TR-A-0006

Evaluation of a spectrum target prediction model in speech perception

音声知覚におけるスペクトルターゲット予測モデルの評価

Masato AKAGI

赤木　正人

1987.6.9

ATR視聴覚機構研究所

概　要 (Abstract)

A model of a phoneme target prediction mechanism based on psychological and physiological knowledge is proposed and evaluated by comparing predicted values with results of auditory psychological experiments. When the trajectories of physical characteristics of speech in spectral space are approximated by a 2nd-order critical damping system, the proposed model can estimate target values using a short period speech wave (50ms) without knowing the onset positions of the spectral transition. Additionally, this model decreases the length of transitional sounds which produce spurious vowels, and recovers vowel characteristics neutralized by co-articulation. Moreover, the model compensates for the transitional part of syllables and achieves the extraction of stable acoustic features for consonants. This model is considered to be a good approximation of the target prediction mechanism in speech perception.

発行時　配付先 (Initial Distribution Specifications)

聴覚研究室員	板倉教授（名古屋大学）
淀川社長	粕谷教授（宇都宮大）
梅田室長	Prof. Sterm (CMV)
樽松社長	Prof. Zue (MIT)
鹿野室長	Dr. Sondhi (Bell研)
桑原主幹員	筧リーダ（NTT 基礎研）
ワイベル客員研究員	古井室長（　〃　）
	河原主幹員（　〃　）

備考(Notes)

本レポートは J. of Acoust. Soc. Am. に投稿済である。

Evaluation of a spectrum target prediction model in speech perception

Masato AKAGI

ATR Auditory and Visual Perception Research Laboratories
Twin 21 Bldg. MID Tower, 2-1-61 Shiromi, Higashi-ku
Osaka, 540 JAPAN
Tel +81-6-949-1835
Fax +81-6-949-1811

1. INTRODUCTION

Analyzation of a continuous speech and extraction of its physical characteristics show that incomplete articulation intervals exist which produce neutralized phonemes or spurious phonemes, and that these are caused by the limitations of the articulatory organ. However, upon hearing a continuous speech, each phoneme neutralized by co-articulation is perceived as if it were uttered clearly, even though its physical characteristics do not reach the target characteristics obtained when the phoneme is uttered separately (recovering phoneme characteristics from co-articulation). For example, if we utter the concatenated vowels /aia/, it will probably be heard as [aia]. However, its physical characteristics are actually [aea], or if [i] does appear, its time interval is short. In addition, spurious phonemes in co-articulation intervals are not perceived even though trajectories of their physical characteristics approach to the characteristics of spurious phonemes. For example, when the diphthong /ai/ is uttered , the spurious phoneme [e] is not perceived even though

the physical characteristics [e] of /e/ appearing in the co-articulation interval between /a/ to /i/.

This phenomenon can be explained by a compensation mechanism which is presumed to exist in the speech perception mechanism (Brady,1961;Lindblom,1967;Kuwabara,1985). If this compensation mechanism can be modeled, it will be applicable to the co-articulation recovery and continuous speech recognition. However, a model which can explain the phenomenon and would be effective in some applications has not yet been found.

In this paper, a model for a prediction mechanism is constructed which assumes that a phoneme target prediction mechanism exists in the speech perception system and that this mechanism can compensate for phonemic characteristics. Although physiological and psychological knowledge of speech perception is necessary for the model, the model is not a faithful model of the physiological organization of the auditory periphery or of the psychological organization of the nerve center; rather it is a global engineering model of the entire system.

Four viewpoints are discussed comparing results of the model with those of psychological experiments. (1)When the trajectories of physical speech characteristics in spectral space are approximated by a 2nd-order critical damping system, the proposed model can estimate the vowel target values using a short period speech wave (50ms) without knowing the onset positions of the spectral transition. (2)A perceptual critical point (Furui,1986), where the following vowel in the Japanese C-V syllable is initially perceived, corresponds well with its model computed physically estimated point. (3)The model compensates for a transitional portion of connected Japanese

vowels and decreases the length of the transitional part which produces the spurious vowel spectrum (These spectra are actually unperceived under any speaking conditions.) (4)The model is also applicable for compensating the transitional part of consonants in Japanese syllables and achieves the extraction of stable acoustic features for consonants. Results indicate that the model corresponds well to the human hearing mechanism.

The balance of this paper is organized as follows. Section 2 describes the outline of the proposed model. Section 3 presents psychological experiments to decide the perceptual critical point. Section 4 evaluates the model using the the above four viewpoints. Section 5 concludes with a summary and suggestions for further investigations. The appendix explains details of the model.

2. OUTLINE OF THE MODEL

A target prediction model predicts the stable spectral target in short term intervals and compensates for phonemic characteristics by using a 2nd-order critical damping model. Physiological and psychological knowledge of speech perception such as, the difference between inner and outer hair-cells (Dallos,1972), or perceptual switching from the preceding to the following vowel (Furui,1986) is considered for the model. However, the model is not completely faithful to the physiological organization of the auditory periphery or to the psychological organization of the nerve center, but is an engineering model for the global periphery to center structure.

Although details of the model are presented in the appendix, algebraic definitions and computer simulated results of the model are introduced in this section.

Assuming that y_n is the spectrum at time n, and $\dot{y}_n$ is the time derivative of spectrum y_n, and k is a feedback factor, then the relationship between y_n and $\dot{y}_n$, and input spectrum to the nerve center, x_n, is as follows;

$$x_n = -k y_n + \dot{y}_n \qquad (1)$$

$$\dot{x}_n = \lambda(x_n + k b_n) \qquad (2)$$

where b_n is the target value and λ is a reciprocal time constant for the transition.

When k, in Eq. (1) is equal to λ, the 1st-order system Eq. (2) is represented by

$$\ddot{y}_n - 2\lambda\dot{y}_n + \lambda^2 y_n = \lambda^2 b_n \qquad (3)$$

which is the 2nd-order critical damping model. The solution of this equation is

$$y_n = a(1 - \lambda n)\exp(\lambda n) + b, \quad n \geq 0. \qquad (4)$$

that is,

$$x_n = -a\lambda\exp(\lambda n) - \lambda b \qquad (5)$$

by substituting Eq. (4) into Eq. (1).

Therefore, let us assume a certain value for λ and calculate the value for the spectral sequence of x_n, applying Eq. (1), such

that

$$x_n = -\lambda y_n + \dot{y}_n. \qquad (6)$$

Then,

$$e(\lambda) = \sum_{n=-N/2}^{N/2} |x_n - \hat{x}_n|^2 \qquad (7)$$

can be minimized using the least mean square error method for

$$\hat{x}_n = -\hat{a}\lambda\exp(\lambda n) - \lambda\hat{b} \qquad (8)$$

by varying the parameters a and b. λ can be optimized by minimizing $e(\lambda)$ as a non-linear optimization problem. The optimum λ, $\hat{a}$, and $\hat{b}$ are obtained as the condition which minimizes $e(\lambda)$.

Previous methods were unable to estimate the parameters of the 2nd-order critical damping model based on a short-term spectrum sequence of about 50ms, because they needed a long-term spectrum sequence that included the onset position of the spectral transition. However, since the method of this paper solves the parameter estimation problem of the exponential function to obtain target b, it does not need to calculate the onset position of the spectral transition.

Simulated results of the proposed target prediction method based on the 2nd-order critical damping model are shown in Fig.1. The thin solid line, dashed line, and thick solid line indicate step function, 2nd-order critical damping function, and predicted target

$\boldsymbol{b}$, respectively. Fig.1 indicates that target $\boldsymbol{b}$ corresponds well to the step function and is satisfactorily predicted by the model.

Speech wave, time functions of the original spectrum $\boldsymbol{y}$ and the target $\boldsymbol{b}$ predicted by the target prediction model are shown in Fig.2 for the diphthong /ai/. This figure clearly indicates that the target is reliably predicted and categorically changes at the phoneme changed position.

3. PSYCHOLOGICAL EXPERIMENTS

To determine the perceptual critical point, which is the point where the preceding vowel is no longer heard or the point where the following vowel is heard, the following two sets of stimuli were presented to listeners.

STIMULI

(a) DATA I: 100 phonotactically possible short syllables of Japanese shown in Table 1 uttered by two trained female speakers were modified by final truncation. The speech wave amplitude, linearly attenuated over a 10ms period, is shown in Fig.3(a). The truncation window shown in Fig.3(a) was varied in 5ms steps continuously over a 50ms interval where it includes the maximum spectral transition position. The total number of stimuli was 1000 (10 conditions x 100 words) for each speaker.

