Copyright ⓒ 2019 KSII

Signal Enhancement of a Variable Rate Vocoder with a Hybrid domain SNR Estimator

Hyung Woo Park
Information and Telecommunication Engineering Department, Soongsil university,
369 Sangdo-Ro, Dongjak-Ku, Seoul, Republic of Korea (06978)
[e-mail: pphw@ssu.ac.kr]

Received July 11, 2018; revised September 5, 2018; accepted September 11, 2018; published February 28, 2019

Abstract

The human voice is a convenient method of information transfer between different objects such as between men, men and machine, between machines. The development of information and communication technology, the voice has been able to transfer farther than before. The way to communicate, it is to convert the voice to another form, transmit it, and then reconvert it back to sound. In such a communication process, a vocoder is a method of converting and re-converting a voice and sound. The CELP (Code-Excited Linear Prediction) type vocoder, one of the voice codecs, is adapted as a standard codec since it provides high quality sound even though its transmission speed is relatively low. The EVRC (Enhanced Variable Rate CODEC) and QCELP (Qualcomm Code-Excited Linear Prediction), variable bit rate vocoders, are used for mobile phones in 3G environment. For the real-time implementation of a vocoder, the reduction of sound quality is a typical problem. To improve the sound quality, that is important to know the size and shape of noise. In the existing sound quality improvement method, the voice activated is detected or used, or statistical methods are used by the large mount of data. However, there is a disadvantage in that no noise can be detected, when there is a continuous signal or when a change in noise is large.This paper focused on finding a better way to decrease the reduction of sound quality in lower bit transmission environments. Based on simulation results, this study proposed a preprocessor application that estimates the SNR (Signal to Noise Ratio) using the spectral SNR estimation method. The SNR estimation method adopted the IMBE (Improved Multi-Band Excitation) instead of using the SNR, which is a continuous speech signal. Finally, this application improves the quality of the vocoder by enhancing sound quality adaptively.

Keywords: CELP, QCELP, Voice codec enhancement, SNR estimation

http://doi.org/10.3837/tiis.2019.02.026 **ISSN : 1976-7277**

1. Introduction

People use sound as one of the basic communication tools in daily life. The human voice is in fact sound radiated to the air through respiratory and vocal organs. Unless filtered through a certain media, this sound cannot be recorded or stored. Differing from letters or pictures, voice cannot be delivered or stored through solid media, however, it is delivered or recorded through media or environments with changing features such as air, communication paths, or recording devices. Although a flexible environment has the advantage of easy utilization it also has the disadvantage of being highly affected by the surrounding environment. This is the case when communicating through voice, and the quantity and quality of information delivered by voice communication varies highly depending on surrounding noise [1-7][23-27].

Today, with the development of IT(Information Technology) in general, the field of sound signal processing and its related technology is also rapidly advancing. Indeed, much research has been carried out in different fields with the aim of enabling smooth conversation between humans and computers, or better communication between people using communication devices [8-12][37-40]. Research into delivering the intentions of a person to a computer through sound analysis and a speech recognition system has shown outstanding potential in quiet places such as a laboratory [13-18]. However, results are less satisfactory in situations where more noises are made. Furthermore, in terms of the process of delivering sound signals a great distance using wire/wireless devices, research on moving a lot of data through a small communication bandwidth like the internet is being carried out. Whether going through a communication line or not, many different methods of improving the quality of sound delivered are being studied and suggested. In contrast, studies on noise and errors occurring in communication lines are carried out in a different way [19-26].

In general, when improving sound signals damaged by noise, it is assumed that noise features are additional factors with no correlation with sound signals. Nevertheless, it is more important to understand the size and type of noises added to improve the overall sound quality. This is especially the case with spectrum subtraction, that subtracts the amount of noise from the spectrum, which makes finding the amount and type of noise critical. Yet, if noise is added to a sound signal, there are in actuality many limitations to estimating the last point of sound period or measuring or assuming the size or type of the added noise [27-33].

To enhance speech quality, Boll(1979) proposed Spectral-Subtraction which uses the voice active detection method [25]. After Boll(1979), other developed speech enhancement mechanisms have been suggested. For instance, MMSE(Minimum Mean Square Estimator) estimated environment noise conditions through statistical analysis [26][27]; MS(Minimum Statistics) minimized noise errors using statistics [11][12]; and HMM (Hidden Markov Model) evaluated noise and voice [13].

In mobile communication systems, or internet voice communication systems , the methods used to effectively change and compress analogue signals into digital signals play the most important role in determining capacity and quality in the communication system. Similarily the function of voice coding technology in a telecommunication system affects the quality of recovered signals and the capacity of the system. The most common method of improving voice coding technology is to enhance the quality of the input signal by using a processor of the coding device. How to improve sound quality has been explained above. Further, one of the most general methods used for improving sound quality is to detect voice period, estimate the noise, and apply spectral subtraction. Among the large amount of exisiting research on sound coding devices, one of the most common reoccuring methods mentioned on VolP or wireless communication is the CELP method. This CELP method enables high quality sound

within a transmission rate of 4.8kbps to be obtained. Furthermore, the standardization of this method in various applications has been completed by many international standards organizations such as ITU-T and TIA/EIA [30-31][36-40].

In this research, we added pre-processing to the ordinary vocoder to improve its quality under an environment with low transmission rates and noisy conditions. In more detail, a sound sample mixed with noise is divided into a sound period with a short term and its SNR is estimated. Then, this SSNR is combined to calculate the SNR of total sound signals, then the spectral subtraction method is applied. This suggested method does not apply VAD, one of the methods used before, and directly utilizes voiced and unvoiced sound information in continuous sound section. In other words, the conventional method can not improve the sound quality by estimating the SNR. However, the method suggested has the advantage of being able to enhance the sound quality, even if the amount and shape of the noise changes between the sound quality enhancement and the voice signal through the SNR estimation [28][29].

This research describes the technology for detecting LPC and pitch for sound generation models, and extracting noise factors in Chapter 2. In Chapter 3, it explains SNR estimation technology. In Chapter 4, it discusses sound coding devices and the suggested pre-processing device of sound coding technology. Finally, it explains the tests and its result, and makes final conclusions in Chapter 6.

