International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:1, No:8, 2007

# On SNR Estimation by the Likelihood of near Pitch for Speech Detection

Young-Hwan Song, Doo-Heon Kyun, Jong-Kuk Kim, and Myung-Jin Bae

***Abstract*—**People have the habitual pitch level which is used when people say something generally. However this pitch should be changed irregularly in the presence of noise. So it is useful to estimate SNR of speech signal by pitch. In this paper, we obtain the energy of input speech signal and then we detect a stationary region on voiced speech. And we get the pitch period by NAMDF for the stationary region that is not varied pitch rapidly. After getting pitch, each frame is divided by pitch period and the likelihood of closed pitch is estimated. In this paper, we proposed new parameter, NLF, to estimate the SNR of received speech signal. The NLF is derived from the correlation of near pitch periods. The NLF is obtained for each stationary region in voiced speech. Finally we confirmed good performance of the estimation of the SNR of received input speech in the presence of noise.

***Keywords*—**Likelihood, Pitch, SNR, Speech.

## I. Introduction

As a communication medium of information, speech is not only used a lot, but it is also the most comfortable. When we have conversation by speech, transmission of the information, which wanted to be delivered, is affected by the noise level. Likewise, the effects of the noise have to be considered on the speech signal processing such as speech recognition, synthesis, and analysis.

Speech is mainly divided into voiced and unvoiced sound [1]. Generally, voiced speech is modeled by quasi-periodic pulse, and unvoiced speech is modeled by white Gaussian noise. The ratio of energy or the zero crossing method is normally used to classify into voiced and unvoiced sound [2]. In the ratio of energy method, voiced sound is decided when the energy is high. In the zero crossing method, voiced sound is decided when the rate of zero crossing is low. Otherwise, it becomes the region of unvoiced sound. Namely, on the v/unvoiced sound decision, voiced sound is decided when periodic energy is high after compared the energy of unvoiced speech signal with the ratio of the periodic energy, but if this ratio is low, it is decided as unvoiced speech signal. At this time, voiced sound has quasi-periodic pitch. When voiced sound is mixed with noise, pitch period is varied randomly, since the AWGN (Additive White Gaussian Noise) is no correlated with the input signal [7]. So we can estimate the noise level, which is mixed with speech signal, by comparing pitch periods in the frame after dividing a frame for each pitch. In section 2, this paper describes the setting up speech analysis section by the detection of energy, and NAMDF (Normalized Average Magnitude Difference Function), which is pitch detection method, to divide a frame for each pitch. Section 3 describes the proposed algorithm which estimates the noise level for arbitrary speech data by analysis of closed pitch in a frame. Section 4 presents the experimental results to show performance of the proposed algorithm. Concluding remarks are drawn in Section 5.

Manuscript received October 12, 2007.

Authors are with Information and Telecommunication Engineering Department, Soongsil University, 1-1 Sangdo 5 dong DongJak-Ku, Seoul, 156-743, Republic of Korea (phone: +82-02-824-0906; fax: +82-02-824-0906; e-mail: song0one@yahoo.co.kr, mjbae@ssu.ac.kr).

## II. The Energy Detection and the Pitch Detection

Energy detection and pitch detection processing are necessary for many speech signal processing systems. Since the effect of different speakers should be decreased in terms of the detection of the pitch period and the location of pitch peaks, the accuracy of speech recognition should be improved with the pitch detection. Likewise for speech synthesis, the naturals and the characteristics are modified and maintained easily [3]. Speech has a property that quasi-periodic pitch period is appeared on the voiced sound which keeps high ratio of energy relatively [4]. In the detected speech section which has energy level higher than threshold, the pitch variation is adjusted on the almost speech analysis-synthesis (vocoder) systems such as speaker recognition, an assistant system for a speech defect person [5].

### *A. Short-Time Energy Detection Method*

Amplitude of speech signal is varied with time. Especially, Amplitude of unvoiced sound is generally much lower than amplitude of voiced sound. The short time energy of speech signal is good parameter which reflects the variation of these amplitudes [6]. The short time energy is defined in (1).