(b) DATA II: 20 different words including every kind of vowel concatenation are shown in Table 2. These were uttered by four speakers, one male and three male, and were modified by

initial and final truncation to remove the influence of the preceding and/or following phoneme. The truncation window shown in Fig.3(b) was varied in 10ms steps continuously over a 400ms interval. The total number of stimuli was 800 (40 conditions x 20 words) for each speaker.

Stimuli were arranged in random order in both DATA I and DATA II.

METHOD

Stimuli were presented over headphones without noise in a soundproof room at a comfortable listening level to four trained listeners. Each listener identified the tokens. Percent correct identification, which was the number of vowels correctly recognized by the listeners, was taken as a function of truncated position relative to "perceptual critical point" shown in Fig.3.

The truncated position at which the vowel identification score is more than 80% for the first time is regard as the truncation position from which following vowel is heard for DATA I and DATA II. The truncated position at which the vowel identification score is less than 80% for the first time is regard as the position from which the preceding vowel is no longer heard for DATA II. These positions are termed "the perceptual critical point". The experimental results show that the standard deviation of the perceptual critical point is 16.0ms for preceding vowels and 16.3ms for following vowels in DATA II.

4. EVALUATION OF THE MODEL

In this section, evaluated results of the proposed model are presented. The evaluated items are optimum frame length for short term target prediction (Section 4.1) and recovery ability of phoneme characteristics from co-articulation (Section 4.2).

4.1 Optimum frame length for target prediction

Although a target prediction method based on a short-term spectrum sequence is proposed in section 2, it is not clear how long the optimum frame length for the target prediction is. This section proposes the optimum frame length by comparing results based on the psychological experiments with results based on the proposed model. 100 Japanese syllables in DATA I uttered by one female speaker are dealt with under five frame length conditions.

Standard deviation of the difference between the perceptual critical point and its physically estimated point is shown in Fig.4. The physically estimated point is the point where the physical distance between target $\boldsymbol{b}$ and the category center of the actual vowel for each speaker initially attains a smaller value than any other vowel. The Euclidean distance of the MEL scale log spectrum is used for the physical measurement. The category center of each vowel (which is the mean of each vowel), /a/, /e/, /i/, /o/ and /u/, is the mean of the spectra of actual vowel uttered in isolation in DATA I for each speaker.

Fig.4 clearly indicates that standard deviation is the smallest at the 50ms frame length. The 50ms frame length

corresponds well with the length of perceptually essential intervals for vowels by Furui (Furui,1986). This indicates in part that the model explains human hearing mechanism well. Based on this result, the model uses the 50ms frame length for the short term target prediction, hereafter.

4.2 Phoneme characteristics recovery from co-articulation

(a) EVALUATION I

In order to evaluate how close the distance is between the perceptual critical point and its physically estimated point, the mean and standard deviation of the differences between them are measured. If the standard deviation is reduced when using the model, it is clear that the model can predict stable features. Correspondingly, if the mean becomes smaller by using the model, it shows that the model can recover vowel characteristics from co-articulation neutralization at the spectral transition position.

For example, sequences of physically identified vowels whose ideal spectra feature smaller spectral distance than the other vowels from spectrum $\boldsymbol{y}$ or from target $\boldsymbol{b}$ at each short period during the utterance of the Japanese word /kiai/ are shown in Fig.5. The distance measure is the Euclidean distance of the MEL scale log spectrum. In Fig.5, the physically estimated point for vowel /a/ is the point where the physically identified vowel changes from [e] to [a] for the first time. The perceptual critical point where the perceived vowel changes from /i/ to /a/, and,

therefore, where vowel /a/ is initially perceived is also shown in Fig.5. It is observed that the perceptual critical point corresponds well to the physically estimated point by using the model.

The above results are only for the word /kiai/. It is not clear if the model decreases the mean and the standard deviation of the difference between perceptual critical point and its physically estimated point for all of DATA I and DATA II.

Experimental results for DATA I and DATA II are shown in Fig.6. Fig.6(a) shows the results for DATA I and Fig.6(b) shows the results for DATA II. These figures indicate that the value of the standard deviation (σ) becomes 1.5 or 2 times smaller in the prediction condition compared with the no-prediction condition.

Additionally, Fig.6(a) indicates that the value of the mean (μ) is smaller in the prediction condition than in the no- prediction condition, specially for the syllables with /j/. On the syllables with /j/, the physical characteristics of the following vowel often do not reach its target characteristics due to neutralization by co-articulation. However, by using the model, the following vowel physical characteristics can be brought close to its target characteristics earlier and the mean μ is made smaller.

These results indicate that the model is able to predict stable phoneme features and to recover vowel characteristics from neutralization. Thus, the model realizes the human hearing function, where each phoneme neutralized by co-articulation in the continuous speech is perceived as if it were uttered separately, even though its physical characteristics do not reach the target characteristics.

(b) EVALUATION II

Results of auditory psychological experiments show that transitional sounds are scarcely perceived. For the word /kiai/ in Table 1, for example, the transitional sound /e/ is perceived on only the 10 ~ 20 ms interval from /i/ to /a/ or from /a/ to /i/ during psychological experiments. Results for the word /kiai/ by physical analysis, which shows the time sequences of vowel nearest to the predicted target spectrum $\boldsymbol{b}$, or to the original spectrum $\boldsymbol{y}$, at each short period are shown in Fig.5. The Euclidean distance of the MEL scale log spectra is used for the physical measurement. In Fig.5(b) spectrum $\boldsymbol{y}$ indicates that the transitional sound /e/ spans 60 ~ 70 ms. However, the length of the transitional sound /e/ in Fig.5(a) for target $\boldsymbol{b}$ becomes much shorter than in Fig.5(b) for spectrum $\boldsymbol{y}$.

These results are only for the word /kiai/. It is not clear that the model decreases the lengths of the transitional sounds for other data. In this section, in order to evaluate the length of the transitional sounds in a prediction condition or in a no-prediction one, the mean and standard deviations of the transitional sound lengths are measured by the data which were uttered under three speaking rate conditions. If the length of the transitional sounds in the prediction condition becomes smaller by using the model, it shows that the model can recover vowel characteristics from neutralization by co-articulation at the spectral transition position under any speaking rate conditions.

Data used for this experiment were six Japanese vowel sets: /aia/, /eoe/, /iai/, /ioi/, /oeo/ and /oio/ in 20 different words

that include every vowel concatenation, $V_1V_2V_1$, in which the transitional sounds are found naturally. These were uttered by one male and female speaker under slow, medium, and fast speaking rate conditions.

In order to measure the length of the spectrum transition part of the above data, spectral transition measure (Furui,1986) was calculated at each short period for each word. The length of the spectrum transition part is measured from the point where the spectral transition measure rises above the threshold value (onset) to where it falls below the threshold value (offset). Relationship between speaking rate conditions and the mean of the syllable lengths or the mean of the lengths of the spectrum transition parts in above data are shown in Fig.7. Fig.7 indicates that the mean of the lengths of the spectrum transition parts under a fast speaking rate condition is shortened 14% compared to that of a slow speaking rate condition. Also, the mean of the syllable lengths under a fast condition is 60% shorter than that under a slow one. Thus, the knowledge (Hiki,1967) that, when speaking rate condition is changed from slow to fast, the lengths of the steady state part are, in the main, shorter and that the lengths of the spectrum transition part are hardly influenced by speaking rate, is confirmed. Additionally, since the lengths of the spectrum transition part are: 45%(slow), 72%(medium), and 93%(fast) of the syllable lengths respectively, the steady state portion hardly exists in these six Japanese words uttered under fast speaking rate conditions.

In order to measure the transitional sound lengths in the above data, sequences of physically identified vowels are

calculated using the same method in Fig.5. The transitional sound length is the time length needed to pass through the different category domains that are between the objective preceding vowel category and the following one. For example, the results of the Japanese word /taian/ uttered by a female speaker under a fast speaking rate condition is shown in Fig.8. That part of /e/ found in the transition part from /a/ to /i/ or from /i/ to /a/ is the transitional sound. Fig.8 indicates that the transitional sound length of target ***b*** clearly becomes shorter than that of spectrum ***y***.

The mean of the transitional sound lengths is shown in Fig.9. Fig.9 indicates that the mean of the transitional sound lengths for spectrum ***y*** is almost always fixed at 70ms under any speaking rate conditions. This suggests that the time length which physical characteristics like spectrum transit from the preceding vowel characteristics to the following one is fixed under any speaking rate conditions. Moreover, if spectrum ***y*** uttered under fast speaking rate conditions is considered, the length in which the objective phoneme characteristics exist is shortened because the mean of the transitional sound lengths occupies more than 60% of the mean of the syllable lengths. This is one reason that the construction of an automatic phoneme recognizer is difficult.

