THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.20, No.3E 2001. 9. pp. 18~24

A Study on Measuring the Speaking Rate of Speaking Signal by Using Line Spectrum Pair Coefficients

Kyung A Jang*, Myung Jin Bae*

*Dept. of Information & Telecommunication Engr, Soongsil University
(Received 10 April 2001; revised 21 September 2001; accepted 11 October 2001)

Abstract

Speaking rate represents how many phonemes in speech signal have in limited time. It is various and changeable depending on the speakers and the characters of each phoneme. The preprocessing to remove the effect of variety of speaking rate is necessary before recognizing the speech in the present speech recognition systems. So if it is possible to estimate the speaking rate in advance, the performance of speech recognition can be higher. However, the conventional speech vocoder decides the transmission rate for analyzing the fixed period no regardless of the variety rate of phoneme but if the speaking rate can be estimated in advance, it is very important information of speech to use in speech coding part as well. It increases the quality of sound in vocoder as well as applies the variable transmission rate. In this paper, we propose the method for presenting the speaking rate as parameter in speech vocoder. To estimate the speaking rate, the variety of phoneme is estimated and the Line Spectrum Pairs is used to estimate it. As a result of comparing the speaking rate performance with the proposed algorithm and passivity method worked by eye, error between two methods is 5.38% about fast utterance and 1.78% about slow utterance and the accuracy between two methods is 98% about slow utterance and 94% about fast utterances in 30 dB SNR and 10 dB SNR respectively.

Keywords: *Speaking rate, Line spectrum pairs*

I. Introduction

Talkers differ in their characteristic speaking rates. Individual talkers will also vary their speaking rate, sometimes even within a single utterance. This variation affects the acoustic patterns of speech by restructuring the relationship between acoustic cues and phonetic categories. The duration of cues change nonlineraly with speaking rate and there may also be changes in spectral patterns. In speech recognition, speaking rate means that variability in speaking rate results in a many-to-many mapping between acoustic properties in speech and the linguistic interpretation of an utterance. In order to recognize the phonetic structure of an utterance, listeners must calibrate their phonetic decisions against the rate at which the speech was produced. Most of the research on rate normalization has investigated the sources of information used by listeners to determine the speaking rate. There is an assumption in much of this research that the normalization process is a passive, automatized filtering process that strips the effects of rate variation away from the signal prior to recognition. However, the algorithm for estimating the speaking rate used at the current is

Corresponding author: Kyung A Jang (kajang@hotmail.com)
Soongsil University, Seoul 156-743 Korea

generally for the speech recognition or synthesis but that algorithm for the performance needs the various database about speech and takes much of the computation complexity. So the algorithm used for recognition or synthesis isn't appropriate way to apply for the vocoder that should transmit the information with low bit rate and be a low complication with simple algorithm. In this paper, we'd like to propose new algorithm using simple parameters to estimate the speaking rate for the vocoder.

II. Speaking Rate

One important source of variation of phoneme duration is the change in speech rate. The rate is a global measure and can be defined as the average number of phonemes per unit of time. For example, in the ARPA WASJ database, sentences are fairly long with an average of 120 phones, so a speaking rate measured over the whole sentence may ignore fluctuations within the utterance. Therefore, we attribute a speaking rate measurement for each phoneme segment that is computed based on the observed duration of a small number of phoneme segments around this phoneme. We define the speaking rate measure as;

