On a Classification of Voiced/Unvoiced by using SNR for Speech Recognition

Jongkuk Kim, Hernsoo Hahn

Department of Information and Telecommunication Engineering, Soongsil University 369 Sangdo-Ro, Dongjak-Gu, Seoul, 156-743, Korea
kokjk@hanmail.net

Abstract - As communication medium of information, speech is not only used a lot, but also is the most comfortable. When we have conversation by speech, transmission of the information, which wanted to be delivered, is affected by the noise level. In speech signal processing, speech enhancement is using to improve speech signal corrupted by noise. Usually noise estimation algorithm need flexibility for variable environment and it can only apply on silence region to avoid effects of speech signal. So we have to preprocess finding voiced region before noise estimation. we proposed SNR estimation method for speech signal without silence region. For unvoiced speech signal, vocal track characteristic is reflected by noise, so we can estimate SNR by using spectral distance between spectrum of received signal and estimated vocal track. The proposed estimation method on voiced speech and the method by using voiced/unvoiced region energy are operated with simple logic as time domain method. And the estimation method on unvoiced region is possible to estimated noise level for narrow-band speech signal by using vocal track properties. It can be applied to rate decision of vocoder and used for pre-processing to decide threshold of noise reduction.

Index Terms - Voiced, Speech production model, White noise, SNR, vocoder, LPC, VAD

1. Speech Analysis

It is often necessary to perform speech enhancement through noise removal in speech processing systems operating in noisy environments. As the presence of noise degrades the performance of speech coders and voice recognition system[10,11]. It is therefore common to incorporate speech enhancement as a preprocessing step in these systems. The other important application of speech enhancement is to improve the perceptual quality of speech in order to reduce listener's fatigue.

The additive noise may be due to the noisy environment in which the speaker is speaking, or it may arise from noise in the transmission media. Furthermore, most of these algorithms only attempt to modify the spectral amplitudes of the noise corrupted speech signal in order to reduce the effect of the noise component while leaving the noise corrupted phase information intact. we study the performance of these filters for the enhancement of speech contaminated by additive white noise. Performance comparisons are accomplished in terms of SNR[1].

Enhancement the speech signal for mobile communication system or signal processing system, which reduces noise has been studied a lot wide side of views. And lots of methods have been used for signal enhancement. And that methods need flexibleness for changeable conditions. In these days some of noise estimation method calculate the noise power when silent region between speech to speech [4].

Using with probability model when noise conditions are changing.

knowledge compilation, and achieved good results. Besides, many researchers applied the extension rule to the model counting problem[18], and many amended it so as to applied it into the TP of modal logic[19] .Still some researchers improved the extension rule, and put forward series of algorithms such as NER, RIER, etc[20,21].

This paper is organized as follows. In section 2, the related extension-rule based TP methods are given. In section 3, the parallel TP method based on the Semi-extension rule is presented. The experimental results of comparing the algorithm proposed in this paper with other algorithms are also presented in section 4. Finally, our work of this paper is summarized in the last section.

2. Speech Analysis

2.1. Speech Feature

Speech sounds can their mode of excitation. The excitation source of unvoiced speech signals is the random noise Generator. The unvoiced speech has no periodicity and appears higher average zero-crossing rate than the voiced signal, because it has the first formant with wide bandwidth at near 3 kHz. Generally, the excitation source of voiced speech is a glottal pulse train that has quasi-periodic pulse and large amplitude. The voiced speech signals have periodicity owing to vibrating of vocal tract[6]. Due to the resonance of vocal tract, the voiced speech has formants with bandwidth. Therefore, the voiced waveforms in a pitch period have damped-oscillation. In frequency domain, the spectrum of voiced speech appears to be multiplied the harmonics of fundamental frequency by formant envelope of vocal tract. Figure 1 is the block diagram of 'Human speech production and machine model' as explained.

