'88 KACC 1988. 10. 21~22

Korean Vowel Recognition in Noise using Auditory Model

°Jae-Seong Shim *, Jae-Hyuk Lee *, Tae-Sung Yoon *,
Seung-Hwa Beack **, Sang-Hui Park*

* Dept. of Electrical Eng., Yonsei Univ.
** Dept. of Electrical Eng., Myongji Univ.

== Abstract ==

In this study, we performed the recognition test on Korean vowel using peripheral auditory model. In addition, for the purpose of objective comparision, the recognition test is performed by extracting LPC cepstrum coefficients from the same data. And the same speech data are mixed with the Guaussian white noise quantitatively, then we repeated the same test, too. So we verified that this auditory model has a adaptability on noise.

1. INTRODUCTION

Human ear is a organ which recognizes speech signal and verificates speaker by collecting and analyzing sounds. In particular, it can interprete main information of input signals in background noise. So the human ear can be considered as a excellent signal analyzer. Recently the study on the mechanism of the ear has actively progressed. The study of human auditory mechanism can be divided into two : One is a study of motion of basilar membrane within cochlea and the other is speech coding in the auditory nerve.

In 1985, Shamma represents speech processing in the auditory system by using /da/ and /ba/ which have the form of consonant + vowel. Deng and Geisler(1987) analyzed speech signal by auditory model considering linear properties and nonlinear properties of basilar membrane. Schroeder and Hall(1974), Geisler and Schwid(1978) present a model for the transduction of mechanical motion of the basilar membrane into activity of auditory nerve. In the study of the auditory nerve, the speech coding and response properties of speech signal in background noise are investigated by Kiang, Maxon(1974), Moller (1977), Rhode(1978), Smith(1979), Sinex(1980), Abbas(1981), Delgutte (1983), etc.

In this study, by using the auditory model on the basis of the previous results, We extracted feature patterns of Korean vowels and executed the recognition test. In addition, we executed the recognition test with the same data in noise.

2. PERIPHERAL AUDITORY SYSTEM MODEL

The organ of hearing consists of the external, middle and inner ear. The external ear has a role of collecting sounds. Sound wave passes through the auditory canal and vibrate the ear drum, the external ear end. The middle ear consists of malleus, incus, and stapes and it transmits the vibration of ear drum to the cochlear fluids in the inner ear. Then the cochlea transduces this mechanical vibration into neural activity.

2-1. The model of external and middle ear

The transfer property of the external ear is concerned with horizontal incident angle and generally its anti-resonance is appeared at 1.5 ~ 7 KHz in frequency domain. Monro presents the second order model[1] of the external ear considering this property.

$$H_E(s) = \frac{1}{(s + 1000)^2 + (7875\pi)^2} \qquad (1)$$

The middle ear has the properties that goes flat from 0 to 3 KHz and discreases from 3 KHz to 20 KHz by 6 dB/octave slope. Monro presents the first order model[1] of the middle ear on the basis of this actual fact.

$$H_M(s) = \frac{1}{s + 6000\pi} \qquad (2)$$

2-2. The inner ear model

Input signal transmitted to the inner ear gives rise to difference in pressure in the fluid across the cochlear partition. The maximum vibration point on the basilar membrane is dependent on frequency components of the input signals. For the high frequency component of input signal, the maximum vibration appears near the stapes and for the lower frequency component, it moves towards the helicotrema. The time delay is zero at the stapes, and increases linearly towards the helicotrema end. Thus, the delay at a point n can be shown as eq.(3).

$$\tau_n = \sum_{i=0}^{n} \Delta i \qquad (3)$$

where, $\triangle_i$ is the delay at the each element of the basilar membrane. The basilar membrane can be described as a series of bandpass filter. The pressure variation to the lymph is in the form of contraction wave and is identical with the input variation wave of the stapes with no delay. The center frequency of band pass filter is determined as a reciprocal of delay time. When the center frequency of transfer function Hn(w) was below 300Hz, we used the second order resonance model[2] presented by Yegnanarayanan, otherwise, Flanagan model[3] is used.

