# Refined Autocorrelation Function for Pitch Detection of Speech

Md. Saifur Rahman, Yosuke Sugiura, and Tetsuya Shimamura
Graduate School of Science and Engineering, Saitama University
Shimo-Okubo 255, Sakura-ku, Saitama 338-8570, Japan
Email: {saifur, sugiura, shima}@sie.ics.saitama-u.ac.jp

***Abstract*—In this paper, a noise robust method for pitch extraction using the autocorrelation function (ACF) and homomorphic deconvolution is proposed. In the proposed method, we extract the vocal source information eliminating the vocal tract information in the autocorrelation domain. The accuracy of pitch extraction is significantly enhanced in an efficient way based on the ACF calculation. Through experiments superior performance of the proposed method relative to some conventional methods is shown.**

***Index Terms*—Pitch, autocorrelation function, homomorphic deconvolution, vocal source, vocal tract.**

## I. Introduction

The pitch is a prominent parameter of speech, and highly applicable for speech-related systems such as speech analysis-synthesis, speech coding, speech enhancement and speaker identification. In the above systems, the system performance is significantly influenced by the accuracy of pitch extraction. Most of the pitch extraction methods can be classified into the time domain approach, frequency domain approach, and both domains approach. They are performed effectively with clean speech [1][2]. In the presence of noise, however, the periodic structure of clean speech is destroyed. Thus, it is more difficult to show the excellent performance of pitch extraction. Among the conventional methods, the autocorrelation function (ACF) [3] and average magnitude difference function (AMDF) [4] show robustness against noise. The ACF uses the same set of input samples of a signal, which correlates with itself by its shifted delay. On the other hand, the AMDF treats the difference between the original speech and its delayed version, which shows similar properties with the ACF. The ACF is, however, affected by the characteristics of vocal tract. For reducing the vocal tract effect, many algorithms have been developed based on a different form of autocorrelation calculation [5]-[7]. For example, YIN [5] focused on the conventional ACF, normalization, and interpolation to reduce the error rates in pitch estimation. Correntropy [6] also provides the similar properties to the ACF. In [6], the authors use the reproducing kernel Hilbert space (RKHS) and the higher order statistics, which preserve the properties of periodic signal to enhance the extraction accuracy of pitch.

In highly noisy environments, the two correlation-based methods; ACF and AMDF, are not good enough compared with the weighted autocorrelation function (WAF) [7]. The WAF also focuses on the ACF, but it is weighted by the inverse of the AMDF.

In the frequency domain, the cepstrum (CEP) method [8] operates the logarithmic arithmetic for separating the periodic components from the vocal tract contribution in speech. Modified CEP (MCEP) [9] uses the liftering and clipping operations on the log spectrum to overcome the defect of the CEP method. In windowless ACF (WLACF) based cepstrum method (WLACF-CEP) [10], firstly the periodicity of the speech signal is emphasized by reducing the noise from the noisy speech signal. Then, the CEP method is applied for reducing the effect of the vocal tract. In low signal-to-noise ratio (SNR) cases, unfortunately, the performances of the above three CEP based methods are still affected by the noise components remained in the quefrency domain.

Recently, BaNa [11] is addressed, which is a hybrid pitch extraction algorithm that selects the first five spectral peaks in the amplitude spectrum of the speech signal. From these spectral peaks, BaNa calculates the ratios of the frequencies with tolerance ranges and extracts the accurate pitch of the speech signal.

In this paper, we propose to use a refined autocorrelation function (RACF) for pitch extraction. Originally, the idea of RACF was found in [12] where the vocal tract information was extracted by using a low-pass lifter (LPL) for the linear predictive analysis [12] of speech. The RACF was generated from the homomorphic deconvolution [13] in the autocorrelation domain. In this paper, we use the RACF to eliminate the effect of vocal tract information utilizing a high-pass lifter (HPL) instead of an LPL in [12]. The ACF provides the convolutive properties of the vocal source and vocal tract information. Therefore, in this paper, we propose the RACF approach to extract more accurate true pitches from the vocal source information and overcome the limitations of the ACF.