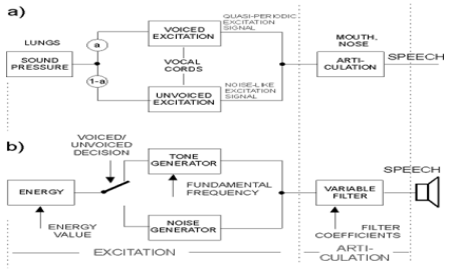

Fig. 1. Speech production model

2.2 SOURCE-FILTER MODEL

Why LPC (Linear prediction code) has been so widely used in speech signal processing ? LPC provides a good model of the speech signal, especially the quasi steady state voiced regions, analysis leads to a reasonable source-vocal tract separation and analytically tractable model (i.e., mathematically precise, simple, and straightforward to implement). The LPC model works well in recognition, coding, transmission, modification applications. Figure 2 is show that LPC model[5].

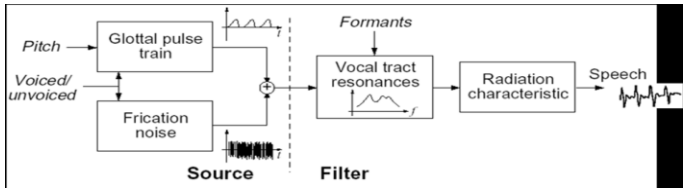

Fig.2. LPC model

The gain of the first formant(F_1) is generally higher 10dB than that of the remain formants, the resonance of the vocal tract can be approximated by envelope of only F1. Therefore, Peak of first positive is more distinguished then other peak in a pitch interval. this peak is consider the glottal peak that effect of glottal is large appear in pitch[9] period interval. In speech signal, the auto-correlation of shot time sample and its close one. we can predict that method of lest mean square is called by linear predict coefficient, and that mechanism is Liner-Prediction-Code(LPC) method. In LPC method speech sound model is can represent by all pole model which LPC analysis with AR-processing. The poles of transfer function are same frequency of formant frequency of voice speech. In this , we studied about basic concept of modeling of speech signals and its representation.

2.3. Noise Signal

To develop speech coder[6,7] that produce good quality, highly intelligible speech at bit rates below 16 kbits/s in a quiet environment, it has been necessary to incorporate more knowledge about the speech production model into the coder itself. Thus, the assumption is that, at the speech coder input, only clean speech and only the speech that one desires to be transmitted is present. one approach to reducing background-noise effects has been to utilize an adaptive filter at the speech coder input, and other approach might be to use multiple microphones and noise cancellation. For the removal of additive white noise, the standard approaches have been spectral subtraction using Wiener filtering or Kalman filtering[3,6].

Since the jointly optimal (here, minimum mean square error) estimation of parameters and filtering of the noisy signal is nonlinear, the joint filtering and parameter estimation problem is typically separated into the cascaded problem of parameter estimation on the noisy input followed by linear filtering using estimated parameters obtained in the first stage.

We now evaluate the performance of the proposed algorithms for speech enhancement along and for coding of noisy speech when the additive noise is white. The objective distortion measure used is the signal-to-noise ratio(SNR) defined by

$$SNR \overset{\Delta}{=} 10\log_{10}\frac{\frac{1}{L}\sum_{n=1}^{L}x^2(n)}{\frac{1}{L}\sum_{n=1}^{L}[x(n)-\hat{x}(n)]^2}\,dB \qquad (1)$$

3. SNR Analysis and Estimation

3.1. Estimation in Speech Signal

We propose new method of SNR estimation of speech sound with noise condition. Such as received sound which is recorded in calm situation or additional noise. The continuous speech has no silence section that only consist of voiced and unvoiced sound. That reason we cannot apply to ordinary voice activity detection(VAD) why is that VAD[4,12] need silence term in speech so that it cannot estimate the noise. But proposed method does not need VAD and it can estimate SNR directly with corrupted data. In this paper, the new SNR estimator classifies speech signal by stable voice section, and unvoiced section for calculate that. And we apply a different method for each section.