Using the model, however, the mean length of the transitional sounds is 20~30ms under any speaking conditions. That is 1/3 of the mean of the transitional sound lengths in the no-prediction condition. Moreover, the mean lengths of the transitional sounds occupy 15% of the mean syllable lengths by using the model under fast speaking rate conditions. Thus, the

time length in which the objective phoneme characteristics exist becomes longer relatively.

Since the proposed model predicts the target by using spectral transition under Eqs.(1) and (2), the information of the spectral transition is regarded as playing an important role in the target prediction model. Actually, the mean length of the spectrum transition part is almost fixed. The mean time length in which physical characteristics like spectrum transit from preceding phoneme characteristics to following ones is also almost fixed under any speaking rate conditions. Thus, the model, using the spectral transition information, is able to predict the target effectively under any speaking rate conditions and to decrease the length of transitional sounds.

These results indicate that the model is able to decrease the transitional sound lengths under any speaking rate conditions. Thus, the model realizes the human hearing function, where spurious phonemes in co-articulation intervals are not perceived even though its physical characteristics approach the characteristics of the spurious phonemes.

(c) EVALUATION III

Although the above sections clarify that the model recovers vowel characteristics neutralized by co-articulation, it is not clear that the model is also able to compensate for the transitional part of consonants and to achieve the extraction of the stable acoustic features of consonants. In this section, in order to evaluate the variation of the predicted spectra or the original spectra in the consonant part, mean and standard deviations of

the sum total of the covariance matrix eigen values for the predicted spectra or the original spectra are measured by analyzing the Japanese plosives, /g,d,b,k,t,p/ in DATA I. If the mean and standard deviation becomes smaller by using the model, then it is clear that the model is also applicable for consonants.

For example, results analyzing the Japanese syllable /ga/ by the model are shown in Fig.10. The solid line in Fig.10(a) indicates the consonant portion which is the interval between two maximum points of the spectral transition measure. Fig.10(b) is the sequence of the original spectrum $\boldsymbol{y}$ and Fig.10(c) is the sequence of the target $\boldsymbol{b}$. Additionally, Fig.10(d) represents the squared distance between $\boldsymbol{b}_n$ and its neighbor, $\boldsymbol{b}_{n-1}$,

$$D = |\boldsymbol{b}_n - \boldsymbol{b}_{n-1}|^2.$$

Figs.10(b) and (c) clearly indicate that target $\boldsymbol{b}$ is more stable in the consonant section than the original spectrum $\boldsymbol{y}$. Moreover, Fig.10(d) shows that the peak of value D corresponds to the boundary between buss-bar and consonant /g/ or between consonant /g/ and vowel /a/.

Fig.11 shows the mean (μ) and standard deviation (σ) of the value S for each plosive consonant in DATA I under no-prediction and prediction conditions. The value S is a measure of variation and is represented as follows:

Assuming that T is a covariance matrix of the original spectrum sequence $\boldsymbol{y}_n$ or the target sequence $\boldsymbol{b}_n$ in the consonant part,

$$T = \sum (\boldsymbol{y}_n - \bar{\boldsymbol{y}})(\boldsymbol{y}_{n+t} - \bar{\boldsymbol{y}})^T \quad \text{or} \quad T = \sum (\boldsymbol{b}_n - \bar{\boldsymbol{b}})(\boldsymbol{b}_{n+t} - \bar{\boldsymbol{b}})^T,$$

where $\bar{\boldsymbol{y}}$ or $\bar{\boldsymbol{b}}$ is the mean spectrum of the consonsnt portion. The value S is the sum total of the eigen values of matrix T, that is,

$$S = \sum \lambda_i.$$

Fig.11 clearly indicates that results are more stable when using the model. Notice that the mean (μ) in the prediction condition is ten times smaller than in the no-prediction condition.

5. CONCLUSION

A model of a phoneme target prediction mechanism based on psychological and physiological knowledge was constructed and evaluated by comparing predicted values with results of auditory psychological experiments. Since the trajectories of physical speech characteristics in spectral space were formulated by a 2nd-order critical damping system, the proposed model could estimate target values using a short period speech wave (50ms) without knowing the onset position of the spectral transition. Additionally, this model decreased the transitional sound lengths which produce spurious vowels and recovered vowel characteristics neutralized by co-articulation. Also, the model compensated for transitional part and achieved the extraction of stable acoustic features not only for vowels but also consonants. These results clarify that the model corresponds well to the human hearing mechanism. In addition, the model

is probably applicable to automatic continuous speech recognition.

Further investigations would include the introduction of new psychological and physiological knowledge, modeling of other mechanisms, for example: contrast and/or assimilation effect and concatenating of these models, and evaluation by auditory psychological experiments.

ACKNOWLEDGMENTS

The author wishes to thank Dr. Sadaoki Furui, Head of Section 4 at NTT Basic Research Laboratories for his helpful suggestions at the start of this study and his continuing guidance. Several helpful discussions and valuable advice, concerning such areas as measures for evaluations, from Dr. Kazuhiko Kakehi, Head of Perception Research Group at NTT Basic Research Laboratories, is gratefully acknowledged. The author would also like to acknowledge the critical reading of this manuscript and the helpful suggestions given by Dr. Yoh'ichi Tohkura, Head of Hearing & Speech Perception Department at ATR Auditory and Visual Perception Research Laboratories.

REFERENCE

Akagi, M. and Furui, S., (1986). "Modeling of Vowel Target Prediction Mechanism in Speech Perception", J. of IECE Japan, Vol.69-A, 10, pp..1277-1285.

Brady, P. T., et. al. (1961). "Perception of Sounds Characterized by a Rapidly Changing Resonant Frequency", J. of Acoust. Soc. Am., Vol. 33, 10, pp..1357-1362.

Dallos, P., et.al., (1972). "Cochlea Inner and Outer Hair Cells: Functional Difference", SCIENCE, Vol. 177, pp..356-359.

Furui, S. (1986). "On the Role of Spectral Transition for Speech Perception", J. of Acoust. Soc. Am., Vol. 80, 4, pp..1016-1025.

Hiki, S., et. al. (1967). "On the Duration of Phoneme in Running Speech", J. of IECE Japan, Vol. 50, 5, pp..849-856.

Klatt, D. H. (1982). "Speech Processing Strategies based on Auditory Models", The representation of speech in the peripheral auditory system (Elsevier Biomedical press), R.Carlson and Granstrom eds., pp..181-196.

Kuwabara, H. (1985). "An Approach to Normalization of Coarticulation Effects for Vowels in Connected Speech", J. of Acoust. Soc. Am., Vol. 77, 2, pp..686-694.

Lindblom, B. E. F. (1967). "On the Role of Formant Transition in Vowel Recognition", J. of Acoust. Soc. Am., Vol. 42, 4, pp..830-843.

Suzuki, H., (1974). "Mutually Complementary Effect between Amount and Rate of Formant Transition in Perception of Vowels, Semivowels and Voiced Stops and a Possible Mechanism for Their Identification", J. of Acoust Soc. Japan, Vol. 30, 3 pp..169-180.

APPENDIX

In this section, details of the proposed model (Akagi,1986) are presented. First, physiological backgrounds of the model are explained in section A-1. Next, basic concepts including psychological evidence for the model are introduced and the algebraic derivations for Eq. (1) and (2) are presented in section A-2. Finally, a method for short term prediction using Eqs. (1) and (2) is proposed in section A-3.

A-1. PHYSIOLOGICAL BACKGROUND

In order to represent the peripheral auditory, the following mechanical organization models of the peripheral auditory system reported by Klatt (Klatt,1982) are used.

[a] Half-wave rectifiers to model the transformation that takes place in the hair-cells.

[b] A log transformation to approximate the loudness phone scale.

[c] A MEL scale to model the basilar membrane.

[d] Lateral suppression circuitry to sharpen peaks seen in the output spectra.

Utilizing [a] and [b], 8kHz speech wave is LPC analyzed incorporating 10 dimensional coefficients, a 32ms frame length and a 1ms frame period. The LPC cepstrum calculated from the LPC coefficients is translated into a log spectrum by inverse-FFT. Using [c], the log spectrum is transformed linear piecewise into a MEL scale. In Fig.A-1, the Basilar Membrane section shows LPC

CEP. (LPC cepstrum), LOG SPEC. (logspectrum), and MEL SCALE (MEL scale) as block diagrams for [a], [b], and [c]. For [d], the following equations for modeling lateral inhibition are adopted;

$$y(x) = f(x) - \int_{-\infty}^{\infty} G(x-x')f(x')dx'$$

$$G(x) = A/2 * \exp(-|x| / a),$$

where $f(\ )$ is an input spectrum to the lateral inhibition model and $G(\ )$ is a weighting function. The lateral inhibition model is shown in the Auditory Nerve section of Fig.A-1.