The remainder of this paper is organized as follows. Section II describes the principle of the proposed method. In Section III, we verify the effectiveness of our method by comparing with some existing methods through experiments. Finally, we conclude this paper in Section IV.

## II. Proposed Method

Let us assume that the clean speech signal, $s(n)$, is corrupted by noise, $v(n)$. Thus, the noisy speech, $x(n)$, is

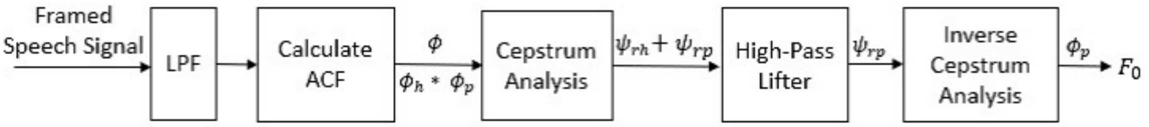

Fig. 1. Block diagram of the proposed method

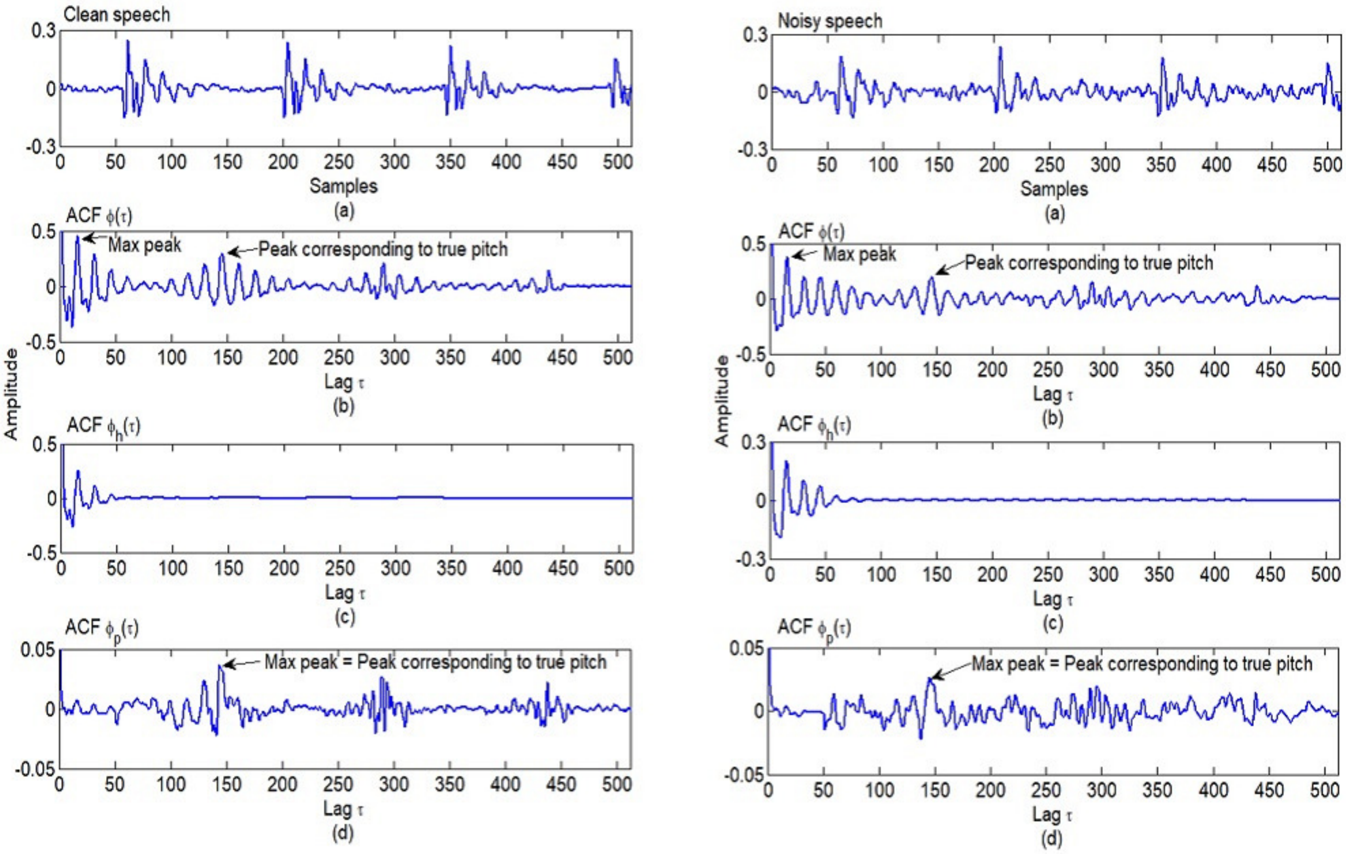

Fig. 2. ACF of male clean speech and noisy speech by using 5ms cut-off quefrency value at SNR=0 dB (white noise)

expressed as

$$x(n) = s(n) + v(n). \tag{1}$$

Figure 1 shows the block diagram of the proposed method. Firstly we apply a low pass filter (LPF) to reduce the effect of noise from the noisy speech signal for increasing the accuracy of pitch extraction. Then, the autocorrelation of the noisy speech signal is computed as

$$\phi(\tau) = R_{ss}(\tau) + R_{vv}(\tau) \tag{2}$$