The first, voice section, we test the correlation of adjoin waveform which distinguished by pitch period. The second, unvoiced region, is using the spectrum-distance-measure method from linear predictive coding parameter to receive formant. The last estimate the SNR of whole speech signal by comparing the energy ratio of voice and unvoiced resign. The figure 3 is simple block diagram which is proposed method that estimates SNR. In the figure 3, the 'speech enhancer' is a preprocessor which does low pass filtering for reduce a error of pitch period corrupt by high frequency parameter of signal and tune the phase for emphasis pitch period. And 'V/UV discriminator' is dividing data to voice and unvoiced section for applying different method to get estimate SNR. In figure, NLF is noise level factor.

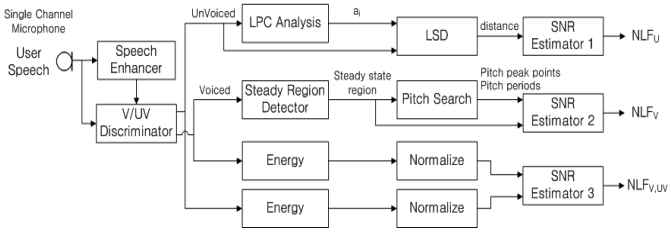

Fig.3. SNR Estimation System

3.2. Estimation In Voiced Sound

In general, in the enhancement of signal degraded by an additive noise, it is significantly easier to estimate the spectral amplitude associated with the original signal than it is to estimate both and phase. In our problem, the disturbing noise is uncorrelated with speech signal. Speech and noise are

modeled as stationary stochastic processes. We can divide the voice region into stable or unstable region. And we use the stable region of the voice speech. Because in this part, signal has not much changeable about a pitch and formant frequency why we make an effort short term speech of raising an accuracy. In stable voice region, we are using a waveform similarity of a pitch period for estimate SNR. And that is important about correct point of a pitch period and periodicity.

In figure 3, V/UV[5] discriminator use a pure received signal, because of exact time processing and that is important of exact pitch period. So that reason needs to normalize speech section. The received signal present by equation (1) that is speech signal with noise as flows,

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

In equation 2, r(n) is received data, s(n) is speech sequence and n(n) is additive noise. Fig. 4(a) represent speech signal and its zoomed data include the pitch period in shot time voiced frame. The design of pitch tracking system for noisy speech is a challenging and yet unsolved issue due to the association of "traditional" pitch determination problems with those of noise processing. It has been demonstrated that prosody can provide the principle cue for resolving some syntactic ambiguities are being developed to include prosodic information into various continuous speech recognition system.

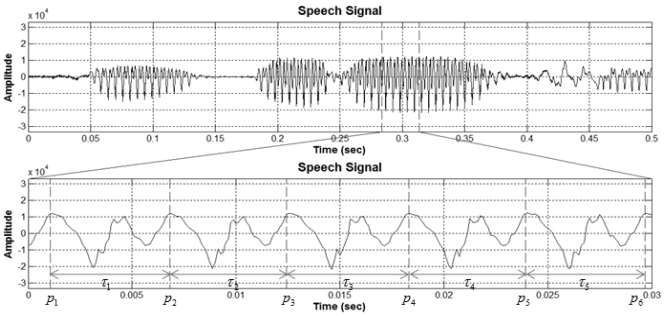

Fig.4. Voiced sound in speech signal

In figure 4, p_i is the start point of pitch and τ_i is sub-frame indicator. Figure 4(b) means one voice frame includes 5 sub-frames and that sub-frames are used for calculate correlation. After the sorting, we can get the coefficient C which represents correlation of signal itself in the frame. The process of getting C represent by equation (5) that is consist of auto-correlation R(t, t+k) which equation (3), and maximum energy V(t, t+k) of that frame.

$$R(\tau_k, \tau_{k+1}) = \sum_{m=0}^{\min(\tau_k, \tau_{k+1})} r(m+p_k) r(m+p_{k+1}) \tag{3}$$

Tow means a pitch period and k is an index of sub-frame.

$$V(\tau_k, \tau_{k+1}) = MAX\left(\sum_{m=0}^{\min(\tau_k, \tau_{k+1})} r^2(m+p_k), \sum_{m=0}^{\min(\tau_k, \tau_{k+1})} r^2(m+p_{k+1}) \right) \tag{4}$$

$$C = \sum_{k=1}^{K-1} \frac{R(\tau_k, \tau_{k+1})}{V(\tau_k, \tau_{k+1})} \tag{5}$$

The C is a sequence of estimated noise parameter. The maximum value of The C is 1, when the signal fame is same from close frame. And the C less than 1, that signal is noise mixed. So we can estimate the SNR by parameter C. Figure 5 is the plot of Estimated SNR and SSNR for compare. SSNR is segmented SNR which get form originally signal to noise ratios in frame.