Dallos (Dallos,1972) has reported that a functional difference between the outer and inner hair-cells exists. That is, the outer hair- cells sense basilar-membrane deviation, while the inner sense its velocity. Farthermore, the outer hair-cells are connected to both afferent and efferent fibers, whereas the inner hair-cells are only connected to afferent fibers. The model also includes these auditory physiological characteristics. This functional difference is represented in Hair-cells section of Fig.A-1.

Although $\dot{\boldsymbol{y}}$ essentially presents the time derivative of spectrum $\boldsymbol{y}$, the 1st-order regression coefficient over a 50ms-long period is used instead of the time derivative because of auditory time resolution and parameter stabilization into consideration. For example, $\boldsymbol{y}$ and $\dot{\boldsymbol{y}}$ which the diphthong /ai/ gets through the model are shown in Fig.A-2.

A-2. BASIS OF TARGET PREDICTION MODEL

In order to develop the target prediction model, following psychological criteria are adopted.

[a] The amount and velocity of the spectral transition compensate each other to preserve the quality of perceived phonemes (Suzuki,1974).

[b] Perceptual switching from the preceding to the following vowel occurs at the maximum spectral transition position (Furui,1986).

[c] A speech wave of approximately 50 ms in duration is used for target prediction in speech perception by human heings (Furui,1986).

Additionally, the following two points are considered.

[d] The dynamics of not only formants but the overall spectral pattern are perceived.

[e] The spectral transition can be represented by a 2nd-order critical damping model.

Although the formant transition is represented by a 2nd-order critical damping model in previous investigations, the spectral transition is also assumed to be represented by a 2nd-order critical damping model in this paper, because the dynamics of the overall spectral pattern are perceived.

For the point [d], the following equations are used for the dynamics of spectrum using y and $\dot{y}$ which are outputs from the model of the lateral inhibition in Fig.A-1.

Assuming that y_n is the spectrum at time n, $\dot{y}_n$ is the time derivative of spectrum y_n and k is a feedback factor, the relation between y_n, $\dot{y}_n$ and input spectrum to the nerve center, x_n, is as follows;

$$x_n = -k y_n + \dot{y}_n. \qquad \text{(A-1)}$$

And the spectrum x_n satisfies the following equation,

$$\dot{x}_n = \lambda(x_n + k b_n), \qquad \text{(A-2)}$$

where b_n is the target value and λ is a reciprocal time constant for the transition. The relation between the equations and the parameters is presented in the Nerve Center section of Fig.A-1. Eqs. (A-1) and (A-2) are the same as Eqs. (1) and (2), respectively.

Eq. (A-1) represents the functional difference between the inner and the outer hair-cells. Eq. (A-2) represents the following:

(a) If λ is large, then transition velocity is small or its amount becomes large, and if λ is small, then the opposite conditions are true. That is, Eq. (A-2) represents a mutually complementary effect. Therefore, Eq. (A-2) satisfies the above point [a].

(b) If $\lambda > 0$, then b is a backward predicted value, and if $\lambda < 0$, then it is a forward one. Moreover, the sign of λ changes at the maximum spectral transition position. Thus, Eq. (A-2) also satisfies psychological point [b].

When k is equal to λ in Eq. (A-1), the 1st-order system Eq. (A-2) is represented by

$$\ddot{y}_n - 2\lambda\dot{y}_n + \lambda^2 y_n = \lambda^2 b_n \qquad \text{(A-3)}$$

which is the 2nd-order critical damping model. Then, Eq. (A-3) satisfies point [e]. The solution of this equation is

$$y_n = a(1 - \lambda n)\exp(\lambda n) + b, \quad n \geq 0, \qquad \text{(A-4)}$$

that is,

$$x_n = -\, a\lambda\exp(\lambda n) - \lambda b \qquad \text{(A-5)}$$

by substituting Eq. (A-4) into Eq. (A-1). Eq. (A-5) is the solution of the Eq. (A-2).

A-3. SOLUTION

Furui reported (Furui,1986) that the ideal vowel spectrum is predictable in the auditory system as the spectral transition target using the dynamic features of the 50ms period. If target b can be predicted from a speech wave of 50ms in duration, a model which satisfies all of the above points can be constructed. In this section, a short-term target prediction method is proposed.

Let us assume a certain value for λ and calculate the value for the spectral sequence of x_n, applying equation (1), such that

$$x_n = -\lambda y_n + \dot{y}_n. \qquad \text{(A-6)}$$

Then,

$$e(\lambda) = \sum_{n=-N/2}^{N/2} | x_n - \hat{x}_n |^2 \qquad \text{(A-7)}$$

can be minimized using the least mean square error method for

$$\hat{\kappa}_n = -\hat{a}\lambda\exp(\lambda n) - \lambda\hat{b} \qquad \text{(A-8)}$$

by varying parameters a and b. λ can be optimized by minimizing $e(\lambda)$ as a non-linear optimization problem. λ is optimized in each unit period because the optimizing problem is not unimodal. Optimum λ, $\hat{a}$, and $\hat{b}$ are obtained as the condition which minimizes $e(\lambda)$. The optimized λ is the reciprocal time constant for the 2nd-order critical damping model and the obtained $\hat{b}$ becomes the target b.

$$\hat{a} = a\exp(\lambda t_0), \qquad \text{(A-9)}$$

where t_0 is the difference between the onset position of the spectral transition in Eq. (A-4) and the central position of the short period for the prediction. This method solves the parameter values of λ, a, and b for the 2nd-order critical damping model by solving the parameter estimation problem of the exponential function which corresponds to the 1st-order model.

It has been impossible to estimate the parameters of 2nd-order critical damping models based on a short-term spectrum sequence about 50ms using previous methods, because they needed a long-term spectrum sequence which included the onset position of the spectral transition. However, the main purpose of the parameter estimation is to obtain only target b. Since our method solves the parameter estimation problem of the exponential function, it does not need to calculate the onset position of the spectral transition.

TABLES

Table 1 The 100 Japanese syllables used for experiments.

CV	pa	pi	pu	pe	po
	ta			te	to
	ka	ki	ku	ke	ko
			tsu		
	sa		su	se	so
	ha	hi	hu	he	ho
	ba	bi	bu	be	bo
	da			de	do
	ga	gi	gu	ge	go
	dza		dzu	dze	dzo
	ra	ri	ru	re	ro
	ma	mi	mu	me	mo
	na	ni	nu	ne	no
	ja		ju		jo
	wa				
CjV	pja		pju		pjo
	kja		kju		kjo
	tʃa	tʃi	tʃu		tʃo
	ʃa	ʃi	ʃu		ʃo
	hja		hju		hjo
	bja		bju		bjo
	gja		gju		gjo
	dʒa	dʒi	dʒu		dʒo
	rja		rju		rjo
	mja		mju		mjo
	nja		nju		njo
V	a	i	u	e	o

Table 2 Short utterances including $V_1V_2V_1$ type vowel concatenation used for the experiments.

central vowel	utterance	$V_1V_2V_1$
/a/	kiai	/iai/
	yokuau	/uau/
	ukeaeru	/eae/
	doao	/oao/
/i/	taian	/aia/
	uiuishii	/uiu/
	eien	/eie/
	koio	/oio/
/u/	kauate	/aua/
	iuishi	/iui/
	meue	/eue/
	itouo	/ouo/
/e/	haearu	/aea/
	kieiru	/iei/
	tsueutsu	/ueu/
	koeo	/oeo/
/o/	kaoawase	/aoa/
	nioi	/ioi/
	inuou	/uou/
	seoe	/eoe/

