka-shi, Kanagawa 238-03, JAPAN

ABSTRACT

A speech synthesis system is developed which directly compiles phoneme wavelet segments selected from a wavelet dictionary containing over 45,000 entries to yield high quality synthesized voice. In ICSLP'90, we proposed the wavelet selection and wavelet concatenate methods used in our system. To realize the system, we establish prosody pattern setting by rules and a wavelet modification procedure to achieve the design goals. Phoneme duration is set according to phoneme environment, and phoneme power is controlled by both pitch frequency and phoneme environment. Tests show the average errors in vowel duration and consonant duration are 28.8msec and 16.8msec respectively, and the vowel power average error is 2.93dB. Wavelet pitch frequency is controlled by an approach based on the pitch synchronous overlap-add method. To avoid abrupt changes in voice spectrum and wavelet shape, an interpolation operation is carried out between voiced wavelets. The synthesized speech has high intelligibility and naturalness, while the original speaker quality is retained.

10.21437/ICSLP.1992-131

1.INTRODUCTION

In text-to-speech conversion, one of the most important problems is making the synthesized voice sound natural. Conventional systems and devices now on the market apply parametric coding techniques to speech synthesis[1]. However, the unconstrained output fails to satisfy the high expectation of users because of its poor naturalness and intelligibility.

In ICSLP'90 we proposed a new method for speech synthesis that directly concatenates phoneme wavelet segments selected from a wavelet dictionary[2]. This method has two strong points. One is that original wavelet segments are used instead of parametric segments, and this creates speech with high intelligibility. The other point is that our method collects a large number of phoneme wavelets (wavelet dictionary) as uttered in words and sentences. Because the dictionary contains wavelets influenced by intonation, speech rate, stress etc., the most appropriate wavelets can be selected from the dictionary according to phoneme context and also phoneme attributes set by rules. The naturalness of the synthesized speech is markedly better than that created with limited units.

Our previous report determined wavelet selection and wavelet pitch modification. This paper develops prosody pattern setting methods by analyzing the relationship between prosody parameter, such as duration and power, and phoneme environment in natural speech. Pitch pattern is generated from accent information and phrase boundary symbols which represent the degree of union between adjoining phrases. Therefore, the proposed system can produce synthetic speech from a limited amount of input data, namely phonetic symbol sequence, accent information and phrase boundary symbols. The synthesized speech quality is further increased by carrying out detailed wavelet modification.

2. SYSTEM OUTLINE

The block diagram of the proposed system is shown in Fig.1. The system accepts phonetic symbol sequence, accent type, and phrase boundary symbols as input data. The types of phrase boundary symbols are shown in Table.1. From this information, prosody patterns such as phoneme power pattern, duration pattern and pitch pattern, are obtained by various rules and tables. This

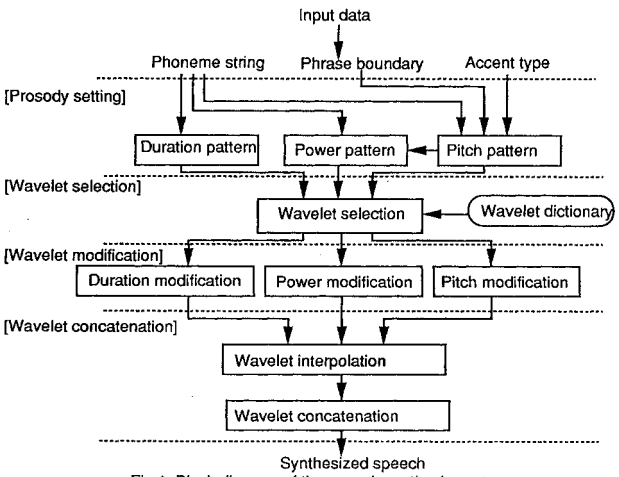

Fig 1. Block diagram of the speech synthesis system

Table 1. Phrase boundary symbols

Symbols	Meanings
*	Strong connection
/	Weak connection
Space	Short pause
,	Middle pause
.	Long pause

process is detailed in section 3. In the wavelet selection stage, wavelets that most closely match the phonetic symbol sequence and the prosody patterns are chosen from the wavelet dictionary by the evaluation function **H** which is defined as follows;

$$H = bn + (1-b)W \quad \text{.....(1)}$$

where,

n=1/eN (e is natural number.)

$$W = w_v||V_p - V_s|| + w_f||F_p - F_s|| + w_t||T_p - T_s|| + w_a||A_p - A_s||$$

Here, ||....|| means the absolute value normalized by the standard deviation of the parameters. The selection parameters are the number of phonemes **N** that coincide between the input string and the phonological context in the dictionary, average pitch **V**, pitch contour **F**, duration **T**, and amplitude **A**. Values identified with suffix **p** are extracted from the wavelet dictionary, and values marked with suffix **s** are the goal values to be generated. **b** is a balance coefficient between the value based on the phonetic environment and that derived from prosody. The weight coefficients among the selection parameters are identified as **w**.