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \tag{1}$$

This equation is rewritten by (2) as follows

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m) \cdot h(n-m) \tag{2}$$

Where

$$h(n) = w^2(n) \tag{3}$$

In (3), it can be presented by Fig. 1. Signal $x^2(n)$ is convolved with $h(n)$ which is impulse response of linear filter like (3).

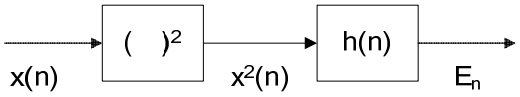


Fig. 1 Block diagram of short time energy detection

### B. Short-time NAMDF Function

Since the importance of pitch detection as remarked above, many methods of the pitch period estimation are proposed. That is classified by Time-dependent method, Frequency-dependent method, Hybrid method. The time-dependent method is that estimates a pitch period by decision logic after an emphasized periodicity of waveform, i.e. The Parallel Processing, AMDF, ACM etc [7,8]. The frequency-dependent method is that detects the fundamental frequency of a voiced sound by measurement of the harmonics interval of the speech spectrum, i.e. Harmonics Analysis [9], Lifter, Comb-filtering etc. On the hybrid method, such as, Cepstrum, Spectrum Comparison method etc, it takes advantages of both the simplicity and the accuracy of pitch period estimation for the time-dependent method and the great capacity for the background noise and the change of phoneme for the frequency-dependent method [9,10].

This paper performs The AMDF which is defined by the absolute value, it is different from the autocorrelation function which is defined by product of x(n) and x(n-k) [8]. (4) presents the AMDF.

$$AMDF(k) = \sum_{m=-\infty}^{\infty} |x(m) - x(m+k)| \tag{4}$$

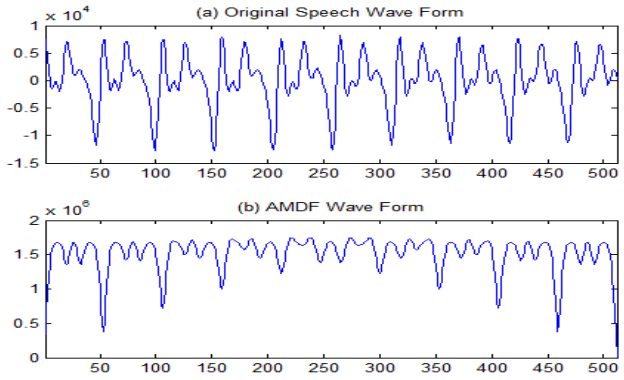


Fig. 2 AMDF for voiced sound

The AMDF function is implemented with subtraction, addition, and absolute value operations, in contrast to addition and multiplication operations for the auto correlation function. Because of the simple operations, The AMDF is faster than autocorrelation method. For this reason, it is normally adjusted in the real time processing. In the autocorrelation method, the function make the maximum value at the multiples of pitch period, but In the AMDF method, the function make the minimum value at the multiples of pitch period. Fig. 2 shows the AMDF method for voiced sound.

To decide the frame whether noise region or stationary region, we use to measure the ratio of average amplitude. $MR(fr)$ presents the ratio of average amplitude for each frame. (5) shows $MR(fr)$

$$MR(fr) = \frac{\sum_{k=N/2}^{N-1} s(n-k)}{\sum_{k=0}^{N/2-1} s(n-k)} \tag{5}$$

Where s(n) is speech signal, variable n is the start index of frame, N is the length of frame, $fr$ is the frame indices. Since the ratio of average amplitude is of closed frame when the frame length is divided by two, it is affected by window function. The reason of dividing a frame for each pitch period is to search a transition region of speech signal by correlation of each pitch section after synchronizing with pitch. The normalized AMDF (Normalized Average Magnitude Difference Function, NAMDF) method is proposed to measure a state of a frame without considering window function [11]. (6) shows NAMDF which is used for pitch detection.

$$NAMDF(d) = \frac{\sum_{n=1}^{N} |s(n) - s(n-d)|}{\sum_{n=1}^{N} \left[ |s(n)| + |s(n-d)| \right]} \tag{6}$$

where, s(n) is speech signal, N is length of window and d is a delay factor.

