Mathematical Modeling of Human Speech Processing Mechanism Based on the Principle of Brain Internal Model of Vocal Tract

Hiromi SAKAGUCHI * and Naoaki KAWAGUCHI **

Still in current speech reseach, has it been a longstanding problem that speech synthesis and recognition techniques are far from reaching a human speech processing level. Man can easily perceive a speaker-independent speech even under noisy enviroment and can simultaneously mimic its articulatory movement. This shows that there are quite systematic relation between perception and production of speech sound. We propose a mathematical modeling of the unified human speech processing system of perception and generation. A speech generation is modeled by the articulatory speech synthesis system G (fAi, s) . fAi shows a vocal tract shape. The response of an artificial phoneme perception is described as $r = f\left(\sum_{\omega} WX - H\right)$. W is the brain internal model of speech generation process, to which the articulatory speech synthesis system function G is applied. X is a speech signal. WX is the linear enhancement filtering for phoneme detection of speech X and the power spectrum of WX represent the correlation between the brain internal model of vocal tract and the actual vocal tract of X. f (•) shows some nonlinear processing. H is a threshold.

1 . Introduction

Various kinds of sophisticated speech processing techniques have been developed since the inception of LPC type analysis / synthesis telephony. Nevertheless, the quality of an artificial speech perception and generation is still inferior to human's ability. This is assumed to be brought by the essential difference in speech processing mechanisms between man and machine. Human speech processing is characterized by having an ability of speech mimic. That is, man can easily recognize a speaker independent speech and simultaneously mimic its utterance. And further,man can easily make a compensatory articulation even under such restriction as fixed mandible with bite block or pipe [1] . These show that speech generation and perception are the

* Associate Professor (Dept. of Mechanical Systems Engineering)
** Student of Master's program (Dept. of Mechanical Systems Engineering)

behavior of a unified processing system of both speaking and hearing with auditory and inner proprioceptive feedback [2]. Then, it can be thought that the ability of speech mimic never consists in signal processing of spectral information but in direct process of coding and decoding of vocal tract characteristics in brain neural networks. From this viewpoint, we have been trying to model mathematically a unified human speech processing system of perception and generation based on the principle of brain internal model of vacal tract. For this purpose, analysis by articulatory speech synthesis is the decisive measures. Articulatory synthesizer is the direct modeling of speech production process and it has been expected to implement high quality speech at low bit rates. But, there has been a long-pending problem of how to estimate the vocal tract shape parameters from natural input speech [3, 4]. This is also the same kind of problem as revealing the human speech mimic mechanism. One of the important objective of our speech research is to clarify the essential nature of phoneme information.

2. Analysis by Articulatory Speech Synthesis

Human speech production process can be directly modeled by an articulatory speech synthesis system provided that an optimum model and its articulatory parameter control are given. But, it is even now hard to implement the articulatory synthesis for a given speech. Because, the human speech processing mechanism have never been quantitatively or mathematically clarified. But the reseach of these mechanisms also depends on the analysis by articulatory speech synthesis shown in Fig.1. Then the optimization of synthesis simulation must be carried out based on the assessment by human auditory criteria.

2.1 Articulatory Speech Synthesizer

The process of human speech generation has been traditionally modeled as the planar wave propagation in vocal tract. This assumption at the first approximation is not denied for telephony frequency region. As to a glottal source, various models are presented. There have

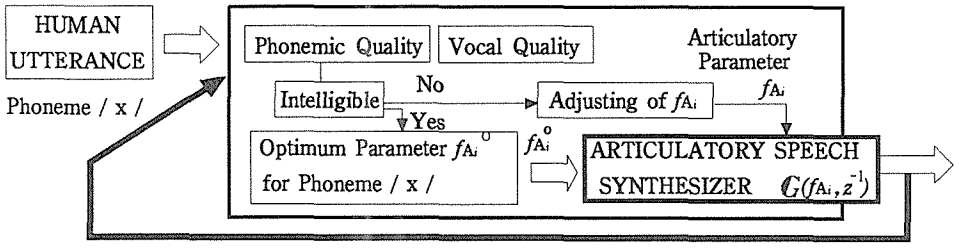


Fig.1 Schematic diagram of Analysis by Synthesis

been several approaches to implementing speech synthesis [4]. Here, we present a newly digitalized system directly transformed from the fundamental wave equation by I(z) transformation (z transform method by means of a digital integrator) [5, 6]. We need to deal with the wave equation as simultaneous formulae of continuity and motion in order to analyze acoustical phenomena faithfully even for nasal and fricative sounds besides vowels. The original wave equation is described as a following pair of formulae.