The wavelet dictionary is constructed from a two hour speech that includes isolated words and sentences uttered by one male speaker. The speech data was passed through a 5.1kHz cut-off low-pass filter and digitized at a 12kHz sampling rate. Acoustic phonetic segments with phonetic labels were determined manually, and then registered into the dictionary with other characteristics such as power, duration, and pitch information. The total number of segments is about forty-five thousand and the most frequent Japanese phoneme /a/ has over five thousand entries.

It is necessary to assign pitch marks to wavelet segments of voiced portions because the pitch synchronous overlap-add method is applied to modify wavelet pitches. Pitch marks are automatically set at the local peaks of the waveform after being passed through a 246Hz low-pass filter for male voice.

Selected wavelets are adjusted to make them fit the goal values more closely. The adjustment is carried out not only on pitch frequency, but also on duration and power. This procedure will be discussed in section 4. Lastly, wavelets are interpolated and concatenated to generate the continuous synthetic speech.

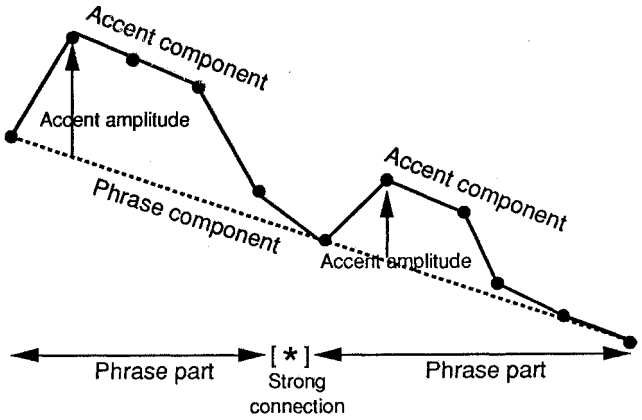

Fig.2 The point pitch pattern assignment model (Black circles mean point pitches)

3. PROSOY SETTING

3-1 Pitch contour generation

The pitch frequency patterns for words and phrases are computed according to a model in which the accent component is added to a gross phrase component as shown in Fig.2[3]. The accent component is formed by high and low point pitches whose locations are assigned at the center of each vowel according to accent type, and the accent amplitudes are determined by the phrase boundary features. The phrase component represents the pitch pattern which starts high and decreases at a constant rate.

3-2 Phoneme duration

Japanese phoneme duration, if the speech rate is constant, is dominated by the preceding and succeeding phonemes[4]. A phoneme duration table, which consists of three phoneme strings, was constructed by analyzing the wavelet dictionary. Figure 3 shows the distribution of vowel /a/'s duration which is defined as the center phoneme in a triphone context. The figure suggests that the vowel duration lain between unvoiced sounds is shorter than between voiced sounds. At the end of the sentence, the vowel duration is found to be obviously lengthened. These characteristics are reflected in the duration table. The input phoneme string is decomposed into a sequence of triphone contexts. The duration data of each decomposed triphone context are extracted from the table and then the duration pattern is formed.

This method was evaluated using twenty-three short sentences. The average errors of vowel duration and consonant duration were 28.8msec and 16.8msec, respectively.

3-3 Segmental power

Because voice power is correlated to pitch frequency and is influenced by context[5], the power of each phoneme is set to the average value of the following two values. One is obtained from

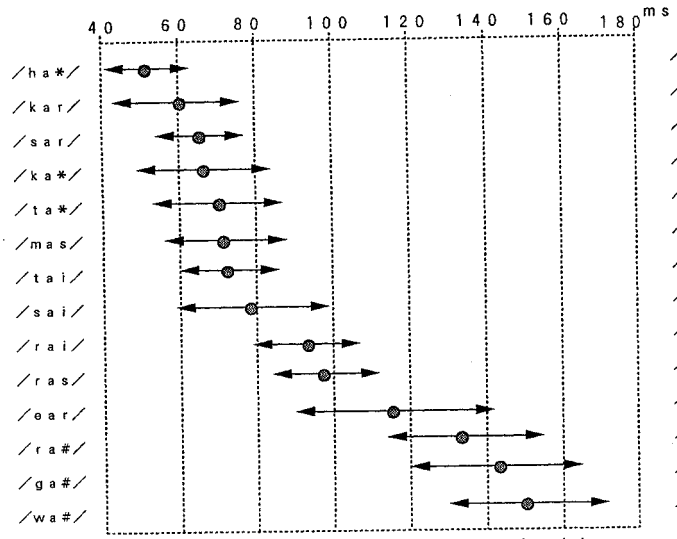

Fig.3 The distribution of /a/'s duration defined in triphones (Average values and standard deviations)