## III. The Estimation of the Noise Level

In this paper, for the first of all for estimation of the noise level, we perform short time energy detection for input speech signal for each frame. And we find pitch period by NAMDF for a frame when the energy is higher than threshold, we divide speech signal with each pitch period for a frame. The variation of phoneme is generally analyzed for a frame since phoneme is varied slower than speech waveform [12]. As the measurement method that the state of a frame can be decided unrelated with window function, NAMDF is used. As we mentioned, to detect the transition region of speech signal by correlation with each pitch which is synchronized with a pitch period [5], we divide a speech signal by each pitch period. To estimate the noise level, we only adjust the stationary region in voiced speech, since on the transition region pitch is varied rapidly.

We obtain the pitch period ($\tau$) by (6), divide input speech by pitch period. We adjust (7) to divided input speech. By (7), we get the correlation coefficient of pitch.

$$Corr_p(k) = \frac{\sum_{n=1}^{\min(\tau_{k-1},\tau_k)} s_{k-1}(n)s_k(n) + \sum_{n=1}^{\min(\tau_k,\tau_{k+1})} s_k(n)s_{k+1}(n)}{\sum_{n=1}^{\tau_k} s_k^2(n)} \quad (7)$$

Where $Corr_p(k)$ is the correlation coefficient of the k-th pitch period with near pitch segments, $s_k(n)$ is divided speech signal by $\tau_k$ which is pitch periods of $s_k(n)$. When the correlation which is divided by pitch period is higher with near pitch segments, the correlation coefficient, which is synchronized with pitch period, becomes closer with 2.

This paper obtains the parameter to estimate the additive noise level by varying the correlation coefficient to be adapted for noise level sensitively. Fig. 3 describes the block diagram of process to detect the region of noise.

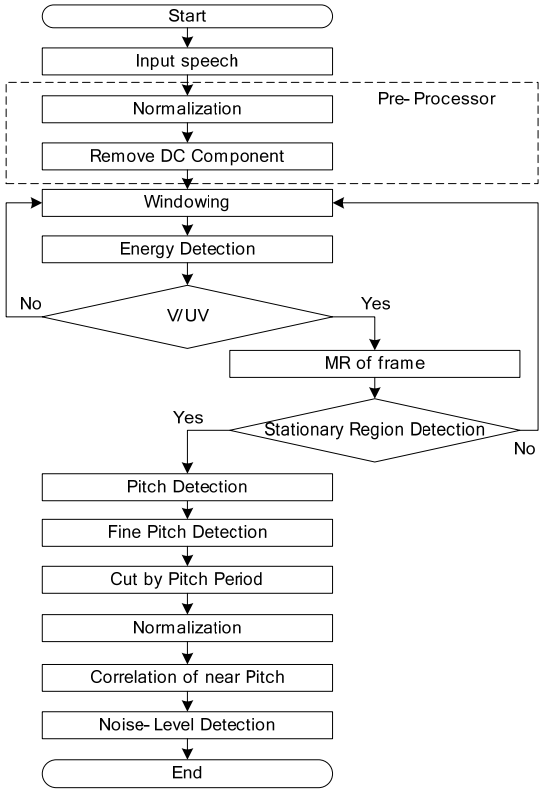


Fig. 3 Block diagram to detect noise level

To divide by a frame, we get the short time energy by Hamming Window. At this time, when the energy is lower than threshold, it is ignored to estimate noise level. In voiced speech, we derive a $MR(fr)$ from (5). We decide the stationary region by $MR(fr)$. And we detect a short time pitch period. After obtained pitch period, Fine Pitch period ($\tau$) is estimated to divide frame by pitch exactly, we divide a frame by this pitch period, and obtain the gain coefficient ($Corr_p$), which is correlation coefficient derived from correlation for the divided speech waveform. When calculate the $Corr_p$, 1st and last segment in frame is not concerned, since it is defected for windowing. This obtained value is closed to 1 on a clean speech, but in the presence of noise, it is varied proportionally by additive noise. To apply this property, we subtracted the correlation coefficient from 1 and squared.