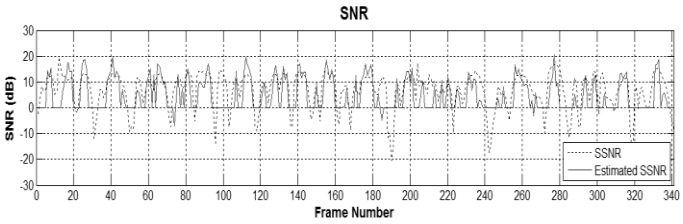

Fig.5. Estimate SNR and SSNR by 10dB Noised

3.3. Estimation in Unvoiced Sound

The signal with additive noise is represented by equation (6). And also can transform into Fourier formulation such as (7).

$$r(n) = e(n) * h(n) + n(n) \tag{6}$$

$$R(\omega) = E(\omega) H(\omega) + N(\omega) \tag{7}$$

The cause of excitation the unvoiced signal is white noise and that suppose to random process N_1. Additive noise also suppose random process N_2. After the assuming N_1 and N_2 we can conclude that equation (8) which is using β that is energy ratio.

$$\log R(\omega) = \log N_1 + \log\left(H(\omega) + \beta\right) \tag{8}$$

In equation (11), the received signal is changing by β. So the spectrum distance which $H(\omega)$ between $R(\omega)$ is influenced and that distance is noise parameter in unvoiced section. We can get spectrum of $H(\omega)$ that using the LPC method. In this paper, using a modified log-spectral distance method for calculate the distance between $H(\omega)$ and $R(\omega)$. The equation (9) show the modified-LSD method[13]

$$D_{\text{mod}} = \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} \left[10\log\left|\hat{H}(\omega)\right| - 10\log\left|R(\omega)\right| \right] d\omega} \tag{9}$$

The figure 6 is estimate SNR plot of unvoiced region. In the figure the estimate SNR flows the SSNR in unvoiced region

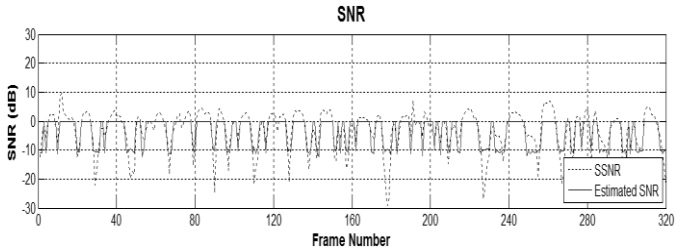

Fig.6. Estimate SNR and SSNR by -10dB noised

3.4. Estimation for Speech by Energy

In ordinary speech signal, the voice section has most of energy. And a noise and an unvoiced section has small amount of energy compare with the voice section. A noise additive all the speech signal but effect of that is different form original signal power. In this paper propose new method calculate the estimate SNR. The method use the energy each part of voice and unvoiced section. The equation (10) is the calculation of that method.

$$NLF_{V,UV} = 10\log_{10}\left(\frac{\frac{1}{N}\sum_{i\in voice}^{M}\sum_{n=1}^{F} r_i^2(n)}{\frac{1}{N-M}\sum_{j\in unvoice}^{N-M}\sum_{n=1}^{F} r_j^2(n)}\right) \qquad (10)$$

The estimator of SNR needs which frame or segment is voice and unvoiced. And in the equation, normalize the estimated SNR by number of frame.

4. Experimental Result

We test the proposed SNR estimator. White Gaussian noise was added to each sentence with an average signal to noise ratio. A noise generator was used for each of the speech files. Consequently, a different white Gaussian noise was added. The reference pitch contour was estimated manually from clean speech. And the continuous speech are recorded by 5 men and 5 women.