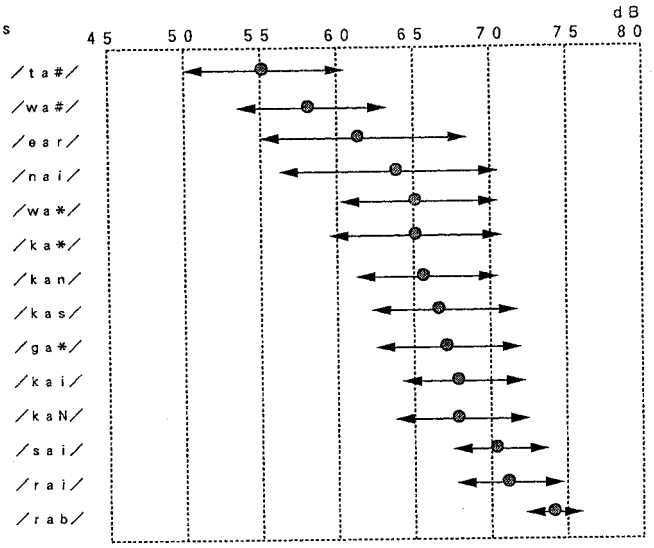

Fig.5 The distribution of /a/'s power defined in triphones (Average values and standard deviations)

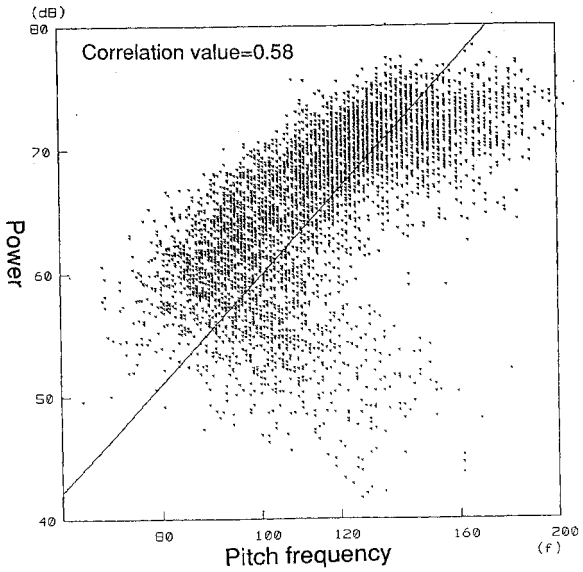

Fig.4 Relationship between pitch frequency and power for vowel /a/

the correlation line on the log-power/log-pitch frequency plane(Fig.4), and the other comes from a phoneme power table constructed in the same way as the duration table.(Fig.5) In Fig.4, some correlation is observed, but the coefficient is 0.58 which is weak. This seems to indicates that it is necessary to investigate the correlation in detail[6].

This power setting method was also evaluated using the same short sentences, and the power average error for vowels was 2.9dB.

4. WAVELET MODIFICATION

4-1 Pitch modification

In the wavelet modification process, pitch modification is the most effective technique to raise speech naturalness. This technique causes two significant effects: the pitch pattern of the synthesized voice matches the goal pattern more closely and the pitch discontinuity between wavelet segments is eliminated. The technique is based on the pitch synchronous overlap-add method[7]. Selected wavelets are cut out by Hanning-type windows whose lengths are twice the synthesis pitch period, and overlap-added according to the synthesis pitch rate. The pitch contour of a wavelet segment is approximated as linear by means of the least square error approach. The segment is normalized by the goal duration in the time scale The pitch modification ratio is defined as the ratio between the goal pitch and the corresponding segment pitch. The modification procedure is given in Fig.6.

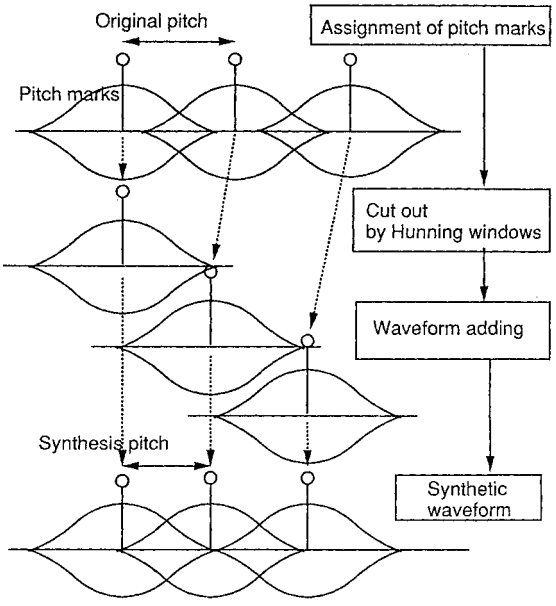

Fig.6 Pitch modification procedure

4-2 Duration control

To realize the required duration, the windowed samples are removed or duplicated according to the differences between selected wavelet length and rule-set duration. In fricative sounds like /s/ and /f/, duration control is performed by waveform suppression or prolongation. When the selected wavelet is suppressed, the waveform data is used as long as the goal duration length, and when the wavelet is prolonged, the waveform data is terminated as the goal duration.

