SPEECH RECOGNITION BASED ON A MODEL OF HUMAN AUDITORY SYSTEM

Takuya Koizumi, Mikio Mori, and Shuji Taniguchi

Dept. of Information Science, Fukui University
3-9-1 Bunkyo, Fukui 910, Japan

ABSTRACT

This paper deals with a new phoneme recognition system based on a model of human auditory system. This system is made up of a model of human cochlea and a simple multi-layer recurrent neural network which has feedback connections of self-loop type. The ability of this system has been investigated by a phoneme recognition experiment using a number of Japanese words uttered by a native male speaker. The result of the experiment shows that recognition accuracies achieved with this system in the presence of noise are higher than those obtained by a combination of frequency spectral analysis by DFT and conventional feedforward neural network and that the cochlea model effectively prevents the deterioration due to noise of recognition accuracy.

1 INTRODUCTION

The superb ability to recognize speech sounds of the human auditory system seems to suggest that modeling the human auditory system properly may lead to a realization of some good scheme of processing speech signals for reliable recognition of speech. The present work is a challenge to this idea.

Since the human auditory system is composed of the auditory periphery and central nervous system comprising infinitely many neurons, modeling the auditory system necessarily requires modeling both the cochlea and neurons. Recently we have developed a novel model of the human cochlea [1] called "a feedback model for cochlea", which can be used as a kind of spectrum analyzer for speech sounds. To develop this model we have noticed specifically two kinds of interaction between the basilar membrane and other constituents of the cochlea. One is an electro-mechanical interaction between the basilar membrane and outer hair cells, and the other is a fluid-mechanical interaction between the basilar membrane and lymph within the cochlea. Taking these interactions properly into account has led to the feedback model for the cochlea, which can not only represent but elucidate some important cochlea characteristics including the sharp frequency selectivity of basilar membrane oscillation.

Recently a simple multi-layer recurrent neural network which has feedback connections of self-loop type has been proposed as an attractive tool for recognizing speech sounds [2]. The cochlea model and the recurrent neural network have been combined to form a reliable phoneme recognizer to be used as a front end of continuous speech recognition system. The ability of this system has been investigated by a phoneme recognition experiment using a number of Japanese words uttered by a native male speaker. The result of the experiment shows that recognition rates achieved with this system are higher than those obtained with other conventional recognition systems. Some important findings about the system include the following:

1. It shows a high phoneme discriminating power not only for vowels but for consonants in the presence of background noise.
2. The cochlea model is capable of effectively preventing the deterioration due to noise of its recognition accuracy.
3. It has a capability of classifying vowel sounds, even when two adjacent formant peaks of them merge into a single peak.

In what follows, first the phoneme recognition system will be described in detail, then the performance of the system in the absence of noise and its robustness for noise will be discussed.

2 THE PHONEME RECOGNITION SYSTEM

2.1 The cochlea model

As mentioned above, the phoneme recognition system is composed of the cochlea model and the recurrent neural network. The cochlea model will be described in brief. To develop this cochlea model we have noticed specifically two kinds of interaction between the basilar membrane and other constituents of the cochlea. One is an interaction due to mechanical-to-electrical and electrical-to-mechanical transductions between the basilar membrane and the outer hair cells, and the other is a fluid-mechanical interaction between the basilar membrane and lymph within the cochlea. The model is able to elucidate some important cochlea characteristics including the sharp frequency selectivity of basilar membrane oscillation due to an active pressure source which is supposed to exist within the cochlea.

Figure 1 depicts the block diagram of a basilar membrane filter which represents the behavior and function of a small section of the basilar membrane. Here, δ and $G_x(s)$ represent, respectively, the displacement and transfer function of the small section which is at a distance x from the oval window. P denotes pressure difference between the scala vestibuli and scala tympani at the position x and $K(x)$ is a gain factor. $G_x(s), K(x)$, and the fluid-mechanical interaction are expressed, respectively, by the following relations:

$$G_x(s) = \frac{1}{sL(x) + R(x) + \frac{1}{sC(x)}} \tag{1}$$

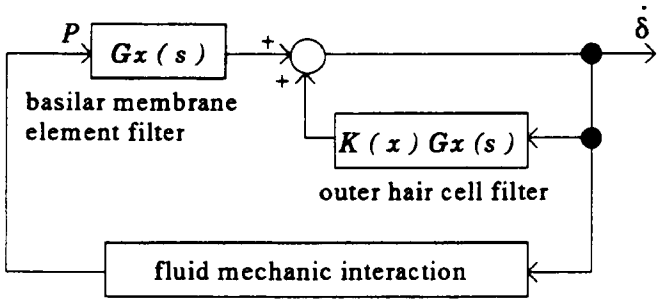


Figure 1: The block diagram of a basilar membrane filter.

