Ann.Bull.RILP(1984)No.18, 113-118

EFFECT OF WINDOWING THE RESIDUAL WAVE ON SYNTHESIZED SPEECH QUALITY

Shuzo Saito and Kenzo Itoh*

1. Introduction

In a speech analysis-synthesis system using the source coding method, two kinds of feature parameters are utilized, that is, excitation source and vocal tract parameters. The excitation source of such a speech analysis-synthesis system is usually composed of a pulse and a white noise generator. When a residual wave signal is used as the excitation source signal instead of such a pulse and a white noise generator, the synthesized speech quality is much improved in terms of linear prediction coding, such as the PARCOR coding system.

Several studies have attempted to reduce the channel capacity of the residual wave signal[1)-3)], as the original residual wave has information contents comparable to the speech input signal. In this paper, the preliminary studies on the effect of time-windowing the residual wave signal on the quality of synthesized speech is described.

2. Experimental Procedure

2.1 Speech analysis-synthesis system

The PARCOR speech analysis-synthesis system was used to extract the PARCOR coefficients of 12 orders and also the residual wave signal. A block diagram of the measuring system is shown in Fig. 1. The speech signal was passed through a low-pass filter of 4.8 kHz cut-off frequency, and its amplitude was then digitized into 12 bits every 100 microseconds of the sampling period. This A/D converted signal was then fed to the PARCOR analyzer, and the PARCOR coefficients of 12 orders were extracted in every frame period of five milliseconds. Then the input speech signal was filtered inversely by the PARCOR coefficients, and the residual wave signal was derived.

Various kinds of time-windowing processing were applied to the residual wave, as will be described below. Such a modified residual wave was fed to the PARCOR synthesizer and used as the excitation source signal. This source signal was fed to the PARCOR ladder type digital filters of 12 stages, which were controlled by the PARCOR coefficients of 12 orders. The synthesized speech output was then used for a preliminary auditory test.

*Musashino Electrical Communication Laboratory, Nippon Telegraph and Telephone Public Corporation

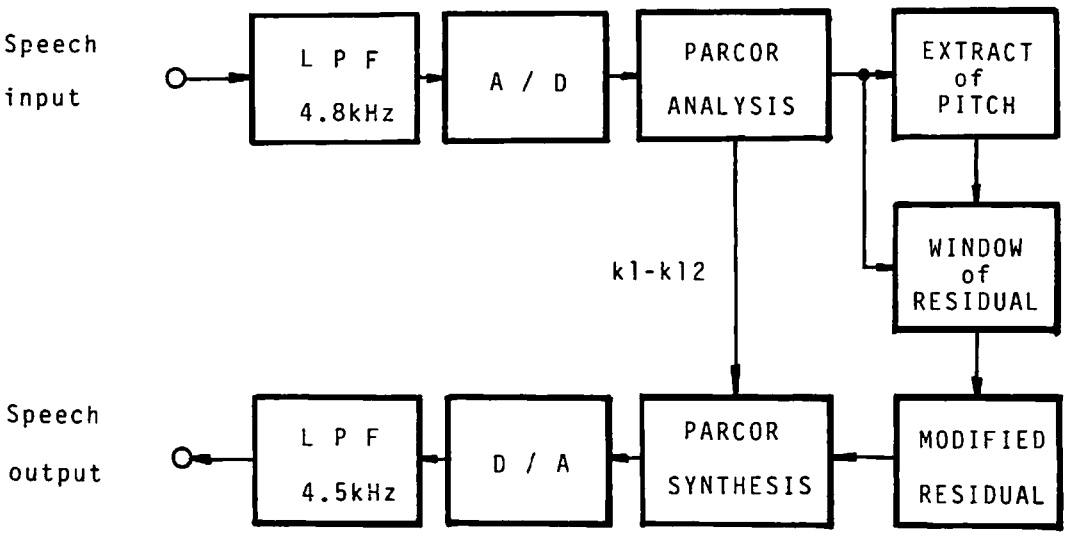

Fig. 1 Block diagram of the speech analysis-synthesis system.

2.2 Time-windowing

Various kinds of time-windows were applied to the residual wave signal as shown in Fig. 2. The time-windowing was synchronized to the pitch period of the residual wave signal. The length of the time-window was shortened gradually and up to a single pulse centering on the maximum value of the residual wave signal at first. Then the position of the time-window was moved back and forth, keeping the length of the time-window at a fixed value.