2. Speech Generation Model and Voice Analysis

2.1 Ordinary speech generation model

Fant(1958) designed a linear model of a voice speech system, which divides voices emitted into the air by excitation source and vocal track. Excitation source and vocal track are mutually independent. This method uses a quasi-period pulse for voiced sounds, a white-Gaussian-noise for unvoiced sounds, and a glottal is a vocal cord model [20][30]. **Fig. 1** is a block diagram of the 'speech generation model' as explained. Excitation source uses impulse train and white sound to present the Voiced/Unvoiced signal by switching mode during the time slot. The pitch frequency has the most crucial role of the excitation source in the voiced section. The vocal track can be modeled according to a resonance tube model known as formant frequency. Formant is the resonant frequency of the vocal track. The resonant frequency varies according to the movement of the tongue and by the movement of the tube in the throat, depending on what is being pronounced. Inversely, if we know the resonant frequency of the voice waveforms coming out of the lips, we will be able to figure out what was pronounced [21].

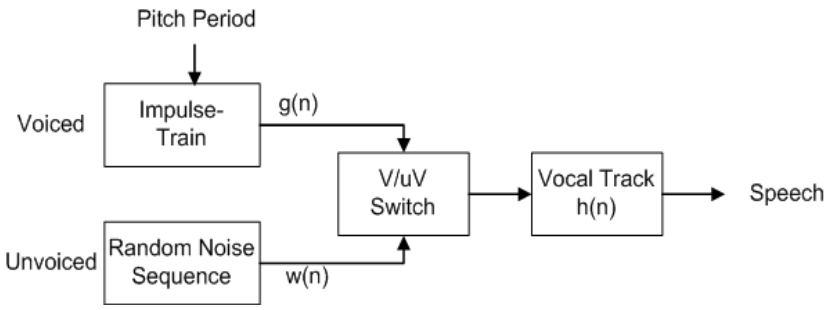

Fig. 1. Speech generation model

2.2 Pitch

Voiced sound happens when air is forced through openings between the glottis, ie, vocal cords [22]. Air from the lungs comes up, and is forced against the closed vocal cord. Then, the vocal cords are opened and closed quasi-periodically by tension in the vocal cords', and air is moved to the vocal track. This quasi-periodical vibration of the vocal cord makes air tremors to excite the vocal track, then voiced. This period, the time from the opening of the vocal cords to their next opening time, is called the fundamental period T0, and the vibrating speed of the vocal cords is called the fundamental Frequency F0. Although the term pitch is often used as having the same meaning as fundamental frequency, there are subtle differences between the two, but in general, pitch is often used to mean the fundamental frequency [23].

2.3 Pitch Detection methods

The wide range of pitch detection methods can be divided into frequency-domain methods and time-domain methods [4] [24]. A time-domain pitch detection method is shown in **Fig. 2**.

Time domain methods estimate the period of a signal, and then invert that value to obtain the frequency. Some of the more complicated methods include the Autocorrelation method (ACM) and the Average-Magnitude-Difference-Function (AMDF), which are parallel processing methods. These methods only involve simple calculation such as addition, subtraction or simple comparison. However, the detection of pitch is complex and error rates increase in transition sections or noise corrupted signals.

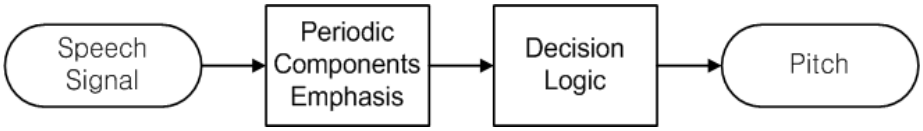

Fig. 2. Pitch Detection Method in Time-Domain

Pitch Detection of the spectrum-domain is shown in **Fig. 3**. A speech signal is given, and then a time domain signal is converted into the frequency domain. Representative pitch detection methods of frequency domain include the harmoinc production spectrum, the frequency lifter method and comb-filtering methods. In general, a spectrum frame has been obtained by the units 20 ~ 40ms, so there is less error in the transition section or from additive background noise. However, transformation into the frequency domain is complex and if the resolution for finding fundamental frequency is increased, processing time becomes longer and becomes dull to the changed characteristics [5].

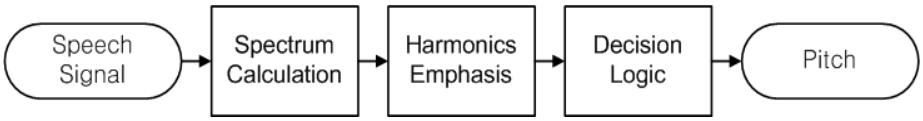

Fig. 3. Pitch Detection Method in Frequency-Domain

2.4 Formant

Formant refers to the frequency peaks of the sound spectrum [22]. Formants are acoustic energy, which is concentrated around a specific frequency of a speech wave. The formant

frequency is determined by the natural frequency of the vocal track. This frequency assumes a simplified model of the vocal track. The formant frequency is measured by searching the emphasized band in the frequency domain [20-23]. Formant is usually assessed "as sound, generated as discussed above, propagates down these tubes, the frequency spectrum is shaped by the frequency selectivity of the tube. This effect is very similar to the resonance effects observed with organ pipes or wind instruments. In the context of speech production, the resonance frequencies of the vocal tract tube" [22: p.88].

3. SNR Estimation of Sound Signal Using Hybrid Domain

3.1 SNR Estimation method on Hybrid domain

This section suggests an algorithm for estimating SNR by analyzing sound signals received from a noise environment in a hybrid domain. In terms of a continuous sound signal, with no silence, it is impossible to calculate SNR by using the output of the existing sound period detector. However, based on the suggested method, one can estimate the amount and type of the noise in a continuous speech signal. This is possible because the suggested method does not divide the received sound into silence or noise sections but computes signal to noise ratio based on the rate of pitch changes, estimations in frequency area, and log-spectrum-distance between the received signals. Thus, methods of detecting sound period such as VAD are unnecessary, and this also allows the assumption of SNR in direct signals [34-35].