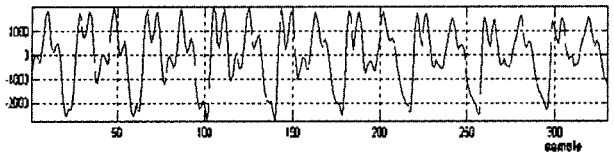
$$r_i = \frac{\sum_{k=-M}^{k=M} d_{i+k}}{\sum_{k=-M}^{k=M} ttl_{i+k}} \quad (1)$$

where r_i is the speaking rate of the *ith* phonetic, d_{i+k} and ttl_{i+k} are the observed duration and the expected duration of the duration probability distribution, respectively, of the *(i+k)* phonetic segment. This definition gives us the flexibility to adjust the speaking rate computation window to any length from one phoneme segment to the whole sentence by appropriately setting the variable M. In the experiments presented, we chose to calculate the speaking rate for speech segments of five phonemes (M=2), based on some preliminary experiments. Given a measure of speaking rate for each phonetic segment, two different duration models were built to take advantage of the speaking rate information. The first approach is similar to[3] and uses the speaking rate information to cluster the training data into different set and compute histogram and probability density functions for the duration models in each set. The second approach assumes that the duration of the phoneme segments is a function of the speaking rate and uses this assumption to generate a single normalized duration model. In both approaches, we need to generate histograms for all the duration models that occur in the training data without taking into account any speaking rate information. These statistics are necessary for the calculation of the mean duration values of the models, which is used in equation 3 in order to compute the phoneme segment[2].

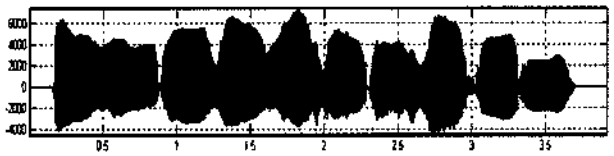
III. Measurement of Speaking Rate by Phoneme Variation Rate

3.1. Estimation of Phonemes Variation in Speech Signal by LSP Parameters

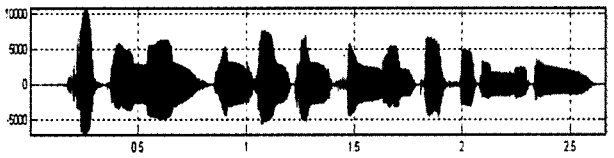
In order to estimate speaking rate how many phonemes is contained in speech signal during limited time. So LSP parameters is used to estimate the phonemes variation rate that LSP parameters represent the information of spectrum. LSP parameter is useful for estimating the distances between neighboring frames because LSP parameters have normalization values. Moreover LSP parameters is also useful that it is very appropriate parameter for applying the speaking rate to vocoder as a parameter as well[16]. First of all, the characteristic of Line Spectrum Pairs parameters depending on speech signal is explained as following. Figure 1 shows the analysis of 10^{th} order Line Spectrum Pairs for voiced speech and Figure 2 represents the analysis of 10^{th} order Line Spectrum Pairs for unvoiced speech. (b) in Figure 1 represents the speech spectrum and LPC envelope and (C) shows LSP and LPC envelope. Formant frequency in speech signal represents a peak point in LPC envelope. It represents a Line Spectrum Pairs



(a) The waveform of voiced speech

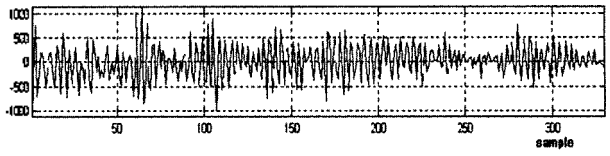


(b) The spectrum of waveform

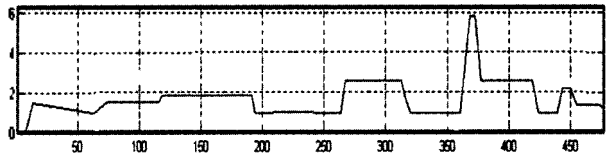


(c) LSP analysis

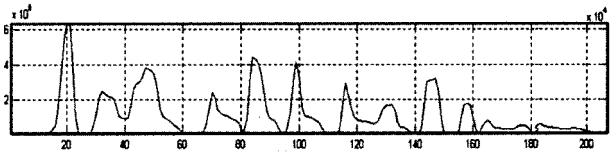
Figure 1. The distribution of line spectrum parameters for the voiced speech.



