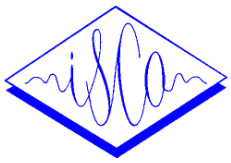


First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Segment Selection and Pitch Modification for High Quality Speech Synthesis using Waveform Segments

Tomohisa Hirokawa *Kazuo Hakoda*
NTT Human Interface Laboratories
Yokosuka-shi,Kanagawa,238-03 JAPAN

ABSTRACT

We propose a new method for speech synthesis that concatenates waveforms selected from a waveform dictionary. The method uses a modified PSOLA technique to alter the pitch of waveforms selected from a dictionary. The limits of acceptable pitch shifts are determined by preference tests. To make segment selection more accurate, we introduce a new factor which considers the spectral continuity across voiced phoneme boundaries. The average spectral difference is reduced from 5.4dB to 2.7dB and the synthesized voice is more fluent.

1. INTRODUCTION

The conventional approach to text-to-speech synthesis uses parametric coding techniques such as linear prediction coding (LPC) for voice synthesis [1][2]. However, the output voices are often listened to be nasal. Some attempts have been made to yield natural sounding synthetic voices from waveforms by circumventing this problem [3][4]. In these cases, however, the number of speech units is too small to adequately represent the complex phonetic variations and coarticulatory characteristics created by context. We proposed a new approach for using a waveform dictionary that yields high-quality synthetic sound [5]. In this speech synthesis method, waveforms are derived from a large volume of naturally uttered words and sentences, and are stored in a dictionary with phonological context and physical characteristics. The total number of entries is forty-five thousand. To further improve naturalness, the pitch contour of each speech unit used must be adjusted for the desired prosody. Charpentier proposed a technique termed PSOLA (pitch synchronous overlap-add method) to modify pitch frequency of waveform [6]. The sound is good, however large pitch shifts degrade voice quality.

In this paper, we first describe a pitch modification method based on PSOLA. The relationship between pitch shift rate and synthetic voice acceptability is confirmed through a hearing test. Another comparative test is conducted to evaluate pitch modification in our speech synthesis method.

It is known that the spectral features of concatenated sounds must be smoothed at their boundaries to avoid discontinuous speech [7][8]. The previous waveform selection function used a combination of factors, some of which evaluated context, average pitch, pitch contour, segment duration, and power. We discuss the introduction of a new factor for waveform selection which represents spectral continuity.

2. METHOD OUTLINE

A schematic overview of the proposed method is shown in Fig.1. First, Japanese textual analysis yields a phonetic symbol sequence and prosody information. From this information, various rules and tables are applied to assign the prosody patterns for each mora of the sounds to be synthesized. Waveforms that most closely match the phonetic symbol sequence and the prosody patterns are selected by the function **H** which is defined as follows;

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

where,

$$W = w_v \| V_p - V_s \|^2 + w_f \| F_p - F_s \|^2 + w_t \| T_p - T_s \|^2 + w_a \| A_p - A_s \|^2$$

$$n = 1/e^N \quad \text{(e is natural number.)}$$

Here, $\| \ldots \|$ means the value normalized by the standard deviation of the various parameters. The

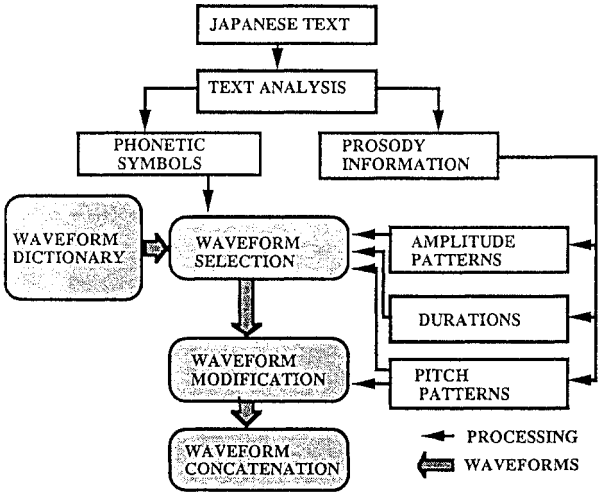


Fig 1. Block diagram of the speech synthesis system based on waveform concatenating method

parameters for selection are the number of coincident phoneme **N**, average pitch **V**, pitch contour **F**, duration **T**, and amplitude **A**. The value with suffix **p** is extracted from the waveform dictionary, and the value with suffix **s** is the goal value to be generated. **b** is a balance coefficient between the value based on the phonetic environment and that derived from prosody.