The general SNR estimation method utilizes the VAD method to delineate between silent and voice sections. Then, it calculates the amount of energy in the silent and voice sections to assume SNR. With the existing method, it is difficult to know the amount and type of noise in sectors where sound is being created continuously or environmental noise changes in the sound period. However, as this method is based on continuous sound signals and parameters in segments of the total signals, including the silent sections, it has the advantage of being able to be used even in cases where environmental noise changes or silent sound sections cannot be detected.

The theory of estimating SNR is to use features of sound signals and noise signals. To begin with, for voiced sound, air generated from the lung goes through the vocal track in a quasi-periodic manner, depending on the opening and closing of the glottis. At this time, pitch in voiced sound is detected. By calculating its frequency, the fundamental frequency of the vocal cords can be calculated. Although sound signals tend to change depending on time, pitch does not vary greatly in this voiced sound period. That is, voiced sound has a high level of correlation with the change in time. In contrast, one of the features of noise is that it has a low level of correlation with the neighboring sectors. By using such features, one can calculate the amount of noise mixed by using changes in pitch in the voiced sound period. If white noise is added to the sound signal, the period section represents one that has noise signal added to the total signal. One can easily calculate the noise by comparing pitch values in the analyzing section. Equation (1) below is for calculating the parameters for estimating the amount of noise in the voiced sound section.

$$ESNR = 10\log_{10}\left(\frac{C}{1-C}\right)^2 \tag{1}$$

Here, C is the value for analyzing corelation in waveform of input signal and this correlation equals parameters normalized in sub-frames. This is calculated by equation(2).

$$C = \sum_{K=1}^{K-1} \frac{R(\tau_k, \tau_{K+1})}{V(\tau_k, \tau_{k+1})} \tag{2}$$

R(,) and V(,) show cross-correlation values and the maximum correlation values of the analyzing sections respectively. That is, this equation calculates changes in standardized pitch by frame and estimates SNR by using calculated values.

Secondly, as explained in the speech generation model, unvoiced sound is excited by random white signals. The speech signals for the unvoiced section can be calculated based on white noise and convolution of the vocal track. Here, noise added with sound signals that equals unvoiced sound has low cross-correlation, however, this noise has similar features with that of voiced sound. By comparing revised log-spectrum distance as shown below, and feature difference between received signals with the noise and estimated LPC parameters, one can compute the gap created from the noise.

$$D_{\mathrm{mod}LS} = \frac{1}{2}\int_{-\pi}^{\pi} \left[\left| 10\log_{10} \hat{H}(\omega) \right| - \left| 10\log_{10} R(\omega) \right| \right] d\omega \tag{3}$$

Here, $\hat{H}(\omega)$ means the frequency response of the received signal and to calculate the frequency response, LPC analysis of the received signal is applied. To understand the comparative difference of $R(\omega)$ in consideration to $\hat{H}(\omega)$, this research omitted the square in LSD(Log spectrum distance) calculation. Also, the absolute value is applied to consider distance. This can be made into an equation as shown below and this modified-LSD is used to calculate the amount of noise added to the received sound signal.

The following diagram shows an algorithm for estimating SNR of a continuous sound signal in a flow chart. At the start, when the sound signal enters, it is segmented into a short time unit that is analyzable and then voiced/unvoiced sounds of each segment frame are determined, which is done by the energy. Then, segments other than voiced sound are categorized as non-voiced segments. Estimation of voiced sound is done by pitch and for non-voiced sound, its SSNR(Segmental Signal to Noise Ratio) is calculated by using LPC and difference in frequency distance of the received signal.

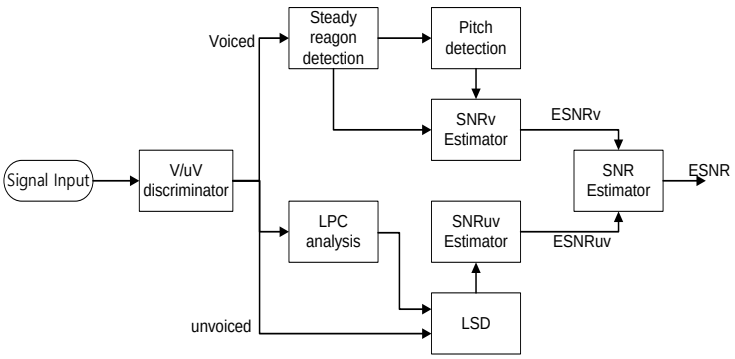

Fig. 4. Block diagram of SNR Estimation Algorithm

3.2 Performance Verification of Hybrid domains SNR Estimator

To verify the function of SNR estimation device, as shown in **Fig. 5**, a sound sample recorded in a studio is accurately measured and a sound signal with added white noise of 10db SNR is applied. The waveform of **Fig. 6** (the upper picture) and the dotted line in the lower picture in **Fig. 6** show SSNR by frame, each calculated by using input signal and added noise. The dotted line show estimated ESSNRv(Estimated segmatal SNR on Voiced) by frame.

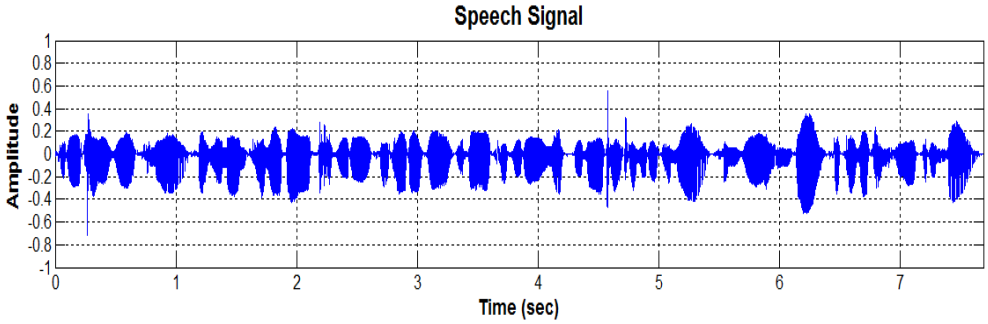

Fig. 5. Clean speech recording signal waveform

As the sound signal has different energy by frame, the correlation coefficient changes depending on the frame. The diagram below shows SSNR curve and $ESSNR_v$ curve by frame. One can assume estimated SNR from the average SNR value of the total sectors.

