QUALITY IMPROVEMENT OF PSOLA ANALYSIS-SYNTHESIS USING PARTIAL ZERO-PHASE CONVERSION

Nobuaki MINEMATSU† *and* *Seiichi NAKAGAWA*‡
mine@gavo.t.u-tokyo.ac.jp *nakagawa@slp.ics.tut.ac.jp*

†Graduate School of Engineering, University of Tokyo
‡Department of Information and Computer Sciences, Toyohashi University of Technology

ABSTRACT

This paper discusses two issues of the quality improvement of *F0* modified speech based upon PSOLA analysis-synthesis. Previous studies[1][2] pointed out that the location of a window of PSOLA influences the quality of synthesized speech and one of them claimed that the center of a window should be located at a pitch pulse in source waveforms. However, pitch pulse detection sometimes fails due to undesired acoustic events. In this paper, several methods are experimentally examined to reduce pitch pulse detection errors. Even when the detection is done correctly, *F0* modified re-synthesized speech sometimes causes "echoes" in the re-arranged waveforms. This is mainly caused by a pitch pulse with small sharpness or by that with two relatively high pulses, not pitch pulses, before and after it. To suppress the echoes with little loss of naturalness, partial zero/π-phase conversion is proposed here. Experiments show the high validity of the proposed methods in improving the quality of re-synthesized speech.

1. INTRODUCTION

We developed a system of generating *F0* modified speech in our previous study[3], which was intended to be used as a speech sample generator for experiments on *F0* contour control of TTS systems or human perception of word accent and sentence intonation. This study describes two issues of the quality improvement of *F0* modified speech by updating several modules in the above system.

The system was based upon PSOLA analysis-synthesis techniques, where PSOLA was performed on source waveforms, not on speech waveforms. This is because a previous study reported that PSOLA conducted directly on speech waveforms sometimes causes spectral distortion and leads to the speech quality degradation[1]. And this spectral distortion is thought to be able to be suppressed by doing the re-arrangement on source waveforms[4]. In our previous study[3], source waveforms were obtained by using LMA (Log Magnitude Approximation) inverse filter, which was designed only by using cepstrum coefficients and could precisely approximate magnitude characteristics of vocal tract in a logarithmic scale[5]. This filter was also adopted as a re-synthesizer, namely, a cepstrum vocoder.

Previous studies pointed out that the location of a window of PSOLA influences the quality of synthesized speech[1][2] and one of them claimed that the center of a window should be located at a pitch pulse in source waveforms. Following this finding, in this paper, PSOLA on source waveforms is done with windows centered at pitch pulse locations. However, pitch pulse detection sometimes fails due to undesired or unexpected acoustic events. These failures inevitably decrease the quality of re-synthesized speech. Firstly in this paper, acoustic factors which causes the detection errors are investigated and analyzed. And based upon the findings obtained, a method is proposed to deal with each factor of the detection errors.

Even when pitch pulses are detected correctly, *F0* modified re-synthesized speech sometimes causes "echoes" in the re-arranged waveforms. This is mainly caused by a pitch pulse with small sharpness or by that with two relatively high pulses, not pitch pulses, before and after it. Although zero-phase conversion on source waveforms can effectively reduce the echoes, the naturalness of synthesized speech is often decreased for other reasons than the echoes. To suppress the echoes with little loss of naturalness, partial zero/π-phase conversion is proposed in this study. Since the sharpness of pitch pulses can be controlled by a parameter in this method, the required modification of the pulses with little quality degradation can be realized.