A waveform dictionary was constructed from a two hour speech that included isolated words and sentences uttered by one male announcer. The voice data was passed through a 6kHz cut-off low-pass filter and digitized at a 12kHz sampling rate. Acoustic phonetic segments with phonetic labels were determined manually and the start and the end points of waveforms were established as the zero cross points preceding the local positive peaks that existed within +5ms and -5ms of the phonetic borders.

Selected waveforms are further adjusted to make them fit the goal pitch pattern more closely. This process is detailed in the following section. Lastly, waveforms are concatenated to generate the continuous synthetic sounds.

3. WAVEFORM PITCH MODIFICATION

Smoothing of pitch contour is important in improving naturalness. We describe in this section a pitch modification method and its evaluation.

(1) Modification Method

Charpentier proposed a pitch modification method termed the PSOLA approach. We utilized this method which is performed in the time domain(TD-PSOLA) because of its simplicity. Pitch marks are set manually at the local peaks in the waveform segments of voiced portions. In the unvoiced portions, neither pitch mark setting nor pitch modification are performed. At the synthesis stage, successive windows are centered around the pitch marks and yield a set of speech samples weighted by the shape of the window. The windows are Hunning types and their length is set at twice the synthesis pitch to decrease the influence of spectral features. The pitch modified speech is obtained by means of overlap-adding the windowed samples which are successively shifted to match the synthesis pitch periods. The pitch modification procedure is described in Fig.2.

Shifting the windowed samples from their original pitch to the synthetic pitch causes an extension or shortening of waveform segment duration. To decrease this time scale distortion, remove or duplication procedures of the windowed samples are performed according to a duration compensation function **P** that is defined as the sum of the differences between **1** and the pitch modification ratio. When function value **P** exceeds **1**, it is regarded that the synthetic speech is one pitch shorter than the original. Therefore the windowed samples at that time, are duplicated. Similarly, when **P** is under **-1**, the windowed samples are removed(Fig.3).

The pitch modification ratio is automatically calculated. The pitch contour of a waveform 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. Then, the pitch modification ratio is defined as the ratio between the goal pitch and the corresponding segment pitch. This scheme has several advantages;

*The pitch frequency is modified uniformly.

*It compensates for the uncertainty of manual pitch mark setting.

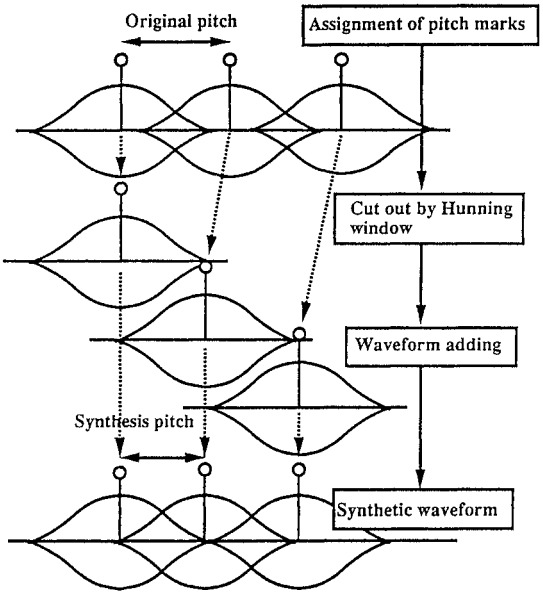


Fig.2 Pitch Modification Procedure

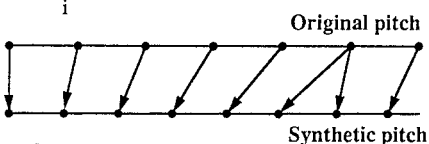
$$\mu pi = \frac{p_{si}}{p_{ai}}$$

μpi :Pitch modification ratio

p_{ai} :Original pitch

p_{si} :Synthetic pitch

$$P = \sum_i (1-\mu pi) > 1 \quad \{Xni\}\ \text{Duplicate}$$



$$P = \sum_i (1-\mu pi) < -1 \quad \{Xni\}\ \text{Remove}$$

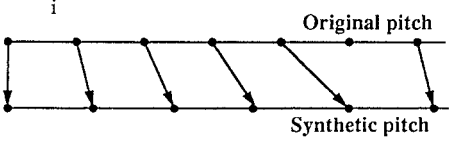


Fig.3 Time Compensation Procedure

(2) Acceptable limits for pitch shift

A big amount of shift from the original pitch degrades voice quality. In order to assess the acceptable limit, pitch shifted sounds were produced from natural sounds. The samples were then subjected to subjective acceptability tests. The test conditions were as follows;

[Pitch shift]
14steps; ±10%,±15%,±20%,±25%,±30%,±40%,±50%.
[Test sentences] 3 short sentences.
[Subjects] 13 males.
[Procedure] Subjects were instructed to judge acceptability against natural sound (YES or NO answer).