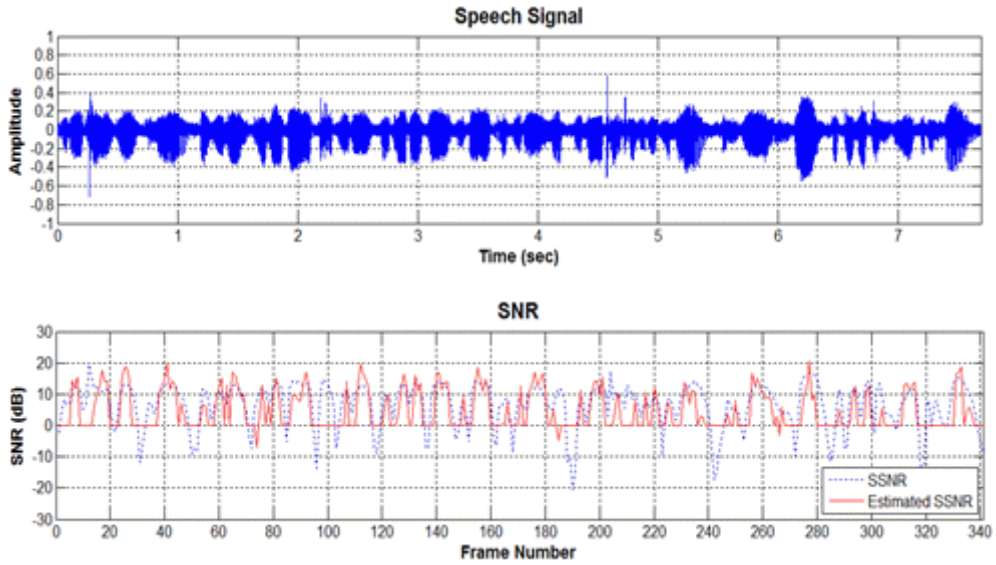

Fig. 6. SSNR and Estimated SSNR plot of 10dB white noise added signal

Value of modified LSD is checked simultaneously to confirm the function of SNR estimating device in the unvoiced sound sector. Similar to the voiced sound, the experiment is done by adding noise to the signal in **Fig. 7** so that exactly 0dB SNR is made.

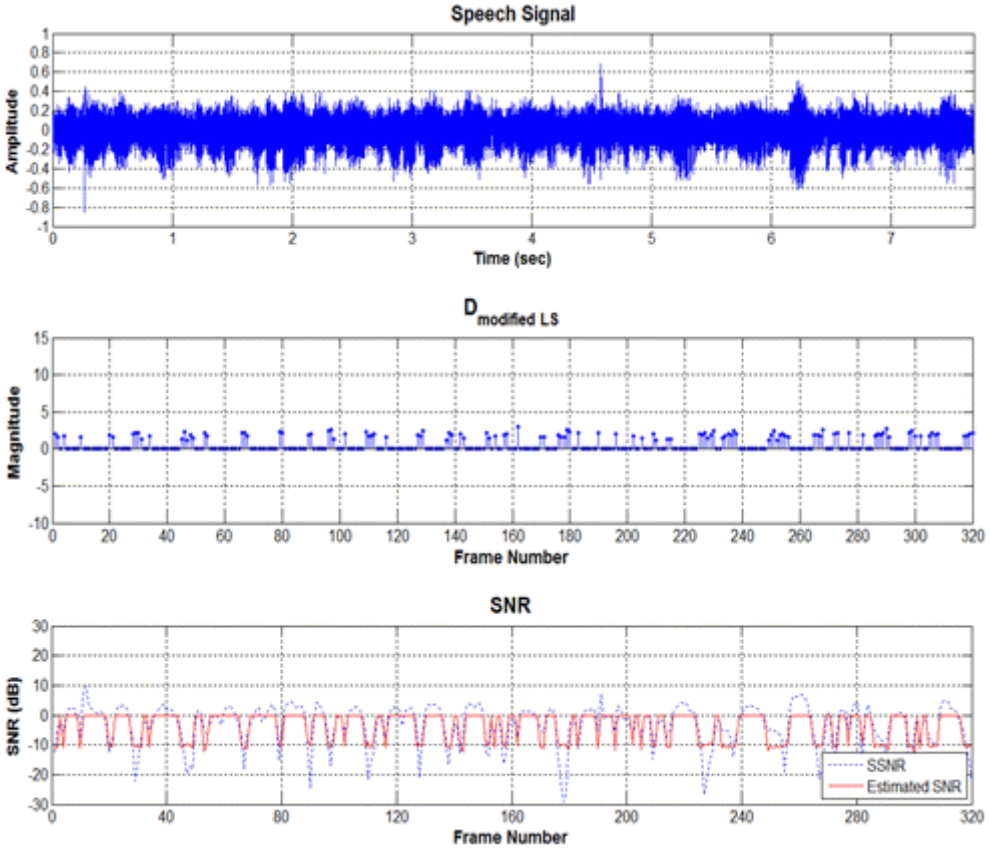

Fig. 7. SSNR and Estimated SSNR plot of 0dB white noise added signal with modified D parameter

The upper part of **Fig. 7** is signal waveform with added white noise at 0dB SNR, and the middle diagram shows D_{modLS} in the unvoiced sector. The lower diagram shows SSNR (dotted line) changes in $ESSNR_{uv}$(Estimated segmatal SNR on UnVoiced, red line), and SNR estimated in unvoiced sound section. Looking at the changes in D_{modLS} in the lower diagram, it was found that noise added to the original sound signal is affecting the estimation of frequency spectrum of the vocal track parameter.