2. REDUCTION OF PITCH PULSE DETECTION ERRORS

2.1. Pitch Pulse Detection Errors

In this paper, pitch extraction is performed in the following manner. Initial values of pitch are calculated by using an autocorrelation-based method, where length of a window is determined proportionally to length of time lag τ[6]. These values are re-calculated and refined by post processing[2]. Using the initial *F0* values and voicing degree of each frame, the post processing is performed. For a given voiced segment, time index the absolute magnitude of a source waveform at which is the largest is firstly obtained, which is referred to as t_0. Then, using t_0 as starting point, autocorrelation of the source waveforms is calculated from between t_0 and $t_0+1/f(t_0)-\epsilon$ to between t_0 and $t_0+1/f(t_0)+\epsilon$, where $f(t_0)$ is an initial value of *F0* at time t_0. Out of the autocorrelation values, the maximum, whose time index is t_1, is selected and $1/(t_1-t_0)$ is adopted as a final *F0* value. This procedure goes on by using t_1 as new starting point. Similar operations are done by calculating autocorrelation from between t_0 and $t_0-1/f(t_0)-\epsilon$ to between t_0 and $t_0-1/f(t_0)+\epsilon$. And final *F0* values are determined throughout the given voiced segment.

Pitch pulse marking for building a waveform database for concatenation-based TTS systems is considered to allow voiced/unvoiced decision errors if the number of errors is quite small compared to the size of the database. In this study, however, we are developing an *F0* modified speech

generator, which shall be used in speech perception experiments. In this case, voiced/unvoiced decision errors are not allowed because *F0* manipulation in a voiced segment is impossible if the segment is judged to be unvoiced. Voicing degree in this study is defined as an autocorrelation value of low-pass filtered source waveforms. To guarantee that a voiced segment is never judged as an unvoiced one, a threshold for the decision has to be set relatively low. Preliminary experiments showed that this requirement resulted in increasing pitch pulse detection errors if the detection method of [2] was used as it was proposed.

Analysis of the pitch pulse detection errors showed that the reasons could be categorized into several acoustic factors.

Inappropriate setting of a starting point t_0
At the edges of a voiced segment of speech waveforms, a large perturbation is easily found and this also causes a large perturbation on source waveforms (**A**). At the locations other than the edges of speech waveforms, small perturbations are easily found. These sometimes causes large perturbations on source waveforms (**B**). If these points are selected as starting points, pitch pulse detection errors should become likely to occur.

Pitch pulse detection over an entire voiced segment
Assume that two different voiced phonemes are uttered and the transitional portion between the two is judged to be voiced by using a low threshold. As told above, pitch pulses are detected by shifting by time interval $1/f(t_0)$ over the entire voiced segment. As a result, the case is sometimes found where the pulses are correctly detected before the transitional portion and are *not* after it (**C**).

Pulse waveform between two consecutive pitch pulses
Some phonemes/speakers show that two consecutive pitch pulses sometimes have a pulse waveform between them. If this pulse has large magnitude, it should be detected as a pitch pulse. This detection error is quite troublesome. A pulse waveform between two pitch pulses often has another pulse $1/f(t_0)$ [sec] away from the first pulse. This means that, once this pulse is detected as a pitch pulse, pitch pulse detection fails repeatedly for a long period (**D**).

Figures 1 to **3** show examples of factors **B** to **D**. In these figures, source waveforms are generated by using LMA inverse filter and their power is normalized.

