Speaker Identification in Noisy Environment with Use of the Precise Model of the Human Auditory System

Tadahiro Azetsu, Masahiro Abuku, Noriaki Suetake and Eiji Uchino

***Abstract*—This paper discusses an approach for speaker identification in noisy environment using the multi-dimensional pulse signals generated from the model of a human peripheral auditory system. The peripheral auditory model employed here consists of a basilar membrane, hair cells, and auditory nerves. The input to this model is a speech signal divided into frames, and the outputs of which are the multi-dimensional pulse signals for each framed signal. The feature vectors based on the post-stimulus time histogram (PSTH) of the pulse signals are used for the speaker identification. In this paper, we propose to set adaptively the threshold of the action potential for pulse generation in the auditory nerve model. In order to verify the performance of noise immunity for the speaker identification, the experiments were conducted for each Japanese vowel spoken by 12 speakers (9 males and 3 females). The effectiveness of using the peripheral auditory model has been verified by comparing with the methods using the conventional LPC spectrum and using the excitation patterns.**

***Index Terms*—Peripheral auditory system, Multi-dimensional pulse signals, Post-stimulus time histogram, Speaker identification, Excitation pattern.**

I. Introduction

THE human auditory system has a high ability to perform complex signal processing tasks. In general, the speech signal of conversation is often corrupted by various noises such as other speech signals, traffic noises, background noises, and so on. However, the human can communicate with each other by paying attention only to the necessary information even under those noises. This is called a cocktail party effect in speech signal processing field [1].

The speech processing system based on the statistical model, which is widely used at present, has indeed a high performance for the normal cases [2]. However, in an extremely noisy environment, it is fairly inferior to the human auditory system which flexibly adjusts to the environment. Therefore, in order to evolve the conventional speech signal processing system into a more useful human interface, it is absolutely necessary to refer to the signal processing way of the human auditory system [3].

In this paper, a precise model of the human peripheral auditory system is firstly composed by combining several conventional component-models, and secondarily the speaker identification by vowel in noisy environment is performed by using the outputs of this model.

In this paper, we propose to set adaptively the threshold of the action potential for pulse generation in the auditory nerve model. A time average of membrane potential and characteristics of the auditory nerve are used.

The effectiveness of the proposed method has been verified by comparing with the methods using the LPC spectrum and using the excitation patterns. The excitation pattern is a spectral information derived by the auditory filter as a function of the center frequency of the filter bank [4]. Experimental results have shown that the proposed method using the human peripheral auditory model has high noise immunity comparing with the other methods.

This work was supported by the Grant-in-Aid for Challenging Exploratory Research of the Japan Society for Promotion of Science (JSPS) under the Contract No. 21650039.

T. Azetsu is with the Yamaguchi Prefectural University, 3-2-1 Sakurabatake, Yamaguchi 753-8502, Japan (e-mail: azetsu@yamaguchi-pu.ac.jp).

M. Abuku, N. Suetake, and E. Uchino are with the Graduate School of Science and Engineering, Yamaguchi University, 1677-1 Yoshida, Yamaguchi 753-8512, Japan (e-mail: {p002vc, nsuetake, uchino}@yamaguchi-u.ac.jp).

E. Uchino is also with the Fuzzy Logic Systems Institute, 680-41 Kawazu, Iizuka 820-0064, Japan (e-mail: uchino@flsi.or.jp).

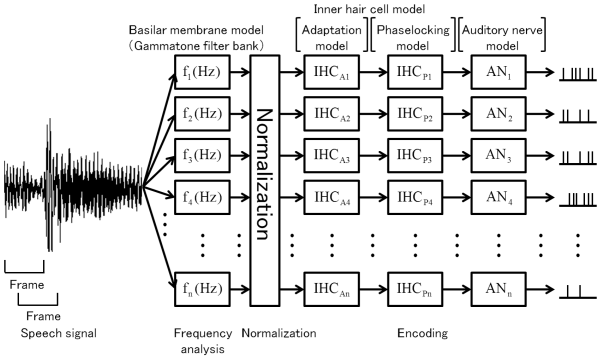

Fig. 1. The schematic diagram of the model of the human peripheral auditory system.

II. Model of human peripheral auditory system

The peripheral auditory system for the speaker identification is modeled here by cascading the models of a basilar membrane, hair cells, and auditory nerves. Figure 1 shows the schematic diagram of this model.

A. Basilar membrane model

The vibration of the basilar membrane is caused by an acoustic stimulation. This vibration can be interpreted as a kind of function to map the frequency of the acoustic stimulation to the position of the basilar membrane.