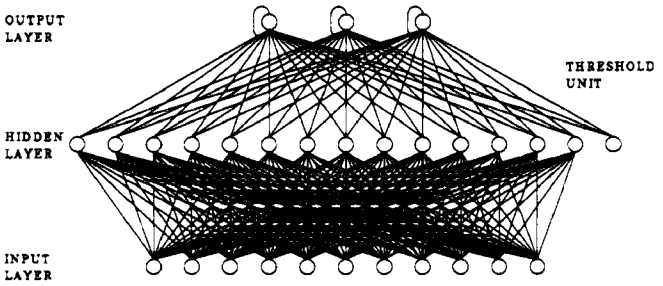


Figure 2: The structure of the three-layered recurrent neural network.

$$K(x) = -407.167 \exp(-1.749x) \tag{2}$$

$$\frac{d^2}{dx^2}P(x) = \frac{2\rho\alpha}{H}\dot{\delta}(x) - \frac{2}{3}\rho\alpha H\frac{d^2}{dx^2}\dot{\delta}(x) \tag{3}$$

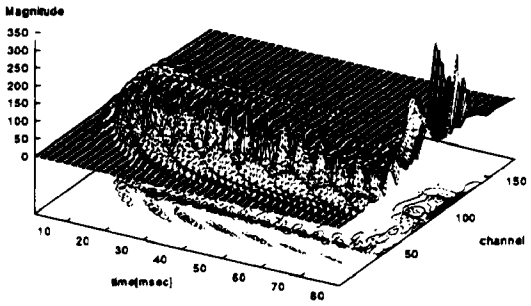
where $L(x)$, $R(x)$, and $\frac{1}{C(x)}$ represent, respectively, the mass, damping, and stiffness of the small section at x, and ρ, H, and α are the density of the lymph, the height of the scala vestibuli, and a constant, respectively.

Since the hair cells are known to produce electric potentials proportional to the velocity $\dot{\delta}$ of the basilar membrane only when the basilar membrane moves toward the tectorial membrane, the cochlea model output is defined as half-wave rectified versions of velocities of 175 equally divided sections of the basilar membrane. The output which the model produces when a speech signal is applied to it is considered to represent a kind of frequency spectrum of the signal in the form of 175-dimensional vector.

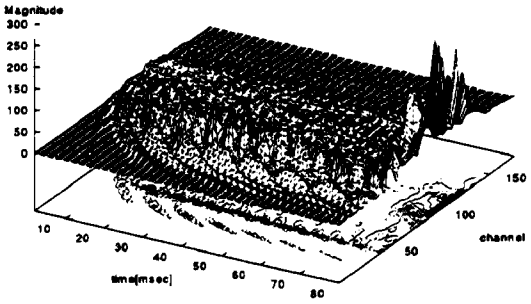
2.2 The recurrent neural network

Figure 2 exemplifies the structure of the three-layered recurrent neural network with feedback connections of self-loop type around output layer units, which is used in the system. This network has been found to surpass other networks of feedforward type and the recurrent network with feedback connections around hidden layer units [3], since this particular feedback structure provides the network with ability to store up incoming time-varying informations.

The cochlea model output for a phoneme input, which is a 175-dimensional vector, is fed to the network to decide on the class of the phoneme.



(a)



(b)

Figure 3: The cochlea model output in the form of sound spectrogram for (a)clean and (b)noisy(SNR 0dB) plosive sounds /b/.

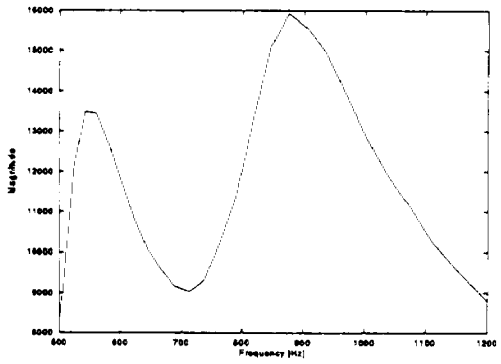
3 SPECTRAL ANALYSIS BY THE MODEL

3.1 Output of the cochlea model

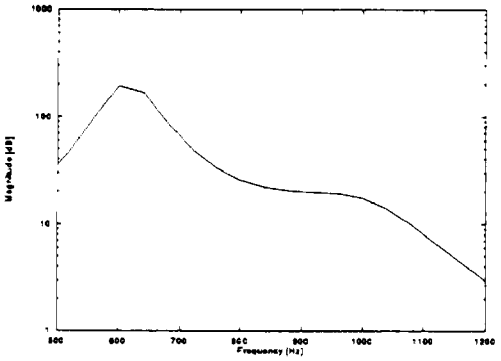
Figure 3 shows the cochlea model output in the form of sound spectrogram for clean and noisy (SNR 0dB) plosive sounds /b/. The effect of noise is found to be very little. This demonstrates the robustness of the model for noise. This robustness can be ascribed to the feedback structure of the model which serves to effectively reduce noise power and to enhance signal power.

