MANIPULATING SPEECH PITCH PERIODS ACCORDING TO OPTIMAL INSERTION/DELETION POSITION IN RESIDUAL SIGNAL FOR INTONATION CONTROL IN SPEECH SYNTHESIS

Toshio Hirai[†‡] *Seiichi Tenpaku*[†] *Kiyohiro Shikano*[‡]
†: Arcadia Inc., Japan ‡: Nara Institute of Science and Technology, Japan
e-mail: thirai@arcadia.co.jp

ABSTRACT

This paper describes the investigation of manipulating positions in a speech pitch when lengthening or shortening the pitch period, that is, lowering or raising fundamental frequency of speech. The experimental results revealed that the preferable positions were at the first half of the pitch period for pitch shortening, and at the second half of it for pitch lengthening. The findings are expected to improve the quality of speech synthesis on pitch modulation.

1. INTRODUCTION

Fundamental frequency (F_0) control is one of the most important fields of research to improve the naturalness of synthetic speech. In order to obtain the F_0 pattern, which is estimated from linguistic information etc. in a text-to-speech system, an approach, in which acoustical parameters (e.g., LPC parameters) are excited with impulse in desired frequency, has been used[1]. In this approach, a speech waveform is predicted from a set of acoustical parameters and an impulse input.

In some research, the input has been replaced with a residual signal sequence corresponding to the difference between the original and the re-synthesized waveforms[2, 3], in order to improve the naturalness of the re-synthesized speech. In this method, to raise F_0, in other words, to shorten the pitch period, some of the residual signal sequence is deleted. On the other hand, to lower F_0, a zero sequence (whose length is appropriate to achieve the desired pitch period) is inserted into the residual signal in order to obtain the desired pitch period in re-synthesized speech.

However, in these past research efforts, there was little (or no) consideration given to the timing of deleting the residual signal sequence or inserting a zero sequence. Therefore, we were motivated to investigate to find out the optimal insertion/deletion position within a pitch of speech. Furthermore, a new method to manipulate the pitch length was proposed and described in this paper.

2. INVESTIGATION APPROACH

As the first step of our investigation, we varied the insertion/deletion position by using a parameter q ($0 \quad q < 1$) during the F_0 modulation process, and then evaluated the quality of the re-synthesized speech in each case by listening. The manipulating position t_m was defined as:

$$t_m = t_{ps} + q \times p_p$$

where

t_{ps} is pitch starting position, and
p_p is pitch period.

The positions are illustrated in **Figure 1**. If the value of q is set near to zero, the manipulating position is found at the beginning of the pitch, and if the value is set near to one, the manipulating position is found near the end of the pitch. The details of each manipulation are described in the next two subsections.

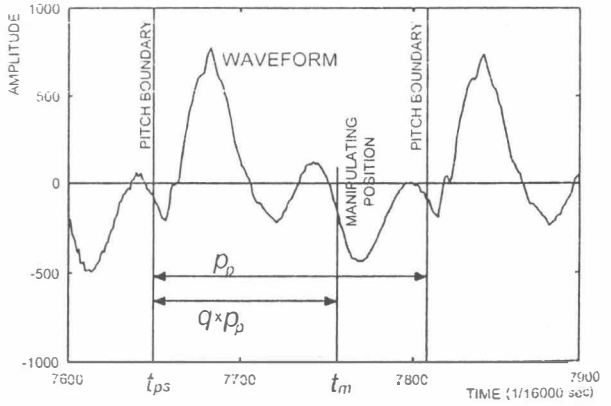

Figure 1: The relationship among pitch starting position t_{ps}, pitch period p_p, q, and manipulating position t_m.

2.1. Insertion Manipulation

As we mentioned in **1. Introduction**, in the previous research, some zeros were inserted into the residual for lengthening pitch[2]. However, zero signal insertion into the residual signal generates unnaturalness in the re-synthesized speech which can be found also in the single impulse LPC synthesis system. Therefore, the residual signal sequence of the original waveform was used instead of the zero signal insertion to lengthen period (p_l) after t_m in this investigation.

In this proposed method, after the use of this original residual signals, the original sequence was used again for the prediction. Therefore, this method can be described as the duplication of residual signal for p_l before t_m. In such concatenation of residual signal sequence, we should consider about the "concatenation gap" at the joining point. However, according to the preliminary re-synthesis experiment of the deletion manipulation, the speech quality is not degraded by the deletion of residual sequence even though there were concatenation gaps of residual signal at the cutting point. Therefore, the duplication of residual signal sequence at the manipulating position was adopted in the investigation. The duplicated portion of the original residual signal segment starts from t_{rs} ($= t_m - p_l$) and ends at t_m.

