Pitch Determination Using Windowless Autocorrelation Based Cepstrum Method

M. A. F. M. Rashidul Hasan, M. Shahidur Rahman and Tetsuya Shimamura
Graduate School of Science and Engineering, Saitama University, Japan
E-mail: {hasan, rahmanms, shima}@sie.ics.saitama-u.ac.jp Tel/Fax: +81-48-858-3776

***Abstract*—In this article a new method for pitch estimation of speech signals in noisy environments is proposed. Cepstrum is a popular method for pitch estimation. The performance of the method is, however, degraded in noisy environments. The proposed method introduces a greatly enhanced cepstrum based pitch estimator which employs the windowless autocorrelation function for producing a noise-reduced version of voiced speech frame without changing its periodic characteristics. Experimental results show that the gross pitch error becomes lower using the proposed method when compared with conventional methods.**

Index Terms: Pitch extraction, cepstrum, windowless autocorrelation function, white and colored noise.

I. Introduction

Pitch (or fundamental frequency, F_0) extraction plays an important role on speech processing and has a wide spread of applications in speech related areas such as speech analysis-synthesis, speech coding, speech and speaker identification systems. For this reason, numerous methods to extract the pitch of speech signals have been proposed. However, in noisy environments, there are very few methods to be able to relatively accurately extract the pitch so far. The properties of speech signals in either time-domain or frequency-domain, or in both, have been utilized for proposing algorithms for pitch determination [1,2,3]. The frequency-domain (FFT based) cepstrum method is one of the traditional methods and has long been considered an accurate and reliable method for determining fundamental frequency, provided that the signal is recorded in a clean environment (typically referred to as studio quality data). The method is able to accurately extract the pitch with little affections of vocal tract [4]. It can extract an accurate pitch of clean speech signals, but is not effective in noisy environments.

The presence of noise complicates the problem of pitch determination and deteriorates the performance of the pitch determination algorithms. Among the reported methods, the autocorrelation based approaches are very popular for their simplicity, low computational complexcity and better performance in noise [5,6,7]. The autocorrelation function (ACF) is, however, the inverse Fourier transform of the power spectrum of the signal. Thus if there is a distinct formant structure in the signal, it is maintained in the ACF. Spurious peaks are also sometimes introduced in the spectrum in noisy or even in noiseless conditions. This sometimes makes true peak selection a difficult task. Additionally, the ACF method is not robust in colored noise. It is therefore expected that if the cepstrum can be made robust against noise, it can be very useful in pitch determination. Towards this end, Andrews et al. proposed a subspace based method to reduce noise effects [8], and Kobayashi et al. discussed a modified cepstrum method for pitch extraction [9]. Ahmadi et al. derived a statistical approach to improve the performance of the cepstrum method [10].

In this paper we propose another modification using the windowless ACF of the signal instead of the signal itself. The windowless ACF of the signal is a noise compensated equivalent of the signal (in terms of periodicity) which leads to accurate estimation even in high noisy conditions. In our proposed cepstrum method, pitch detection is robust in colored noise and it is competitive with the ACF in white noise case.

This article is organized as follows. Section II describes the problem of the conventional cepstrum method. Section III explains the proposed method. In Section IV, we verify the effectiveness of our method by comparing with some conventional methods based on experimental results. Finally in Section V, we conclude the paper.

II. Problem Description

The cepstrum of the speech signal $x(n)$ is usually defined as the inverse Fourier transform of the logarithm of the Fourier transform of $x(n)$ as

$$C(n) = F^{-1}\{log|F\{x(n)\}|\} \tag{1}$$

where $C(n)$ denotes the cepstrum at integer index n, and F and F^{-1} denote forward and inverse Fourier transform, respectively. The independent variable n in the cepstrum domain is in time units and has been termed as quefrency. The voiced speech is modeled by convolving the vocal tract response with the periodic excitations. The cepstrum technique operates by separating the periodic components from the vocal tract contribution. The formant effect is thus ignored in the cepstrum method. The cepstrum method is a well-known pitch extraction approach but it yields unsatisfactory results for noisy speech. The accuracy of the cepstrum method depends on the clearness of periodicity on the log spectrum. In case of clean speech, the periodicity of waveform on the log spectrum is clear. For example, the pitch peak is accurately estimated by the cepstrum method in clean speech as shown in Fig. 1. However, when the speech is affected by additive noise as shown in Fig. 2. it fails to detect the true peak.