where $R_{ss}(\tau)$ and $R_{vv}(\tau)$ correspond to the autocorrelation functions of the clean speech signal and noise, respectively, and $\tau$ is the lag number.

In (2), $\phi(\tau)$ can also be written as

$$\phi(\tau) = \phi_h(\tau) * \phi_p(\tau) \tag{3}$$

where $\phi_h(\tau)$ and $\phi_p(\tau)$ are the autocorrelation functions of the vocal tract and the vocal source, respectively, and $*$ denotes the convolution. Eq. (3) indicates that the $\phi(\tau)$ is highly influenced by the vocal tract information $\phi_h(\tau)$, which makes difficult to detect more appropriate pitches.

Next, we employ the homomorphic deconvolution technique which consists of CEP analysis, high-pass lifter and inverse CEP analysis in the autocorrealtion domain. The CEP of the ACF can be defined as

$$\psi_{rn}(\tau) = \mathrm{IDFT}[\sigma(f)] \tag{4}$$

where

$$\sigma(f) = log(|\mathrm{DFT}[\phi(\tau)]|). \tag{5}$$

DFT[$\cdot$] and IDFT[$\cdot$] mean discrete Fourier transform and inverse discrete Fourier transform, respectively. In (5), $\sigma(f)$ is the logarithm spectrum, which preserves the additive property