Followings are the summary of the manipulation. The flow is illustrated in **Figure 2** (the entitled numbers correspond to each other):

1. Residual signal after t_{rs} is calculated.
2. Speech signals before t_m are slid forward for the period p_l as the seed for LPC re-synthesis processing. The length of the slid signal sequence is equal to the order of the LPC processing (N).
3. New waveform is estimated from the slid original speech signal sequence, LPC parameters of the pitch, and the residual sequence according to the LPC re-synthesis calculation:

$$y_i = r_i + \sum_{j=1}^{N} b_j \times y_{i-j}$$
$$(i = t_{rs}, t_{rs}+1, t_{rs}+2, \cdots)$$

where

r_i is residual signal at $t = i$,
N is the order of LPC,
b_j is j-th LPC parameter value, and
y_i is slid original speech signal
($t_{rs} - N \le i < t_{rs}$) or
predicted waveform ($t_{rs} \le i$).

4. The difference between the original waveform and the estimated one is calculated. The point with the minimum difference is the transferring position from the estimated signal to the original signal.

2.2. Deletion Manipulation

In order to shorten the pitch period, a part of the residual signal sequence was deleted in this investigation. In this method, some residual signals are deleted for shortening length (p_s) after the manipulating position t_m. Then, new waveform is predicted according to the preceding waveform, the residual signal sequence after the deleted one, and the LPC parameters.

By the way, it is considered that the deletion effect is large if q is large, since the residual sequence located at the pitch boundary, which contains much information, is deleted with the shortening manipulation. Therefore, in this investigation, the manipulating position was set at $t_m - p_s/2$. Then the residual signals from this position for p_s was deleted. Consequently the effect of eliminating the residual at the pitch end, was expected to be suppressed.

Table 1: Speech attributes

sampling freq.	16kHz
talker	male, unprofessional narrator
utterance	"koohiini mirukuo iremasuka?" (Do you need milk for coffee?)
duration	1.88 sec

Table 2: LPC analysis conditions

order	16
window length	2 pitches
window center	center of the pitch
window type	Blackman

3. EXPERIMENTS

3.1. Speech Data

One speech sample was used for the manipulation of the pitch period. Each pitch was bounded at the sudden descent positions of the Electro Glott-Graph (EGG) signal which was attached to the waveform data[4]. Roughly speaking, the boundaries were located at the positions of the residual signal with the maximum amplitude. Compared to the EGG signal, a speech signal delays due to the propagation of speech sound wave from lips. The attributes of the analyzed/re-synthesized speech are shown in **Table 1**.

3.2. Speech Analysis Tool and Analysis Parameters

As an speech analysis tool, the Speech Signal Processing Toolkit[5] was used. The conditions of the LPC analysis are shown in **Table 2**.

3.3. Manipulation Procedure of Lengthening/Shortening Pitch Period

Manipulation was performed on all pitch periods of the speech signals of from -50 samples to $+50$ samples in 10 samples steps. The manipulation corresponded to the F_0 modification from $+120$Hz to -40Hz approximately for the used speech. The duration of each re-synthesized speech was changed depending on whether lengthening or shortening. The manipulating positions were located at $q = 0.2, 0.3, \cdots, 0.8$.