$$Var_p(k) = \left|1 - \frac{Corr_p(k)}{2}\right|^2 \quad (8)$$

And then the proposed parameter $Var_p$ is derived from output. (8) shows $Var_p$. This describes emphasized difference between clean speech and noisy speech. In the clean speech, $Var_p$ becomes low value closed to 0, and in the presence of noise it varies and depends on the noise level. After obtained $Var_p$, we performed summation of all $Var_p$ s for each frame, and derived the average by division into the number of $Var_p$ s for speech section. So, we are able to estimate the additive noise level, which is mixed to the speech signal, by NLF (Noise-Level Factor). It is defined in (9).

$$NLF = \frac{\sum_{n=1}^{N} Var_p(\mathrm{n})}{N} \quad (9)$$

Where $Var_p$ is derived parameter from the correlation coefficient of speech section which is divided by pitch period, N is the number of pitch segments for the stationary region in voice speech signal. The noise level is estimated by NLF. NLF is obtained for each stationary region.

## IV. Experimental Result

Computer simulation was performed to evaluate the proposing algorithm using a personal computer (AMD Athlon x2 chipset) interfaced with the 16-bit AD/DA converter. To measure the performance of the proposing algorithm, we used the following speech data. The speech data was sampled at 8 kHz and was quantized by 16bits. With each of the 30 people men and women for 30 seconds, the data was recorded in a quiet room.

To simulate natural condition, for clean speech, we make the data source, to which the noise is added, the corresponded SNR is 30dB, 20dB, 10dB, 0dB individually. At this time, we use pink noise and white noise for additive noise

Fig. 4 shows the correlation coefficients of speech signal which is synchronized with pitch period. In Fig 4, (a) shows the original input speech. (b) describes the pitch variation as the correlation coefficient (likelihood) of closed speech section which is divided by pitch period for original input speech without noise. (c) shows the $Corr_p$ s that the speech signal is mixed with white noise with 20dB SNR, (d) is the $Corr_p$ s that the speech signal is added with pink noise with 20dB SNR. When pitch is same as closed pitch, $Corr_p$ becomes 1. At the

original speech, $Corr_p$ is nearly 1, on the other hand, in the presence of noise, the $Corr_p$ shows huge variation much more than of the $Corr_p$ of the original speech.

Fig. 5 describes NLF which is gained from 30 of speech source with different noise levels. The square points are the NLF of received speech signal with pink noise, and the triangle points are the NLF of received speech signal with white noise. From the result we can confirm the NLF is raised by increasing the additive noise.

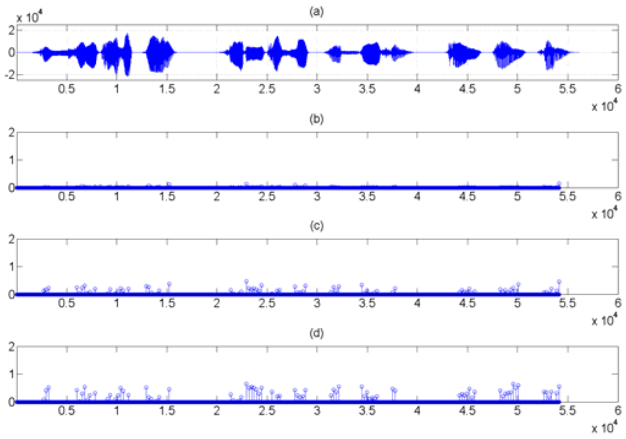


Fig. 4 Speech signal and $Var_p$

(a) Input speech signal
(b) $Var_p$ of clean signal
(c) With 20dB SNR
(d) With 10dB SNR