$$\left.\begin{aligned} -\frac{\partial p}{\partial x} &= \frac{\rho}{S(x)} \cdot \frac{\partial u}{\partial t} + \frac{\rho Re}{S(x)} u \qquad \text{(formula of motion)} , \\ -\frac{\partial u}{\partial x} &= \frac{S(x)}{\rho c} \cdot \frac{\partial p}{\partial t} \qquad \text{(formula of continuity)} , \end{aligned}\right\} \quad (1)$$

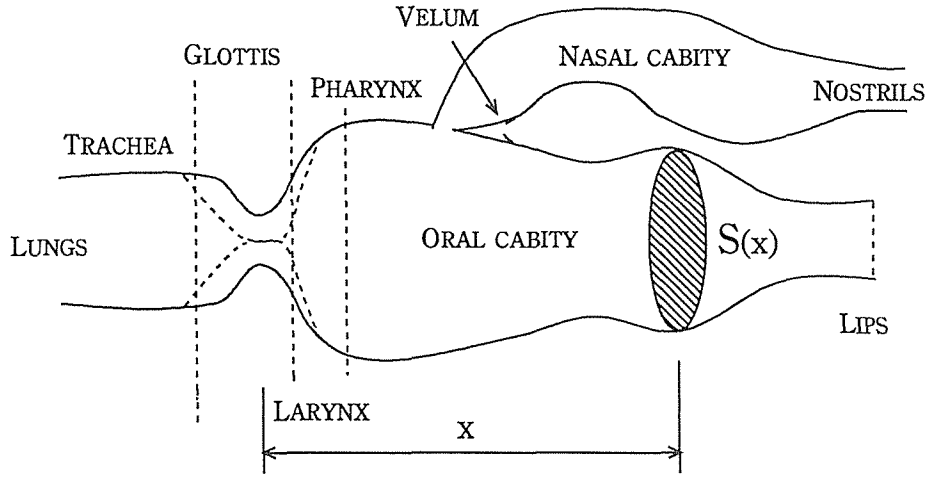
where,

x : distance from glottis — S(x) : vocal tract area

p : sound pressure — u : volume flow velocity

ρ : air density — c : sound flow velocity

Re : damping coefficient .



Schematic representation of the vocal system

The transform of the above analog system Eq.(1) to a digital sytem can be described as follows. At first, the vector differential equation is derived as shown in Eq.(2) by I(z) transformation for space variable of Eq.(1). Next, the final difference equation for time variable of Eq.(2) can

be derived by the same z transformation. This difference equation (Eq.(3)) can be utilized for an articulatory speech synthesis.

・ Vector differential equation

$$\left.\begin{aligned} &A\begin{pmatrix}\dot{u}\\ \dot{p}\end{pmatrix}+B\begin{pmatrix}u\\ p\end{pmatrix}=c_1\dot{u}_g+c_2\ddot{u}_g\ ,\\ &p(t)=C^T\begin{pmatrix}u\\ p\end{pmatrix}\quad\text{(output speech signal)}\ ,\end{aligned}\right\}\qquad(2)$$

$$\left\{\begin{aligned} &u : \text{volume flow vector}\\ &p : \text{sound pressure vector}\\ &A,B : \text{matrices decided by vocal tract shape}\ .\end{aligned}\right.$$

・ Difference equation

$$\left.\begin{aligned} &\begin{pmatrix}u\\ p\end{pmatrix}_n=P\begin{pmatrix}u\\ p\end{pmatrix}_{n-1}+d_1u_{gn}+d_2u_{gn-1}\ ,\\ &p_n=C^T\begin{pmatrix}u\\ p\end{pmatrix}_n\ .\end{aligned}\right\}\qquad(3)$$