For make an accuracy result, remove long term silent section. And whole data sampling at the 8 kHz and. 16 bit. Experiment frame length is 32 msec. at that time 1 frame consist 256 samples. Figure 7 is additive White Gaussian noise by eq.(1). And figure 8-10 is result of estimate SNR plot that change SNR. Horizontal axis means SNR that is amount noise energy, and vertical axis means result at that time.

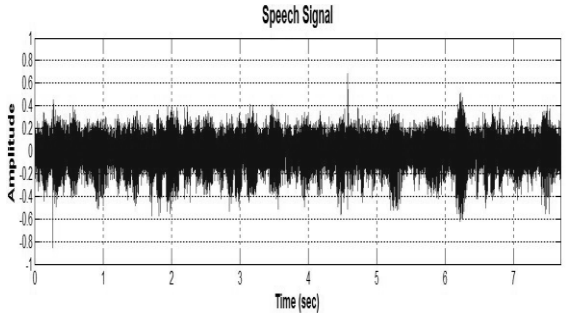

Fig.7. Additive White Gaussian noise

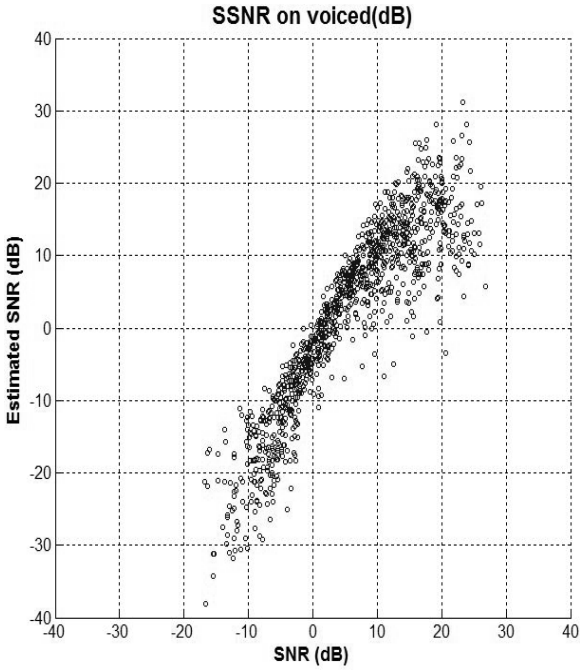

Fig.8. SNR of voiced by NLF in white noises

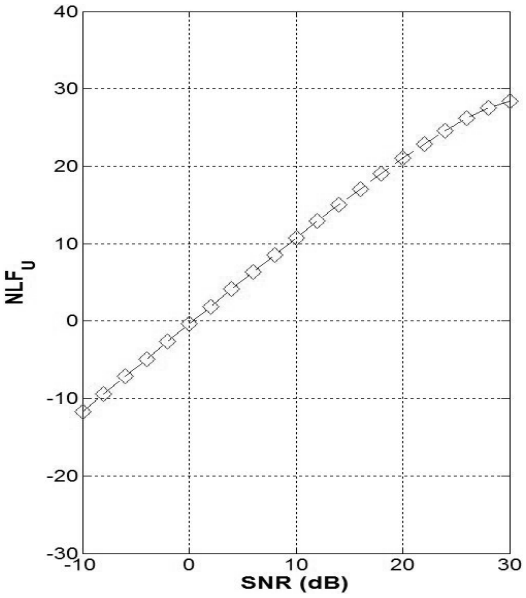

Fig.9. SNR of unvoiced by NLF in white noises

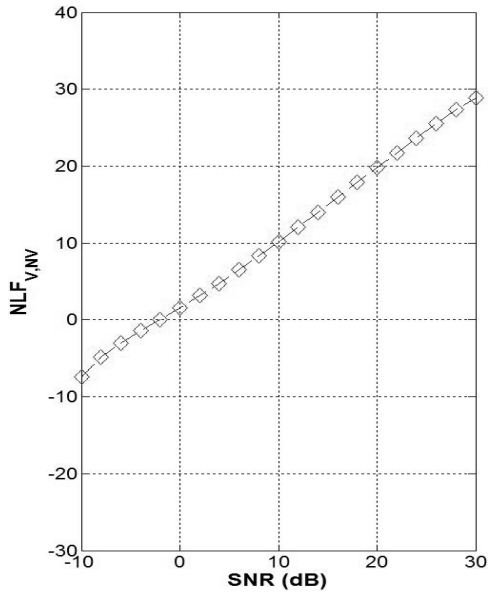

Fig.10. SNR of speech by NLF in white noises

For stationary region of voiced speech signal, waveform is very correlated by pitch period since voiced speech is quasi-periodic signal. So we can estimate the SNR by correlation of near waveform after dividing a frame for each pitch. For unvoiced speech signal, vocal track characteristics reflected by noise, so we can estimate SNR by using spectral distance between spectrum of received signal and estimated vocal track. Lastly, energy of speech signal is mostly distributed on voiced region, so we can estimate SNR by the ratio of voiced region energy to unvoiced.

5. Conclusions

In speech signal processing, it is very important to detect the pitch exactly in speech. If we exactly pitch detect in speech signal, In the analysis, we can use the pitch to obtain properly the vocal tract parameter without the influences of vocal cord. It can be used to easily change or to maintain the naturalness and intelligibility of quality in speech synthesis and to eliminate the personality for speaker-independence in speech recognition. We have proposed in this paper a synthesis of some efficient methods we have developed for enhancement speech in additive white Gaussian noise. however, was that the optimization of the parameters was a very difficult and tedious task when altering the noise and speech condition.