2.3 Speech material

The speech material used for the preliminary, auditory tone quality test was utterances of a Japanese sentence lasting about 3 seconds produced by a male and a female speaker.

2.4 Subjects

Two male listeners with normal hearing engaged in the tone quality test under a monaural listening situation. A subjective rating method was used for this tone quality test.

3. Results

3.1 Effect of the length of the time-window

The effect of the length of time-windowing the residual wave signal on synthesized speech quality was measured by varying the

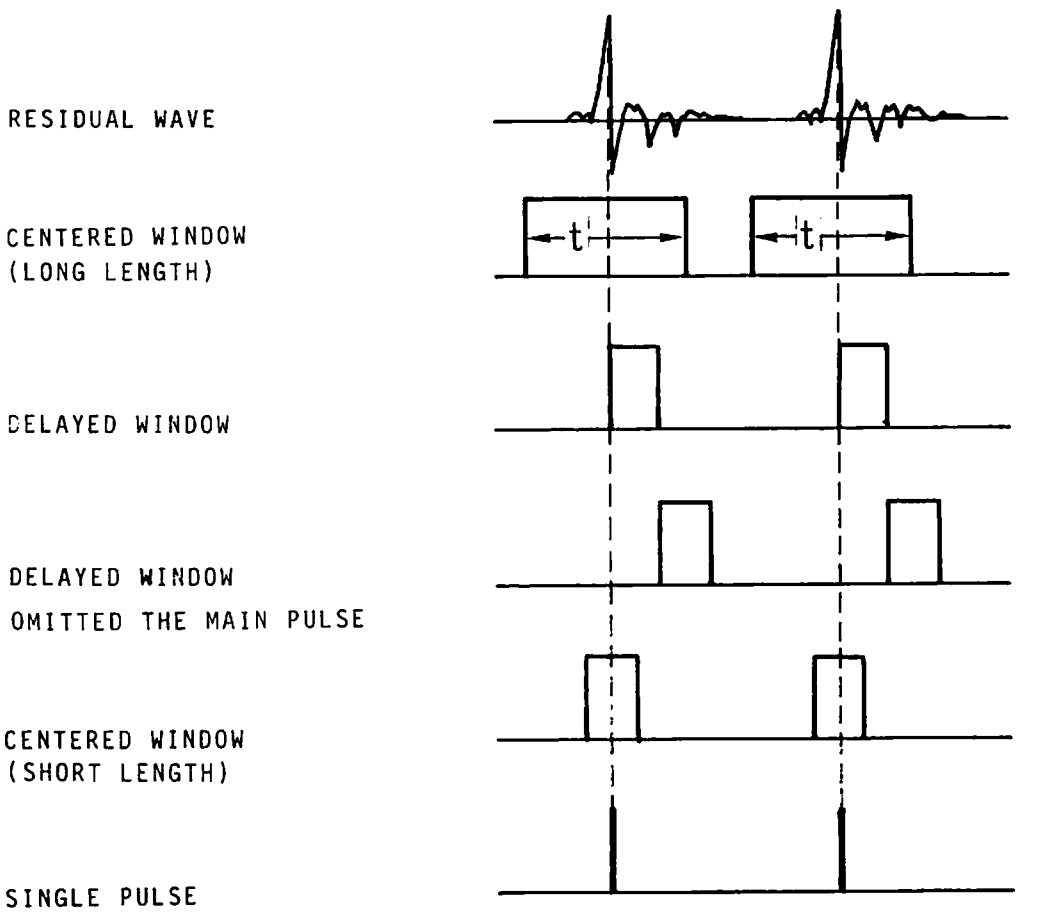

Fig. 2 Schematic illustration of the time-window applied to the residual wave signal.