For convenience of system analysis and mathematical modeling, synthesis system functions are derived as follows

$$G(f_{Ai},s)=\frac{C^T\cdot\mathrm{adj}(sA+B)\cdot(c_1+sc_2)}{|sA+B|}\ ,\qquad(4)$$

$$G(f_{Ai},z^{-1})=z\text{ transform of }G(f_{Ai},s)\ .\qquad(5)$$

2.2 Nature of phoneme information

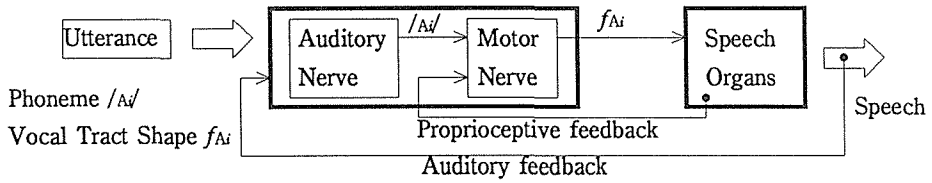
Phoneme is an acoustic unit of speech language. There are two kinds of quantitative features for the expression of phoneme information. One is a formant or pole / zero location. The other is a vocal tract shape or its movement. This is, we insist, the direct feature of phoneme for human speech processing mechanism. From the results of both speech synthesis simulation and its pole/zero tracking analysis of Eq.(4), the behaviors of these location are found to be too complicated to detect or estimate from a given speech sound, especially nasal consonants [6].

And further, formant structures are speaker-dependently varied and are made ambiguous by environmental noise. But, the fundamental vocal tract shape or articulatory movement for a given phoneme is assumed to be speaker-independently analogous. With regard to an ability of instantaneous speech mimic, we assume that human auditory nerve and articulatory motor nerve are systematically related by the brain internal model of vocal tract. Namely, in auditory nerve system, there exist hypothetically the learned and memorized neural networks which react selectively to each common vocal tract shape for speaker-dependently varied spectral informations. We call this neural network as an brain internal model of vocal tract. To this internal model, an articulatory speech synthesis system as shown in Eq.(4) is also applicable.

3 . Artificial Speech Mimic

Speech mimic is a representative function of the human speech processing ; such an unified process of hearing and speaking as shown in Fig.2. Phoneme perception is considered to be an inverse process of speech generation. An articulatory speech synthesizer is the forward model of speech generation, but it can be also effectively utilized as the inverse model which detect the articulatory information of a given speech sound.

[1] Scheme of Speech Mimic Processing



[2] Modeling

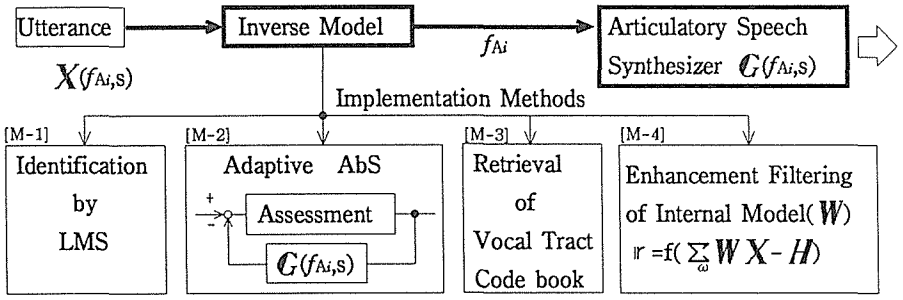


Fig.2 Schematic diagram of Speech Mimic processing

3.1 Inverse Model

In Fig.2-[2] are shown several approaches to solving the inverse problems of obtaining articulatory parameters from spoken utterances. Since a spectral information does not always represent articulatory characteristics, it is not appropriate to identify them by the least mean squares method for the spectral errors. Then, as the inverse model, we can not help depending on the forward model, i.e. an analysis by synthesis procedure. An articulatory parameter setting and an assessment of the synthesis quality must be adaptively carried out by both spectrum and auditory criteria [M-2] [3, 4]. Model [M-3] shows the recently presented speech mimic system [7]. In this model, the vocal tract shape movements are retrieved from a large codebook of previously prepared vocal tract shapes. This results in high computational load and large memory requirements, but the process of this retrieval has nothing to do with a human processing. We present the model [M-4] as more appropriate procedure for human speech perception.