There certainly remains considerable future work to be done towards a more significant improvement in mobile communication which remains a complex environment, mainly in non-stationary conditions and low SNR. It can be applied to rate decision of vocoder and used for pre-processing to decide threshold of noise reduction.

Acknowledgements

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2010-(C1090-1021-0010)).

References

[1] J. Sohn, N. S. Kim, and W. Sung, A statistical model-based voice activity detector, IEEE Signal Processing Lett., 6,1(1999).

[2] Y. D. Cho and A. Kondoz, Analysis and improvement of a statistical model-based voice activity detector, IEEE Signal Processing Lett., 8, 10(2001).

[3] Ing Yann Soon, Soo Ngee Koh, Chai Kiat Yeo, Noisy speech enhancement using discrete cosine transform, www.elsevier.nl, Speech communication 24(1998).

[4] Jerry D. Gibson, Speech coding in mobile radio communication, processing of the IEEE, 86, 7(1998).

[5] A. J. Accardi and R.V.Cox, modular approach to speech enhancement with an application to speech coding, J. Acout. Soc. Am, 10, 3(2001).

[6] T. Agarwal and P. Kabal, Pre-processing of noisy speech for voice coders, in Proc. IEEE Workshop on Speech Coding(2002).

[7] I. Cohen, Relaxed statistical model for speech enhancement and a priori SNR estimation, IEEE Trans. Speech Audio Processing, 13, 5(2005).

[8] M. Kleinschmidt, J. Tchorz, and B. Kollmeier, Combining speech enhancement and auditory feature extraction for robust speech recognition, Speech Commun., 34, 1-2(2001).

[9] Y. L. Cho, J. K. Kim, and M. J. Bae, A study on Improvement upon Mixed Voices Pitch-Detection System to Frequency, ASK, Proceedings of Autumn Season, 23,2(s)(2004).

[10] A. Nogueiras. etc, Speech emotion recognition using Hidden Markov Models, Proc. of Eurospeech 2001, 4(2001).

[11] Hoffmann, H, Kernel PCA for novelty detection, Pattern recognition, 40(3)(2007).

[12] Ioannou S, Caridakis G, Karpouzis K, Kollias S, Robust feature detection for facial expression recognition, EURASIP J Image Video Process, 16(2007).

[13]Naden, C.,Macho, D, & Hermando. L, Frequency and time filtering of filter-bank energies for robust HMM speech recognition, Speech Communication, 34(2001).