Proceedings of the Second APSIPA Annual Summit and Conference, pages 514–517,
Biopolis, Singapore, 14-17 December 2010.

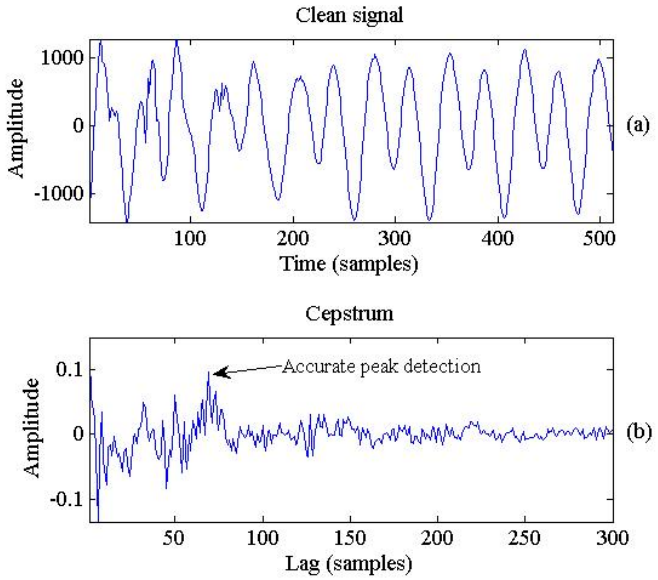

Fig. 1. (a) Clean speech frame, (b) Cepstrum in clean speech (True F_0=134 Hz and estimated F_0=134 Hz).

The error in Fig. 2 comes from the log operation which is used to deconvolve the multiplicative processes of the vocal tract and excitation. The nonlinear log operation introduces speech correlated noise products which change the algebraic structure assumed in the cepstrum processing.

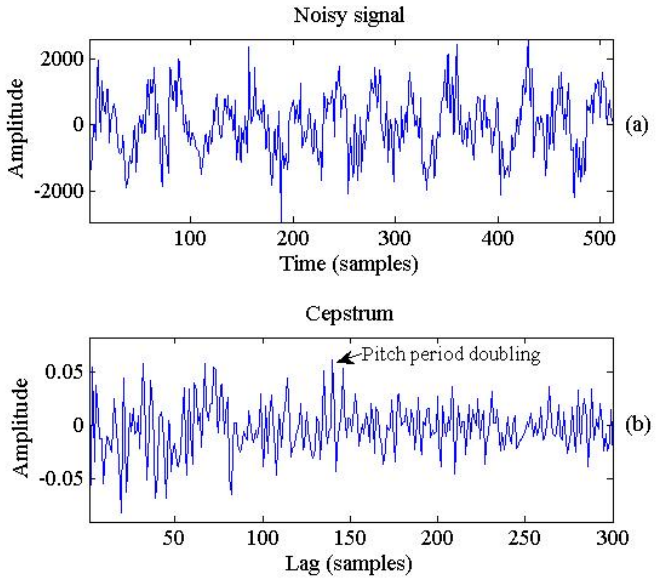

Fig. 2. (a) Noisy speech frame, (b) Cepstrum in noisy speech (True F_0=134 Hz and estimated F_0=69 Hz)