The frequency characteristics of the basilar membrane can be realized technologically by the filter bank. In this paper, the Gammatone filter [5], which is specially designed to describe the auditory activities of the human basilar membrane, is used.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

The impulse response of the Gammatone filter is given by:

$$g(t) = at^{n-1}\exp(-2\pi bt)\cos(2\pi f_c t + \phi), \quad (1)$$

where a, b, and n are the parameters which determine the shape of the impulse response. f_c and ϕ are a center frequency and a phase, respectively. The frequency characteristic of this filter is that the bandwidth becomes wide as the frequency goes high.

B. Hair cell model

The Meddis inner hair cell model is a neural transduction model between the hair cell and the auditory nerve, which are located in the cochlea [6][7]. Comparing with other hair cell models, the Meddis' model is described by simple mathematical expressions, and it can reflect the activities of the auditory system better than other models, e.g., an adaptation effect is well embedded in the Meddis' model.

In this paper, for that reason we employ the Meddis inner hair cell model. This model consists of three reservoirs and one factory. Three reservoirs are called free transmitter pool, synaptic cleft and reprocessing store.

Let $q(t)$, $c(t)$, $w(t)$ be amounts of transmitters in the free transmitter pool, the synaptic cleft, and the reprocessing store, respectively. Figure 2 shows a signal flow of the Meddis inner hair cell model. The numbers in Fig.2 correspond to the numbers of the following explanations of the function of the Meddis inner hair cell model.

1. The acoustic stimulation $S(t)$ is applied to the model.
2. The permeability $k(t)$ is determined depending on $S(t)$.
3. The transmitters in the free transmitter pool are released into the synaptic cleft according to the value of $k(t)$.
4. The transmitters in the synaptic cleft stimulate the auditory nerves. In this model, it is assumed that the firing probability is proportional to $c(t)$. $h \cdot c(t)$ is an output of this model, which is a generation probability of pulse per second.
5. Some of the transmitters in the synaptic cleft are lost, and the remainder are collected in the reprocessing store.
6. The transmitters in the reprocessing store are transferred to the free transmitter pool after reprocessed.
7. The free transmitter pool is replenished with the transmitters from the factory depending on the amount of the transmitters lost in the synaptic cleft.

The Meddis inner hair cell model is described by the following three differential equations and by one probabilistic function:

$$k(t) = \begin{cases} \frac{g(S(t)+A)}{S(t)+A+B} & \text{for } S(t) > -A \\ 0 & \text{for } S(t) \le -A, \end{cases} \quad (2)$$

$$\frac{dq}{dt} = y\,(1 - q(t)) + xw(t) - k(t)q(t), \quad (3)$$

$$\frac{dc}{dt} = k(t)q(t) - lc(t) - rc(t), \quad (4)$$

$$\frac{dw}{dt} = rc(t) - xw(t), \quad (5)$$

$$\text{Prob(event)} = hc(t). \quad (6)$$

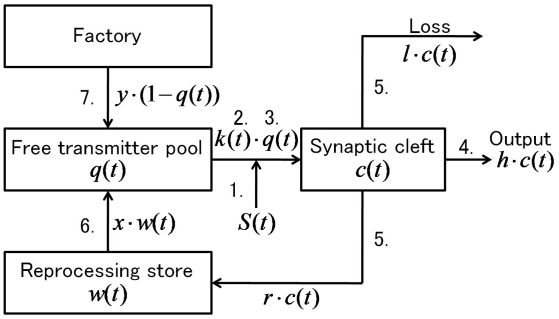

Fig. 2. The signal flow of Meddis inner hair cell model.

This model has seven parameters of y, x, l, r, g, A, and B. $-A$ is the lowest acoustic stimulation at which the transmitter is released.

The phase-locking model [8] is followed by this Meddis inner hair cell model, because Meddis' model doesn't have a phase-locking property.

C. Auditory nerve model

The membrane potential of the auditory nerve increases as the transmitters are released into the synapse between the inner hair cell and the auditory nerve. The auditory nerve generates a neural pulse when the potential exceeds a certain threshold. And after that, this potential decreases with time. The auditory nerve has a refractory period with a certain length (about 1ms), incapable of action after the neural pulse is generated [9].

The above procedures of the membrane potential generation are realized by the simple functions as follows.

1) Modeling of membrane potential generation: In this paper, the model proposed by [8] is used as the generation model of the membrane potential. The generation of the membrane potential is modeled by the simple function te^{-t}, which decreases right after the rapid increase of potential.