3.2 Vowel spectra

Figure 4(a) shows a frequency response of the model to a vowel /o/ and Figure 4(b) is an LPC spectrum of the same vowel. In the frequency response of the model the first and second formant peaks can be clearly seen to be separated, while in the LPC spectrum the two formant peaks merge into



(a)



(b)

Figure 4: (a)A frequency response of the model to a vowel /o/. (b)An LPC spectrum of the same vowel.

a single peak. This demonstrates the formant discriminating power of the cochlea model.

3.3 Fast spectral analysis

To obtain the frequency response of the model requires solving a great number of differential equations simultaneously. So computation time is a serious problem.

There is, however, a way of calculating the frequency response fast enough to render a real time spectral analysis of speech sounds possible. As is clear from Eq.1, the model output is given as the convolution in time domain of model input and the impulse response, i.e., the inverse Laplace transform of $G_x(s)$. If the model input is decomposed into a number of frequency components by means of the fast Fourier transform, then it will be possible to analytically evaluate the convolution integrals, reducing considerably computation time. Through the use of this method one can calculate very quickly the frequency response of the model.

4 PHONEME RECOGNITION EXPERIMENT

To evaluate the ability of the system a phoneme recognition experiment was performed on the proposed phoneme recognition system. The method and result of the experiment will be described below.

4.1 Phoneme data

A set of phoneme tokens was derived from a Japanese word database provided by ATR Interpreting Telephony Laboratories, Kyoto, Japan. This database contains 5,240 Japanese words digitized at 20kHz, which were uttered by a single male speaker in a quiet environment.

Figure 5 demonstrates how phoneme vectors are generated from power spectra of phoneme tokens which the cochlea model puts out as 175-dimensional vectors every 2ms. By replacing every five successive components of each of such vectors with their average these vectors are reduced to 35-dimensional power spectral vectors. A time window whose length is 26ms is shifted over those 35-dimensional vectors, 8ms at a time. Thirteen windowed vectors are added up to yield a new 35-dimensional vector. Thus this window produces a 35-dimensional vector at each position and 7 such vectors altogether for each phoneme token. Three such vectors are concatenated to yield a 105-dimensional vector. This process produces five 105-dimensional vectors altogether for each phoneme token and those vectors which express the spectral variation of the phoneme token under consideration are used as input vectors for the recurrent neural network.

4.2 The phoneme recognition experiment

The phoneme recognition experiment was carried out to evaluate the phoneme discriminating power of the system in the absence of noise and its robustness for noise, using clean and noisy phoneme tokens. Table 1 shows a result of the experiment performed using clean phoneme tokens of 5 Japanese vowels and 18 consonants and compares the average recognition rate (%) of the system comprising the cochlea model and the recurrent neural network with that of two similar phoneme recognition systems in which the cochlea model is replaced with spectral analysis by the discrete Fourier transform (DFT). In one of those systems the recurrent neural network (RNN) is also replaced with a conventional feedforward neural network (NN). All the details of those neural networks are given in Table 2.

Table 1 clearly indicates that the RNN surpasses the NN in phoneme recognition accuracy. This may suggest that feedback loops within the RNN function as memories for storing up time-varying spectra of speech sounds and improve appreciably the phoneme discriminating power of the network.

It was found that the system with the cochlea model is able to correctly recognize vowels, even when two adjacent formant peaks of them merge into a single peak.

Figure 6 shows a result of the recognition experiment performed for voiced plosives /b/, /d/, and /g/ in the presence of white noise. It is clear from the figure that compared with the system using the spectral analysis by the DFT, the system using the cochlea model achieves not only higher recognition accuracies but a less deterioration due to noise of recognition accuracy. Under the SNR of 10dB, the deterioration of recognition rate is 3.1% with the system using the cochlea model, while 44.9% with the system using the DFT, and the recognition rate obtained with the former is 27.9% higher than the recognition rate obtained with the latter. One of the causes for this greater deterioration of recognition rate with the system using the DFT seems to be preemphasis which is needed to provide the spectral analysis by the DFT with a proper high frequency enhancement.