$$H(s) = \begin{cases} \dfrac{(w_n/3)s}{s^2 + (w_n/3)s + w_n^2}, & f_n \le 300\text{Hz} \\ C_1 w_n^4 \left(\dfrac{2000\pi\, w_n}{w_n + 2000\pi}\right)^{0.8} \left(\dfrac{s}{s+w_n}\right) \left(\dfrac{1}{(s+\alpha_n)^2 + w_n^2}\right)^2, & f_n > 300\text{Hz} \end{cases} \quad (4)$$

where, $f_n = \frac{1}{\tau_n}$: the center frequency to which the point n distance from the stapes

$C_1 w_n^4 \left(\frac{2000\pi\, w_n}{w_n + 2000\pi}\right)^{0.8}$: gain factor

2-3. The haircell/auditory nerve model

In 1974, Schroeder and Hall proposed haircell/auditory nerve model[4] for the transduction of mechanical motion of the basilar membrane and the pressure of the lymph into action potential in the auditory nerve. The neural transmitter of the hair cell, "Quanta", generated in the haircell at a fixed average rate r, fire the afferent nerves in proportion to the number n(t) and permeability p(t) of Quanta and some of them disappear without causing nerve firing.

Basic equations of the model are as follows.

$$p(t) = p_o\{1/2y_n(t) + [1/4y_n^2(t) + 1]^{1/2}\} \quad (5)$$

$$r = \overline{n(t)p(t)} + \overline{n(t)g} \quad (6)$$

$$f(t) = n(t)p(t) \quad (7)$$

where, eq.(5) represents the permeability function of quanta in permeating through the cell wall, eq.(6) is the generation rate and eq.(7) is the firing probablity per unit time.

3. EXPERIMENTS, RESULTS AND DISCUSSION

3-1. Recognition system

In this study, Korean single vowels /아/, /애/, /어/, /오/ and /우/ are used as speech data for speaker-dependent and speaker-independent test. The data are obtained from 375 vowels of 5 male speaker in noiseless recording room of the Educational Broadcasting Station of Yonsei Univ. And to observe the characteristics of the model to noise, Gaussian white noise is added to the 375 phonetic data. The noise is changed quantitatively with signal to noise ratio(SNR) 20dB, 10dB and 5dB.

For extraction of the feature, we selected 44CH among 18CH, 22CH and 44CH because 44CH shows the best recognition rate and excluded 100CH, 200CH because of their very low recognition results. In 44CH, the range of center frequencies of the the inner ear filter is 100 - 5000 Hz. To compare with the auditory model's recognition test, another test using LPC(linear prediction coefficients) and LPC cepstrum coefficients in eq(8)[5] is carried out.

$$Co = \ln \varepsilon_{min}$$
$$Ci = -a_i - \frac{1}{i}\sum_{k=1}^{i-1} kC_k a_{i-k} \qquad 1 \le i \le p \quad (8)$$

where, C_i: LPC cepstrum coefficient

a_i: linear prediction coefficient

p : linear prediction coefficient order

The order of linear prediction coefficients and the number of LPC cepstrum coefficients are fixed on 28, and the least square mean error(ε min) is normalized to 1. The distance is calculated by Euclidian distance measuring method and the final recognition pattern is determined as reference pattern that has minimum distance .

3-2. Recognition test

Fig.1 represents the auditory model output in 44CH for the vowel /아/. If the 1st channel is the stapes and the 44th channel is the helicotrema, we can see that the time delay increases towards the helicotrema. In this output, the periodicity of the original wave is observed and some local peaks are concentrated on specific channels. To observe these features in detail, the response of the auditory model in 44CH is represented in Fig.2(a). The peak values are found near 250Hz, 833.3Hz and 2857.1Hz. This feature is called the formant frequency which is one of the most important information of vowels. Fig.2(b) represents the log power spectrum of the same data. Overlapping of Fig.2(a) and (b) shows that the places holding their peak values are nearly coincident. This shows that the auditory model reflects the frequency information of input signal well.

