Noise Whitening-Based Pitch Detection for Speech Highly Corrupted by Colored Noise

Kyung Jin Byun, Sangbae Jeong, Hoi Rin Kim, and Minsoo Hahn

ABSTRACT— Pitch estimation is important in various speech research areas, but when the speech is noisy, accurate pitch estimation with conventional pitch detectors is almost impossible. To solve this problem, we propose a new pitch detection algorithm for noisy speech using a noise whitening technique on the background noise and obtain successful results.

I. INTRODUCTION

The importance of a reliable and accurate pitch detection algorithm is well recognized in the speech processing area because such an algorithm can provide the more accurate spectral and prosody information needed in all speech research fields, such as speech synthesis, voice color conversion, speech coding, and speech recognition [1], [2]. To estimate pitch frequency, a simple average magnitude difference function (AMDF) or an autocorrelation function is generally used [3]. Some studies have proposed wavelet- and waveform similarity measure-based pitch detection algorithms [4], [5]. All these detectors are known to produce rather successful pitch detection results when the signal-to-noise ratio (SNR) is above 20 dB. However, when the SNR is considerably lower, robust pitch detection is difficult because pitch doubling or halving frequently occur. Median filters are usually adopted to suppress this but the results are still far from satisfactory. Among the different kinds of algorithms proposed in [3], [4], and [5], the one based on autocorrelation is the most robust against noise but at the cost of increased computational complexity.

Manuscript received Oct. 16, 2002.

Kyung Jin Byun (phone: +82 42 860 5831, e-mail: kjbyun@etri.re.kr) is with Mobile Telecommunication IC Design Team, ETRI, Daejeon, Korea.

Sangbae Jeong (e-mail: sangbae@icu.ac.kr), Hoi-Rin Kim (e-mail: hrkim@icu.ac.kr), and Minsoo Hahn (e-mail: mshahn@icu.ac.kr) are with School of Engineering, Information and Communications University, Daejeon, Korea.

In general, the spectra of real environmental noises are not white but colored, and the accuracy of pitch prediction decreases more rapidly with colored noise than with white noise. To solve this problem of pitch estimation, we propose a more accurate noise-robust pitch detection algorithm with a noise-whitening procedure.

II. CONVENTIONAL PITCH ESTIMATION ALGORITHMS

Pitch detection algorithms can be classified into two general categories. One includes non-event detection pitch estimators while the other includes event detection estimators. For non-event detection estimators, algorithms based on an AMDF or autocorrelation function are generally used. In an AMDF-based method, for a given signal $x(n)$, the pitch estimation function $\gamma_x(k)$ in (1) is estimated for every analysis frame and pitch values are evaluated by checking the nearest minimum smaller than a prefixed threshold.

$$\gamma_x(k) = \sum_{m=-N}^{N} |x(n+m) - x(n+m+k)|. \tag{1}$$

In autocorrelation-based algorithms, the autocorrelation function of an input signal frame is evaluated as in (2).

$$\begin{aligned} \Phi_x(k) &= E[x(n)x(n+k)] \\ &\cong \frac{1}{2N+1} \sum_{m=-N}^{N} x(n+m)x(n+m+k). \end{aligned} \tag{2}$$

For the P-periodic $x(n)$, $\Phi_x(k) \cong \Phi_x(k+P)$. Thus, we can find the pitch period simply by checking the nearest maximum of the function. Autocorrelation-based methods are

more noise-robust than AMDF-based ones but this is at the cost of increased computations.

For event detection pitch detectors, [4] and [5] reported fairly good performance for considerably clean speech with the area information- and Dyadic Wavelet Transform (DyWT)-based algorithms. The DyWT of the given signal tends to show local maxima around the discontinuities of the original signal. Consequently, glottal closure instants, the interval of which can be interpreted as the pitch, can be easily found by detecting the local maxima.

III. PROPOSED PITCH ESTIMATION ALGORITHM

In noisy environments, we assume that noises are additive, not convolutive. In this case, the autocorrelation function of noisy speech, $y(n)$, can be represented as in (3), where $x(n)$ and $v(n)$ denote clean speech and noise, respectively.

$$\begin{aligned} \Phi_y(k) &= E[y(n)y(n+k)] \\ &= \Phi_x(k) + \Phi_v(k) \\ y(n) &= x(n) + v(n). \end{aligned} \tag{3}$$