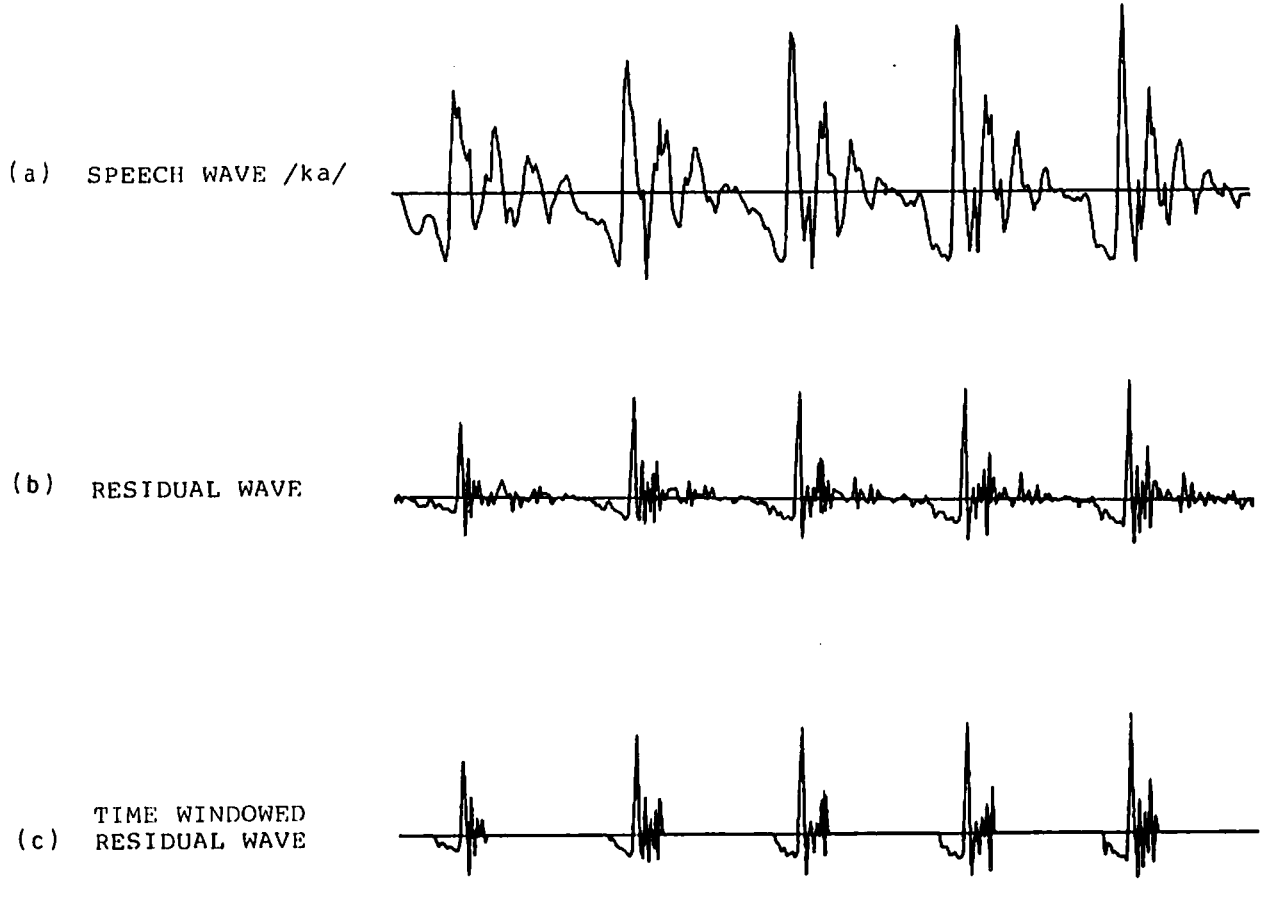

Fig. 3 Illustrations for the Japanese monosyllable /ka/ of (a) the original speech wave form; (b) its residual wave form after PARCOR analysis; and (c) its time-windowed residual wave form.

length of time-window. An example of this time-windowing is illustrated in Fig. 3. Figs. 3 (a) and (b) are partitions of the input speech wave of the Japanese monosyllable /ka/ and its residual wave, respectively. Fig. 3 (c) is a partition of a time-windowed residual wave of 2.5 milliseconds length.

In the auditory test of the synthesized speech quality, the length of the time-window was varied from about 5 milliseconds to about 0.5 milliseconds. The shorter the length of the time-window, the more the synthesized speech quality deteriorated. It was estimated that the synthesized speech quality was acceptable provided that the length of the time-window was greater than about one-third of the pitch period. When the length of the time-window was shortened to a single pulse--that is, a pulse set at the maximum level of a residual wave within one pitch period as shown in Fig. 4--the synthesized speech quality was worse than that produced by an ordinary PARCOR synthesizer using both a pulse and a white noise generator. It appears that the average time processing of the excitation source signal in an ordinary PARCOR synthesizer brings about such a result.

