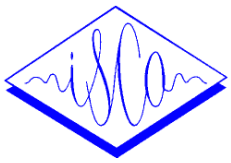

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

FORMANT EXTRACTION MODEL BY NEURAL NETWORKS AND AUDITORY MODEL BASED ON SIGNAL PROCESSING THEORY

Kazuaki OBARA **Hideyuki TAKAGI**

Central Research Laboratories,
Matsushita Electric Industrial Co., Ltd.
3-15, Yagumo-Nakamachi, Moriguchi, Osaka 570 JAPAN
TEL <+81>6-909-1121, Facsimile: <+81>6-906-0177
e-mail: obara@atr-hr.atr.co.jp takagi@it4.crl.mei.co.jp

ABSTRACT

This paper discribes a formant extraction model constructed by an auditory model for preprocessing and a neural networks (NN) that extracts formant. This auditory model consists of the first stage based on physiological foundings and the second stage based on a hypothetic function. The first stage is consisted of a basilar membrane model of 190ch critical band pass filter bank (CBF), hair cell model of half-wave (HW) rectification and satulating property, and synapse or axon model of low pass (LP) characteristics. The second stage has hypothetical function that has homomorphic processing. Finally, a NN for formant extraction and its experiment are described. This NN extract formant from tge output of auditory model. The resuls displayed in this paper shows that our formant extraction model performs well for not only supervised data but also unsupervised data.

I. INTRODUCTION

In the field of speech recognition, the researches on methods of HMM and NN have been growing. In any methods, the performance depends on feature parameter. There are two research directions of feature parameter; that is, the parametric method represented by LPC parameter and another method represented by auditory model. The former method features mathematical representation. On the contrary, it is very difficult to find the breakthrough if the model reaches the limit, as the previous speech recognition research shows. The latter method has reliability on its direction in the sense that it can proceed the analysis on the basis of human who has the mechanism to recognize speech. The difficulty is to keep up with the physiological and psychological analyses.

This paper intends to find the breakthrough of speech process from the latter method's standpoint. At this standpoint, can we solve the problem that we just follow the physiological and psychological findings? The basis of this paper is that positive introduction of a hypothesis for physiological function analysis leads the progress of feature extraction model. Additionally, the auditory model research may be of benefit to physiology as the relation between theoretical physics and experimental physics.

Based on this concept, this paper proposes and evaluates the model adopting the hypothetic function reasoned from physiological findings and signal processing theory. We take formant extraction model in the concrete. That is, we study physiological findings from auditory peripheral and central nervous system, then assume lifter function in signal processing field between peripheral and central as the hypothetic function. Comparing the model that includes the hypothetic function with the one that does not include it shows how the hypothetic function has an effect on formant extraction. Section 3 describes the experiment of formant extraction NN using the output of the proposed auditory model. It can be considered that this NN has lateral inhibition function to detect peak on spectrogram. By the output of this NN, the effectiveness of the hypothetic function is proved.

II. AUDITORY MODEL DESIGNED ON SIGNAL PROCESSING THEORY

This section explains the auditory model that executes homomorfic analysis by considering the auditory peripheral function from the viewpoint of signal process theory. The auditory model consists of two stages. Stage I is based on physiological findings, while stage II is the hypothetical function.

The physiological findings of Stage I are followings: frequency analysis function of basilar membrane, HW rectifier and saturating of hair cell, and LP characteristics of synapse or axon.

Stage I is modeled by these physiological findings. First, basilar membrane is represented by 190ch critical band filters [1]. Decay characteristics of each filter is asymmetric in low-frequency side and high-frequency side.

Second, HW rectifier of hair cell, saturating characteristics, and LP characteristics of synapse or axon are modeled. This LP characteristics is a model of reduction of synchrony in nervous responses as stimulus frequency is increased. This reduction is said to caused in synapse because of ion diffusion process [2], and is also said to caused in axon because the Hodgkin-Huxley Equ. has LP characteristics [5]. Typical model modeled them one by one. On the other hand, we consider that LP characteristicsof synapse after HW rectifier corresponds to power calcuration. Then the squired aquired power is calculated instead of (HW rectifier of hair cell + LP characteristics of synapse). That is, we define stage I of auditory model as (CBF + the squired power + saturating characteristics) instead of (CBF + HW rectifier + LP filter + saturating characteristics) represented by the model of S.Seneff [2].