If the additive noise $v(n)$ is white, the corresponding autocorrelation function can be expressed as $\Phi_v(k) = \sigma_v^2 \delta(k)$, where σ_v^2 is the noise power. Therefore, the autocorrelation function of the noise-corrupted speech $y(n)$ is the same as that of the clean speech except at $k = 0$. Because pitches are determined by searching the nearest maximum value exceeding a prefixed threshold, the performance of the pitch detection algorithm based on autocorrelation is irrelevant to the SNR when the noises are ideally white. However, for various real environmental noises, such as car noises, that are not white but colored, a more severe performance degradation in pitch estimation is expected with an increased noise level. To cope with this type of performance degradation, noise whitening is helpful. Among the many whitening methods, we utilize the fact that the forward prediction error of the noise-like input signal is almost white. In other words, the 5th order linear predictive coefficients are calculated in silence or pause intervals and noisy speech is filtered by using these coefficients. This order is chosen because it gives fairly successful noise whitening results with less computational complexity. More details of this noise whitening procedure can be found in [6]. After this whitening procedure, the autocorrelation function-based pitch estimation is performed.

Before the whitening procedure, we apply a highpass filter with a cutoff frequency of 80 Hz because the environmental noises for our test came from running cars and they have high power below 80 Hz. Finally, a median filter is applied to the evaluated pitch contour to reduce possible pitch doubling and halving. Figure 1 gives a block diagram for our overall algorithm.

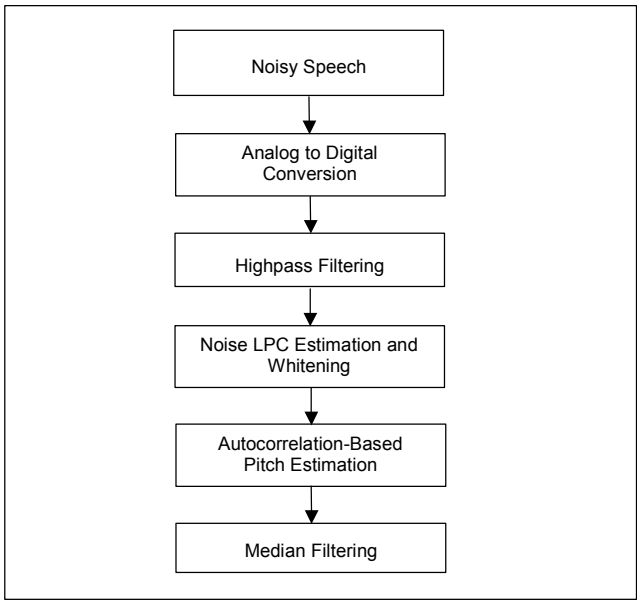

Fig. 1. Overall pitch estimation procedure.

IV. EXPERIMENTAL RESULTS

We used sentence speech signals of 4 male and 4 female speakers in their 20s and each speaker uttered 4 sentences. The speech signals were recorded on a digital audio tape and then sampled at 8 kHz with a 16 bit resolution. The test sentences were about 2 seconds long. Pitch estimation functions were calculated for 160-point, i.e., 20 ms windows while pitch values were evaluated for every 10 ms by sliding the analysis window with this amount. The 5th order linear predictive coefficients needed for the background noise whitening were calculated with the Levinson-Durbin algorithm [7] and finally, pitches were estimated with the autocorrelation method after noise whitening. Lastly, the 5th order median filter was adopted for the estimated pitch contour smoothing. For calculating pitch values only for voiced speech frames, we decided whether the frame was voiced or unvoiced by utilizing log energy and a zero crossing rate before applying the autocorrelation-based pitch estimation procedure. By adding test noises to clean speech, we obtained the SNR values of –5 dB, 0 dB, 5 dB, 10 dB, and 15 dB for our experiments.

The performance index for our test was based on Rabiner's method [8]. The pitch estimation error $e(n)$ is defined by (4) where $F_t(n)$ and $F_e(n)$ are the true and the estimated pitch

in the sample number, respectively.

$$e(n) = F_t(n) - F_e(n). \quad (4)$$

If $|e(n)| > 10$, the estimated pitch value at frame n is considered to be the gross pitch error (GPE). We evaluated the performance of pitch estimators by calculating the GPE probability. The true pitch contours of the test sentences were obtained by applying the pitch estimator based on autocorrelation to the clean speech signals with manual correction of possible pitch doubling or halving.

We tested our algorithm for two types of colored noises: one was real car noises and the other was artificial noises obtained by filtering white noises with a power spectrum-shaping filter, $1/(1 - 0.8z^{-1})$. Figures 2 and 3 summarize our experimental results for the two types of noises.