(a) The waveform of voiced speech



(b) The spectrum of waveform



(c) LSP analysis

Figure 2. The distribution of Line spectrum parameters for the unvoiced speech.

that generally constitute a narrow width with a pair of lines. So when the phonemes have a variation the formant frequency of each phoneme has changed as much as a phoneme changes and line spectrum pairs also changed. Therefore we make a number of the variation of LSP location to obtain a variation rate of phonemes with a numerical vaule. Generally, Line Spectrum Pairs being located narrowly represents the formant frequency. But 10th number of line spectrum pairs in figure (c) doesn't signify the formant frequency at all so the error can be caused by a rest of LSP parameters that don't represent a formant frequency. Figure 2 shows that the LSP of unvoiced speech has a difference clearly with the LSP of voiced speech. The spectrum information of LSP uses for estimating the starting point which is with the unvoiced speech. The starting with the unvoiced speech has a variety more than one with the voiced speech so when we obtain the information about that in advance it can be a important information in vocoding or recognition processing.

IV. Measurement of Speaking Rate by the Estimation of LSP Distance

4.1. Extraction of Silence Period Data to Detect a Speech Period

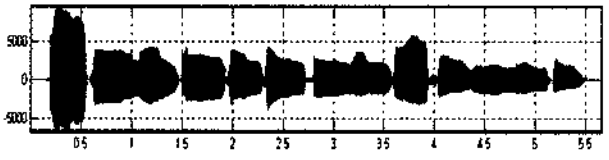
Speaking rate in this paper gets rid of a silence period from speech signal. If the silence period is considered when the speaking rate is measured then the different result is caused even about the fast utterance. So the speech period should be detected first to measure an applicable speaking rate. In this paper, first the energy and the value of LSP parameters use for detecting a speech period and the silence period to extract a parameter for speech period detection is from some frames of first part in utterance. The data extracted in this some frames for detecting a speech period is compared with the data of a speech period. The signal during first 150 msec sampled 8 kHz is assumed as a silence period and the data of energy and LSP parameters is extracted. 200% of energy obtained in former processing is set to the energy threshold.

4.2. Estimation of the LSP Distance in a Neighboring frame

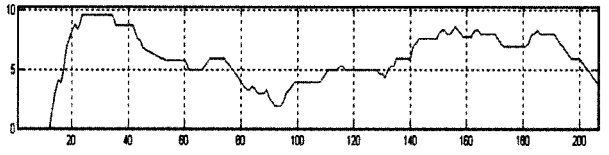
The spectrum of speech signal doesn't change much but unvoiced speech because of speech production model before measuring the LSP distance. So the LSP distance is measured by an average LSP value during 60 msec and the method for measuring the LSP distance is Euclidean distance measurement.

$$D(n)=\frac{1}{P}\sum_{i=0}^{P}\left|LSP_n(i)-LSP_{n+1}(i)\right|^2 \quad (1)$$

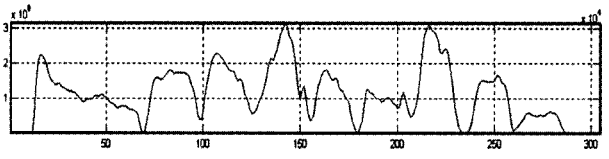
(D) is the distance between the analysis period of nth and *(n+1)th* and P is the analysis order of LSP. *LSPn* is a average *LSP* of *nth* analysis period and is a average LSP of *(n+1)th*. Figure 3 is a variation of LSP distance for speech signal. (a) is a waveform in a time domain about speech signal and (b) is a spectrogram of speech signal. The horizontal axis is a value in a time domain and ordinates axis is a value in a frequency domain. (c) is a measurement of LSP distance like equation 1. Utterance is / Ah Aha Uh Ueo Oh Yo Uh U Eu Ee/ which is a vowel as voiced speech. Compare with figure (c).