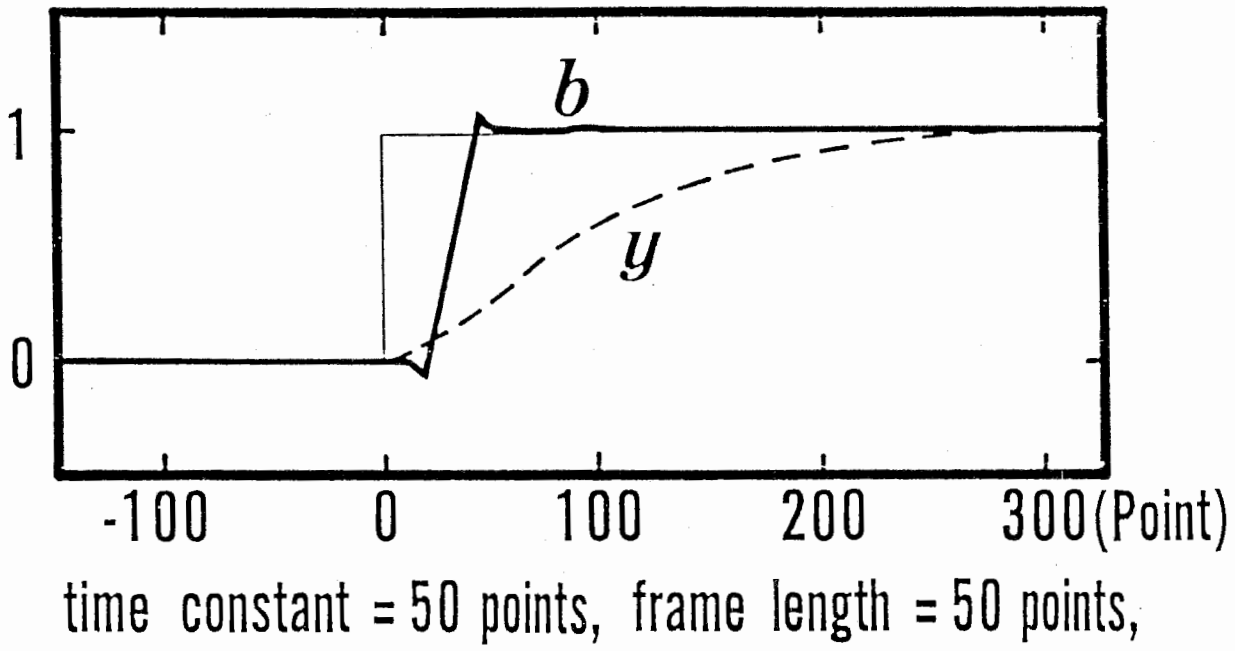


Fig.1 Simulated results of the proposed prediction method based on the 2nd-order critical damping model.

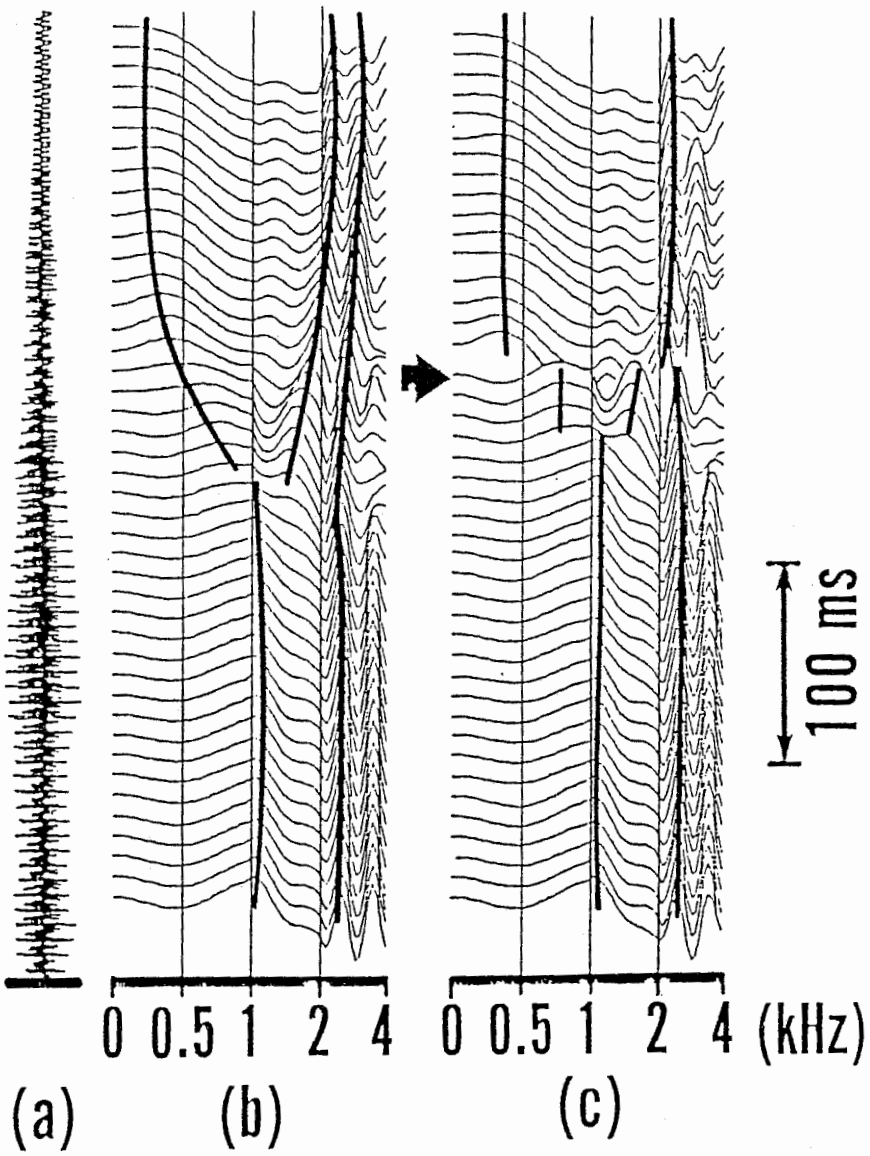


Fig.2 Output spectrum sequences associated with the proposed model for the diphthong /ai/. (a) wave form, (b) spectrum sequence, $\boldsymbol{y}$, (c) estimated spectrum sequence, $\boldsymbol{b}$.

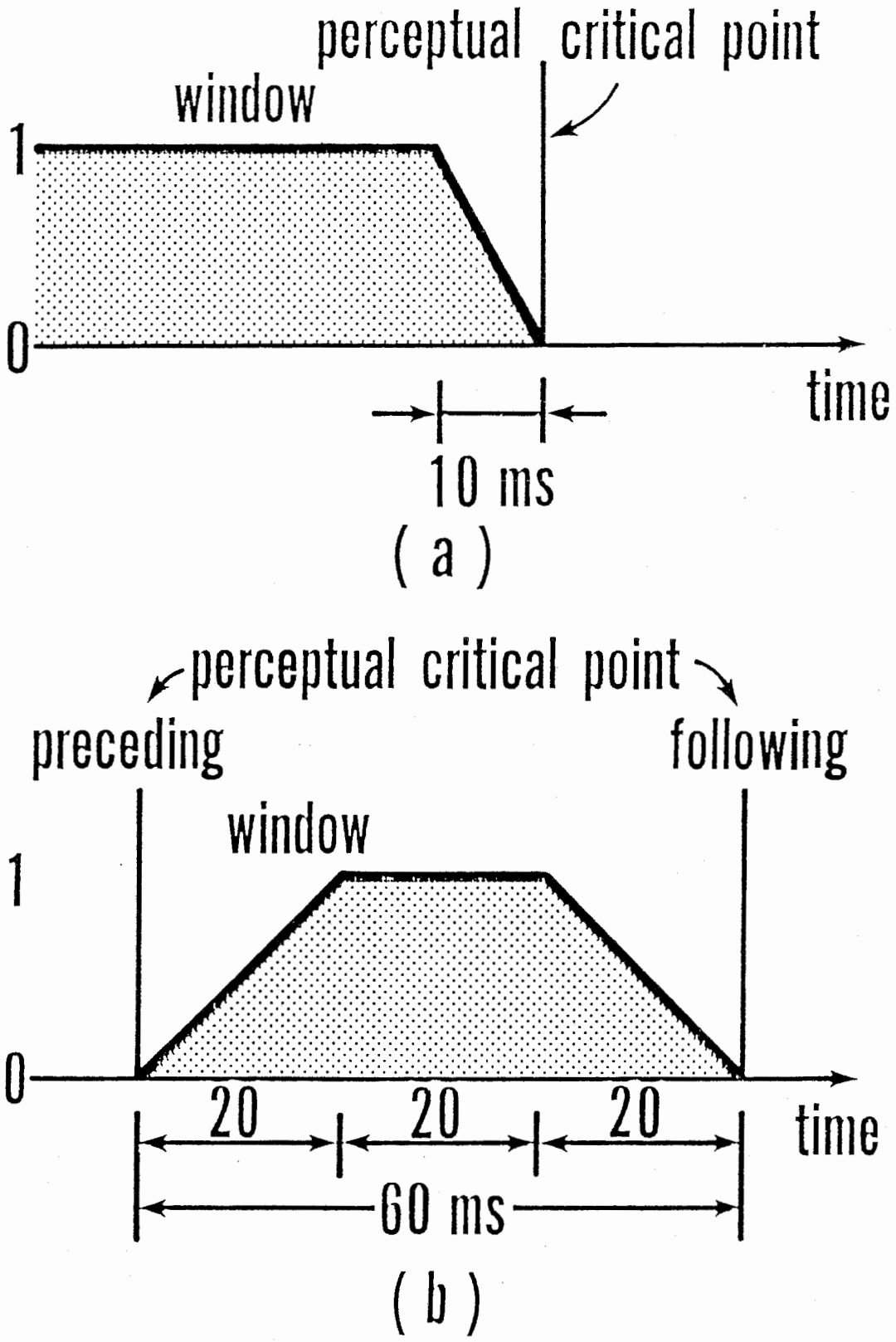


Fig.3 Truncation windows used for the listening experiments and perceptual critical points used for evaluation of the proposed model. (a) window for DATA I, (b) window for DATA II.