4. Vocoder and proposed preprocessor of vocoder

4.1 Vocoder

In speech waveform coding technology, there is a method for transmitting information about the waveform itself and a method for compressing and transmitting, by using the characteristics inherent in the speech signal. In particular, the vocoder represents the second mechnism, where the transmit end models the characteristics of a voice signal and sends the various parameters, and restores(synthesizes) the voice signal using the received parameters at the receiving end. Vocoder is an acronym for voice coding and decoding. The vocoder is also a category of a voice codec for analyzing and synthesizing human voice signals for compression, multiplexing of voice and sound signals, voice encryption, voice conversion, and voice synthesis. The incipient vocoder type, the channel coder, developed for 1930s

communications systems, was proposed to reduce the bandwidth for multiple transmissions and receptions [35-40].

In a vocoder, a method for reducing band capacity is to decrease the redundancy occurring in the transmission process. If the encoder and decoder of the vocoder use mutually agreed parameters, it is possible to synthesize the undelivered sound in the process of reconstructing at the decoder. In addition, the redundancy of the speech signal represented by the pitch, formant, and perceptual parameters in a short interval, means that the transmission rate can be saved by using that data. Generally, a vocoder operating at a data rate of 4 ~ 10 Kbps has a somewhat lower sound quality than a 16 ~ 64 Kbps waveform coding, but is superior to waveform coding at a bandwidth lower than 10 Kbps. On the other hand, the vocoder using a speech generation model of the human voice , music, environmental sounds, and wide-band sound signals can not be accurately reproduced, and the signal processing delay and power consumption increase [35-40].

The standard speech coder starts from the 64kbps PCM (Pulse Code Modulation) method adopted in ITU-T recommendation G.711 in 1972 and standardized to 32kbps ADPCM (Adaptive DPCM) and 16kbps LD-CELP (Low-Delay CELP). In ITU-T, the standardization of the 8kbps voice encoder with lower transmission rate that can be used in the wireless environment was carried out in 1996 by using CS-CELP (conjugated structure algebraic CELP) as G.729 and as ACELP / MP-MLQ (algebraic CELP / Multipulse Maximum Likelihood Quantization) 5.3 / 6.3kbps dual rate as G.723.1 draft recommendation. The commercial and standardized vocoder types include 4.8 Kbps /2.4 Kbps Improved Multi Band Excitation coder, 4.8 ~ 8 Kbps Department of Defense Code Excited Linear Prediction coder, 8 Kbps Vector Sum Excited Linear Prediction coder, 8 Kbps Qualcomm CELP coder and 16 Kbps Low Delay CELP coder [35-40]. Low transmission rate vocoders compress and transmit voice model parameters without transmitting human voice as it is, which makes for efficient use of channel capacity [35-40].

4.2 Proposed preprocessor of vocoder

Speech signal refers to “time-variant-signal, which includes pitch and formant” [4:p.187]. However, the signal involves quasi-periodical components such as formant, pitch, and group energy enveloped in short period analysis of voice. In noise corrupted voice signals, it is difficult to decide the start and end points of the speech, and to detect the voice section and the noise section [5]. In particular, when the amount of noise is large and the special colored noise is mixed, the SNR of the voice signal is low as is well known. However, it is difficult to estimate SNR’s value and it is more difficult to improve the sound quality.

However, if we know that the SNR is low and if the type of noise is predictable, we can use sound quality enhancement by spectral subtraction with a noisy signal in the frequency domain. Unfortunately, estimating the SNR and the type of noise is difficult when a signal is continuous and noisy, as shown in previous studies.

The proposed preprocessor of speech coder adds SNR estimation and adaptive noise to enhance the speech quality of existing variable rate vocoder. First, it estimates the SNR for the continuous speech input signal using the SNR estimation method of the hybrid domain described in Section 3.1 Then, based on this, the noise is removed adaptively in a short interval in continuous speech signal, and then used as the input of the speech codec for ordinalty. **Fig. 8** is a block diagram of a variable rate vocoder including the proposed preprocessor. The proposed block diagram is explained again. The SNR of the signal input from the preprocessor

is estimated and adaptive noise canceling is performed. Then, the process analyzes the LPC analysis and the transmission rate, which are general transmission processes of the vocoder, and re-extracts the pitch information necessary for the vocoder. Then, the received signal is compressed and transmitted using a codebook.

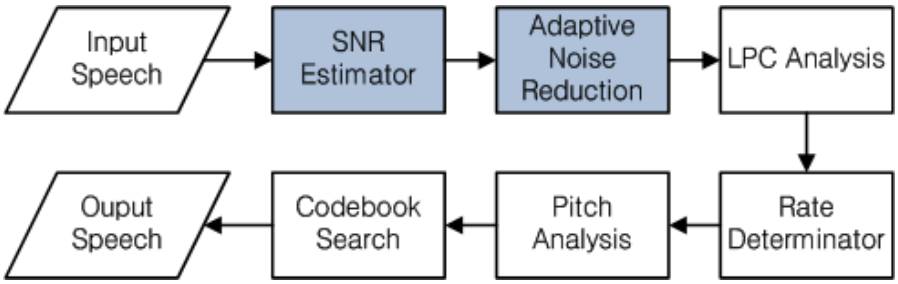

Fig. 8. Block Diagram for Proposed Sound Coding Device

In the proposed method, the ESSNR estimated by the hybrid domain SNR estimation method is inputted to the adaptive noise canceller, and the μ calculated from the ESSNR and equation (4) is used as a factor in the adaptive noise canceller. Here, λ is obtained experimentally as a constant representing the maximum improveable SNR. μ is used to obtain the mean noise magnitude and then the noise is subtracted from the original signal to improve the sound quality.

$$\mu = \frac{\lambda - ESSNR}{2} \tag{4}$$

5. Experiment and Result

5.1 Environment for Test and Data Collection

For functional evaluation of the suggested algorithm, objective tests were carried out in various noise environments including white noise. The experiment was undertaken to affirm the usability of noise cancellation in the changes of SNR. The sample for the experiment was the National Chart of Education, and the data was prepared after 8 KHz sampling, 16 bits per sample quantization. The sample was read repeatedly by five men and five women. The result values were obtained by adding the noise to the sampling and measuring the SNR and PESQ. The voices were recorded in an anechoic room.