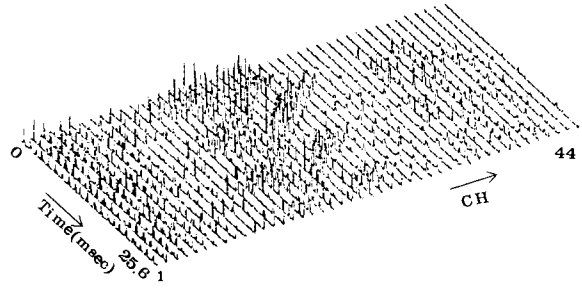
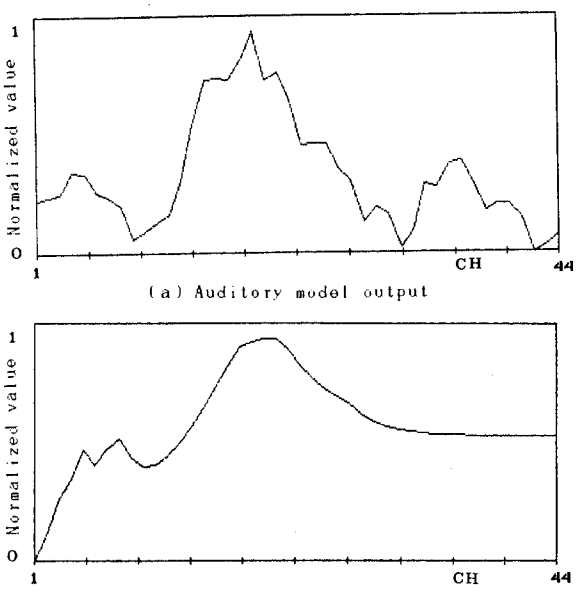


Fig.1 Output of auditory model for vowel /아/ in 44CH

To select the number of channels, we carried out speaker-dependent and independent test for 18CH, 22CH and 44CH. In speaker-independent test, 125 data of Korean single vowels /아/, /애/, /어/, /오/ and /우/, articulated 5 times by 5 male speakers, and in speaker-dependent test, we used 25 data of a speaker RKK. The reference template is obtained by averaging the feature vector of the first vowel data of each speaker in speaker-independent test and is determined by selecting the first data for 5 vowels of speaker RKK in speaker-dependent test.

As a result, recognition rates for 18CH,



(a) Auditory model output

(b) Log power spectrum

Fig.2 Auditory model and log power spectrum output for /아/

22CH and 44CH are 90.4%, 91.2%, 96.8% in speaker-independent test and 92%, 100%, 100% respectively in speaker-dependent test. In the more enlarged channel, 100CH and 200CH, the recognition rates were very low and recognition time was so long that we excluded them. On the basis of these results, we used 44, the number of channels for the model.

Recognition tests are executed with speaker-independent test and speaker-dependent test for 375 Korean vowels, and to investigate the characteristics of the model for noise, we mixed original data with Gaussian white noise. The signal to noise ration, SNR is varied as 20, 10 and 5 dB. The speaker-independent test is executed by two way, according to how to take reference template.

Fig.3 is the waveforms of the original Korean vowel /오/ and the noise-vowel mixed signals. The mixed signals' SNR are 20, 10 and 5 dB respectively. Fig.4 is wave forms overlapped feature vectors on 5 Korean vowels that are articulated 5 times by a speaker KDJ. For the same articulation, feature vectors are similar, so this is suitable for the recognition test. Table I shows the recogniton results.