(a) Speech signal



(b) Spectrogram of speech signal



(c) Variation of LSP distance

Figure 3. A measurement of LSP distance.

4.3. The Measurement of Speaking Rate

In order to estimate a speaking rate of speech signal first, we decide whether a present frame is a silence frame or not. A decision for a silence is used by an average energy threshold and LSP parameter. When the energy of a present analyzing period is lower than a threshold value and LSP distance of silence is also lower than a threshold value of LSP distance obtained before, a present frame is decided as silence frame. The silence frame assumed by former processing excepted from further processing for estimating a speaking rate. After a silence decision the LSP distance is estimated between a present frame and neighboring frame. When the estimated distance value is over LSP threshold value we decide that a present frame has a variation of phonemes and the time is counted since last period has a variation of phonemes.

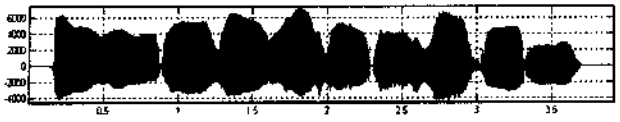
$$SPR=\frac{F_s}{VST(n)-VST(n-1)} \quad (2)$$

SPR means the speaking rate representing the number of phoneme changed per one second and *VST*(n) is the time when the phoneme has changed at the current, *VST(n-1)* is the time when the phoneme has changed in former times. *VST* is measured with a unit of analysis period. For example, when 30 msec (240 smaples) of the utterance sampled 8 kHz as analysis period if 10^{th} and 15^{th} of analysis period has happened a variation of phonemes the speaking rate can be measured as following.

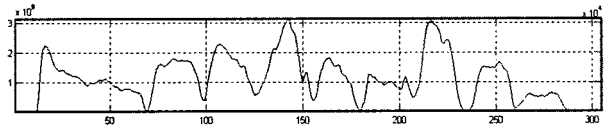
$$SPR=\frac{8000}{240\times(15-10)}=6.67 \quad (3)$$

4.3.1. Processing when the Period is Starting with Unvoiced Speech

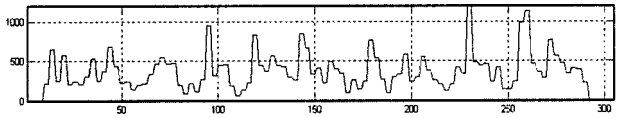
The period starting with unvoiced speech is shorter than the period with voiced speech so unvoiced speech period should be considered with voiced speech. In this paper, the phonemes starting with unvoiced speech is tied with voiced speech and is estimated a speaking rate. For instance, in case of /chu/ + /ah/ or /sa/ + /ah/, we estimate a speaking rate in the period when the sound of /ah/ finished not about /chu/. In order to estimate speaking rate



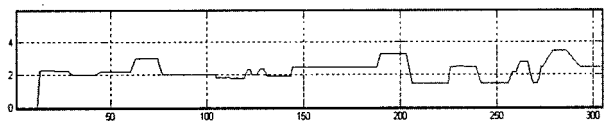
(a) Speech waveform



(b) Energy Variation

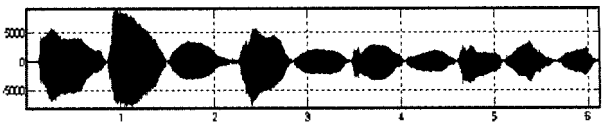


(c) Variation of the distance of LSP

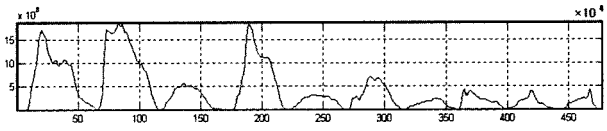


(d) Speaking rate

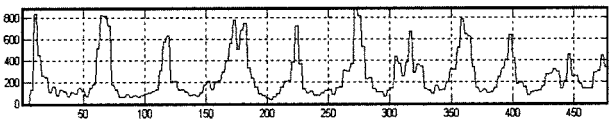
Figure 4. Speaking rate in case of uttering fast /Ah Yah Yuh Yeo Uh Yuh U Ee/.