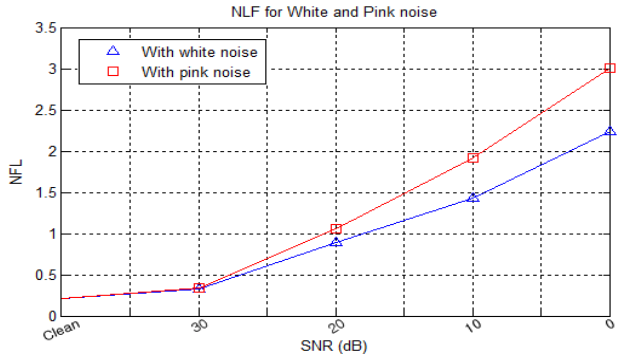


Fig. 5 NFL for White and Pink Noise

## V. Conclusion

On the speech processing, if we can estimate the SNR of speech signal, we can confirm the performance of input system, vocoder, channel etc. Utterance has the stable pitch level which is used when people say something generally. However this pitch should be changed irregularly in the presence of noise. So it is useful to estimate SNR of speech signal by pitch. In this paper, we proposed new method to estimate the SNR of speech signal by this property.

The likelihood of the closed pitch is presented by new parameter. And the method which can estimate the noise level of input speech signal is proposed. We also get good performance through this method. The pitch variation coefficient was nearly 1 (one) if there is no noise. On the other hand, in the presence of noise the NFL was increased by addition of noise level proportionally, even though it had a little difference depending on the kinds of noise. On this wise, we derived the NFL from the likelihood of near pitch, and the NLF reflects the SNR of speech.

## References

[1] J.S. Han, *Speech Signal Processing*. Seoul: Osung Media, 2000. ch . 2.
[2] A. Caruntu, G. Toderean, A. Nica, "Automatic Silence/Unvoiced/Voiced Classification of Speech Using a Modified Teager Energy Feature," *WSEAS*, pp. 62-65, Nov., 2005
[3] WangRae Jo, JongKuk Kim, and Myung Jin Bae, "A Study on Pitch Detection in Time-Frequency Hybrid Domain," *Springer-Verlag, Lecture Notes in Computer Science*, vol. LNCS 3406, pp.437-440, Feb 2005.
[4] Hans Werner Strube , "Determination of the instant of glottal closure from the speech wave," *J., Acoust., Soc., Am*, Vol. 5, No. 5, pp. 1625-1629, November 1974.
[5] J.K. Kim, D.S. Na, M.J. Bae "On a pitch alteration technique in transformation domain of speech signals," DCDIS, 2007. pp. 522-526
[6] M. Bae, J. Rheem, and S. Ann "A Study on Energy Using G-peak from the Speech Production Model," *KIEE, Korea*, Vol. 24, No. 3, pp. 381-386, May 1987.
[7] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech signals*. New Jersey: Englewood Cliffs, Prentice-Hall, 1978, ch 6, 7.
[8] P. E. Paparnichalis, *Practical Speech Processing*. New Jersey: Prentice-Hall, Inc, Englewood Cliffs, 1987.
[9] S. Seneff, "Real Time Harmonic Pitch Detection," *IEEE Trans. Acoust. Speech, and Signal Processing*, Vol. ASSP-26, pp. 358-365, Aug. 1978.
[10] S. D. Stearns & R.A. David, *Signal Processing Algorithms*. New-Jersey: Prentice-Hall, Inc, Englewood Cliffs, 1988.
[11] M.J. Bae, "On Detecting the Steady State Segments of Speech Waveform by using the Normalized AMDF," *IEEK*, Vol.14, No.1, pp. 600-603, June, 1991.
[12] S. J. Kim. "A Segmentation Algorithm of the Connected Word Speech by Statistical Method," *IEEK*, Vol.26, No. 4, pp. 151-162, Apr., 1989.
[13] M. J. Bae, and S. Ann, "Fundamental Frequency Estimation of Noise Corrupted Speech Signals Using the Spectrum Comparison," *J., Acoust., Soc.*, Korea, Vol. 8, No. 3, June 1989.

**Y. H. Song** (M'07) became a Member (M) of ASK in 2007. He is from Republic of Korea and was born in 1981. Next, He received the B.S. degree in Electronic Engineering from Soongsil University in 2007. He is currently the under M.S. degree at Soongsil University in Seoul, Korea.

He received the military service from Jun, 2001 to Aug, 2003. He researched "The Identification of Sound Source for Infrasound" with the department of national defense. His research interests include speech signal processing, speech synthesis, speech recognition, speech coding, and audio coding.