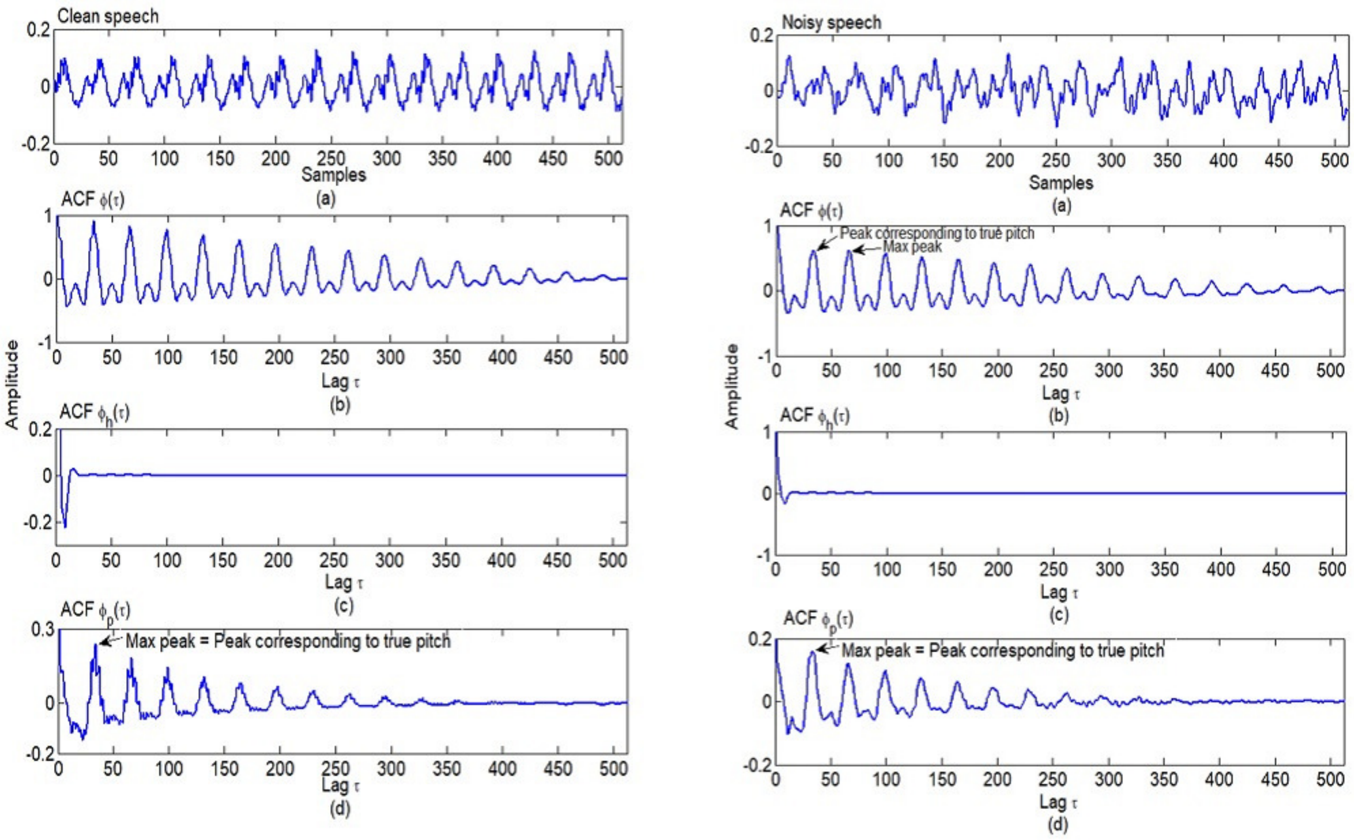

Fig. 3. ACF of female clean speech and noisy speech by using 1ms cut-off quefrency value at SNR=0 dB (white noise)

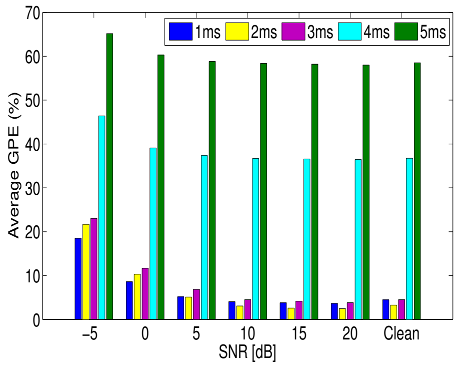

Fig. 4. Relation between cut-off quefrency level of HPL and GPE at different SNRs (female speakers).

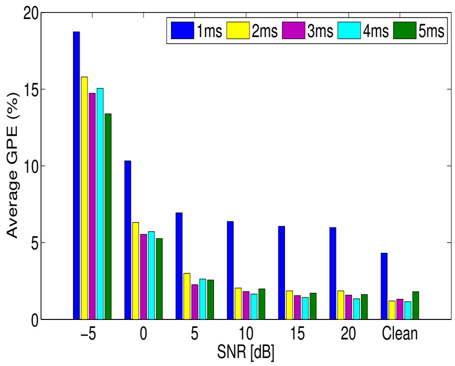

Fig. 5. Relation between cut-off quefrency level of HPL and GPE at different SNRs (male speakers).

of the logarithm spectrum of the vocal source and that of the vocal tract. Thus, its IDFT operation separates the contribution of the ACF of the vocal tract from that of the source in the CEP domain. We apply a high-pass lifter (HPL) on the CEP of the ACF. The HPL is used to eliminate the effect of the vocal tract. The vocal source information $\psi_{rp}(\tau)$ on the CEP can be obtained by multiplying a HPL, $L(\tau)$, as

$$\psi_{rp}(\tau)=\psi_{rn}(\tau)L(\tau) \tag{6}$$

$$L(\tau)=\begin{cases}0, & for \quad 0\le\tau\le L^{'}, M-L^{'}\le\tau\le M\\ 1, & otherwise\end{cases} \tag{7}$$

where $L^{'}$ represents the cut-off quefrency level. It should be noted here that the symmetric property of real CEP is considered.