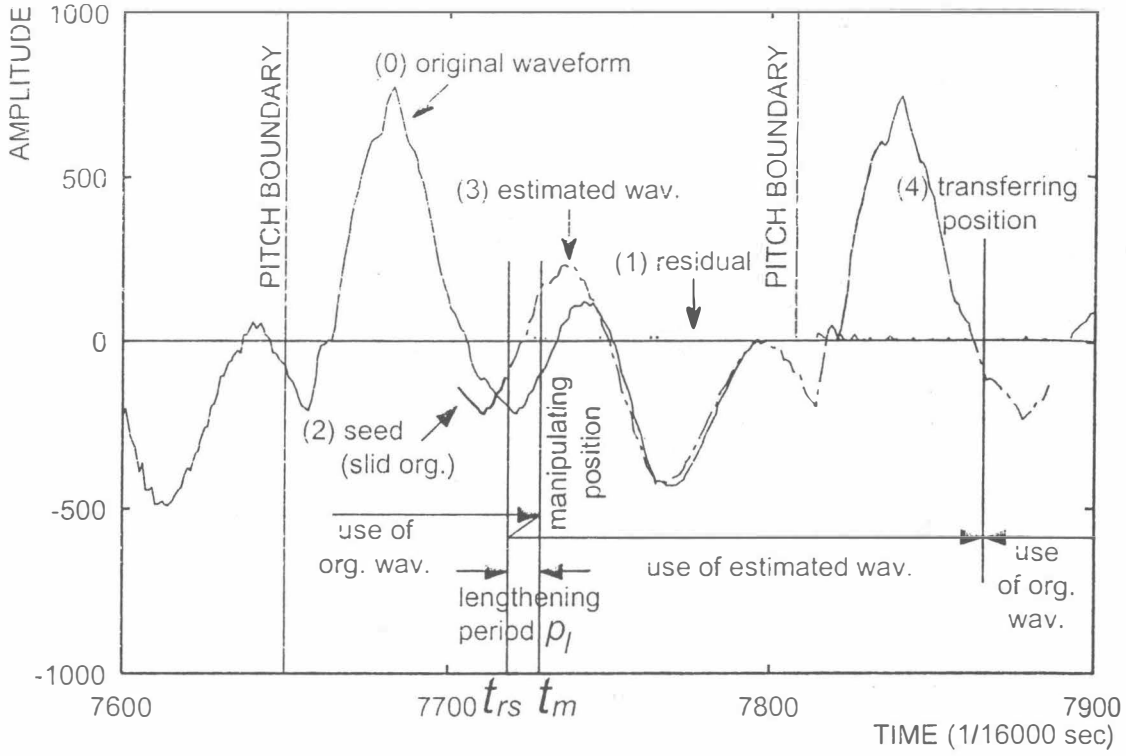

Figure 2: Illustration of the pitch lengthening manipulation

4. RESULTS AND DISCUSSIONS

The listening test that evaluated the quality of the F_0 modulated speech revealed that a large q for insertion manipulation and a small q for deletion were preferable, respectively. These tendencies were intensified by the increase of the manipulating length in both cases. The degradation of the manipulation of deletion was greater than that of the manipulation of the insertion.

The reason why a smaller q was preferred in the manipulation of deletion is that the local maximum of the residual signal, which was roughly corresponding with the pitch ends, would be deleted if the value of q was large. In other words, the deletion of the residual signal around the end of the pitch might cause the degradation of re-synthesized speech as described in **2.2. Deletion Manipulation**.

On the other hand, the reason why larger q was preferred in the case of lengthening the pitch period seems that the residual signal has more information, which can't be described by LPC parameters, at the first half of the pitch period than the second half of it. Therefore, if the location of manipulation is set in the second half of the pitch (that is, q is large), the duplicated residual doesn't have much information itself, so the re-synthesized speech wouldn't get great damage.

5. CONCLUSIONS

The investigation of the manipulating position in a speech pitch for F_0 modulation was executed. Many manipulating positions for insertion/deletion were tried out during the re-synthesizing process which were applied to the speech waveform by using LPC parameters and the modified residual signal sequence. The results of the experiments revealed that the preferable positions were at the first half of the pitch period for pitch shortening (F_0 raising), and at the second half of it for pitch lengthening (F_0 lowering). The findings are scheduled to be employed into a speech synthesis system.

6. RE-SYNTHESIZED SPEECH SAMPLES

The re-synthesized waveforms in this research are accessible through the internet site:

http://www.arcadia.co.jp/thirai/icslp2000/

7. ACKNOWLEDGEMENT

The authors appreciate Prof. H. Kawahara and Mr. Y. Atake for their offering the speech data with the EGG signal.

8. REFERENCES

1. J.D. Markel and A.H. Gray. *Linear prediction of speech*. Springer-Verlag, 1976.
2. H. Kuwabara and T. Takagi. Speech quality control by the analysis-synthesis method and an application to the enhancement of abnormal speech. In *Trans. of the Committee on Speech Research, Acoust. Soc. Jpn.*, SP86–57, 1986. In Japanese.
3. N. Minematsu, S. Nakagawa, and K. Hirose. Prosodic manipulation system of speech material for perceptual experiments. In *Proc. ICSLP*, pages 2056–2059, 1996.
4. H. Kawahara and Y. Atake. Vocal fold closure and speech event detection using group delay. In *Trans. of the Committee on Speech Research, Acoust. Soc. Jpn.*, SP99-171, 2000. In Japanese.
5. K. Tokuda. Reference manual for Speech Signal Processing Toolkit ver. 2.0, 2000. http://kt-lab.ics.nitech.ac.jp/~tokuda/SPTK/.