Next, we discuss stage II based on the hypothetic function for the proposed auditory model. The neuron that selectively reacts to formant is found in auditory cortex of cats [3]. This is regarded that there is a function to facilitate formant extraction between auditory peripheral and central nervous system. On the basis of the above signal process theory, our proposed auditory peripheral model corresponds to (Fourie transform (CBF) + power (HW rectifier + synapse) + saturating characteristics). If we appropriate this saturating characteristics as logarithmic, the process to facilitate formant extraction in auditory pass can be explained as following hypothesis: "The function to faclitate formant extraction is LP filter of frequency axis, that is, LP lifter, and formant is extracted by homomorphic processing between auditory peripheral and central nervous system".

We assume the hypothetic LP lifter function in auditory system and define the model added this hypothetic function shown in Fig. 1 as the auditory model of formant extraction.

Fig.2 shows how much the hypothetic function of stage II effects formant extraction. Compared with the output only from Stage I shown in Fig.2(a), Fig.2(b) shows the output of model added the hypothetic function facilitates formant extraction with removing pitch and its harmonics.

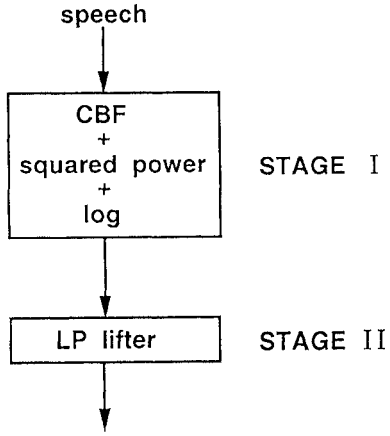

Fig.1 Proposed auditory model

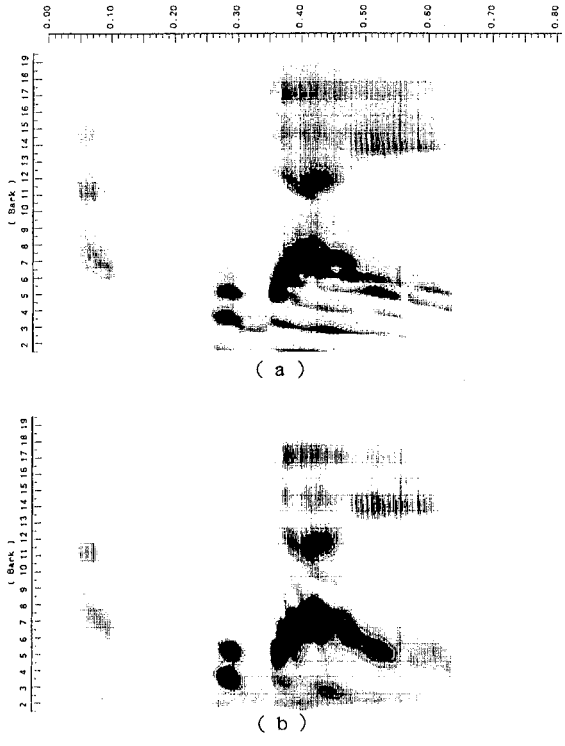

Fig.2 Output of auditory model
(a) output of Stage I
(b) output of Stage II

III. FORMANT EXTRACTION BY NN

We previously conducted formant extraction experiment with three NN models [4]. This experiment proved the necessity of time information, usefulness of relative spatio-temporal information and effect of prewriting of NN [4]. In this paper, we conduct an experiment considering the result of the above experiment and applying formant extraction of independent speaker. The model for this experiment is NN that only uses relative spatio-temporal information [4]. The size of spatio-temporal is expanded more than the one of the previous experiment. The NN that uses relative spatio-temporal information learns to determine whether the center is formant or not only by input spatio-terminal information. Namely, special frequency component to each unit in input layer is not allocated, and auditory model output of spatio-temporal is swept. This NN process corresponds to lateral inhibition because it selectively emphasizes only peak of spatio-temporal pattern.