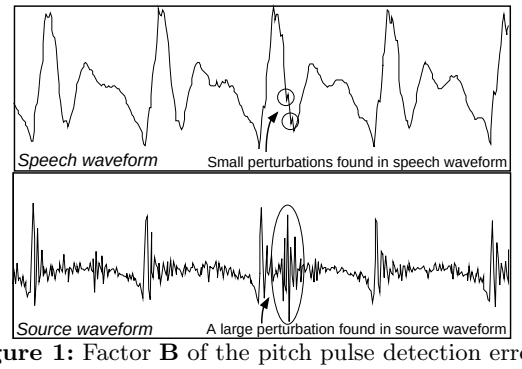

Figure 1: Factor **B** of the pitch pulse detection errors

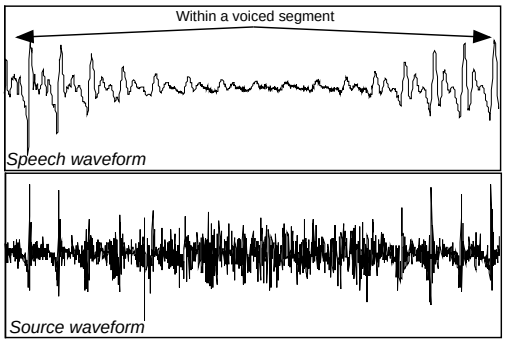

Figure 2: Factor **C** of the pitch pulse detection errors

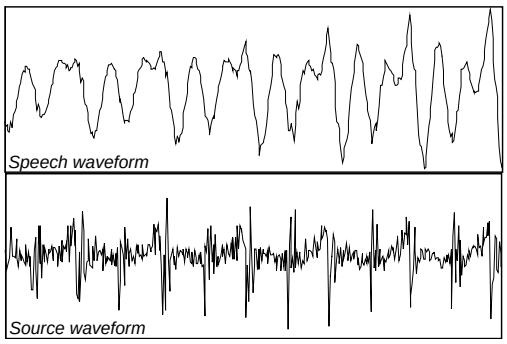

Figure 3: Factor **D** of the pitch pulse detection errors

2.2. Reduction of the Detection Errors

Errors caused by factor **A** can be easily avoided by not using either edge of a voiced segment as an initial position of pitch pulse detection (**A'**). As for factor **B**, considering that perturbation on source waveforms can never be found periodically, the following method can reduce errors caused by **B**. The initial position of pitch pulse detection should be determined by using the magnitude of autocorrelation at $t_0-1/f(t_0)$ and $t_0+1/f(t_0)$, namely, preceding and following pitch pulses, in addition to the position of the largest magnitude on source waveforms, i.e. t_0 (**B'**).

Pitch pulse detection should be done in individual voiced segments separately but does not have to be done with an interval of $1/f(t_0)$ repeatedly throughout a segment. In other words, it is allowed to be done after dividing a voiced segment into sub-segments. If the division is possible at transitional portions in the segment, errors by factor **C** can be reduced. **Figure 4** shows two types of autocorrelation (voicing degree). They are obtained by using two different cut off frequencies of LPF, which is performed on source waveforms. Voicing degree by low cut off frequency (500 Hz) and that by high cut off frequency (1500 Hz) will be called vd_L and vd_H respectively in the rest of the paper. Portions which has low vd_H and high vd_L at the same time correspond well to the transitional portions between two consecutive voiced phonemes. Detail procedures for dealing with errors by factor **C** is as follows.

A voiced segment is defined by using vd_L and its threshold, in which pitch pulses should be detected. And voiced segment n is divided into sub-segments $\{n_i\}$ by using vd_H and its threshold. The threshold to divide segment n is calculated as a mean value of vd_H over segment n. Boundary between sub-segments n_i and n_{i+1}, $t_{i,i+1}$, is calculated as

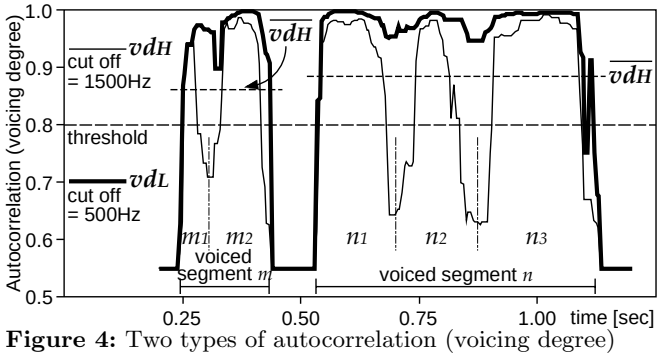

Figure 4: Two types of autocorrelation (voicing degree)

$$t_{i,i+1} = \frac{\sum_{t \in n_i \sim n_{i+1}, vd_H(t) < \overline{vd_H(t)}} w(t) \times t}{\sum_{t \in n_i \sim n_{i+1}, vd_H(t) < \overline{vd_H(t)}} w(t)} \quad (w(t) = 1 - vd_H(t)).$$

By detecting pitch pulses on individual sub-segments separately, errors by factor **C** could be reduced (**C'**).

