Weighted Autocorrelation for Pitch Extraction of Noisy Speech

Tetsuya Shimamura, *Member, IEEE*, and Hajime Kobayashi

***Abstract*—In this paper, we propose a modified version of the autocorrelation pitch extraction method well known to be robust against noise. Utilizing that the average magnitude difference function (AMDF) has similar characteristics with the autocorrelation function, the autocorrelation function is weighted by the reciprocal of the AMDF. By simulation experiments, it is shown that the proposed pitch extraction method is useful in noisy environments.**

***Index Terms*—Autocorrelation, noisy speech, pitch extraction.**

I. Introduction

PITCH period (or fundamental frequency) extraction plays an important role on speech processing and has a wide spread of applications in speech related areas. For this reason, many methods to extract the pitch of speech signals have been proposed [1]. However, performance improvement in noisy environments is still desired. For example, this is particularly true in speech enhancement systems, because in such systems the accuracy of pitch extraction is directly related with the quality of speech after the operations of enhancement. Also, speech communication systems often transmit pitch information. To do this, we have to extract the pitch of speech signals in practical noisy environments. Unfortunately, we do not have a single method reliable and accurate for pitch extraction in noisy environments.

Correlation based processing is known to be comparatively robust against noise. The autocorrelation function method (AUTOC) [5] is classified into this category, and may be one which provides the best performance in noisy environments [2], [3]. In [7], an integrated method for the AUTOC has been proposed. Correlation based processing also includes the AMDF method [6].

In this paper, we propose a new pitch extraction method which uses an autocorrelation function weighted by the inverse of an AMDF. The characteristics of the AMDF are very similar with those of the autocorrelation function. The AMDF produces a notch, while the autocorrelation function produces a peak. However, both functions essentially have the same periodicity. The proposed method utilizes the feature that in a noisy environment, the noise components included in the autocorrelation function and AMDF behave independently (and are uncorrelated each other). This feature will be validated in this paper. By such uncorrelated properties, the peak of the autocorrelation function is emphasized in a noisy environment when the autocorrelation function is combined with the inversed AMDF. As a result, it is expected that the accuracy of pitch extraction for the AUTOC is improved.

The remainder of this paper is organized as follows. Section II describes the principle of the proposed method. In Section III, we first show the results of preliminary test for the proposed autocorrelation function. After that, we confirm the effectiveness of our method by comparing with some conventional methods based on several experimental results. Finally, we conclude this paper in Section IV.

II. Proposed Method

A. Principle

Autocorrelation function $\phi(\tau)$ is calculated by

$$\phi(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+\tau) \tag{1}$$

where

$x(n)$	speech signal;
τ	lag number;
n	time for a discrete signal.

The characteristic of $\phi(\tau)$ is that $\phi(\tau)$ has a large value when $x(n)$ is similar with $x(n+\tau)$. If $x(n)$ has a period of P, then $\phi(\tau)$ has peaks at $\tau = lP$ where l is an integer.

Essentially, $\phi(0)$ gives the largest value among $\phi(\tau), \tau = lp$, for $l = 0, 1, 2, \ldots$. The second value is given by $\phi(P)$. Other peaks of $\phi(\tau)$ usually decrease as τ increases. Therefore, we can estimate the pitch period P from the location of the peak at $\tau = P$.

Let us assume that $x(n)$ is a noisy speech signal given by

$$x(n) = s(n) + w(n) \tag{2}$$

where $s(n)$ is a clean speech signal and $w(n)$ is additive white Gaussian noise. In this case, we have an autocorrelation function given by

$$\begin{aligned}\phi(\tau) &= \frac{1}{N}\sum_{n=0}^{N-1}(s(n)+w(n))(s(n+\tau)+w(n+\tau)) \\ &= \frac{1}{N}\sum_{n=0}^{N-1}(s(n)s(n+\tau)+s(n)w(n+\tau) \\ &\quad + w(n)s(n+\tau)+w(n)w(n+\tau)) \\ &= \phi_{ss}(\tau)+2\phi_{sw}(\tau)+\phi_{ww}(\tau)\end{aligned} \tag{3}$$

Manuscript received September 23, 1999; revised June 4, 2001. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Peter Kabal.

T. Shimamura is with the Department of Information and Computer Sciences, Saitama University, Saitama 338-8570, Japan (e-mail: shima@sie.ics.saitama-u.ac.jp).