III. Proposed Method

The autocorrelation function $R(k)$ of a speech signal $x(n)$ is defined by

$$R(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+k) \tag{2}$$

for $x(m)$, m = 0, 1, 2, . . . , N-1, where N is the length of the underlying speech frame and k is the lag number. It is known that the periodicity of $x(n)$ and that of $R(k)$ become equivalent. Moreover, since the autocorrelation of a signal is obtained by an averaging process, it can be treated as a noise-compensated version of the speech segment in terms of periodicity. However, in case of the conventional autocorrelation, a finite length of windowed speech segment is involved in the computation. As the lag number increases, there is less and less data involved in the computation leading to reduction in amplitude of the correlation peaks. In the windowless condition, the signal outside the window is not considered as zero. Thus the number of additions in the averaging process is always common. This results in almost similar amplitude correlation peaks even as the lag number increases. Thus the windowless autocorrelation sequence is a more appropriate noise-reduced equivalent of the speech segment in terms of periodicity. We propose to use a windowless version of the ACF of the speech signal instead of speech signal in the DFT based cepstrum method, which can be given by

$$C(k) = F^{-1}\{log|F\{R_{wl}(k)\}|\} \tag{3}$$

where $R_{wl}(k)$ is the windowless autocorrelation function of the speech signal $x(n)$, which is defined as

$$R_{wl}(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+k) \tag{4}$$

for $x(m)$, m = 0, 1, 2, . . . , $2N$-1. An N length sequence of $R_{wl}(k)$, k = 0, 1, 2, . . . , N-1 is obtained in (4) and all of them are utilized in (3) (for the cepstrum method, an N length signal of $x(n)$, N = 0, 1, 2, . . . , N-1 is used in (1)). For the ACF in (2), when $(n+k) > N$, $x(n+k)$ becomes zero. However, in (4), $x(n+k)$ is not zero outside N. This modification makes $R_{wl}(k)$ more stronger in periodicity. The pitch peak is emphasized as seen in Fig. 3.

Application of the obtained sequence with cepstral deconvolution is found to result in an significant improvement in pitch estimation for both clean and noisy environments. The cepstrum derived using the windowless ACF of the noisy speech in Fig. 2(a) is shown in Fig. 4. This time the true peak is accurately determined. Fig. 5 represents a block diagram of the proposed method.

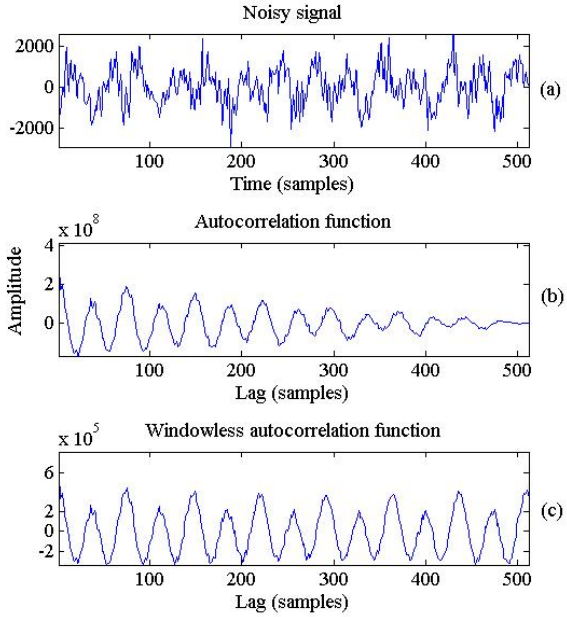

Fig. 3. Speech segments of (a) Noisy speech frame, (b) Conventional ACF, (c) Windowless ACF.

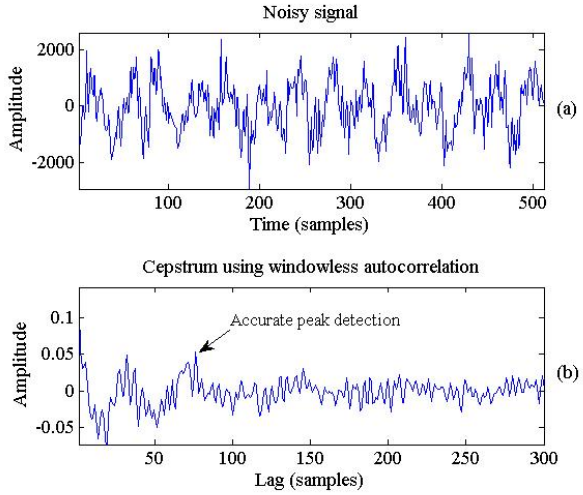

Fig. 4. (a) Noisy speech frame, (b) Proposed cepstrum in noisy speech (True F_0=134 Hz and estimated F_0=134 Hz).