Usually, a frame length of 20~40ms is used in the time section but this research applied 64ms in consideration of the FFT window size. The total number of samples by frame is 512. For window overlap, it shows a good function under 50% when estimating SNR of sound signal and this research applied 25% overlap to measure SNR by frame. Hamming Window 512 sample was used for dividing the frame. Lastly, white noise was added to the final signal recording to make it 0, 10, and 20dB SNR.

The test results were verified using objective function evaluation by using signal improvement and PESQ through direct comparison between SNR estimation and original sound signal. Perceptual Evaluation of Speech Quality is defined as “the result of several years

of development and is applicable not only to speech codecs but also to end-to-end measurements" [40:p.7].

5.2 SNR Estimator experiment result

Fig. 9 shows the test results of verifying functions of the SNR estimation device under a white noise environment. At this time, the input signal is calculated to have a 2dB interval SNR starting from -20dB to 20dB by adding the noise. It was found that the result changes linearly under the white noise environment depending on SNR changes. There was a change in gradient near the area where SNR value was -2dB and at the area with signal over 0dB, degree of estimation became slow. Overall, SNR estimation was higher than the input SNR. Average of estimated error is 5.84dB and standard deviation of the error value is 3.71.

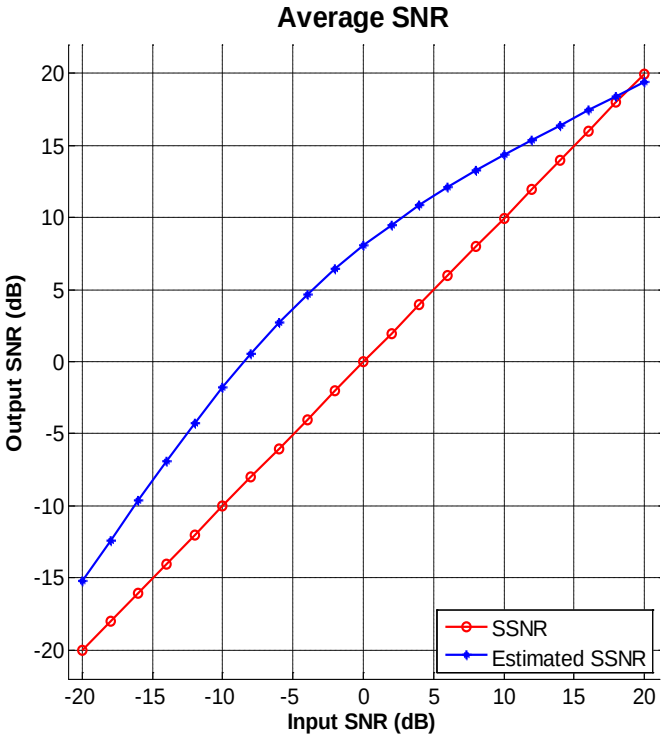

Fig. 9. Estimate SNR plot of white noise environment

Fig. 10 shows test results for verifying the function of the SNR estimating device under a noise environment inside a car. At this time, input signal is calculated to have 2dB interval SNR starting from - 20dB to 20dB by adding the noise. Estimated test results changed around the - 3dB area, and output SNR compared to input sustained a linear shape. Overall, similar to the white noise, it had a relatively lower SNR than input but its ratio varies around an SNR of 18dB and 20dB. What is more, it was found out that there was a change in estimated test results in SNR areas with under - 8dB. The average of estimated errors is 2.59dB and standard deviation of the error amount is 3.30.

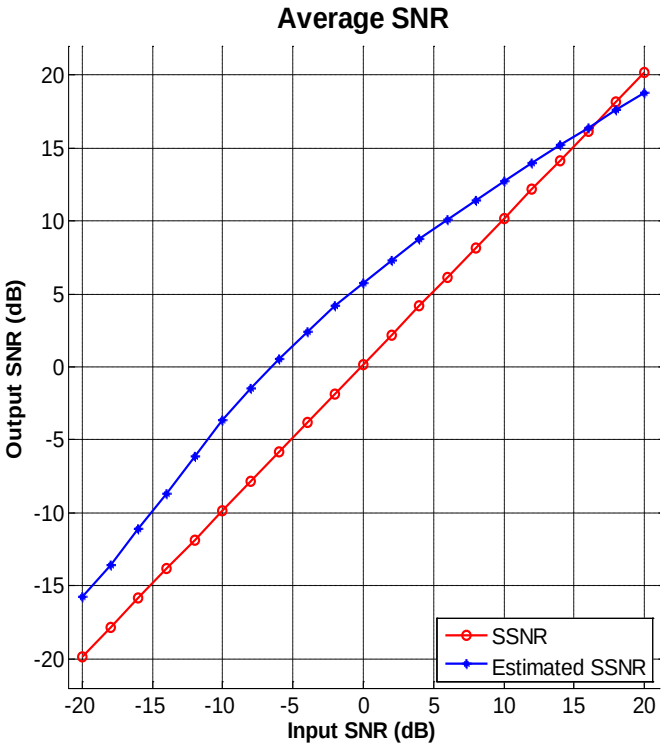

Fig. 10. Estimate SNR plot of car interior noise environment

5.3 Functional Test of Pre-processor of Vocoder

With SNR estimation value verified in 4.2, a test was carried out based on the spectral-subtraction method. The output value obtained from this process is the ANC output value. Afterwards, SNR values going through the vocoder and coding devices that went through the ANC are compared. **Table 1** shows SNR values based on ANC and improved SNR value. Looking at the average signal value with input signal of 0dB SNR, ANC output was enhanced by 5.94dB. Once this signal goes through the vocoder, there was an approximate –2dB decrease in sound quality of the vocoder and output of the vocoder from ACN output turned out to be 4dB SNR signal. It was found that enhancement of average 4dB SNR was made for 0dB SNR. Also, average amounts of improved signal value of 0, 10, 20dB were made by 2.75dB. In consideration of the fact that there was a decrease of an average of 6dB SNR in sound quality in G.723.1 transmission vocoder, it is estimated that an average of 8dB improvement in sound quality was made.