H. Kobayashi was with the Department of Information and Computer Sciences, Saitama University, Saitama 338-8570, Japan. He is now with Pioneer Corporation, Tsurugashima 350-2288, Japan.

Publisher Item Identifier S 1063-6676(01)08234-7.

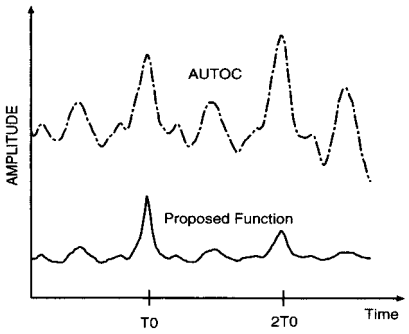

Fig. 1. Autocorrelation function and proposed function. T_0 corresponds to the true pitch period.

where

$\phi_{ss}(\tau)$ autocorrelation function of $s(n)$;
$\phi_{sw}(\tau)$ cross-correlation function of $s(n)$ and $w(n)$;
$\phi_{ww}(\tau)$ autocorrelation function of $w(n)$.

For large N, if $s(n)$ does not correlate with $w(n)$, then $\phi_{sw}(\tau) = 0$. Furthermore, if $w(n)$ is uncorrelated, then $\phi_{ww}(\tau) = 0$ except for $\tau = 0$. In such a case, the relations

$$\phi(\tau) = \phi_{ss}(\tau) + \phi_{ww}(\tau)(\tau = 0) \tag{4}$$

$$\phi(\tau) = \phi_{ss}(\tau)(\tau \neq 0) \tag{5}$$

are valid. Based on these properties, the AUTOC provides robust performance against noise.

The autocorrelation function with the period of P has some peaks at the locations of lP. Although the maximum peak is located at $\tau = P$ except for the case of $\tau = 0$, in some cases, the peak located at $\tau = 2P$ becomes larger than that located at $\tau = P$, as shown in Fig. 1. Then, a half pitch error occurs. On the other hand, as also shown in Fig. 1, a peak is often made at $\tau < P$. This situation, in some cases, leads to a double-pitch error. For such reasons, if unnecessary peaks of autocorrelation function as shown in Fig. 1 are suppressed somehow, then it is expected that the accuracy of pitch extraction becomes higher.

For the purpose of emphasizing the true peak the AUTOC makes, we propose an autocorrelation function weighted by an inversed AMDF. The AMDF is described by

$$\psi(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} |x(n) - x(n+\tau)|. \tag{6}$$

The AMDF has the characteristic that when $x(n)$ is similar with $x(n+\tau), \psi(\tau)$ becomes small. This means that if $x(n)$ has a period of $P, \psi(\tau)$ produces a deep notch at $\tau = P$. Therefore, $1/\psi(\tau)$ makes a peak at $\tau = P$. Furthermore, the additive noise $w(n)$ included in $\psi(\tau)$ behaves independently with that included in $\phi(\tau)$ (see Appendix). Hence, using the autocorrelation function weighted by $1/\psi(\tau)$, it is expected that the true peak is emphasized, and as a result the errors of pitch extraction are decreased.

The proposed function is given by

$$\eta(\tau) = \phi(\tau)/(\psi(\tau) + k) \tag{7}$$

where k is a fixed number $(k > 0)$. The AMDF in (6) provides at $\tau = 0$

$$\psi(0) = 0 \tag{8}$$

which invokes a divergence of the directly inversed AMDF. For this reason, the denominator in (7) is stabilized by adding the number k.

Fig. 1 shows the autocorrelation and proposed functions obtained for a speech signal corrupted by noise. In this case, by picking the maximum amplitude of each function, the proposed function leads to the true pitch, while the autocorrelation function does an erroneous one.

B. Implementation

Pitch of the segmented speech is estimated by searching the peak of the weighted function. However, if we use the weighted function directly, the accuracy of pitch extraction is not so accurate. Therefore, the proposed system uses interpolation based on 3 points around the detected peak. It is known that such interpolation on the autocorrelation function is useful for improving the accuracy of pitch extraction [8]. The interpolation operation used in this paper is based on Lagrange's method. The region for searching the pitch peak is set to be from 50 Hz to 400 Hz, which corresponds to the region of the fundamental frequencies of most men and women.