Detection errors caused by factor **D** is handled by the following method. For sub-segment n_i obtained by **C'**, $N(n_i)$ candidates for an initial position for pitch pulse detection is generated by **B'**. Number $N(n_i)$ is determined proportionally to length of segment n_i. After that, for each of the initial position candidates, pitch pulses are detected. And a set of pitch pulses whose averaged magnitude is the largest is adopted as a final set of pitch pulses of segment n_i and they are used for PSOLA analysis-synthesis.

2.3. Evaluation Experiments

To examine the validity of the above methods **A'** to **D'**, several experiments were designed and carried out. 100 different sentences spoken by 5 male speakers and 5 female speakers were prepared. They were digitized with 10 kHz and 16 bit sampling. *F0* and LPC cepstrum coefficients (1 to 24 dimensions) were calculated with 5 msec frame rate. With the speech material, pitch pulses were firstly detected without using methods **A'** to **D'**. And re-synthesized speech was generated with the detected pitch pulses. In the experiments, all the *F0* contours for the re-synthesis were generated by manually approximating the observed contours with Fujisaki model[7]. The manual operations often fixed pitch extraction errors. It should be noted that locations of pitch pulses were different between original source waveforms and those for the re-synthesis. Next, another set of re-synthesized speech were generated by applying the proposed methods **A'** to **D'** to correctly detect pitch pulses of the same 100 sentence speech. The only difference between two sets of re-synthesized speech was the locations of pitch pulses used for the re-synthesis.

To five male subjects, three sentence speech samples, original, X, and Y, were presented in one session. To X and Y, re-synthesized speech without the proposed methods and that with the methods for the same sentence were randomly assigned. After hearing the three samples, subjects were asked to judge which of X and Y was more natural and how large the difference was between the two samples.

Table 1: Naturalness improvement by decreasing pitch pulse detection errors
P/Q : without/with the proposed methods **A'** to **D'**

	P≫Q	**P>Q**	**P≥Q**	**P=Q**	**P≤Q**	**P<Q**	**P≪Q**
male	0.0	0.0	4.4	33.2	6.0	4.2	2.2
female	0.0	0.0	2.4	27.2	7.2	8.4	4.8

These two tasks were done simultaneously by selecting an answer out of seven candidates, i.e. $X \gg Y$, $X > Y$, ..., $X < Y$, and $X \ll Y$, which corresponded to seven levels of the differences. **Table 1** shows the naturalness improvement obtained by using the proposed methods. In the table, the probability of each candidate being selected as an answer is indicated in the form of percentage. Clearly seen here, the proposed methods are quite effective to improve the naturalness and the improvement is larger for female speech. This tendency is because pitch pulse detection errors are more likely to occur in female speech.

3. PARTIAL ZERO/π-PHASE CONVERSION ON SOURCE SIGNALS

3.1. Partial Zero/π-phase Conversion

The previous section realized the naturalness improvement for re-synthesized speech with little *F0* modification by reducing pitch pulse detection errors. Even when pitch pulses are detected correctly, however, echoes are sometimes generated with *F0* modification. And preliminary analysis indicated that the generation of the echoes was mainly attributed to source waveforms. For example, a pitch pulse with small sharpness and that with relatively high pulses around it (see **Figure 3**) were likely to produce the echoes. To suppress the echoes, it is be effective to conduct zero-phase conversion on source waveforms, which can increase the sharpness. However in this case, while the echoes are suppressed, the naturalness of the synthesized speech is decreased for other reasons than the echoes.