Table 1. SNR value comparing that the ANC use

input	ANC output	normal vocoder output SNR	ANC output vocoder SNR	SNR difference of Noise canceling
0	5.94	0.00	4.00	4.00
10	13.12	7.23	10.43	3.20
20	20.13	13.42	14.48	1.06

We confirmed improved value through SNR by checking the degree of enhancement by physical quantity of the signal, then PESQ analysis was undertaken to confirm any improvements in quality of sound heard. The highest score of PESQ is 5 points and no unit exists. For 0dB input sample, there was no big improvement found but for 20 and 30dB, there was an approximate average of 0.37 points improvement in PESQ. **Table 2** shows the average values of PESQ test results. This test was done to test PESQ of general vocoder output and ANC output. In terms of 0dB SNR signal, there was a 0.06 point increase in the PESQ value in consideration to a decrease in sound quality of the vocoder. However, for the 10, 20dB SNR signals, there was only an approximate increase of 0.3 points. The total average value was improved by 0.37. This equates to a listener being able to hear words that he/she could not hear before, and still being able to recognize the signal even if continuity of the sound has decreased by 30%.

Table 2. PESQ value comparing that the ANC use

Input (PESQ)	ANC output	normal vocoder output PESQ	ANC output vocoder PESQ	PESQ difference of Noise canceling
1.62	1.80	1.24	1.30	0.06
2.20	2.42	1.78	2.21	0.43
3.01	3.25	2.69	2.99	0.30

6. Conclusion

This research has examined estimation methods for the amount of noise in a continuous sound signal and tested its function by pre-processor of vocoder. The previous noise estimating algorithm can only calculate noise in a section where no sound signal exists to minimize the influence of the sound signal. However, for this method, it is necessary to use sound section extraction to find periods with noise only. If no silent periods exist in the sound signal then noise estimation in the silent period through sound extraction and SNR method cannot be applied. Further, if the size of the noise changes, it becomes more difficult to estimate the amount of the noise. For sound signals depleted by noise, as silence and noise have similar shapes, it is not easy to extract the silent sector. Nevertheless, it is necessary as well as important to estimate the volume of the noise to improve sound signals in various environments.

This research suggests a way to estimate SNR in the received sound signal period in cases where a sound signal is damaged by noise. Depending on the origin of the sound signal, it is divided into non-silence and silence sounds. In each section, this research tries to estimate SNR based on special features of sound and noise signals. When this SNR estimation device is applied, it provides the big advantage of not only calculating noise energy in the silent period through sound section extraction like VAD, but of also being able to estimate noise levels in continuous sound signal.

To check functionality of the suggested method, tests were carried out in various noise environments, including white noise. Depending on the amount of the noise, mostly, linear responses were made. However, there were some cases where SNR estimation varies depending on the type of the noise. This is because shapes shown in the silence sector, and the time and frequency scopes of the noise look similar, which results in having almost the same

shapes in the noise and silence sound. Moreover, in a functional evaluation through vocoder, sound improvement using noise canceling had an average 2.75dB increase in SNR and 0.37 in PESQ. In the future, research on the following issues shall be done; tests on a wider variety of samples, determination on the rate that changes variably and its influence on sound quality, and any improvements that could be made in SNR estimations and algorithms for improving sound quality.

References

[1] Hyung Woo Park, “Enhancement of the Variable Rate Vocoder with the Spectral SNR Estimate,” *KSII The 11th Asia Pacific International Conference on Information Science and Technology(APIC-IST)*, 2016.

[2] Li, Jingchao. "A New Robust Signal Recognition Approach Based on Holder Cloud Features under Varying SNR Environment." *KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS*, Vol.9, No.12, pp.4934-4949, 2015. Article (CrossRef Link)

[3] Li, Jingchao. "A Novel Recognition Algorithm Based on Holder Coefficient Theory and Interval Gray Relation Classifier." *KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS*, Vo.9, No.11, pp.4573-4584, 2015. Article (CrossRef Link)

[4] Hyung Woo Park, Myung-Sook Kim and Myung-Jin Bae, “Improving Pitch Detection through Emphasized Harmonics in Time-Domain,” *Computer Applications for Database, Education, and Ubiquitous Computing. EL 2012, DTA 2012. Communications in Computer and Information Science, CCIS*, vol 352, Springer, 2012. Article (CrossRef Link)

[5] Hyung woo Park, A-Ra Khil and Myung-Jin Bae, “Pitch Detection based on Signal to Noise Ratio Estimation and Compensation for Continuous Speech Signal,” *Convergence and Hybrid Information Technology. ICHIT 2012. Communications in Computer and Information Science, CCIS*, Vol.310, pp.767-774, Springer, 2012. Article (CrossRef Link)

[6] Hyung-Woo Park, Seong-Geon Bae and Myung-Jin Bae, “Pitch Gross Error Compensation in Continuous Speech,” Springer, *LNCS*, 2012. Article (CrossRef Link)

[7] H.W Park and M.J Bae, “IMBE Model Based SNR Estimation of Continuous Speech Signals,” *The Acoustical Society of Korea*, Vol.29, No.2, pp.148-153, 2010.

[8] Hong Kook Kim and Richard C. Rose, “Cepstrum-Domain Model Combination Based on Decomposition of Speech and Noise Using MMSE-LSA for ASR in Noisy Environments,” *IEEE Tran. ON Audio, Speech AND Language Processing*, VOL. 17, NO. 4, pp. 704-713, 2009. Article (CrossRef Link)

[9] Myung-Suk Song, Chang-Heon Lee, Seok-Pil Lee and Hong-Goo Kang, "Performance Analysis of a Class of Single Channel Speech Enhancement Algorithms for Automatic Speech Recognition," *Journal of Acoustical Society of Korea*, Vol.29, No.2E, pp.86-99, 2010.