For generating the RACF, the inverse CEP operation is applied to the $\psi_{rp}(\tau)$ as

$$\phi_p(\tau)=\mathrm{IDFT}(exp(X_p(f))) \tag{8}$$

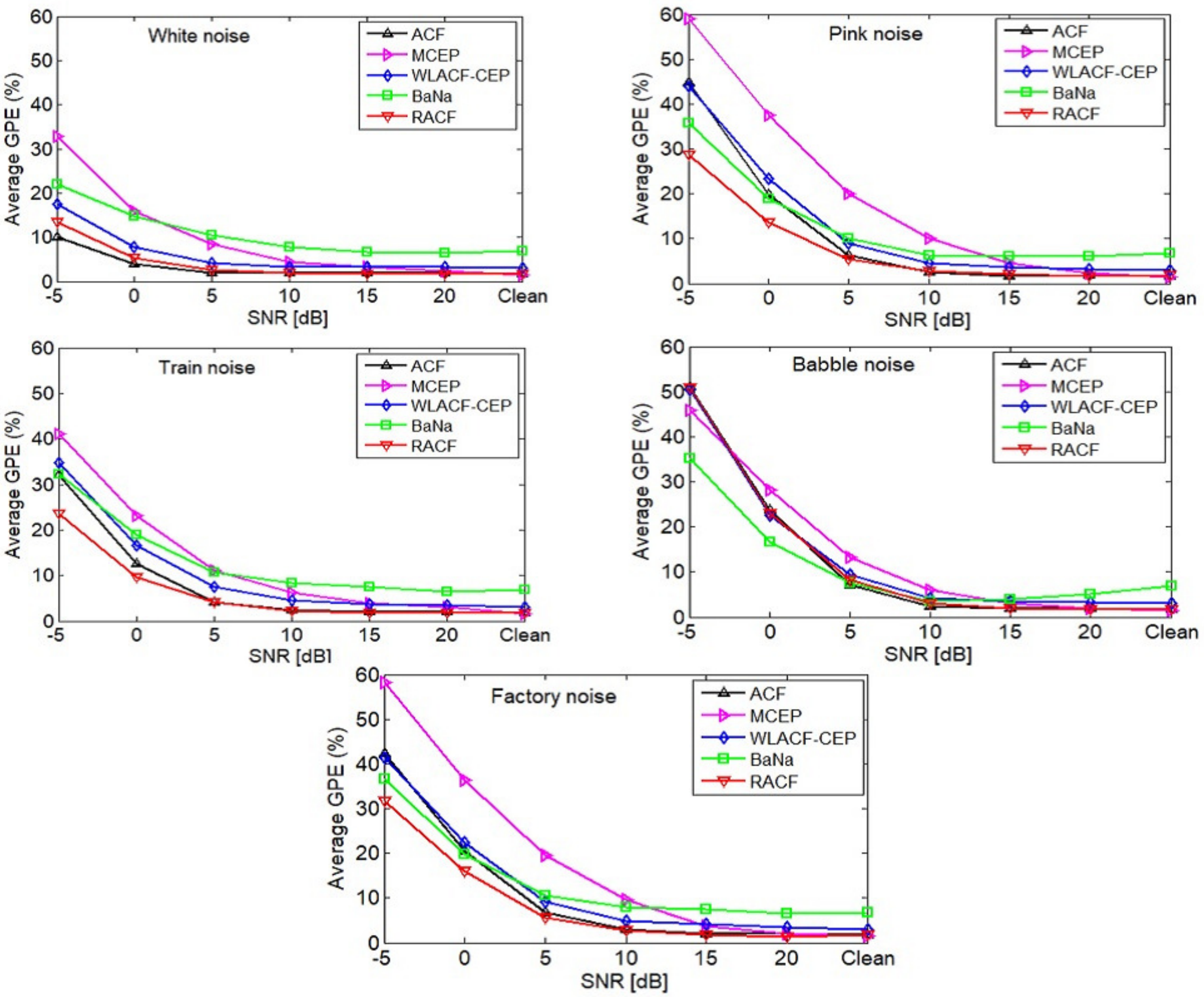

Fig. 6. GPE for four male speakers with different types of noise under different SNR levels.

where

$$X_p(f) = \mathrm{DFT}[\psi_{rp}(\tau)]. \tag{9}$$

Finally, from the resulting ACF $\phi_p(\tau)$, the pitch period is determined by finding the maximum peak in a range of the possible pitch period as done in the ACF method.