3.2 Effect of the position of the time-window

The effect of the time-window position within the residual wave on the synthesized speech quality was measured, keeping the length of the time window at 2.5 milliseconds.

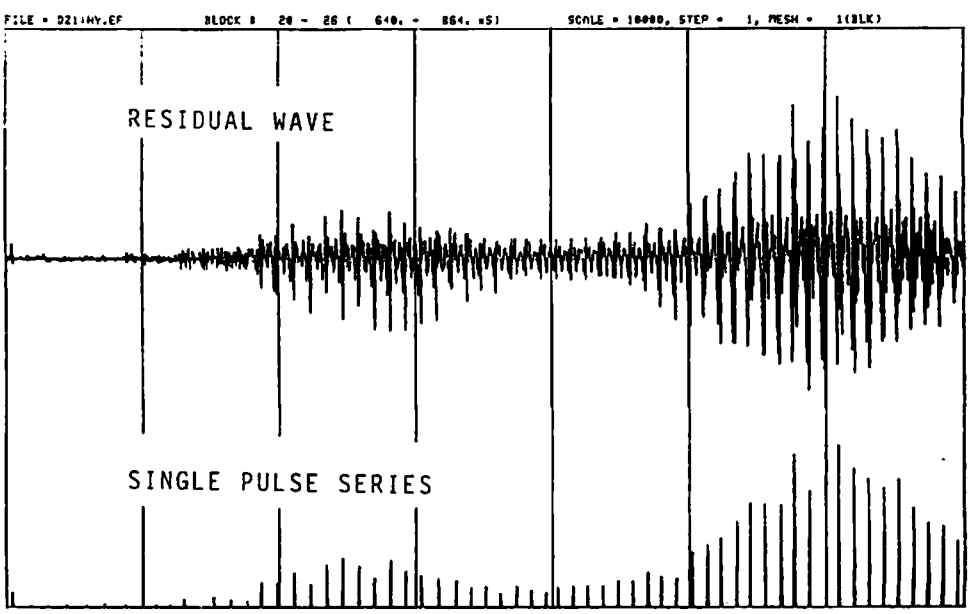

Fig. 4 An example of a time-windowed residual wave having a single pulse at its maximum level within a single pitch period.

It was observed that when the time-window was moved back and forth slightly, there was no influence on the synthesized speech quality, but a severe deterioration of the synthesized speech quality occurred if the time-window was collapsed with the main pulsive wave having the maximum level in the residual wave. It was concluded that the time-window should be centered on the main pulsive wave of the residual wave, denoted as A in Fig. 3 (c).

When the time-window was moved further and the main pulsive wave of the residual wave was omitted from the excitation source signal of the PARCOR synthesizer, it was observed that the synthesized speech quality became as good as the situation with the time-window centered on the main pulsive wave, although the level of the synthesized speech was lowered.

3.3 Effect of a delay in the time-windowed residual wave

The effect of a relative delay of the time-windowed residual wave on the PARCOR coefficients fed to the digital filters of the PARCOR synthesizer was also measured.

It was obsereved that there was no deterioration in the synthesized speech quality, provided that the delay did not exceed about 7.5 milliseconds, which was a value comparable (set at exactly 1.5 times) to the frame interval time in the PARCOR analysis of the present study.

4. Conclusions

The following may be concluded from the present study.
(1) About one-third of the residual wave is needed to synthesize speech of acceptable quality.
(2) Centering the time-window on the main pulsive wave is also needed.
(3) Omitting the main pulsive wave from the time-windowed residual wave does not cause a deterioration in the quality of the synthesized speech, but leads to a decrease of the synthesized speech level.

References

1. Atal, B.S. and V. Stover (1975); Voice excited predictive coding for low-bit-rate transmission of speech, J. Acous. Soc. Amer., 57, Suppl. 1, S35.
2. Itoh, K. and S. Saito (1975); Study on excitation source signal of a PARCOR speech synthesizer available for compound sound signal, Report of Annual Meeting of the Institute of Electronics and Communication Engineers of Japan, No. 1231, 1138. (March - 1975)
3. Atal, B.S. and J.R. Remde (1982); A new model of LPC excitation for producing natural-sounding speech at low bit rates, Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 614. (April - 1982)