Table 1: Comparison of average recognition rates(%).

	cochlea model	DFT	
	RNN	RNN	NN
5 vowels	99.4	99.1	97.0
18 consonants	85.3	85.1	63.3

Table 2: Description of the neural networks.

		cochlea model	DFT	
		RNN	RNN	NN
dimensionality of input vector		105 (35x3)	112 (16x7)	112 (16x7)
number of units	input layer	105	112	112
	hidden layer	40	40	26
	output layer	5,18	5,18	5,18

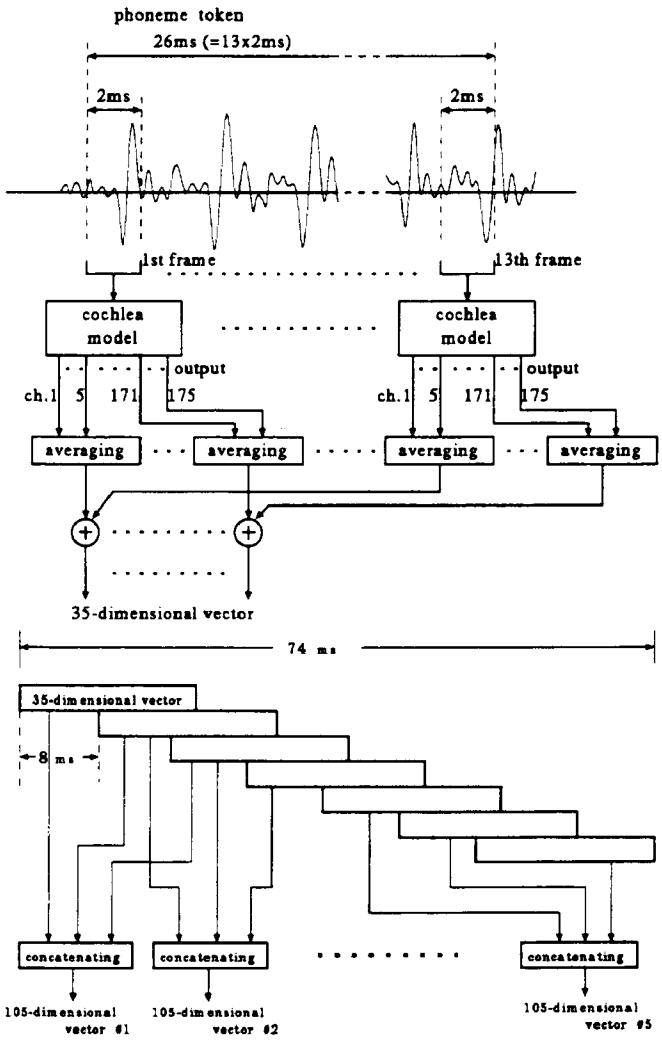


Figure 5: Generation of 105-dimensional phoneme vectors from phoneme tokens.

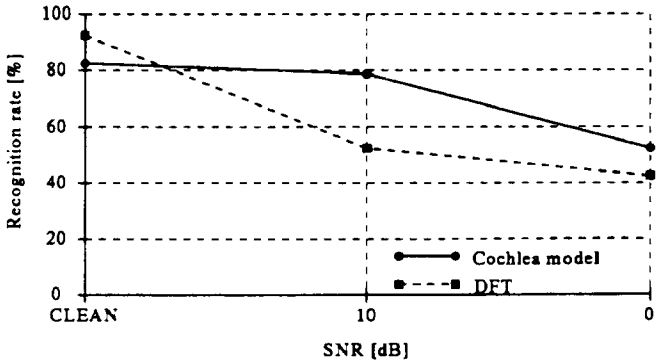


Figure 6: Average recognition rates of the system in the presence of noise.

5 CONCLUSIONS

A new phoneme recognition system based on a model of human auditory system has been described above. For the purpose of evaluating the ability of this system a phoneme recognition experiment was performed using a number of phoneme tokens derived from a Japanese word database. Findings from the result of the experiment can be summarized as follows:

1. The cochlea model is capable of resolving merged formant peaks of vowels. This leads to a correct recognition of vowels which have such formant structures.
2. The cochlea model has a higher formant resolving power than the DFT, which leads to higher recognition rates of the system for vowels.
3. The system using the cochlea model achieves higher recognition rates in the presence of white noise and a less deterioration due to noise of recognition rate than the system using the DFT.
4. A fast algorithm is available for calculating the frequency response of the cochlea model to render a real time spectral analysis of phoneme tokens possible.

REFERENCES

[1] Taniguchi, S., Koizumi, T., and Shimizu, R. "A Nonlinear Feedback Model for Cochlea," *J. Acoust. Soc. Japan*, 47: 259-267, 1991.

[2] Koizumi, T., Mori, M., and Taniguchi, S. "Recurrent Neural Networks for Phoneme Recognition," *Proc. Int. Conf. Spoken Language Processing*, 1996.

[3] Watrous, R.L., Ladendorf, B., and Kuhn, G. "Complete gradient optimization of a recurrent network applied to /b/, /d/, /g/ discrimination," *J. Acoust. Soc. Am.*, 87: 1301-1309, 1990.