IV. Experiments And Results

To evaluate the proposed method, natural speeches spoken by three Japanese female and three male speakers are examined. Speech materials are 11 sec-long sentences spoken by every speaker sampled at 10 kHz rate taken from NTT database [11]. The reference file is constructed by computing the fundamental frequency every 10 ms using a semi-automatic technique based on visual inspection. The simulations were

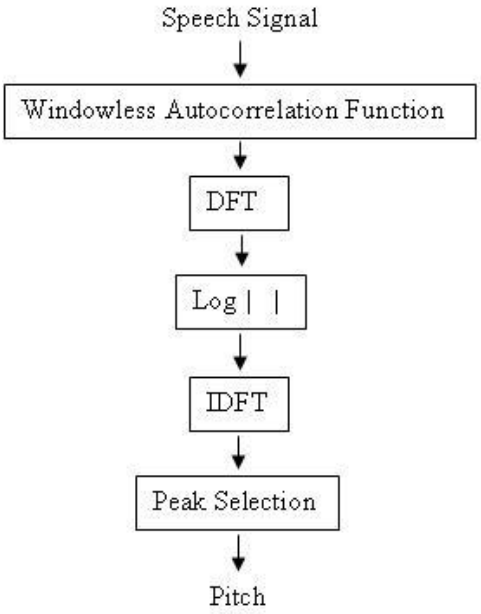

Fig. 5. Block diagram of the proposed method.

performed after adding different types of noise to these speech signals. Pitch estimation error is calculated as the difference between the reference and estimated fundamental frequency. If the estimated pitch for a frame deviates from the reference by $> 20\%$, we recognize the error as a gross pitch error (GPE) [5,6]. Otherwise, we recognize the error as a fine pitch error (FPE). The index GPE is often expressed in percentage denoted as %GPE. The true pitch values are obtained from the original database. The GPE and FPE are commonly used as a measure of errors in estimating the pitch frequency. The possible sources of GPE is pitch doubling, having, inadequate suppression of formants to affect the estimation. The experiment conditions are tabulated in Table I.

TABLE I
Conditions of Experiments

Sampling frequency	10 kHz
Window size	51.2 ms
Frame shift	10 ms
Number of FFT points	1024
SNRs(dB)	$\infty, 20, 15, 10, 5, 0$

We used speech signals corrupted by babble, exhibition, train and white noises. The number of GPEs found in determining the pitch using autocorrelation method (ACF), conventional cepstrum method (CEP) and proposed method (WACEP) for female and male speakers are shown in Figs. 6 and 7, respectively. From these figures, in most of the cases the percentage GPE is reduced in our proposed method compared to the other conventional methods. In all cases the proposed method gives far better results than the cepstrum method in different types of noise condition. Especially at SNR= 5dB and SNR= 0dB, the proposed method gives better

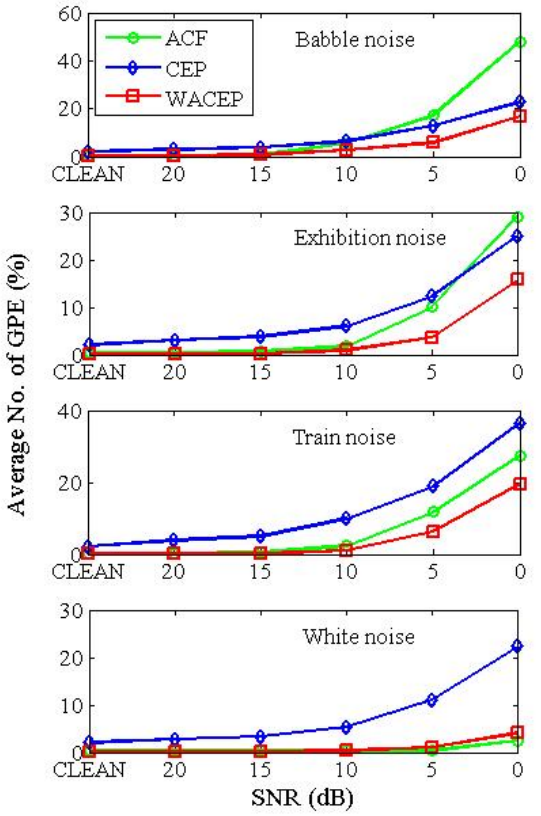

Fig. 6. Average gross pitch error (GPE) for three female speakers (Total frame: 1735) at various SNR conditions.