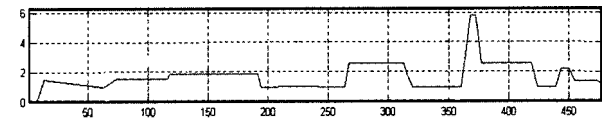
(a) Speech waveform



(b) Energy Variation

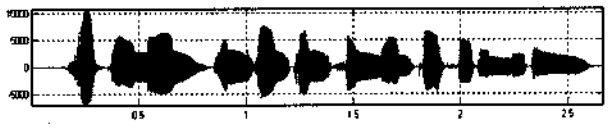


(c) Variation of the distance of LSP

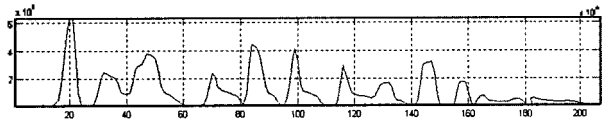


(d) Speaking rate

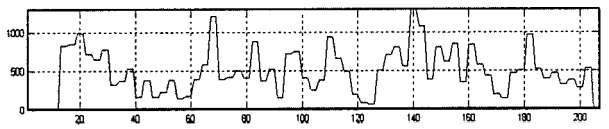
Figure 5. Speaking rate in case of uttering slowly /Ah Yah Yuh Yeo Uh Yuh U Ee/.



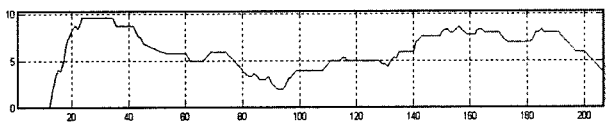
(a) Speech waveform



(b) Energy Variation

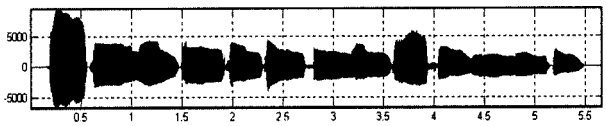


(c) Variation of the distance of LSP

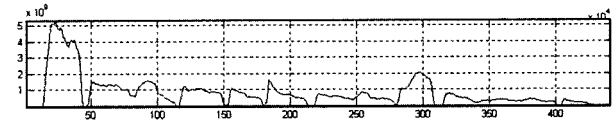


(d) Speaking rate

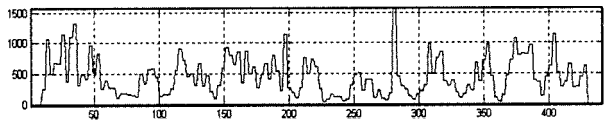
Figure 6. Speaking rate in case of uttering fast /Yeogieun Umseongtongshin Yeongusileenida/.



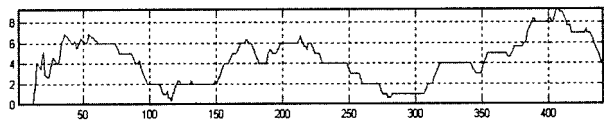
(a) Speech waveform



(b) Energy Variation



(c) Variation of the distance of LSP



(d) Speaking rate

Figure 7. Speaking rate in case of utterring slowly /Yeogieun Umseongtongshin Yeongusileenida/.

like that first, we should make a decision whether the analysis period is an unvoiced speech or voiced speech. The decision is worked and made by detecting a location of Line Spectrum Pairs which has the most narrow distance among line spectrum pairs. If the location of Line spectrum pairs which has the most narrow distance is in high frequency domain such like over 2 kHz we assumed that is an unvoiced speech. However, making a decision about silence in advance is performing all of analysis period so the analysis period not silence one takes an applicable process.