The NN of this experiment has 3.5 bark × 50 ms spatio-temporal input layer (7 × 5 units). NN is feed-forward type that fires if formant exists in the center of spatio-temporal pattern. In this experiment, 173,170 training data from 80 words of one male speaker is input to the auditory model explained in section 2, then the output is input to NN for formant extraction.

Fig.3 shows the output of open data for the NN. In Fig.3, (a) is DFT spectrogram, (b) is the output of the proposed auditory model, and (c) is the output of formant extraction NN. Although the channel resolution of the auditory model is 0.1 Bark, one of this NN is 0.5 bark. Therefore, display accuracy for extraction result is five times as fuzzy as the former case. Considering this difference, formant extraction NN proves high capability of extraction to open data as well as training data. Fig. 4 shows the result of applying formant extraction NN to open data of words of a female. Pitch harmonics remains because cutoff cefrency of Stage II in Fig.1 is still for male speech. However, power peak including pitch harmonics is extracted almost accurately. This experiment proves formant extraction NN is available for female speech that has different formant information. This is the effect of NN model that determines formant only from relative spatio-temporal information.

IV. DISCUSSION

There are two features of the proposed auditoy model. One of them is the modeling concept. Authers have modeled the physiological findings by signal processing theory based on functional interpretation. Besides the conventional methods modeled the auditory physyological findings one by one. In the proposed model, HW rectifying of hair cell and LP filtering of synapse are interpreted as power extraction function. Then we proposed the hypothesis of homomorfic analysis in auditory peripheral.

The other feature is that we constructed the model from the viewpoint of homomorfic analysis by assuming the hypothetic function called LP lifter to facilitate formant extraction. This function is based on the existence of neuron reacting to formant in auditory cortex and findings in peripheral. We must wait for the progress of physiological study to determine if the hypothetic function is really existent, but there is possibility that the field of auditory physiology can take benefit from this hypothetic function.

It is possible to expand the proposed model not only to formant extraction but also other functions by changing the hypothetic function of Stage II. For example, HP lifter is available for Stage II in Fig. 1 in case of pitch extraction model. By introducing other homomorfic analysis, the auditory model that extracts other features can be realized.

V. CONCLUSION

In this paper, we assumed the hypothetic function for LP liftering and constructed the auditory model by physiological findings of auditory peripheral and formant selective neuron in auditory cortex. The model output that seems to facilitate formant extraction has been obtained. Using the output of this model, the NN extracted formant. As a result, NN has extracted formant as a lateral inhibition model. Furthermore, the experiment proved fine generalization by NN structured for independent speaker.

REFERENCES

[1] T.Komakine, et al., "A filter-bank design for simulating cochlear frequency analysis function", IEICE Tech. Report SP-87-45, 1986 (Japanese)

[2] S.Seneff, "A joint synchrony/mean-rate model of auditory speech processing", J. Phonetics, 16, 1, pp.55-76, 1988

[3] N.Maruyama, et al., "Unit responses of the cat's auditory cortex to synthesized formants", Proc. Japan Acad., 55, Ser. B, 1979

[4] H.Takagi, et al., "Pseudo-Formant Extraction by Neural Net", Spring meeting of the Acous. Soc. of Japan 3-P-11, pp.249-250, 1988 (Japanese)

[5] Y.Horikawa, "Filtering Properties due to Velocity Dispersion on an Axon", Trans. of IEICE, J72-D-II, 4, pp.621-629 (1989)