Figures 2 and 3 show examples of how to determine the pitch period by using the proposed method. In Figs. 2 and 3, we consider the aliasing in the ACF of male and female speakers for clean speech and noisy speech, respectively. In the ACF, the effects of the vocal source and vocal tract are convolved with each other. Therefore, sometimes false peaks arise in the ACF. In Fig. 2, the low pitched male speaker provides a long pitch period. For this reason, the pitch extraction is less affected by the convolutive behavior of vocal source and vocal tract. We used an LPL and an HPL with the cut-off quefrency value 5ms, to generate the ACF of the vocal tract information, $\phi_h(\tau)$, as well as the ACF of the vocal source information, $\phi_p(\tau)$. In Fig. 3, the high pitched female speaker has a short pitch period. This results in that the vocal source information is highly overlapped to the vocal tract information. We used an LPL and an HPL with the cut-off quefrency value to generate the ACFs of vocal source and vocal tract information. In Fig. 3, it is shown that the ACF of the vocal source, $\phi_p(\tau)$ provides higher accuracy to identify the true peak.

## III. Experiments

### A. Experimental Condition

Experiments were conducted on speech signals, spoken by four Japanese male and four female speakers, which were sampled at a rate of 10 kHz. The speech materials are 11 sec long sentences taken from a database developed by NTT Advanced Technology Corporation [14]. To generate the noisy speech, we added different types of noise to the speech signals. White noise, pink noise, babble noise, and factory noise were taken from the NOISEX92 database [15], while train noise