From Table I we can see that the recognition rate decreases according as the clean data is mixed with 20dB, 10dB and 5dB respectively, except the speaker-dependent test on a few speaker. Comparing the auditory model and the LPC cepstrum coefficients, when we used the auditory medel, the recognition rate curve is more slowly down than the LPC cepstrum coefficients as SNR goes down to 5dB. So the auditory model is more useful for recognition test in noise. When the recognition rate of original signal is fixed at 100% and the SNR of noise is changed to 20dB, 10dB and 5dB, the recognition rates of the auditory model are 100%, 96.4% and 92.3%, while

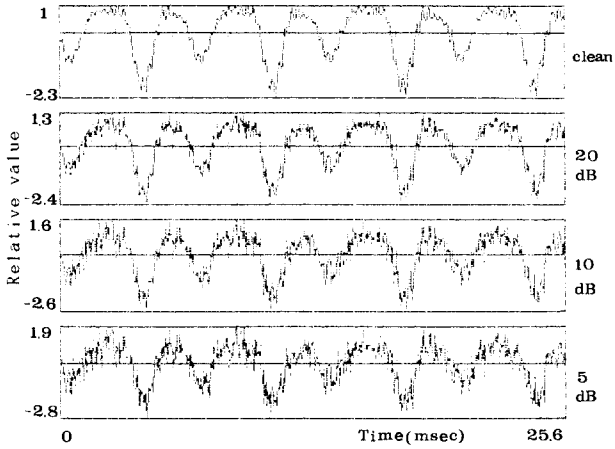


Fig.3 Waveforms of Korean vowel /오/ and noise-vowel mixed signal that its SNR is 20dB, 10dB and 5dB

Table I Result of recognotion test on 5 Korean vowels

feature parameter	SNR (dB)	speaker	recognition rate(%) speaker-dependent test (I)	(II)	speaker-independent test
model output (44CH)	∞	JSW	86.7	97.3	
		KDJ	100	100	
		LSJ	100	100	96.8
		NOK	97.3	98.7	
		RKK	100	100	
	20	JSW	86.7	97.3	
		KDJ	100	100	
		LSJ	100	100	96.8
		NOK	94.7	98.7	
		RKK	100	100	
	10	JSW	89.3	97.3	
		KDJ	100	100	
		LSJ	100	98.7	93.3
		NOK	96	98.7	
		RKK	100	100	
	5	JSW	92	96	
		KDJ	98.7	100	
		LSJ	97.3	98.7	89.3
		NOK	96	100	
		RKK	94.7	100	
LPC cepstrum coeff. (28)	∞	JSW	90.7	98.7	
		KDJ	98.7	100	
		LSJ	98.7	98.7	86.4
		NOK	98.7	97.3	
		RKK	97.3	100	
	20	JSW	90.7	82.7	
		KDJ	85.3	84	
		LSJ	80	84	82.4
		NOK	90.7	92	
		RKK	98.7	96	
	10	JSW	70.7	78.7	
		KDJ	84	81.3	
		LSJ	84	81.3	78.9
		NOK	84	82.7	
		RKK	90.7	92	
	5	JSW	76	78.7	
		KDJ	65.3	70.7	
		LSJ	78.7	81.3	77.9
		NOK	80	77.3	
		RKK	77.3	76	

those of LPC cepstrum coefficients are 95.4%, 91.3% and 90.1% respectively. These results show that the auditory model considering physiological characteristics of the ear is somewhat adaptable for noise. And in the speaker-dependent test, the result is similar to that of speaker-independent test, too.