3.2 Principle of Internal Model

For the extraction of a speaker-independent articulatory information from spoken utterance,LMS method is not available, but it is useful for our approach to consider the following functional in frequency domain:

$$J(f_{Ai})=\int_{\omega}[|\mathbf{X}|-|P_j\cdot\mathbf{G}(f_{Ai},L_j)|]^2 d\omega \quad , \tag{6}$$

$$=\int_{\omega}\left[|\mathbf{X}|^2-2|P_j\cdot\mathbf{G}||\mathbf{X}|+|P_j\cdot\mathbf{G}|^2\right]d\omega \quad , \tag{7}$$

where, $P_j\cdot\mathbf{G}$:the model for a given speech $\mathbf{X}$

P_j :equivalent glottal model

f_{Ai} :vocal tract shape of phoneme /Ai/

L_j :vocal tract length .

A speech spectrum $\mathbf{X}$ contains various informations such as personality, emotion and phoneme. They are expressed by L_j , P_j and f_{Ai}, but these parameters are not always independent; systematically related each other in physiological mechanisms. This makes it hard to find a minimum point of the functional J for various cases of f_{Ai} . Then, we pay attention to the crossterm in Eq.(7), namely,

$$\int_{\omega}|P_j\cdot\mathbf{G}||\mathbf{X}|d\omega=\int_{\omega}|P_j\cdot\mathbf{G}(f_{Ai},L_j)\cdot\mathbf{X}|d\omega \, . \tag{8}$$

$G \cdot X$ means a filtering of the system function G for an input speech signal X and it also has a meaning of cross power spectrum which represents the correlation between model and acutual vocal tract. Among the various cases of G's with fA_i varied, if a certain fA_i is regarded to be equivalent to the vocal tract of X, the filtering response of this case is enhanced. This shows a strong correlation of vocal tract shape i.e. a selectivity of phoneme. It is thought that making Eq.(6) minimum is equivalent to making Eq.(8) maximum. But, it is even hard to find its maximum, because Eq.(8) contains personal information P_j. Then, as an application of the above enhancement phenomenon, we introduce a set of G's denoted by $\{W_i\}$ or W, and propose a mathematical model of the human phoneme perceptibility as follows.

$$r = f\left(\sum_{\omega} WX - H\right), \tag{9}$$

where, we call W a set of brain internal model of vocal tract. The notation f(••) represents somenonlinear processing. H means a threshold.