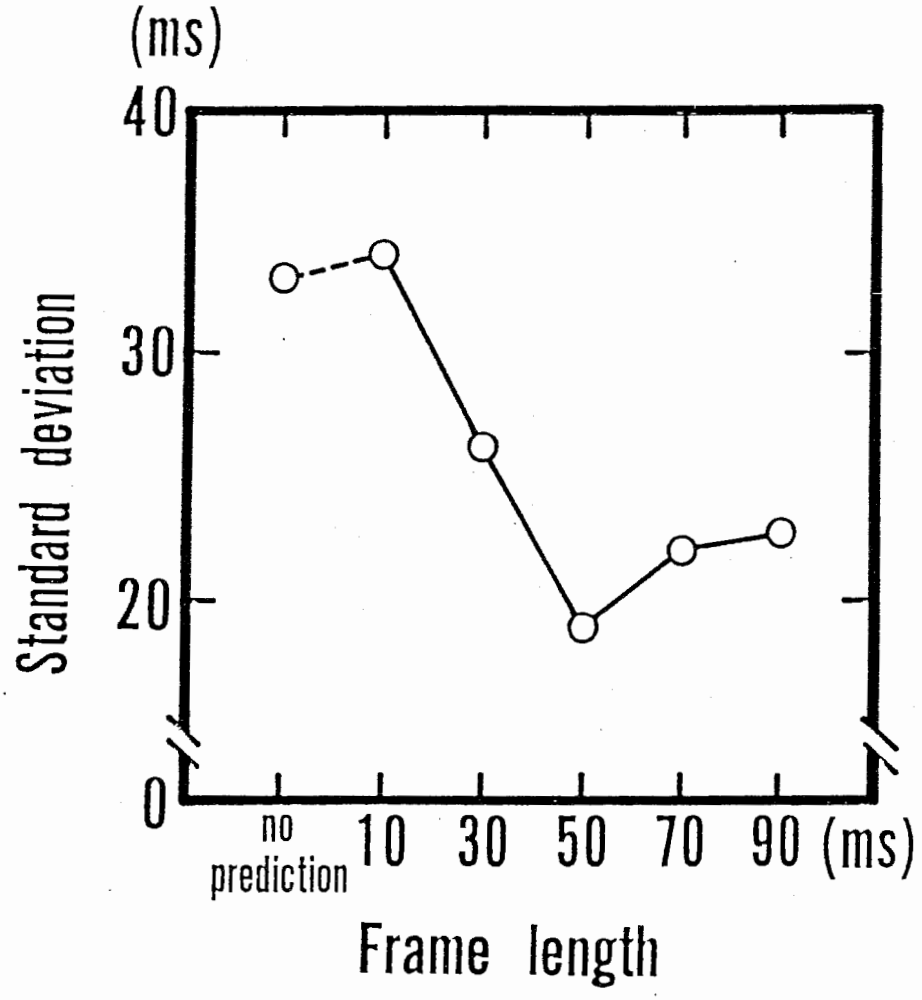


Fig.4 Standard deviation of the difference between perceptual critical point and its physically estimated point obtained based on the proposed model under five frame length conditions.

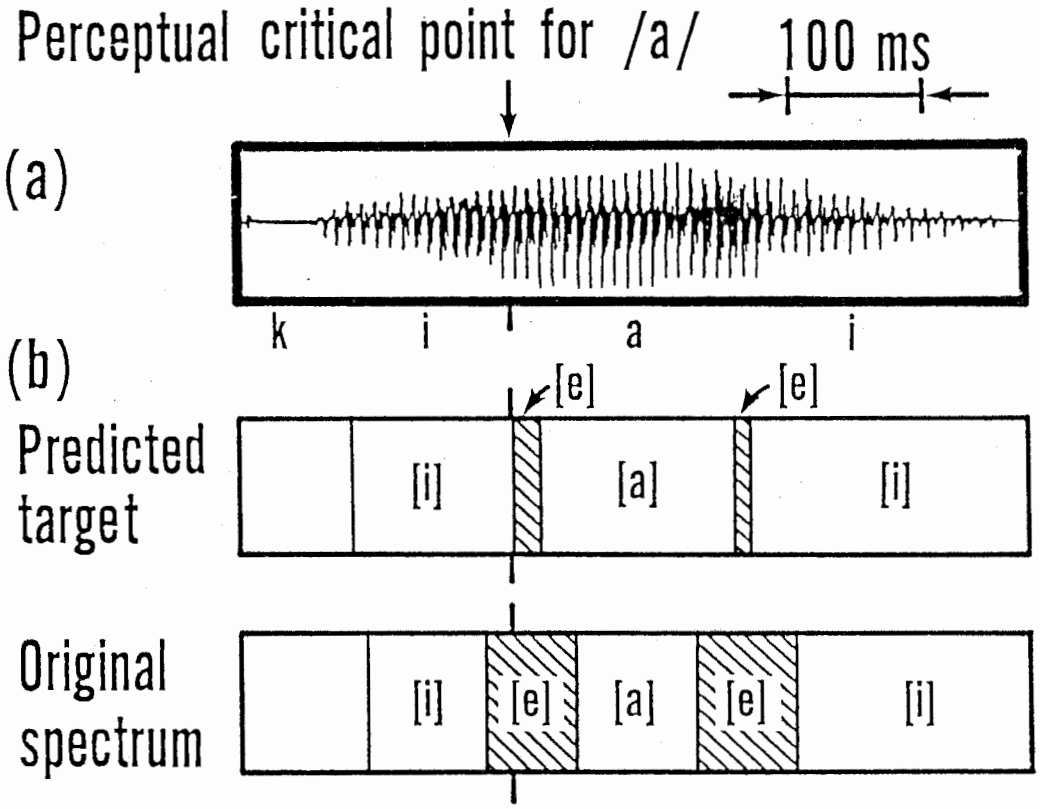


Fig.5 (a) Speech wave and (b) time sequences of vowel nearest to the predicted target spectrum, $\boldsymbol{b}$, or to the original spectrum, $\boldsymbol{y}$, at each short period in the Japanese word /kiai/.