2) Modeling of pulse generation of auditory nerve: The pulse of the auditory nerve is generated when the membrane potential exceeds a certain threshold, which is not of course in the refractory period. This process is expressed as follows:

$$S_i(t) = \begin{cases} 1 & V_i(t) \ge U_i \\ & \text{and } S_i(t') = 0 \text{ for } t' \in [t - t_r, t] \\ 0 & \text{otherwise} \end{cases}, \quad (7)$$

where V_i is a membrane potential at the i-th channel (auditory nerve), U_i is a threshold, and t_r is a refractory period. The pulse signal $S_i(t)$ expressed by Eq.(7) is the output of this peripheral auditory model.

The frequency axis is represented by [ERB-number] which is based on the characteristics of the human auditory system [10]:

$$\text{ERB} = 24.7(0.00437F + 1.0)[\text{Hz}], \quad (8)$$

$$[\text{ERB-number}] = 21.4\log_{10}(0.00437F + 1.0), \quad (9)$$

where F is a center frequency [Hz]. The channel is placed between [ERB-number]=3 (87.39 Hz) and [ERB-number]=41.48 (19,780 Hz) at equal intervals. The number of the channels is set to 331 in this paper.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

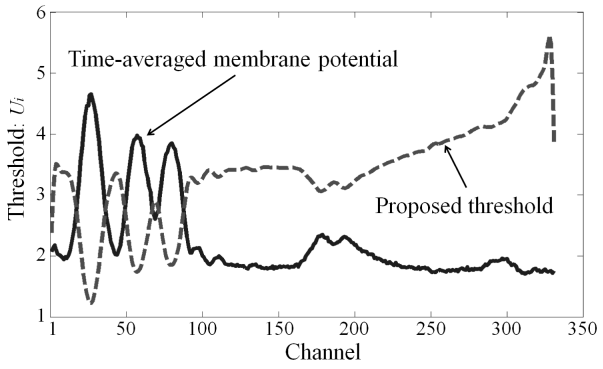

Fig. 3. The proposed threshold obtained from the time-averaged membrane potential.

D. Post-stimulus time histogram (PSTH)

In the raster display of pulse signals, the time axis is divided into a short interval with a constant width. The representative value of the post-stimulus time histogram is the average of the number of pulses in each interval [11]. It is generally abbreviated as PSTH, which is a typical statistic information for pulse signal.

III. Proposed setting of threshold U_i using membrane potential

The threshold U_i of the action potential for the pulse generation in the auditory nerve model is an important factor influencing the patterns of the multi-dimensional pulse signals generated. In this paper, U_i is determined by the following three steps using a time average of the membrane potential and the characteristics of the auditory nerve.

A. Time average of membrane potential

First, the time average of the membrane potential is calculated for each channel. The following R_i, which is mainly composed of the time average $< V_i(t) >$, is introduced to determine U_i:

$$R_i = - < V_i(t) > + 2C \frac{1}{n} \sum_{i=1}^{n} < V_i(t) >, \tag{10}$$

where $< \cdot >$ is a time average operator and n is the number of the channels. C is a constant value, which is set to 1.2 in this paper. R_i has the effect to generate pulses more frequently at the channel having a large average of the membrane potential.

B. Lateral inhibition of auditory nerve

The auditory nerve has the characteristics of the lateral inhibition. The lateral inhibition can be represented by the DOG (Difference Of two Gaussians) filter defined by the following equation:

$$g_i^{DOG} = \frac{1}{2\pi\sigma_e^2} e^{-\frac{i^2}{2\sigma_e^2}} - D \frac{1}{2\pi\sigma_d^2} e^{-\frac{i^2}{2\sigma_d^2}}, \tag{11}$$

where σ_e, σ_d, and D are the parameters of the DOG filter, whose values are 2.0, $1.6\sigma_e$, and 1.0, respectively. The DOG filter is used in the determination of U_i, which improves the frequency resolution.