4-3 Power control

Power control is simple. The power control coefficient is defined as the ratio between the goal power and the selected wavelet power. The wavelet samples are multiplied by the coefficient.

4-4 Wavelet segment interpolation process

The selected wavelets are interpolated between the last pitch waveform of the preceding segment and the top pitch waveform of the succeeding segment to avoid abrupt changes in voice spectrum and waveform shape. Hearing tests confirm that one or two pitch waveform interpolation is effective to reduce noise occurrence. An example of an interpolated waveform is displayed in Fig.7 as well as the adjacent synthesized waveforms. The synthesized speech is characterized by smooth movements that are comparable to those of continuously uttered speech.

5. CONCLUSION

We have proposed a new speech synthesis system which utilizes just phoneme string, accent information and phrase boundary symbols as input data. The system concatenates wavelet segments to generate continuous speech waves. This paper has proposed several methods for determining prosody pattern. Prosody tables which consists of phoneme string triplets was obtained by analyzing the wavelet dictionary. Prosody patterns are set according to the table value that matches one of the three phoneme strings. In the case of power, the relationship to pitch frequency is also utilized. These methods were quantitatively evaluated and confirmed to be effective.

Wavelet modification significantly increases synthesized speech quality. We discussed pitch frequency, duration and power modification methods, and wavelet segment interpolation to ensure smooth concatenation. In particular, pitch modification and wavelet segment interpolation are efficient in eliminating discontinuities from the output speech.

ACKNOWLEDGEMENT

The authors would like to thank Dr. Yukio Kobayashi, Dr. Sadaoki Furui, and Dr. Noboru Sugamura for their encouragement during this work.

References

[1]H.Hakoda et.al. "A new Japanese text-to-speech synthesizer based on COC synthesis method", Proc.ICSLP-90,1990.

[2]T.Hirokawa, K.Hakoda "Segment selection and pitch modification for high quality speech synthesis using waveform segments", Proc.ICSLP-90,1990.

[3]H.Hakoda, H.Sato "Prosodic Rules in Connected Speech Synthesis", Systems, Computers, Controls, Scripta Electronica Japonica 3, Vol. 11(1980).

[4]N.Kaiki, K.Takeda, Y.Sagisaka "Phoneme Duration Setting in Sentence Utterances", Trans.ASJ,SP90-2,1990(in Japanese).

[5]K.Mimura, N.Kaiki, Y.Sagisaka et.al. "Analysis and control of temporal patterns of speech power using statistical methods", Trans.ASJ,SP91-4,1991(in Japanese).

[6]K.Itoh, T.Hirokawa, H.Sato "Segmental Power Control for Japanese Speech Synthesis", these proceedings.

[7]F.Charpentier, E.Moulines "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones", Proc.Eurospeech'89,1989.

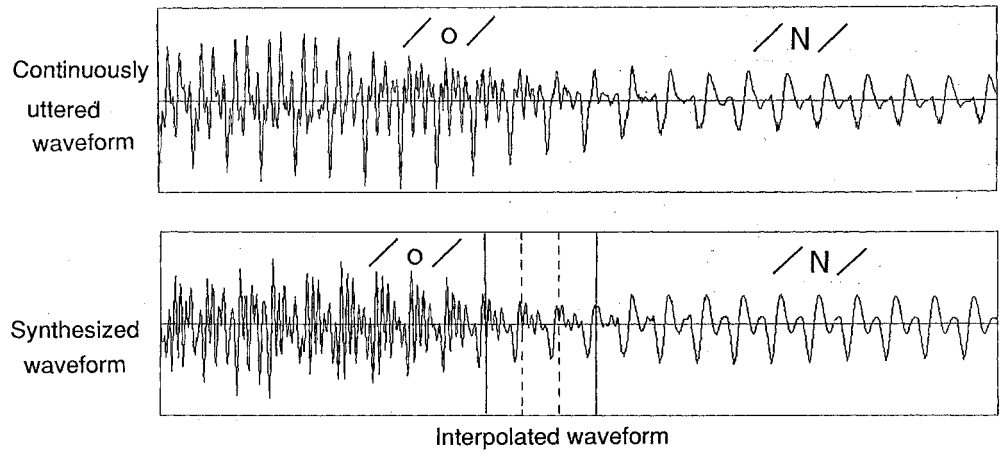

Fig.7 An example of synthesized waveforms
(A part of "bakuon")