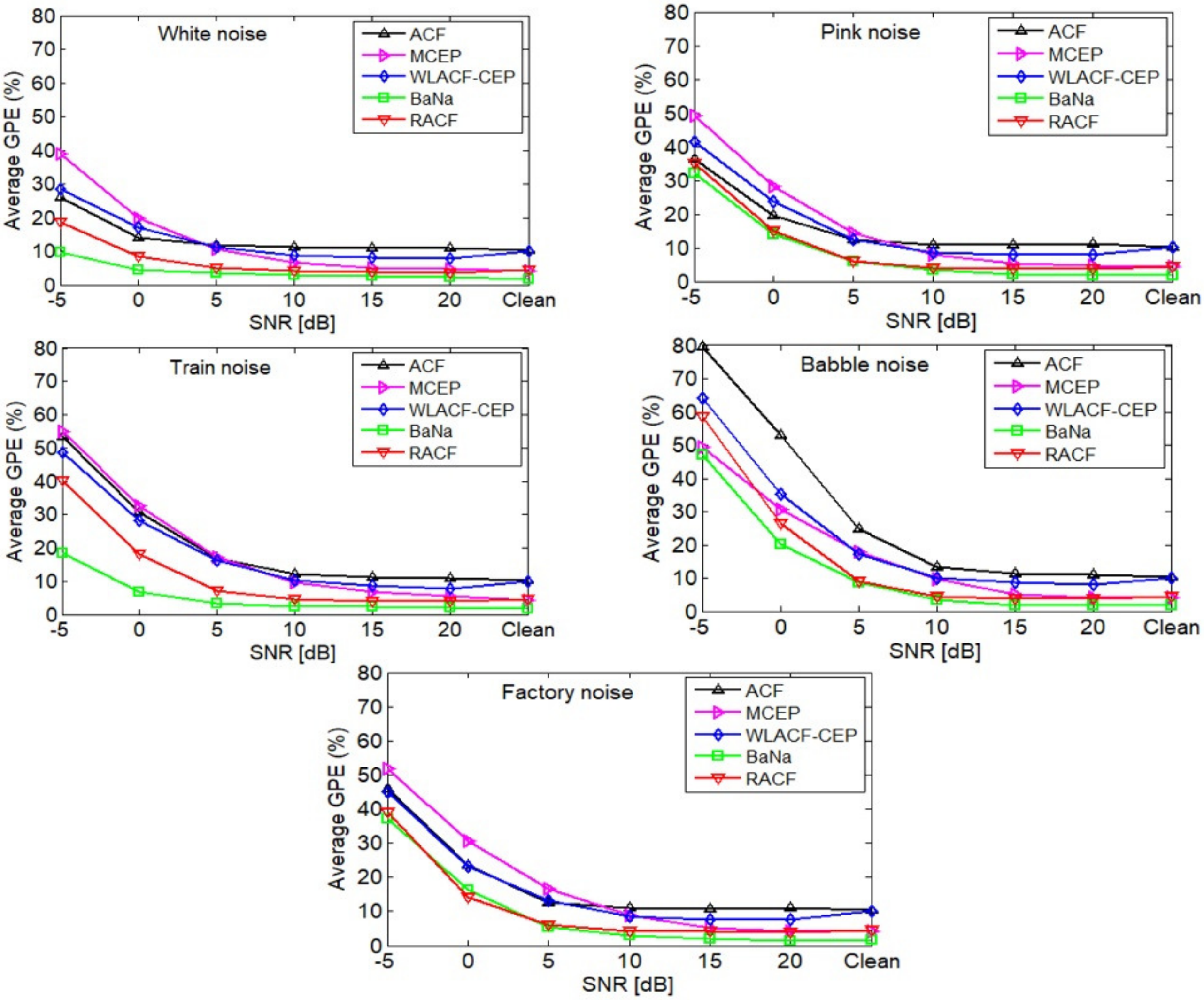

Fig. 7. GPE for four female speakers with different types of noise under different SNR levels.

was taken from the Japanese Electronic Industry Development Association (JEIDA) noise database [16]. The SNR was set to -5, 0, 5, 10, 20, $\infty$ [dB] and the other experimental conditions were frame length of 51.2 ms except for BaNa, frame shift of 10ms, band limitation of 3.4 kHz and DFT (IDFT) length, $M$, of 1024 points.

The following error $e(l)$ was used for the evaluation of fundamental frequency extraction accuracy based on Rabiner's method [2]:

$$e(l) = F_1(l) - F_2(l), \quad for \quad l = 1, 2..., k \tag{10}$$

where $k$ corresponds to the number of frames in the utterance, $F_1(l)$ and $F_2(l)$ are the fundamental frequency extracted from the noisy speech and the true fundamental frequency at the $l$-th frame, respectively. In (10), $e(l)$ indicates an extraction error. If $|e(l)| \geq 10$ Hz, we recognized the error as gross pitch error (GPE). We detected and assessed only voiced parts of sentences for the extraction of the pitch.

### *B. Preliminary Experiments*

Figures 4 and 5 represent the selection of more appropriate cut-off quefrency levels of the HPL in male and female speakers, respectively. From Fig. 4, we observe that the 1ms HPL of RACF shows lower GPE rates at low SNR for female speakers. On the other hand, we see that the 5ms HPL of RACF shows lower GPE rates at low SNRs (-5dB and 0dB) for male speakers in Fig. 5. In Figs. 4 and 5, the cut-off quefrency levels of 2ms and 4ms HPL provide the lowest error rate in clean speech, but these are competitive with the cut-off quefrency levels of 1ms and 5ms HPL, respectively. From this point of view, we select the threshold levels of HPL as 1ms and 5ms for female speech and for male speech, respectively, in the proposed method.

### *C. Performance Comparison*

The pitch extraction performance of the conventional and proposed methods was investigated. In white noise, the power spectral density is constant across the entire frequency spectrum, while the power spectral density is not uniformly distributed in color noise. All parameters of the conventional methods are the same as those in the proposed method, except for the frame length for BaNa. Specifically, for BaNa the frame length was set as 60 ms according to the suggestion in [11]. The source code to implement BaNa was collected from [17].