The relation of the acceptance rate to pitch shift step is shown in Fig.4. If the acceptability limit is 75%, the pitch shift limit must be ±0.2octave. This value is relatively small for speech synthesis based on concatenating waveforms, because conventional methods usually employ just a few waveform segments for phonemes or diphones. In our method, however, a number of segments are prepared for each phoneme. Therefore, it is easy to adjust the waveform pitch to the required pitch without sound quality deterioration.

(3) Evaluation of pitch modification

A preference test was conducted to evaluate the effect of pitch modification.

[Test sounds]
one:synthetic sound with pitch modification.
two:synthetic sound without pitch modification.
[Test sentences] 16 short sentences.
[Subjects] 7 males.
[Procedure] Judgement of "Which has the better quality?"

The preference scores and pitch modification rates are shown in Fig.5. As can be clearly seen, pitch modification is effective in almost all sentences. The contour of the pitch modification rate curve demonstrates a tendency similar to that of the preference score. This indicates that voice quality improves as pitch modification rate increases. No subjects objected to the voice quality change in pitch modified sounds and this is interpreted as indicating that the pitch modification rates of the test sounds are within the above-mentioned acceptability limit value.

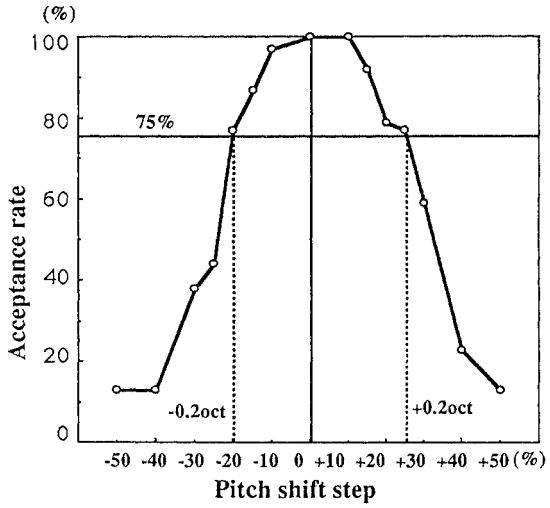


Fig.4 Relation between acceptance rate and pitch shift step

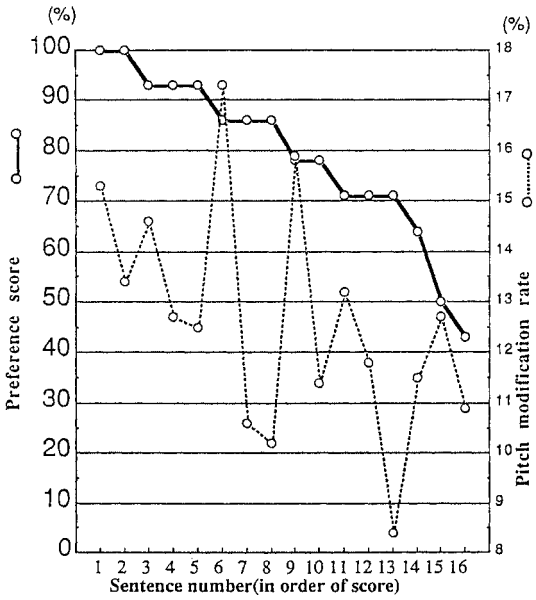


Fig.5 Preference score and pitch modification rate

4. INTRODUCING OF SPECTRAL CONTINUITY IN THE WAVEFORM SEGMENT SELECTIVE FUNCTION

In the previous function **H** for waveform selection, the first term simply represents the coincidence of phonological environment. A new additional factor **G** which expresses spectral continuity, is introduced to smooth the spectral transition.

$$G(dB)=/V_i-V_{i+1,k}/$$

/../; Difference between LPC spectral envelops
V; Waveform segment to be selected
i; Phoneme number
k; Candidate number

A new function for waveform selection is defined as follows;

$$Hnew=(1+G)H \quad \text{.....(2)}$$

Here, **1dB** is added so as to not overestimate the influence of spectral continuity, and the value was determined from informal hearing tests. Using this function, waveform segments are selected according to the following procedures;

*By the previous function **H**, the high ranking **n** candidates are picked up.
*At the voiced phoneme boundaries, waveform segments are selected from among the candidates by the new function **Hnew**.
*At the unvoiced phoneme boundaries, the top candidate is selected.

*These steps are applied in the right-to-left direction on input data.
These procedures were used to process five short sentences and the average spectral difference decreased nearly by half, that is from 5.44dB to 2.66dB.