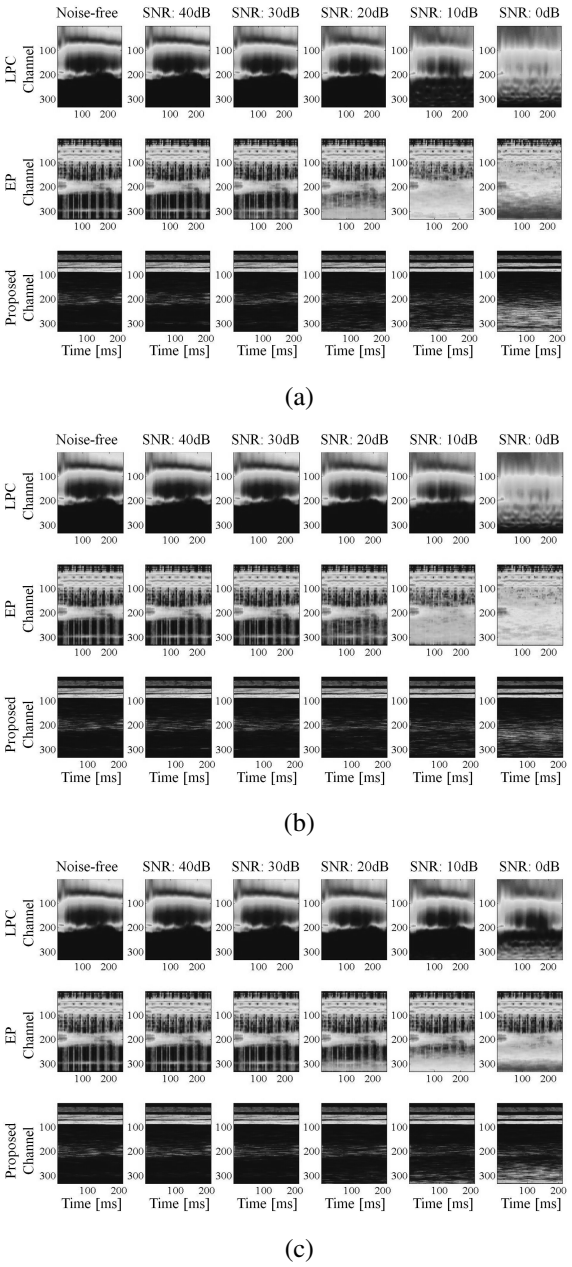

Fig. 4. Output patterns of Japanese vowel "e" of a certain speaker obtained from each method under the following noises. (a) White noise. (b) Pink noise. (c) Blue noise. LPC: LPC spectrum, EP: excitation patterns, and Proposed: proposed human auditory model.

C. Suppression of pulse generation in high frequency region

In case of the fixed U_i, it is experimentally confirmed that the pulse generation in the high frequency region occurs more frequently than in the low frequency region. The following W_i is then introduced to suppress the pulse generation in the high frequency region, by using the high order polynomial with respect to the channel number i:

$$W_i = \frac{1}{2} \left(\frac{i}{n} \right)^a + 1, \tag{12}$$

where a is a constant, which is set to be 6 in this paper. The threshold U_i is finally determined by the following equation:

$$U_i = W_i (R_i * g_i^{DOG}). \tag{13}$$

Figure 3 shows the threshold obtained from the time-averaged membrane potential by the above mentioned pro-

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE I
SPEAKER IDENTIFICATION RATES BY EACH METHOD IN THE LOW FREQUENCY REGION UNDER WHITE NOISE [%]

LPC (LPC spectrum)	EP (Excitation patterns)	Proposed (Human auditory model)
11.3	55.3	75.0

cess.

IV. EXPERIMENTAL RESULTS

In order to verify the noise immunity of the proposed method for the speaker identification, we have performed the experiments using 5 sets of 5 Japanese vowels spoken by 12 speakers (9 males and 3 females).

The vowels are corrupted by three noises, i.e., white noise, pink noise, and blue noise. The speaker identification accuracy is evaluated for each vowel. The subspace method [12] is used as a pattern recognition method.

For comparisons, the speaker identification by using the LPC spectrum and by using the excitation patterns are also performed. Figure 4 shows the output patterns of Japanese vowel "e" of a certain speaker obtained from each method under noises. From those results, it is observed that the proposed method has a higher noise immunity than the other methods.

Figure 5 shows the speaker identification rates versus SNR by each method under noises. In case when the noise level is low, the human peripheral auditory model has a less performance than the other methods. However it is better than the other methods as SNR decreases.

Further from Fig.4, it is observed that the PSTH of the peripheral auditory model is less affected by noise in the low frequency region. Thereupon, Table 1 shows the speaker identification rates using the low frequency region (from [ERB-number]=3 (87.39 Hz) to [ERB-number]=24.6 (3,015 Hz) when SNR is 0 dB under white noise. It is confirmed the superiority of the peripheral auditory model.