V. Experiment and Results

Computer simulation was performed to evaluate the proposed algorithm using IBM computer (300 MHz) interfaced with the 16-bit AD/DA converter. To measure the performance of the proposed algorithm, we used following speech data. Speech data was sampled at 8kHz and was quantized at 16bits. Following sentences are uttered by 5 male and 5 female speakers five times who are the middle or later 20 and utterance is spoken twice with the different speaking rate respectively. Those sentences were used as speech data.

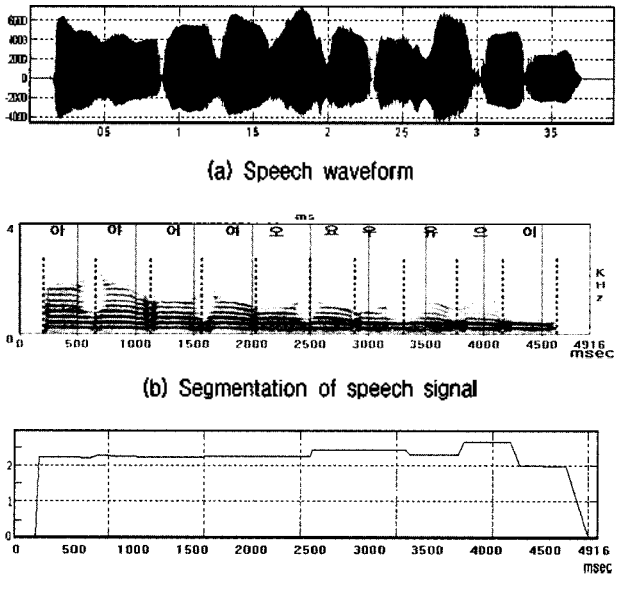
1) "/ Ah Yah Yuh Yeo Uh Yuh U Ee/"
2) "/ Yeogieun Umseongtongshin Yeongusileenida/"
3) "/Il Ee sam Sa Oh Yuk Chil Pal Gu Ship/"
4) "/Ahrumdaun Gaulimnida/"
5) "/Danmalgi Saang Cheodae Kyumoeeda/"

The proposed algorithm using in this paper is implemented C-language. The time required for each utterance and the variation of phoneme for each sentence is measured respectively. The process for getting each parameter to measure the speaking rate is in Figure 4. (a) is the waveform of input speech signal. We can know that the same sentence is uttered fast or slow with index in Figure 4 and one phoneme is delayed much by uttering the sentence longer. (b) is the energy of speech signal and (c) is the distance of LSP coefficient between neighboring analysis periods. The distance of LSP coefficient has fixed value and the variation value phoneme is also less in terms of maintaining one phoneme but the distance of LSP coefficients is much higher definitely in case of changing the phoneme.

(c) and (d) are the speaking rate got by the variation of LSP coefficient. Former speaking rate is used in case of maintaing the same variation of phoneme. In Figure 4, the speaking rate has the same value when one phoneme continues longer like (c). But in case of uttering fast, the measurement of speaking rate has higher value than one of utterance spoken slowly. So we realize that the proposed method for measuring the speaking rate presents the real speaking rate.

5.1. Measurement of the Accuracy of Speaking Rate

To measure the accuracy of speaking rate value resulted in the proposed algorithm in this paper, we compared with two which are estimated values by the proposed algorithm and passivity method segmented by eye. Passivity method segmented accurately and directly through spectrogram first and confirmed the position of segmentation via listening with ears.



(a) Speech waveform

(b) Segmentation of speech signal

(c) Speaking rate

Figure 8. Speaking rate in case of uttering fast /Ah Yah Yuh Yeo Uh Yuh U Ee/.

Table 1. Comparison of average speaking rate measured.