III. Experiments

A. Experimental Details

Speech data in the experiments was taken from "20 Countries Language Database (NTT Advanced Technology)." The speech signals used were uttered by four male and four female speakers. Each of the speech signals consisted of about 10 s Japanese sentences, which were sampled by a rate of 10 kHz with a band limitation of 3.4 kHz. The experiments were conducted by adding white Gaussian noise to the speech signals. The other details in the experiments are as follows:

- window size: 25.6 ms;
- frame shift: 10 ms.

Each SNR of the noisy speech signals used in the experiments was ∞ dB, 10 dB, 5 dB, 0 dB, -5 dB.

B. Evaluation Method

Based on Rabiner's method [4], the following was used for the evaluation of pitch extraction accuracy:

$$e(n) = F_t(n) - F_e(n) \tag{9}$$

where

$F_t(n)$ true fundamental frequency;
$F_e(n)$ extracted fundamental frequency;
$e(n)$ extraction error.

If $|e(n)| \geq 10$ Hz, we recognized the error as a gross pitch error (GPE). Otherwise, we recognized the error as a fine pitch error (FPE). For the GPE, proportion of the errors was calculated. For the FPE, standard deviation of the errors was calculated.

By inspection of clean speech waveforms in the time domain, the true pitch period was obtained and then $F_t(n)$ was calculated by inverse operation.

C. Preliminary Test

The proposed function $\eta(\tau)$ in (7) contains the stabilized factor k in the denominator. We need to know how the setting of

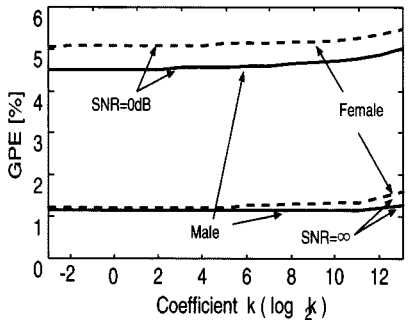

Fig. 2. Dependency on the constant parameter for the proposed function.

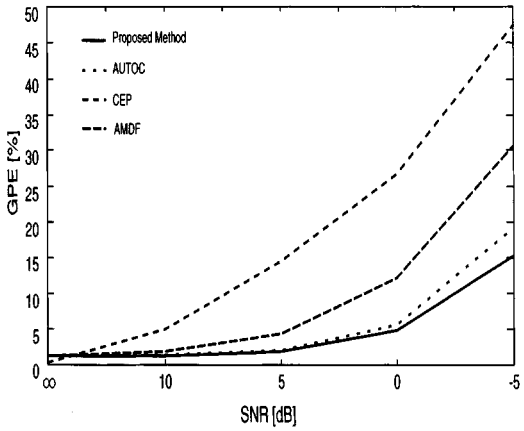

Fig. 3. Gross pitch error.

k influences the accuracy of pitch extraction. Therefore, we conducted the preliminary test to investigate the performance dependency on k. Fig. 2 shows the result. Fig. 2 shows how the value of k relates with the resulting GPE. From this figure, we notice that the accuracy of pitch extraction is almost constant if k is set to be less than $2^7(=128)$.

D. Performance Comparison

To investigate the accuracy of the proposed pitch extraction method, we conducted experiments which compare it with three conventional methods. The conventional methods are the autocorrelation function method(AUTOC), the cepstrum method (CEP) and the AMDF method. The procedure of pitch extraction for the conventional methods is commonly as follows:

1) speech signal is segmented by 256 point Hamming window;
2) each function for the pitch extraction methods is calculated from the segmented speech;
3) pitch period is estimated based on the location of the functional peak after interpolation.

The CEP uses the fast Fourier transform (FFT) to transform the speech waveform into the frequency domain. The FFT was calculated by 2048 points after 1792 zero padding to the segmented speech to increase frequency resolution.

The coefficient k of the proposed function was set to 1. This is because for this setting, the weighted function in (7) is normalized as

$$0 < \frac{1}{\psi(\tau)+k} \leq 1. \tag{10}$$

In general, normalization is desired for implementation.