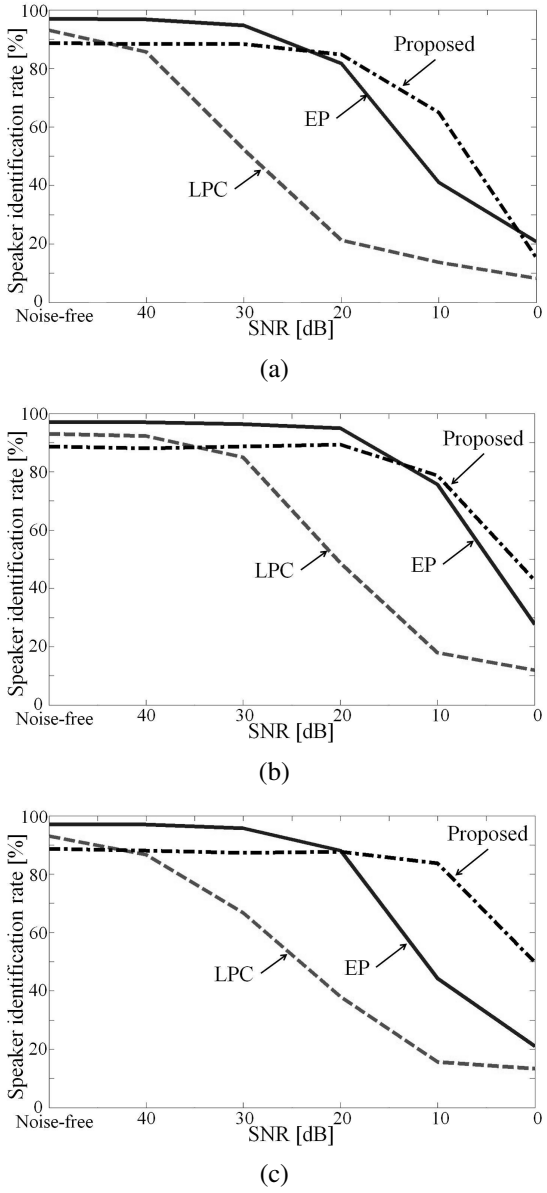

Fig. 5. Speaker identification rates versus SNR by each method under the following noises. (a) White noise. (b) Pink noise. (c) Blue noise. LPC: LPC spectrum, EP: excitation patterns, and Proposed: proposed human auditory model.

V. CONCLUSION

In this paper, the approach for the speaker identification in noisy environment by vowel, using the multi-dimensional pulse signals generated from the model of the human peripheral auditory system, was discussed. The experiments were conducted for each Japanese vowel spoken by 12 speakers (9 males and 3 females). The effectiveness of the proposed method was verified by comparing with the methods using the LPC spectrum and using the excitation patterns. The proposed method has a high noise immunity.

The future works are to conduct the experiments using the actual environmental noises.

REFERENCES

[1] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol.17, no.9, pp.1875–1902, 2005.
[2] T. Azetsu, E. Uchino and N. Suetake, "Blind separation and sound localization by using frequency-domain ICA," *Soft Computing*, vol.11, no.2, pp.185–192 2007.
[3] M. Abuku, T. Azetsu, E. Uchino and N. Suetake, "Application of peripheral auditory model to speaker identification," *Proceedings of the 2nd World Congress on Nature and Biologically Inspired Computing (NABIC 2010)*, pp.666–671, 2010.
[4] B. C. J. Moore, B. R. Glasberg and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *J. Audio Eng. Soc.*, vol.45, no.4, pp.224–239, 1997.
[5] R. D. Patterson and J. Holdsworth, "A functional model of neural activity patterns and auditory images," *Advances in Speech, Hearing and Language Processing*, vol.3, pp.547-563, 1996.
[6] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *J. Acoust. Soc. Am.*, vol.79, no.3, pp.702–711, 1986.
[7] R. Meddis, "Simulation of auditory-neural transduction: Further studies," *J. Acoust. Soc. Am.*, vol.83, no.3, pp.1056–1063, 1988.
[8] K. Maki, M. Akagi and K. Hirota, "Functional model of auditory peripheral system: Modeling phase-locking properties and spike generation process of auditory nerves," *J. Acoust. Soc. Jpn.*, vol.65, no.5, pp.239–250, 2009 (in Japanese).
[9] J. O. Pickles, *An Introduction to the Physiology of Hearing, 2nd ed.*, Academic Press, 1988.
[10] B. C. J. Moore, *An Introduction to the Physiology of Hearing, 5th ed.*, Emerald Group Publishing Ltd, 2003.
[11] W. A. Yost, *Fundamentals of Hearing, 5th ed.*, Academic Press, 2006.
[12] Y. Ariki, S. Tagashira and M. Nishijima, "Speaker recognition and speaker normalization by projection to speaker subspace," *Proceedings of 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, vol.1, pp.319–322, 1996.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)