3.3 Enhancement Filtering

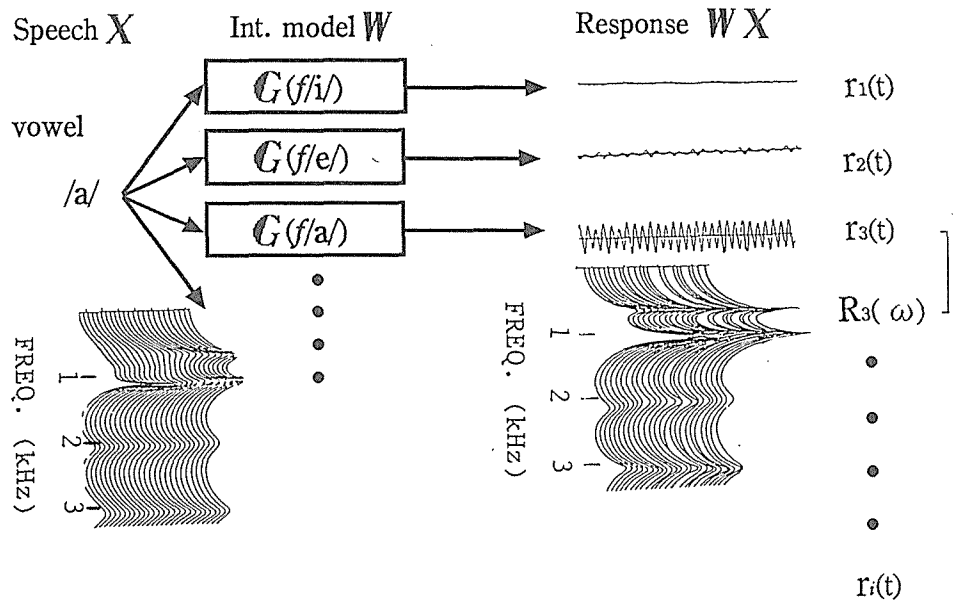
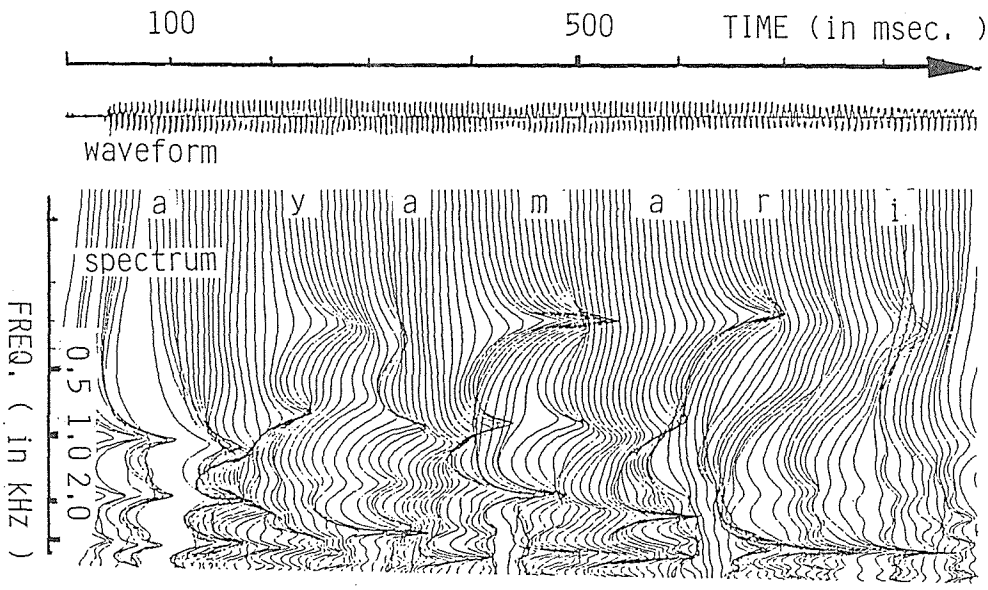


Fig.3 Schematic representation of enhancement filtering. For an example of input signal "vowel /a/", responses $r_i(t)$'s except $r_3(t)$ are inactive. $r_3(t)$ shows an amplitude enhancement of the time response of WX. $R_3(\omega)$ shows its frequency enhancement which is assumed to be a mathematical modeling of auditory nerve activity.

[I] Original speech signal. Japanese word /ayamari/



[II] Filtering response WX for $W = G$ (*f*/a/)

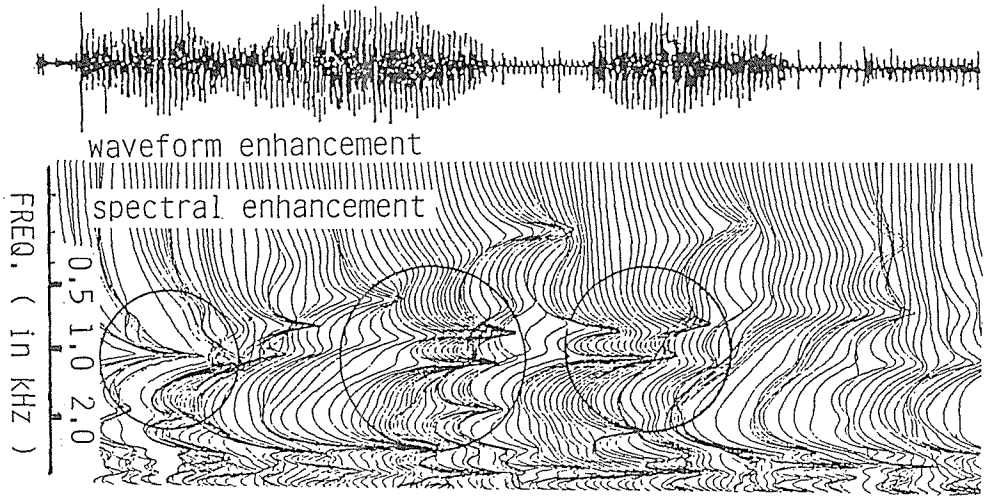


Fig.4 Examples of enhancement filtering characteristics. Speech signal X is Japanese word " ayamari " from a sample of the speech signal data base [Tohoku Univ. Matsushita] .