To suppress the echoes with little loss of naturalness, partial zero/π-phase conversion is introduced here. In the method, a window of source waveforms was partially zero-phase converted in the following equation.

$$\begin{aligned} x_r(k) &:= (-1)^k \sqrt{x_r(k)^2 + x_i(k)^2} && (k < K) \\ x_i(k) &:= 0.0 && (k < K), \end{aligned}$$

where the complex form of spectrum $(x_r(k), x_i(k))$ is converted only within the range of $k < K$. Applying this conversion on source waveforms concentrates their power on the window center and the sign of the waveform at the center comes to be positive. By converting $(-1)^k$ term into $(-1)^{k+1}$, π-phase conversion is defined, where the power is also concentrated on the center but the sign of the waveform at the center comes to be negative. **Figure 5** shows partial zero-phase conversion in the upper figures and partial π-phase conversion in the lower figures. In the figures, various values are assigned to K. Which conversion to perform is determined not by referring to the original sign of a pitch pulse at the window center but by referring to the signs observed more often in the voiced segment.

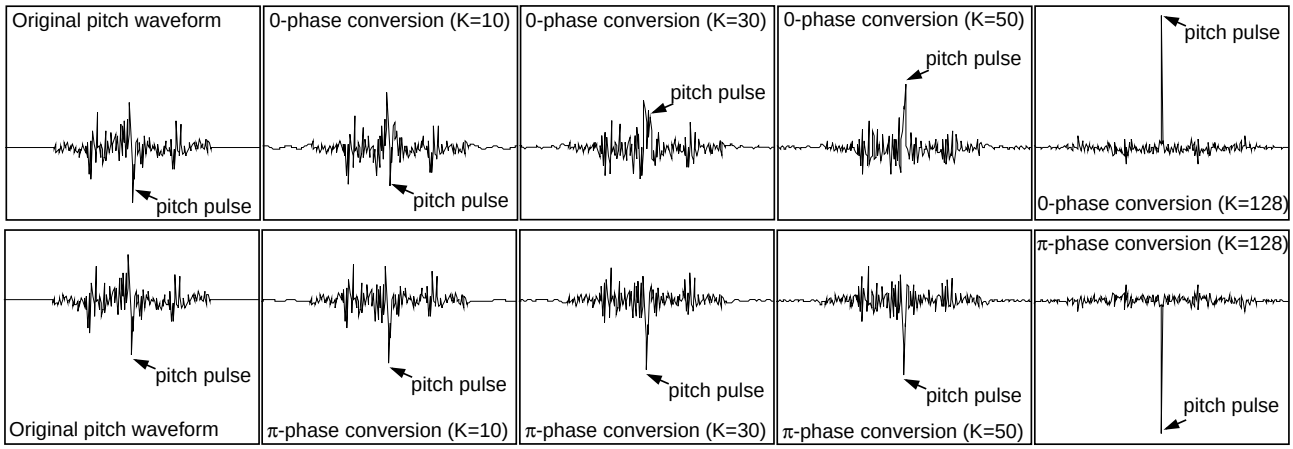

Figure 5: Partial zero-phase conversion (the upper figures) and partial π-phase conversion (the lower figures)

Table 2: Naturalness improvement by performing partial zero/π-phase conversion
P : with no conversion
Q : with partial zero/π-phase conversion

	$\mathbf{P \gg Q}$	$\mathbf{P > Q}$	$\mathbf{P \geq Q}$	$\mathbf{P = Q}$	$\mathbf{P \leq Q}$	$\mathbf{P < Q}$	$\mathbf{P \ll Q}$
male(+0.6)	0.0	0.0	1.4	22.6	11.0	12.6	2.4
female(+0.6)	0.0	0.0	0.4	6.8	14.8	18.4	9.6
male(−0.6)	0.0	0.0	2.0	45.2	2.8	0.0	0.0
female(−0.6)	0.0	0.0	0.4	32.4	12.6	4.2	0.8