[10] Arun Narayanan and DeLiang Wang, "A CASA-Based System for Long-Term SNR Estimation," *IEEE. Trans. on Audio, Speech, and Language processing*, Vol. 20, No.9, 2012. Article (CrossRef Link)

[11] Colin Breithaupt and Rainer Martin, “DFT–based Speech Enhancement for Robust Automatic Speech Recognition,” *ITG-Fachtagung*, Aachen, Germany, 2008.

[12] Colin Breithaupt and Rainer Martin, "Analysis of the Decision-Directed SNR Estimator for Speech Enhancement With Respect to Low-SNR and Transient Conditions," *IEEE Trans. on Audio, Speech, and Language processing*, Vol.19, No.2, pp.277-289, 2011. Article (CrossRef Link)

[13] D. Y. Zhao, W. Bastiaan Kleijn, A. Ypma and Bert de Vries, “Online Noise Estimation Using Stochastic-Gain HMM for Speech Enhancement,” *IEEE. Trans. on Audio, Speech and Language Processing*, Vol. 16, No. 4, pp. 835-846, May. 2008. Article (CrossRef Link)

[14] A. J. Accardi and R. V. Cox., “A modular approach to speech enhancement with an application to speech coding," *J. Acout. Soc. Am*, Vol. 10, No. 3, pp. 1245, Sep. 2001. Article (CrossRef Link)

[15] E. Nemer, R. Goubran and S. Mahmoud, "Speech Enhancement Using Fourth-Order Cumulants and Time-Domain Optimal filters," *Sixth European Conference on Speech Communication and Technology*, 1999.
[16] G. M. Davis, *Noise Reduction in Speech Applications*, CRC Press, Chapter 1, Chapter 6, 2002.
[17] Gang Wang, Chunguang Li and Le Dong, "Noise Estimation Using Mean Square Cross Prediction Error for Speech Enhancement," *IEEE Tran. on Circuits and Systems-I*, Vol. 57, No. 7, 2010. Article (CrossRef Link)
[18] HeaKyung Jung, YuJin Kim, and JaeHo Chung, "Formant-broadened CMS Using the Log-spectrum Transformed from the Cepstrum," *Journal of Acoustical Society of Korea*, Vol.21, No.4, pp.361-373, 2002.
[19] Y. H. Song, J. H. Ahn, and M. J. Bae, "On the noise detection from correlation of near pitch waveforms," GESTS Society, *GESTS Int'l Trans. Computer Science and Engineering*, Vol.44, No.1, pp.45-54, Jan. 2008.
[20] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. New Jersey: Englewood Cliffs, Prentice Hall, 1978.
[21] MyungJin Bae, Sanghyo Lee, *Digital Voice Analysis*, Dongyoung press, 1998.
[22] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*, PEARSON Education, 2011.
[23] L. R. Rabiner and R. W. Schafer, *Introduction to Digital Speech Processing*, Foundations and Trends in Signal Processing, 2007.
[24] M. BAE, H. YOON, S. ANN, "On Altering the Pitch of Speech Signals in Waveform Coding -Alteration Method by the LPC and Pitch Halving," *Journal of the Acoustical Society of Korea*, Vol.10, No.5, pp.11-19, 1991.
[25] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Tran. on Acoustics, Speech and Signal Processing*, Vol.27, No.2, pp.133-120, 1979. Article (CrossRef Link)
[26] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE. Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 443-445, Apr. 1985. Article (CrossRef Link)
[27] Y. Ephraim and I. Cohen, "Recent advancements in speech enhancement," *The Electrical Engineering Handbook*, 3rd ed. Boca Raton, FL: CRC Press, to be published [Online].
[28] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE. Trans. Speech Audio Processing*, Vol. 11, No. 5, pp. 498-505, Sep. 2003. Article (CrossRef Link)
[29] S. Maithani and R. Tyagi, "Noise Characterization and Classification for Background Estimation," in *Proc. of IEEE, International Conference on Signal Processing Communications and Networking*, pp. 208-213, Jan. 2008. Article (CrossRef Link)
[30] Fant, C. G. M., Acoustic theory of speech production. Royal Institute of Technology, Division of Telegraphy - Telephony, Report No. 10 (Stockholm), 1958.
[31] Sadaoki Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Tran. on ASSP*, Vol.29, No.2, pp.254-272, 1981. https://doi.org/10.1109/tassp.1981.1163530
[32] Seong-Geon Bae, Hyung-Woo Park and Myung-Jin Bae, "On a New Enhancement of speech Signal using Non-uniform Sampling and Post Filter," Springer, *LNCS*, 2012. Article (CrossRef Link)
[33] W. Jiang, W. K. Lo and H. Meng, "A New Voice Activity Detection Method Using Maximized Sub-band SNR," *IEEE. ICALIP2010*, pp.80-84, 2010. Article (CrossRef Link)
[34] WangRae Jo and MyungJin Bae, "On a Fast Pitch detection using the Cepstrum Analysis," GESTS Society, *GESTS International Transactions on Acoustic Science and Engineering*, Vol.2, No.1, pp.1-8, 2004.
[35] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE. Tran. on ASSP*, Vol.35, pp.947-954, 1987. Article (CrossRef Link)

[36] A. M. Kondoz, Digital Speech -Coding for Low Bit Rate Communications Systems, John Wiley & Sons, 1994.
[37] IS-733 draft, TIA/EIA.
[38] D. W. Griffin and J. S. Lim, "Multiband Excitation Vocoder," *IEEE. Tran. on Acoustics, Speech and Signal processing*, Vol. 36, No. 8, August 1988. Article (CrossRef Link)
[39] *IMBE VOCODER DESCRIPTION*, Digital Voice System, 1993.
[40] ITU-T Recommendation P.862, ITU-T.

Hyung Woo Park received a Ph.D., an M.S., and a B.S. in Electrical Engineering from Soongsil University. He is an assistant professor at the Information and Technology Department at Soongsil University, Seoul, Korea. His current research interest includes sound signal processing, big data analysis, voice analysis, noise reduction system, wave field synthesis, railway noise, and Internet of Things.