In order to evaluate the effect of the spectral continuity factor, another comparative test was performed.
[Test sounds] 4 kinds of sounds (Table.1).
[Sentences] 5 short sentences, same in the estimation of spectral difference reduction.
[Procedure] Judgement of "Which has the better quality?"

Table 1 Test Sounds

	Spectral Continuity	Pitch Modification
Sound 1	off	off
Sound 2	on	off
Sound 3	off	on
Sound 4	on	on

The results shown in Fig.6 confirm an interesting fact. The spectral continuity factor is effective, however, the strength of the effect is much less than that of pitch modification. Two reasons for this are possible: one, voice quality is mainly attributed to pitch sensation and two, in some cases, it appears that segments with lower coincident degree of the physical characteristics among the candidates are selected because of the spectral continuity factor. Consequently, the adoption of this factor must be treated carefully. Figure 7 plots the relation between the preference score of sounds selected by the new function **Hnew** and spectral difference reduction. From this figure, it can be suggested that the spectral continuity factor should be introduced when the spectral difference reduction exceeds 3dB.

5. CONCLUSION

For the purpose of improving synthetic voice quality, two techniques, pitch modification and spectral continuity, were investigated. A pitch modification method based on PSOLA was proposed. Tests confirm that the pitch shift limit is ±0.2 octave and waveform pitch modification is effective in increasing sound quality.

A new waveform selection function involving a spectral continuity factor is proposed in this paper. Spectral continuity is also effective, however, its effect is comparatively smaller than the pitch modification effect and its adoption must be based on the spectral difference reduction value.

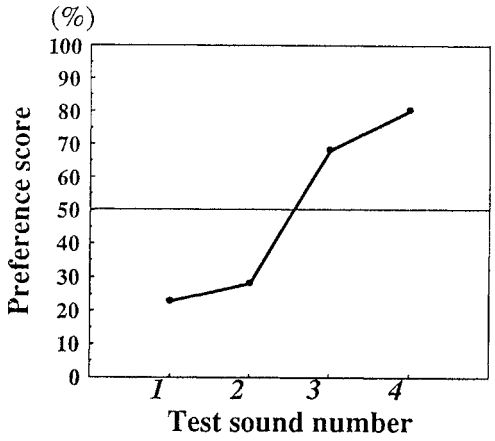


Fig.6 Results of preference test

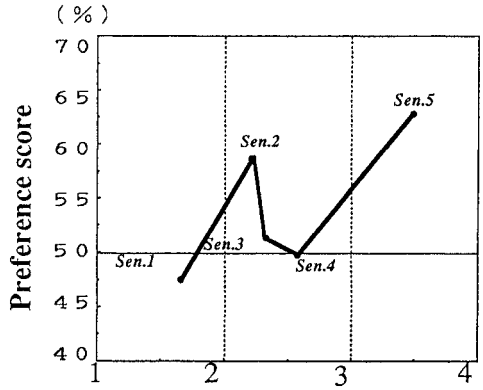


Fig.7 Relation between preference score and spectral difference reduction

ACKNOWLEDGEMENT

The authors wish to express their gratitude for the guidance and encouragement received from Dr. Sadaoki Furui and Dr. Hirokazu Sato. They also wish to express their thanks to the members of his group for participating in the hearing tests.

References

[1] K.Hakoda et.al."Japanese Text-to-Speech Synthesizer Based on Residual Excited Speech Synthesis" Proc.ICASSP,1986
[2] S.Nakajima,H.Hamada "Automatic Generation of Synthesis Units Based on Context Oriented Clustering" Proc.ICASSP,1988
[3] T.Yazu, k.Miki, M.Morito, K.Yamada "Speech Synthesis Method Using Symmetric Segmental Waveform" Trans. Commit. Speech Res.,ASJ,s83-67,1983(In Japanese)
[4] Y.Mitome, K.Fushikida "A Speech Synthesis Method for Unrestricted Words Using Pitch Synchronous Interpolation between CV-VC Waveforms" Trans. Commit. Speech Res.,ASJ,sp82-06,1982(In Japanese)
[5] T.Hirokawa "Speech Synthesis Using a Waveform Dictionary" Proc.Eurospeech'89,1989
[6] F.Charpentier,E.Moulines "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones" Proc.Eurospeech'89,1989
[7] K.Abe,Y.Sagisaka "A synthesis Unit Selection Method Adapting to an Input Phoneme" Proc.Spring Meet.ASJ,1-1-23,1988(in Japanese)
[8] T.Nomura,H.Sato "Representation of Continuous Speech by the Concatenation of Phonetic Segments for Speech Synthesis" Proc.Spring Meet.ASJ,1-7-21,1989(in Japanese)