Speaking Rate	Proposed algorithm		Passive algorithm with the segmentation by eys	
	Fast	Slow	Fast	Slow
Utterance(1)	2.879	1.564	2.26	1.52
Utterance(2)	4.949	3.187	5.12	3.27
Utterance(3)	3.472	2.129	3.53	2.20
Utterance(4)	5.432	2.103	4.84	2.56
Utterance(5)	4.622	2.545	5.23	2.53

Table 2. Error rate between the measurements by proposed method and by passive method.

Utterance	Error rate(5)	
	Fast	Slow
(1)	4.6	1.2
(2)	3.3	1.7
(3)	3.8	1.4
(4)	7.5	2.3
(5)	7.7	2.3
Average	5.38	1.78

VI. Conclusion

Speaking rate represents how many phonemes in speech signal have in limited time. It is various and changeable depending on the speakers and the characters of each phoneme. However, Speech vocoder hasn' t considered the speaking rate till now. The conventional speech vocoder decides the transmission rate for analyzing the fixed period no regardless of the variety rate of phoneme but if the speaking rate can be estimated in advance, it is very important information of speech to use in speech coding part as well. It can be applied to each different vocoding method between the part of slow speech and fast speech respectively. In this paper, we propose the method for presenting the speaking rate as parameter in speech vocoder. To estimate the speaking rate, the variety of phoneme is estimated and the Line Spectrum Pairs is used to estimate it. Neighboring LSP of speech signal are compared for limited frames. Utterances are recorded with regular speaking rate and fast or slowly on a purpose. As a result of comparing the speaking rate performance with the proposed algorithm and passivity method worked by eye, error between two methods is 5.38% about fast utterance and 1.78% about slow utterance and the accuracy is 98% and 94% about utterances in 30 dB SNR and 10 dB SNR respectively.

References

1. M. J BAE, "Digital Speech Analysis," Dong young Publication, pp. 95-120, 4, 1998.
2. Anastasios Anstasakos, Richard Schwartz, Han Shu "DURATION MODELING IN LARGE VOCABULARY SPEECH RECOGNITION," *Proc. IEEEInt. Conf. Acoustics, Speech and Signal Processing*, pp. 628-631, 1995.
3. M. Jones and P. C. Woodland, "Using relative duration in large vocabulary speech recognition," *Proceeding EUROSPEECH '93*, Berlin, Germany, pp. 311-314, September 1993.
4. A. N. Ince, Digital Speech Processing (speech oding, synthesis, and recognition), Kluwer cademic Publishers, 1992.
5. J. R. Deller, Jr., John G. Proakis, John H. L. Hansen, "Discrete-Time Processing of Speech Signals," Maxwell Macmillan International, pp. 124-125, 1993.
6. S. Furui, "Digital Speech Processing, Synthesis, and Recognition," pp. 129, MARCEL DEKKER, INC. 1991.
7. A. M. Kondoz, "Digital Speech," John Wiley & Sons Ltd, pp. 84-92, 1994.
8. B. S. Atal and J. R. Remde "A New Model of LPC Exitation for Producing Natural-Sounding Speech at Low Bit Rates," *Proc. Int. Conf. on Acoust., Speech and Signal Processing*, pp. 614-617, Apr. 1982.
9. K. Samudravijaya, S. K. Sin and P. V. S. Rao, "Pre-recognition measures of Speaking Rate," *Speech Communication*, Vol. 24, pp. 73-84, 1998.

[Profile]

• Kyung A Jang

KyungA Jang recived the B.S and M.S degree in Information & Telecommunication Engineering from Dong-Shin university and SoongSil University in 1998 and 2001. Since 2001, she is working the Ph.D degree at SoongSil university. Her current research interests are in the area of speech Coding and Analysis.

• Myung Jin Bae

The Journal of the Acoustical Society of Korea, Vol. 14, No. 1E.