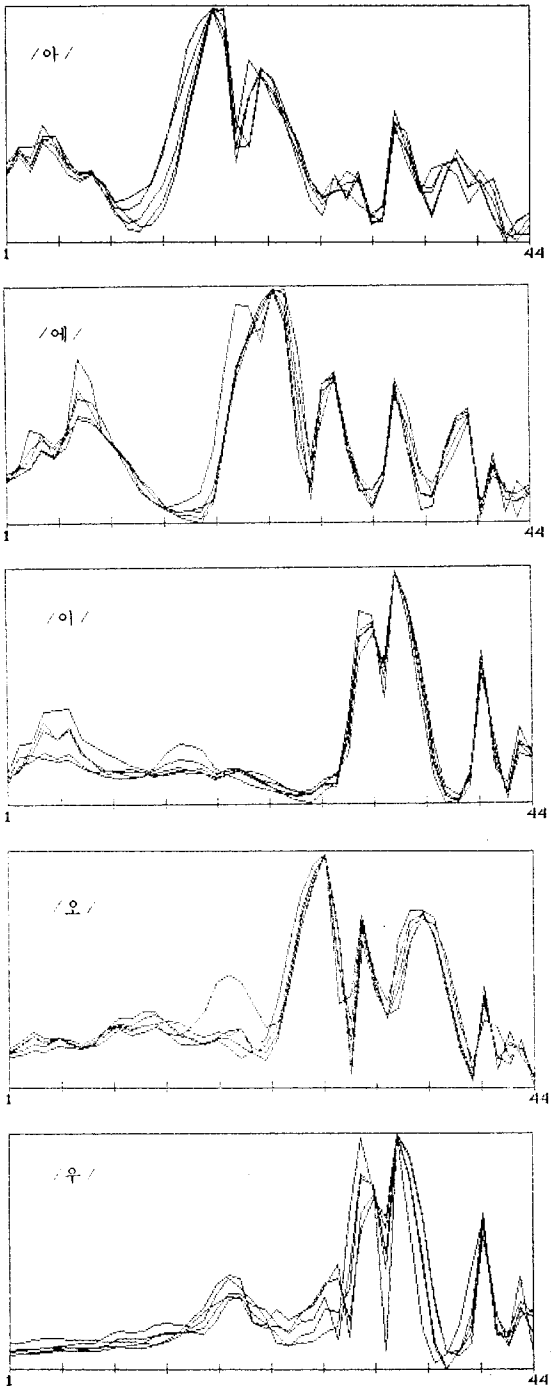


Fig.4 Feature patterns of Korean vowels

4. CONCLUSION

In this study, we modelled the human peripheral auditory system and we executed recognition test on 5 Korean vowels. And the same speech data are mixed with the Guaussian white noise quantitatively, then we repeated the same test. So we verified that this auditory model is adaptable for noise. The results obtained in this study are summarized as follows.

1) The auditory model used in this study reflects well the information of time and frequency in input vowel signal.

2) The recognition tests for Korean vowel show that the recognition rate by auditory model outputs is higher than by LPC cepstrum coefficients.

3) Especially as for the data mixed with noise, the recognition rate by auditory model outputs is much higher than by LPC cepstrum coefficients. The difference of the recognition rates between two feature parameters is above 10%.

REFERENCE

1. Monro, D. M., "Computer modelling of the peripheral mechanical response of the auditory system," in Auditory Investigation : the scientific and technological basis, edited by H.A. Beagley, Clarendon Press, pp. 431-450, 1979.
2. Yegnanarayanan, G., "A new model of hearing and its performance in pitch perception," Ph.D. thesis, Delaware Univ., 1985.
3. Flanagan, J.L., "Speech analysis, synthesis and perception," Springer Verlag, pp.109-112, 1972.
4. Schroeder, M.R. and Hall, J.L., "Model for mechanical neural transduction in the auditory receptor," JASA 55, pp.1055- 1060, 1974.
5. Saito, S. and Nakata, K., "Fundamentals of speech signal processing," Academic Press, 1985.
6. S.D. Stearns and R.A. David, "Signal processing algorithms," Frentice Hall, pp.52-55, 1988.
7. S.M. Kay, "Modern spectral estimation:theory and application," Prentice Hall, pp145-147, 1988.
8. J.H. Lee,"The study on the speech recognition using auditory model," Master thesis, Yonsei Univ., 1987.
9. T.S. Yoon,"A study on Korean monosyllabic word recognition using auditory model," Ph. D. thesis, Yonsei Univ., 1988.
10. D.T. Haar,"Mechanisms of speech recognition," Pergamon Press, pp.12-15, 1985.