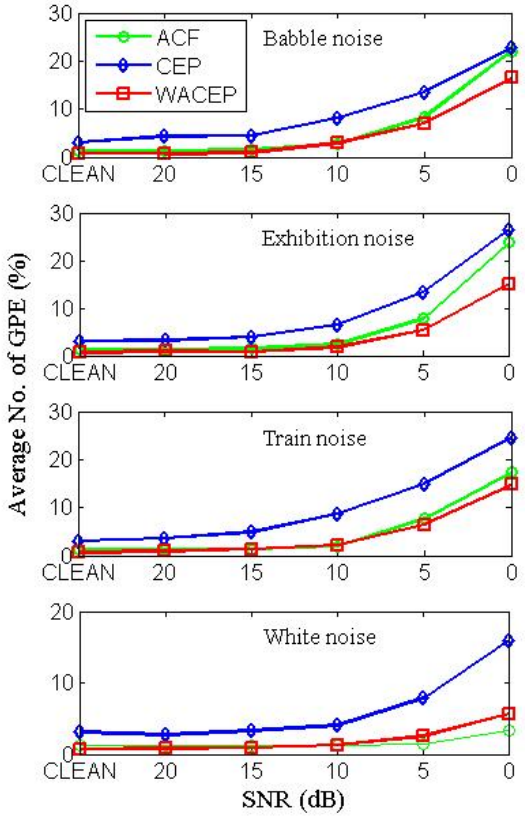

Fig. 7. Average gross pitch error (GPE) for three male speakers (Total frame: 1658) at various SNR conditions.

results than the autocorrelation method except for white noise case. The proposed method performs much better for female voice than for male voice because of wider harmonics in the frequency domain. As a whole, we see that the proposed method is an efficient method to extract the pitch in noisy environments.

V. Conclusions

Perfect pitch estimation is a tricky problem in speech analysis particularly in noisy environments. In this article, we proposed a modified cepstrum method by utilizing windowless autocorrelation function. Simulation results verified that our method is competent to extract the pitch of speech signals accurately in noisy environments.

References

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

[2] L. R. Rabiner, R. W. Schafer, *Theory and Applications of Digital Speech Processing,* 1st ed., Prentice Hall, 2010.

[3] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," IEEE Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-24, no. 5, pp. 399-418, October 1976.

[4] A. M. Noll, "Cepstrum pitch determination," J. Acoust. Soc. Am., vol. 41, no. 2, pp. 293-309, February, 1967.

[5] A. Cheveigne, and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., vol. 111, no. 4, pp.1917-1930, 2002.

[6] M. K. Hasan, S. Hussain, M. T. Hossain, and M. N. Nazrul, "Signal reshaping using dominant harmonic for pitch estimation of noisy speech," Signal Processing, vol. 86, pp. 1010-1018, 2006.

[7] T. Shimamura, and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," IEEE Trans. on Speech and Audio Processing, vol. 9, no. 7, pp. 727-730, October 2001.

[8] M. S. Andrews, J. Picone, and R. D. Degroat, "Robust pitch determination via SVD based cepstral methods," Acoustics, Speech and Signal Processing, ICASSP-90, vol. 1, pp. 253-256, 1990.

[9] H. Kobayashi, and T. Shimamura, "A modified cepstrum method for pitch extraction," Proceedings of IEEE Asia-Pacific Conference on Circuits and Systems, pp. 299-302, November 1998.

[10] S. Ahmadi, and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," IEEE Trans. on Speech and Audio Processing, vol. 7, no. 3, pp. 333-338, 1999.

[11] Multilingual Speech Database for Telephometry, NTT Advance Technology Crop., Japan, 1994.