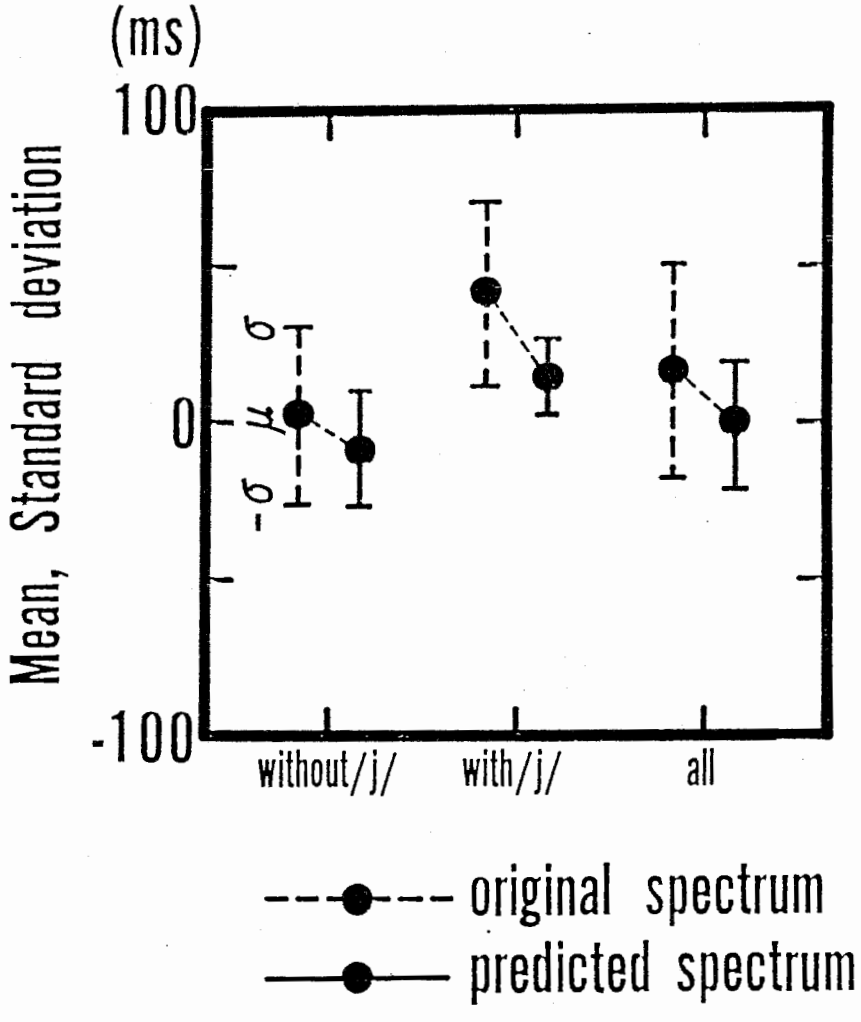
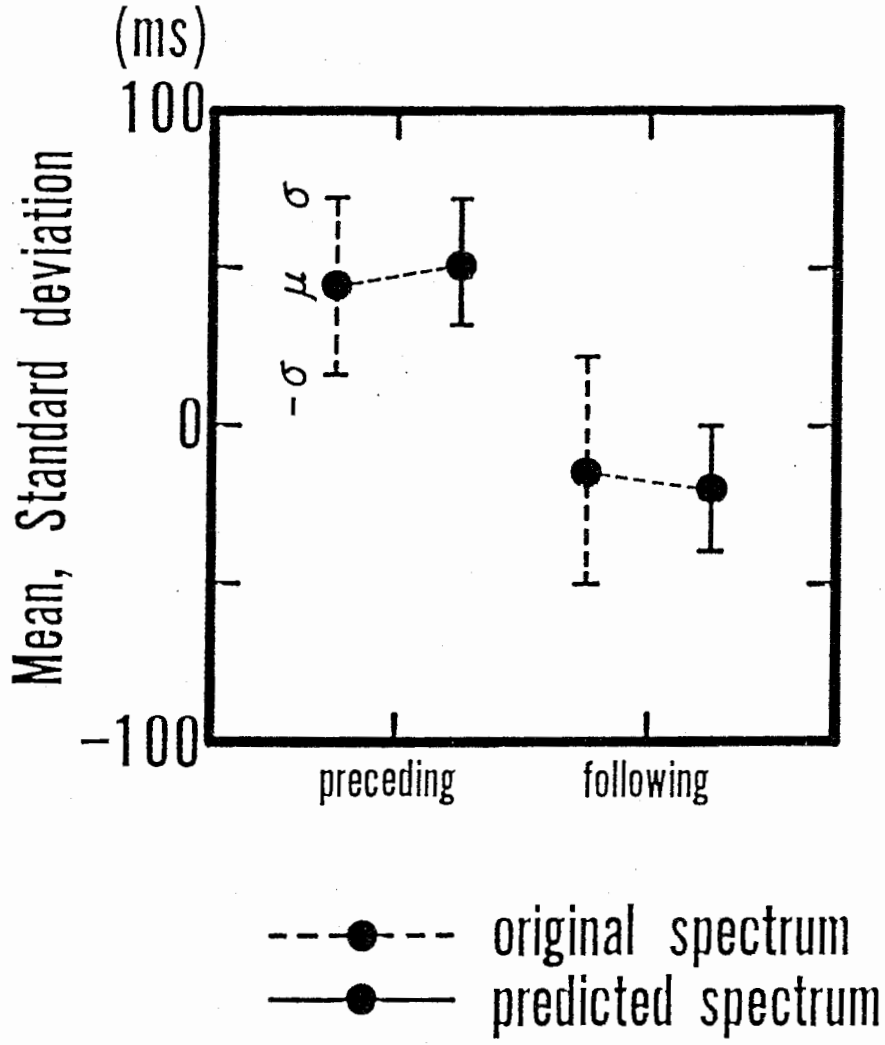


Fig.6 Mean (μ) and standard deviation (σ) of the difference between the perceptual critical point and its physically estimated point. (a) DATA I, (b) DATA II.


(ms)
100
0
−100
Mean, Standard deviation
σ
μ
−σ
preceding
following
original spectrum
predicted spectrum

(b)

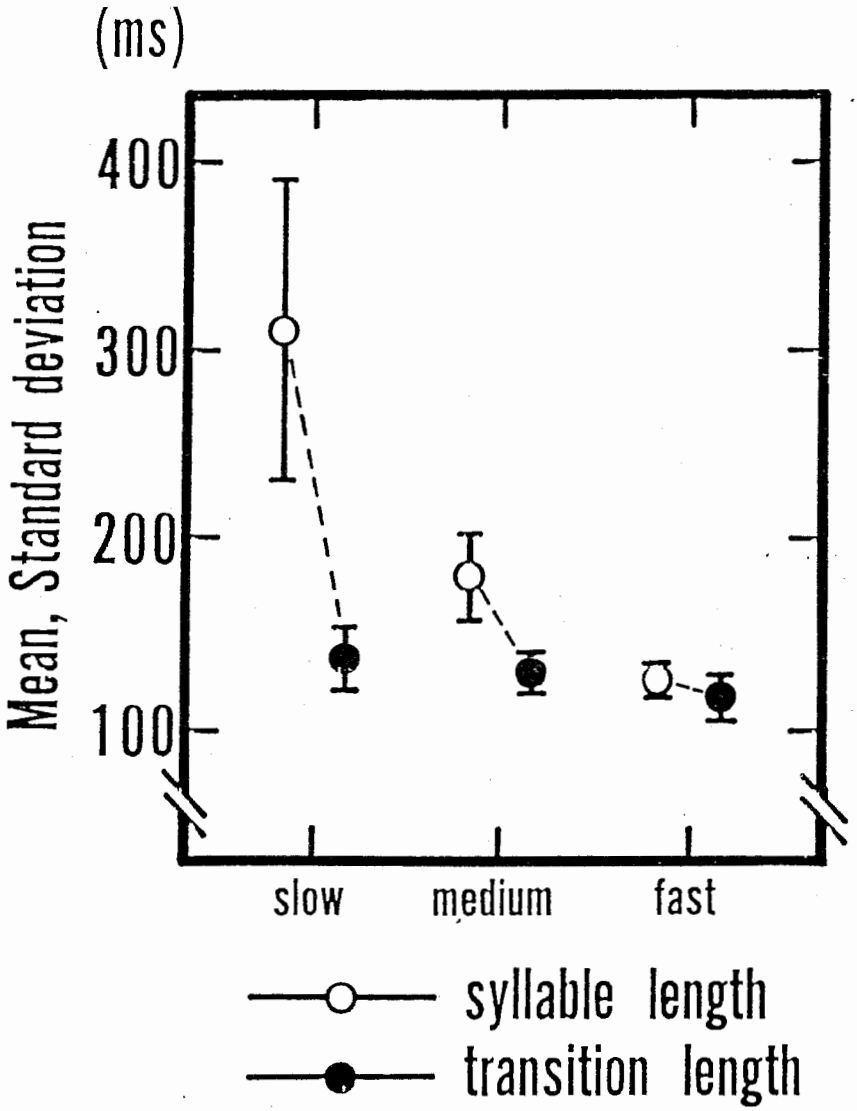


Fig.7 Mean (μ) and standard deviation (σ) of syllable lengths and lengths of the spectrum transition parts under three speaking rate conditions.