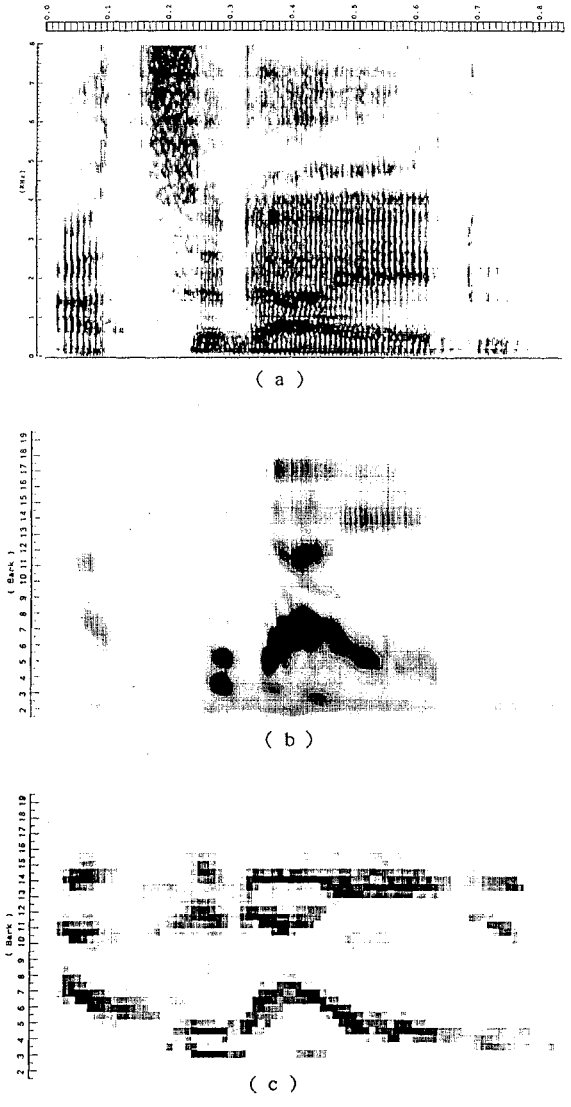

Fig.3 Formant extraction for open word male spoke
(a) DFT spectrogram
(b) output of proposed auditory model
(c) output of the formant extraction NN

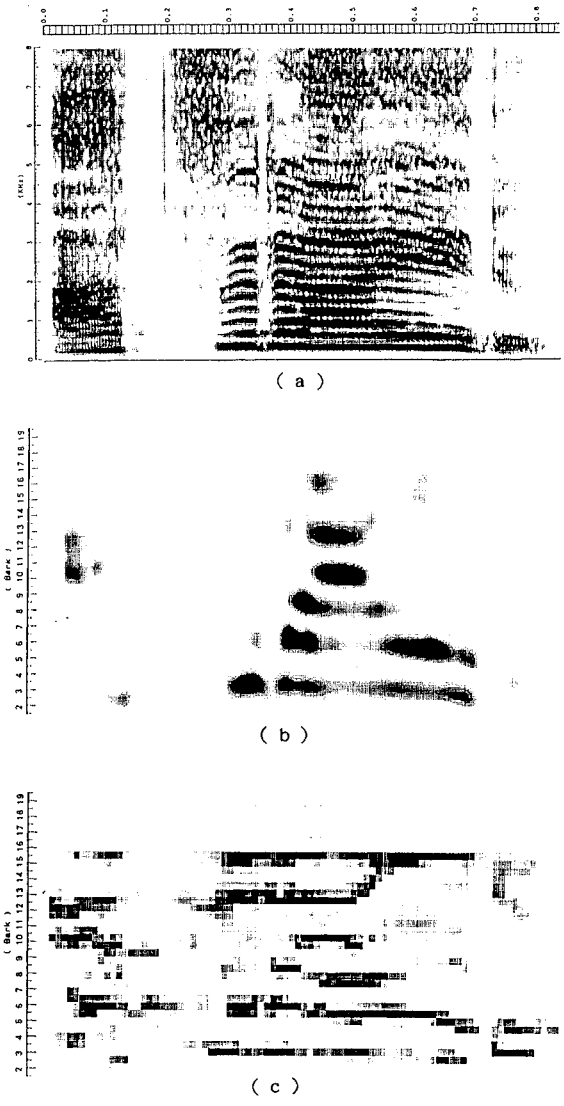

Fig.4 Formant extraction for open word female spoke
(a) DFT spectrogram
(b) output of proposed auditory model
(c) output of the formant extraction NN