As these figures show, our proposed algorithm improves the pitch detection performance more for the real car noises than the artificial ones and the amount of improvement is greater than 10% for below 0 dB SNR environments. The DyWT-based algorithm seems somewhat robust against noise for the artificial noises but very poor for the real car noises. There is a very slight performance degradation of the proposed method compared with the DyWT-based algorithm above 10 dB SNR values but it is almost negligible.

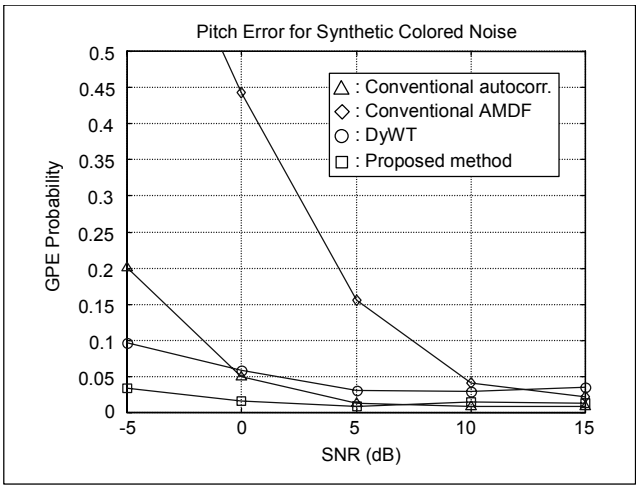

Fig. 2. Performance with synthetic colored noises.

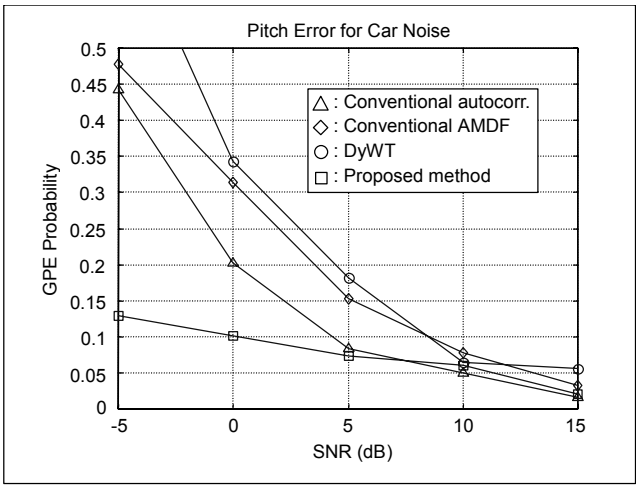

Fig. 3. Performance with real car noises.

V. CONCLUSION

In this paper, we proposed a new pitch detection algorithm, which is robust against additive background noises. As our experiments demonstrate, the proposed algorithm can noticeably improve the performance index represented by the GPE probability in severely noisy conditions having SNR values below 5 dB, while it shows almost equivalent performance for rather high SNR values. Based on these results, we strongly recommend use of our algorithm for various high quality speech processing algorithms including speech recognizers and coders when they are expected to operate in severe noise environments.

REFERENCES

[1] Ki-Seung Lee, Richard V. Cox, "A Low Bit Rate Speech Coder Based on a Recognition/Synthesis Paradigm," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, 2001, pp. 482-491.

[2] Sanghun Kim, Youngjik Lee, and Keikichi Hirose, "Unit Generation Based on Phrase Break Strength and Pruning for Corpus-Based Text-to-Speech," *ETRI J.*, vol. 23, no. 4, Dec. 2001, pp. 168-176.

[3] L.R. Rabiner and R.W. Shafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978, pp. 141-161.

[4] S. Kadambe and G.F. Boudreaux-Bartels, "Application of the Wavelet Transform for Pitch Detection of Speech Signals," *IEEE Trans. on Information Theory*, vol. 38, no. 2, 1992, pp. 917-924.

[5] M. Hahn and D.G Kang, "Precise Glottal Closure Instant Detector for Voiced Speech," *Electronics Lett.*, vol. 32, no. 23, Nov. 1996, pp. 2117-2118.

[6] S. Jeong, and M. Hahn, "Speech Quality and Recognition Rate Improvement in Car Noise Environments," *Electronics Lett.*, vol. 37, no. 12, 2001, pp. 800-802.

[7] M.H. Hayes, *Statistical Digital Signal Processing and Modeling*, John Wiley & Sons, 1996, pp. 216-263.

[8] 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 ASSP.*, vol. 24, no. 5, 1976, pp. 369-377.