Figures 6 and 7 show the average GPE results of four male and four female speech data, respectively, with different types on noise. The speech database used, NTT Advanced Technology database, has no ground truth of fundamental frequency, thus we manually determined the fundamental frequency for each frame by sight carefully. From Fig. 6, it is evident that the performance of the proposed method achieves the lowest average GPE rate [%] among all methods to be compared at every SNR, except for the low SNR case in white noise where the ACF provides better performance. The ACF in white noise gathers the affects on the first lag, while the other lags lead to them vanishing. For babble noise in the male speaker, the proposed method performs better in high SNR environments, but BaNa performs better at low SNR. For the female speaker in Fig. 7, it is observable that the average GPE [%] of the proposed method is significantly superior to the other conventional methods over all SNR cases, except for BaNa. BaNa performs much better for female speech than for male speech, because wider harmonics in the frequency domain are used. However, it should noted here that the computational complexity of BaNa is extremely high. Table 1 shows the processing time per one-second data for each method (averaged for five trials in each method). The computer we used was a PC with Intel (R) Core(TM) i5-6400K, 4 [GHz] clock speed of CPU and 8 [Gigabytes] of memory.

TABLE I
PROCESSING TIME PER SECOND OF SPEECH

| ACF | MCEP | WLACF-CEP | BaNa | RACF |
|---|---|---|---|---|
| 1.277 | 0.284 | 2.533 | 28.769 | 1.398 |

## IV. CONCLUSION

In this paper, we proposed a simple and noise robust method to extract the pitch of speech more accurately utilizing the ACF of the vocal source. Experimental results showed that the proposed method is an efficient and effective method to extract the pitch even in several noise environments.

## REFERENCES

[1] W. Hess, “Pitch determination of speech signals,” Springer-Verlag, 1983.

[2] 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. 24, no. 5, pp. 339-417, Oct. 1976.

[3] L. R. Rabiner, “On the use of autocorrelation analysis for pitch detection,” IEEE Trans. Acoust., Speech Signal Process, vol. 25, no. 1, pp. 24-33, Feb. 1977.

[4] M. J. Ross, et al, “Average magnitude difference function pitch extractor,” IEEE Trans, Acoust., Speech Signal Processing, vol. 22, no. 5, pp. 353-362, Oct. 1974.

[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] J. W. Xu and J. C. Principle, “A pitch detector based on a generalized correlation function,” IEEE Trans. on Audio, Speech and language Processing, vol. 16, no. 8, pp. 1420-1432, Nov. 2008.

[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, 2001.

[8] A. M. Noll, “Cepstrum pitch determination,” Journal on Acoust. Soc. Am., vol. 41, no. 2, pp. 293-309, 1967.

[9] H. Kobayashi and T. Shimamura, “A modified cepstrum method for pitch extraction,” Proc. on IEEE Asia-Pacific International Conference on Circuits and Systems Microelectronics and Integrating Systems (APCCAS), 1998.

[10] M. A. F. M. R. Hasan, M. S. Rahman and T. Shimamura, “Windowless-autocorrelation-based cepstrum method for pitch extraction of noisy speech,” Journal of Signal Processing, pp. 231-239, May 2012.

[11] N. Yang, H. Ba, W. Cai, I. Demirkol and W. Heinzelman, “BaNa : A noise resilient fundamental frequency detection algorithm for speech and music,” IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol. 22, no. 12, Dec. 2014.

[12] M. S. Rahman and T. Shimamura, “Linear prediction using refined autocorrelation function,” EURASIP Journal on Audio, Speech, and Music Processing, vol. 2007, pp. 1-9, 2007.

[13] A. Oppenheim and R. Schafer, “Homomorphic analysis of speech,” IEEE Trans. on Audio and Electroacoustics, vol. 16, no. 2, pp. 221-226, 1968.

[14] “20 Countries Language Database,” NTT Advance Technology Corp., Japan, 1988.

[15] A. Varga and H. J Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” Speech Communication, vol. 12, no. 3, pp. 247-251, 1993.

[16] S. Itahashi, “Creating speech copora for speech science and technology,” IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences, vol. E 74, no. 7, pp. 1906-1910, 1991.

[17] Wcng, wireless communication networking group, [Online]. Available: http://www.ece.rochester.edu/projects/wcng/code.html