Speech signal X is modeled as $P_j \cdot G(fA_i)$, but P_j and fA_i are unknown. W is denoted by $\{G(fA_k)\}$, and fA_k's are given as a code book of vocal tract shape. Quantitative meaning of enhancement filtering is as follows. The response "$G(fA_k) \cdot G(fA_i)$" is strengthened when $fA_k=fA_i$, then unknown $G(fA_i)$ can be selected from other filtering of WX.

Experiments of enhancement filtering have been carried out for synthesis speech and speaker-independent natural speech. The enhancement characteristics can, visually be displayed on spectral patterns or time response wave patterns, some example of enhancement filtering characteristics are shown in Fig.4. However, quantitative decision of phoneme identification is a pending ploblem. There has been still no decisive approach to the nonlinear processing method for Eq.(9). But, from some experiments [8, 9], these responses of Eq.(9) shows an ideal characteristics as shown in Fig.5. Phoneme selectivity in its categorization is comparably sharp and speaker-dependently broad.

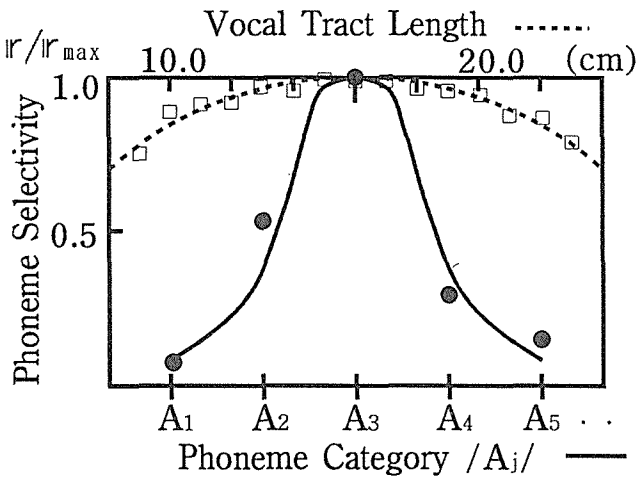


Fig.5 The fundamental characteristics of phoneme selectivity shown in Eq.(9). Phoneme selectivity versus category (lower abscissa) is characterized by full line curve or filledcircled experimental data [9]. This shows speaker-independently sharp selectivity. Phoneme selectivity versus vocal tract length (upper abscissa) for a fixed phoneme is characterized by broken line or open squared data [9]. This shows speaker-independently broad selectivity.

4. Conclusion

The essential feature of phoneme information is not formant pattern but vocal tract

characteristics. It is the inverse problem to estimate vocal tract shape or its movement from spoken utterance. Since a speech spectrum does not always represent articulatory characteristics, it is not appropriate to identify them by LMS type method. Phoneme perception is considered to be an inverse process of speech generation. An articulatory speech synthesizer is the forward model of speech generation, but it can be also effectively utilized as the inverse model which detect the articulatory information of a given speech sound. We propose the inverse model [M-4] in Fig.2 as more appropriate procedure for human speech perception.

5 . References

1)B.Lindblom, J.Lubker, T.Gay: Journal of Phonetics , **7** , 147(1979).

2) W.J.Levelt: Speeking, (MIT Press. Cambridge,1989).

3) J.Flanagan, K.Ishizaka, L.Shipley : JASA,**68**, 780(1980).

4) M.Sondhi, J.Schroeter : IEEE, ASSP. **35**, 955(1987).

5) H.Sakaguchi,T.Kirihata:IECE of Japan.**J68-A**,9, 795(1985).

6) H.Sakaguchi, S.Kobayashi :Tech.Rep. of IEICE.**EA88-55**, 27(1988).

7) M.Rahim, C.Goodyear, W.Klein, J.Schroeter,M.Sondhi:JASA.**93**,1109(1993).

8) H.Sakaguchi,N.Kawaguchi :Tech.Rep. of IEICE.**EA93-51**,1(1993).

9) H.Sakaguchi,J.Yokoyama : Tech.Rep. of IEICE.**EA93-52**,9(1993).