Q : with partial zero/π-phase conversion
R : with complete zero-phase conversion

	$\mathbf{R \gg Q}$	$\mathbf{R > Q}$	$\mathbf{R \geq Q}$	$\mathbf{R = Q}$	$\mathbf{R \leq Q}$	$\mathbf{R < Q}$	$\mathbf{R \ll Q}$
male(+0.6)	0.0	0.0	0.0	12.2	17.0	18.2	2.6
female(+0.6)	0.0	0.0	2.2	12.0	18.4	14.2	3.2
male(−0.6)	0.0	0.0	0.4	23.0	16.2	8.0	2.4
female(−0.6)	0.0	0.0	0.0	13.4	19.8	12.2	4.6

3.2. Evaluation Experiments

Speech samples used in Section **2.3.** were also used here. After pitch pulses were detected from these samples with the proposed methods in Section **2.**, each of two types of F_0 modification, +0.6 [oct] and −0.6 [oct] conversion, was done by using three types of source waveforms, with no conversion, with partial zero/π-phase conversion, and with complete zero-phase conversion. Evaluation was done similarly to that in Section **2.3.** except that original samples were not presented here. Upper bound K in partial zero/π-conversion was set proportional to the target F_0 value. And it was determined experimentally as follows.

$$K = \frac{700}{\text{target fundamental period [sample]}}$$

Table 2 shows results of the experiments separately for each of male and female speech samples and for each of two types of F_0 modification. The tables indicate that the re-synthesized speech generated with partial zero/π-phase conversion is the most natural of the three types of source waveforms examined, and that irrespective of gender of the speakers and of direction of the F_0 modification. Comparison between results of male speech and those of female speech shows that the improvement is larger in female speech as in Section **2.3.**. This is because pitch pulses with small sharpness are found more often in female speech. Another comparison between directions of the F_0 modification indicates that the improvement is larger when F_0 is increased. This is simply due to inevitable overlaps of re-arranged waveforms when F_0 is converted to be higher.

4. CONCLUSIONS

This paper realized the quality improvement of F_0 modified speech based upon PSOLA analysis-synthesis. Firstly, the reduction of pitch pulse detection errors was examined. Here, the errors were categorized into several groups and, for each group, an effective method was proposed. Even when pitch pulses are detected correctly, F_0 modified speech sometimes causes echoes in re-arranged waveforms. This is mainly caused by pitch pulses with small sharpness. To suppress the echoes with little loss of naturalness, partial zero/π-phase conversion was secondly proposed in this paper. Evaluation experiments showed that both of the proposed methods were quite effective in improving naturalness of F_0 modified speech and that especially in the cases of female speech and/or increasing F_0.

REFERENCES

1. H. Kawai *et al.*, "A study of a text-to-speech system based on waveform splicing," Technical report of IEICE, SP93-9, pp.49–54 (1993, in Japanese).
2. Y. Arai *et al.*, "A study on the optimal window position to extract pitch waveforms," Technical report of IEICE, SP95-8, pp.53–59 (1995, in Japanese).
3. N. Minematsu *et al.*, "Prosodic manipulation system of speech material for perceptual experiments," Proc. ICSLP'96, pp.2056–2059 (1996).
4. F. Charpentier *et al.*, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Proc. EUROSPEECH'89, pp.13–19 (1989).
5. S. Imai, "Log Magnitude Approximation (LMA) Filter," Trans. IEICE, J63-A, 12, pp.886–893 (1980, in Japanese).
6. H. Fujisaki *et al.*, "Proposal and evaluation of a new scheme for reliable pitch extraction of speech," Proc. ICSLP'90, pp.473–476 (1990).
7. H. Fujisaki *et al.*, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J. Acoust. Soc. Jpn.(E), vol.5, no.4, pp.233–242 (1984).