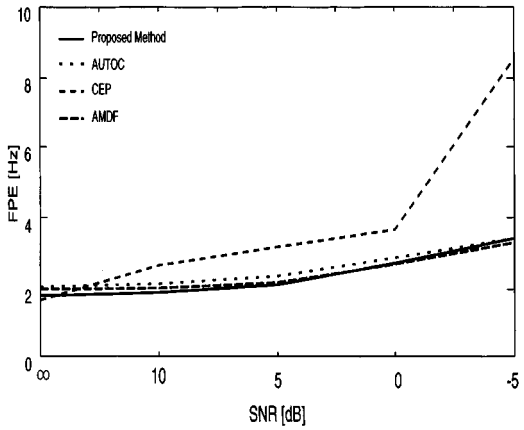

Fig. 4. Fine pitch error.

Figs. 3 and 4 show the results of the compared performance evaluations. From these figures, it is observable that the proposed method provides better performance than the conventional methods in noisy environments. Although the FPE of the proposed method is competitive with that of the AUTOC as well as that of the AMDF method, the proposed method, at very low SNR, provides a moderate reduction in the GPE relative to the AUTOC as well as a significant reduction in the GPE relative to the AMDF method. On the other hand, for clean speech, the proposed method is inferior to the CEP known as the best pitch extraction method in noiseless environments. However, in practical situations we face noisy speech in most of cases. As a result, the proposed method's robustness against noise may be useful.

IV. Conclusion

We have proposed a modified version of the autocorrelation method for pitch extraction. Based on the experimental results it has been shown that the proposed method is useful in noisy environments. Especially, it has been asserted that the proposed method provides a moderate improvement relative to the conventional autocorrelation method at very low SNR.

Appendix

Noise independency between autocorrelation function and AMDF: Inserting (2) into (6), we have

$$\begin{aligned}\psi(\tau) &= \frac{1}{N}\sum_{n=0}^{N-1}|s(n)+w(n)-s(n+\tau)-w(n+\tau)| \\ &= \frac{1}{N}\sum_{n=0}^{N-1}|s(n)-s(n+\tau)+w(n)-w(n+\tau)| \\ &\leq \frac{1}{N}\sum_{n=0}^{N-1}|s(n)-s(n+\tau)| \\ &\quad + \frac{1}{N}\sum_{n=0}^{N-1}|w(n)-w(n+\tau)| \\ &= \psi_{ss}(\tau)+\psi_{ww}(\tau)\end{aligned} \tag{11}$$

where $\psi_{ss}(\tau)$ is an AMDF of $s(n)$ and $\psi_{ww}(\tau)$ is an AMDF of $w(n)$. Considering a scaling parameter $\alpha(\tau)$ determined by the value of τ in a stationary case, we can derive the following:

$$\psi(\tau) = \alpha(\tau)(\psi_{ss}(\tau) + \psi_{ww}(\tau)). \tag{12}$$

The noise component in (12), $\psi_{ww}(\tau)$, which is described by

$$\psi_{ww}(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} |w(n) - w(n+\tau)| \tag{13}$$

is obviously independent on that on the autocorrelation, $\phi_{ww}(\tau)$, which is described by

$$\phi_{ww}(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} w(n)w(n+\tau). \tag{14}$$

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their helpful comments and suggestions.

REFERENCES

[1] W. J. Hess, *Pitch Determination of Speech Signals*. Berlin, Germany: Springer-Verlag, 1983.

[2] K. A. Oh and C. K. Un, "A performance comparison of pitch extraction algorithms for noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1984, pp. 18B4.1–18B4.4.

[3] W. J. Hess, *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel-Dekker, 1992.

[4] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 5, pp. 399–417, Oct. 1976.

[5] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust,, Speech, Signal Processing*, vol. ASSP-25, no. 1, pp. 24–33, Feb. 1977.

[6] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust,, Speech, Signal Processing*, vol. ASSP-22, pp. 353–362, Oct. 1974.

[7] D. A. Krubsack and R. J. Niederjohn, "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-39, pp. 319–329, Feb. 1991.

[8] J. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367–377, Dec. 1972.

Tetsuya Shimamura (M'91) received the B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1986, 1988, and 1991, respectively.

In 1991, he joined Saitama University, Saitama, Japan, where he is currently an Associate Professor. His interests are in digital signal processing and its applications to speech processing and communication systems.

He is a member of EURASIP and IEICE.

Hajime Kobayashi was born in Tokyo, Japan, in 1973. He received the B.E. and M.E. degrees from Saitama University, Saitama, Japan, in 1997 and 1999, respectively. While at the University, he belonged to the Shimamura Laboratory and researched pitch extraction in noisy environments.

He is now with Pioneer Corporation, Japan, where he is engaged in development of speech recognition algorithms.