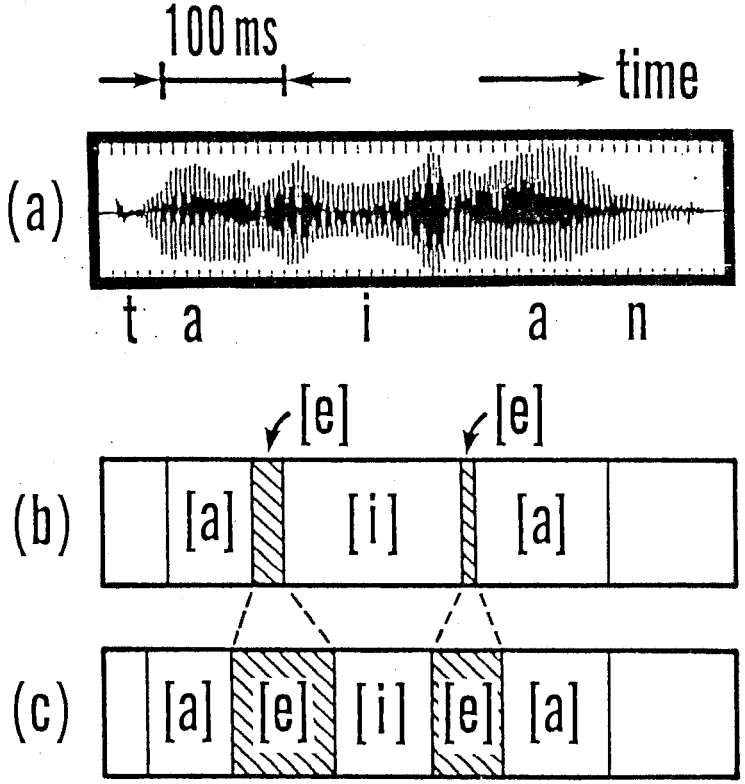


Fig.8 (a) Speech wave and (b) a time sequence of vowels nearest to the predicted target spectrum, $\boldsymbol{b}$, or (c)that to the original spectrum, $\boldsymbol{y}$, at each short period in the Japanese word /taian/ uttered under the fast speaking rate condition.

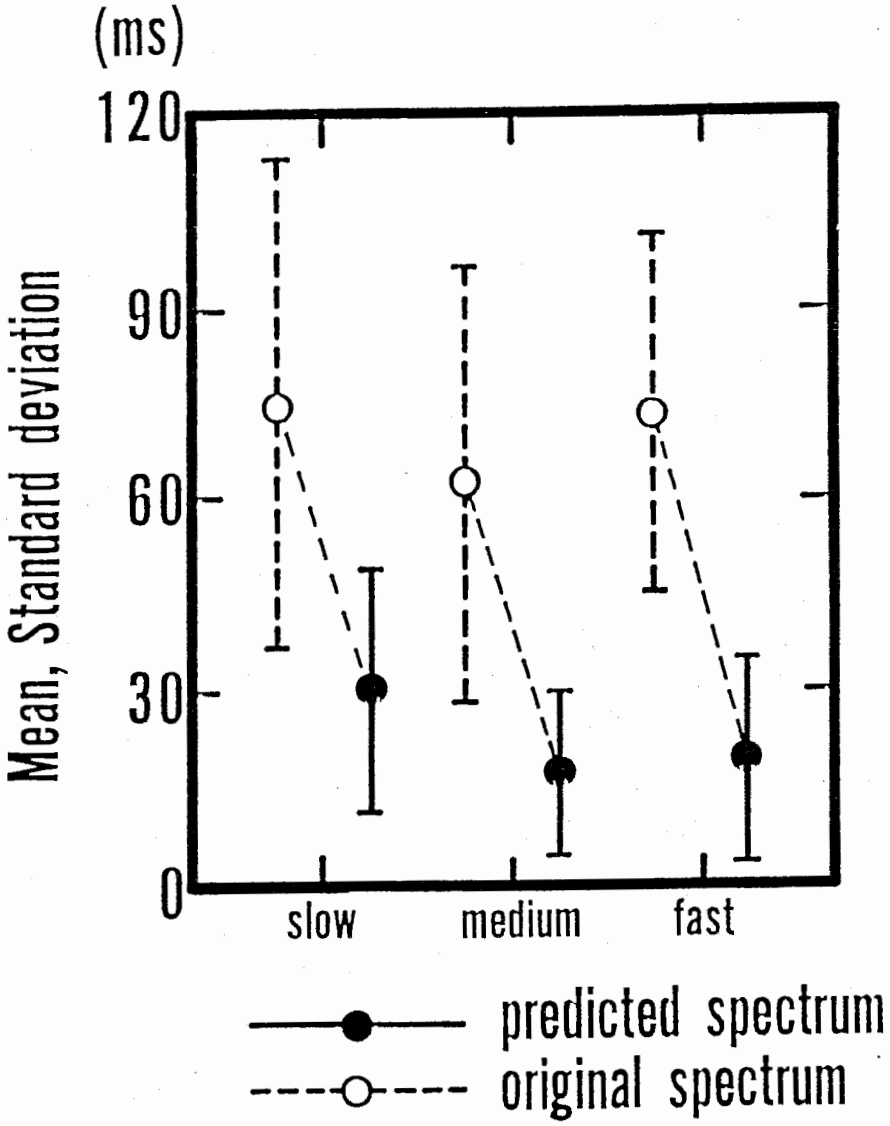


Fig.9 Mean (μ) and standard deviation (σ) of the transitional sound lengths under three speaking rate conditions.

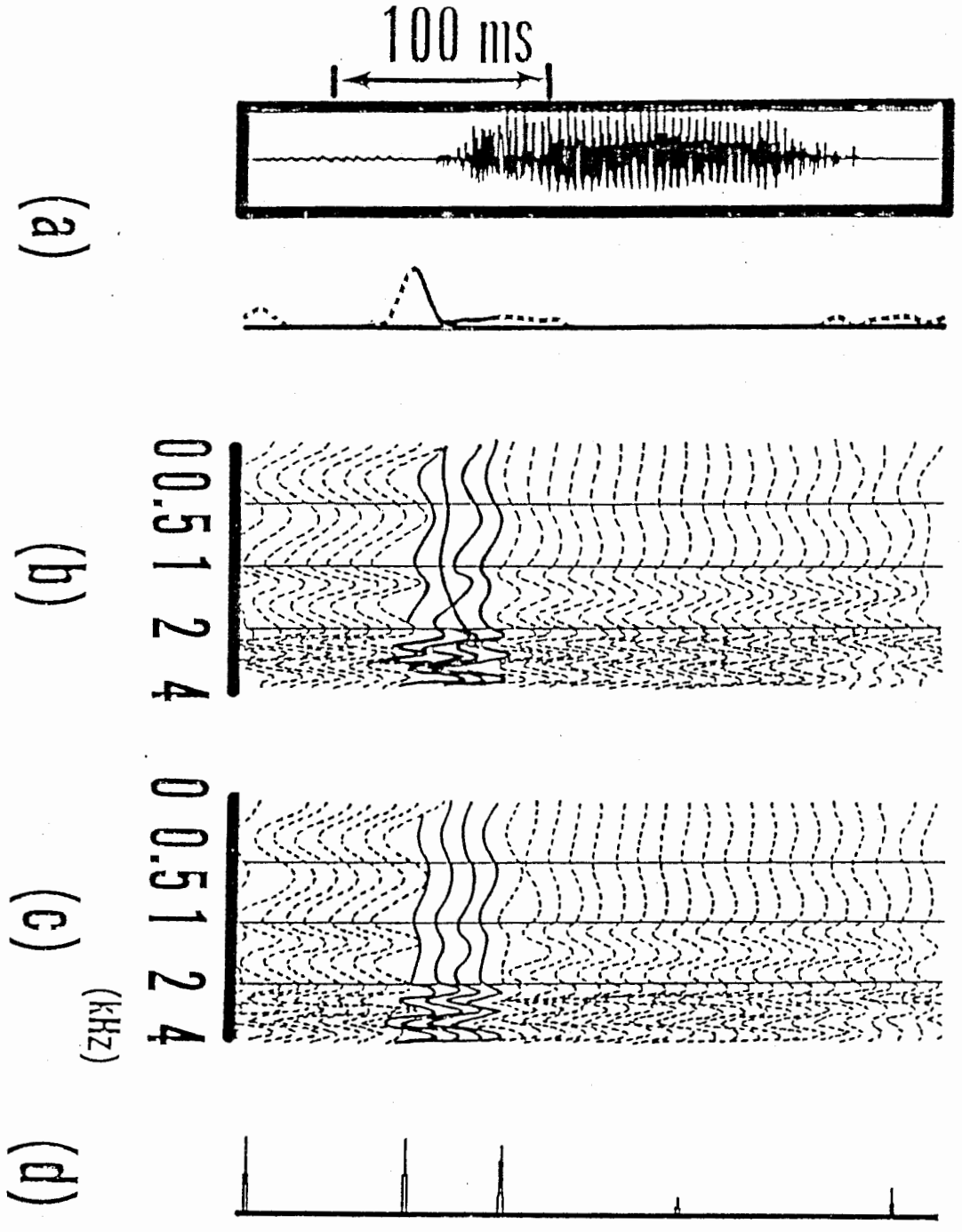


Fig.10 Output spectrum sequences associated with the proposed model for the Japanese syllable /ga/. (a) wave form and spectral transition measure, (b) spectrum sequence, $\boldsymbol{y}$, (c) estimated spectrum sequence, $\boldsymbol{b}$, and (d) measure of variation, $\mathbf{D}$. The parts of solid lines indicate the consonant parts.

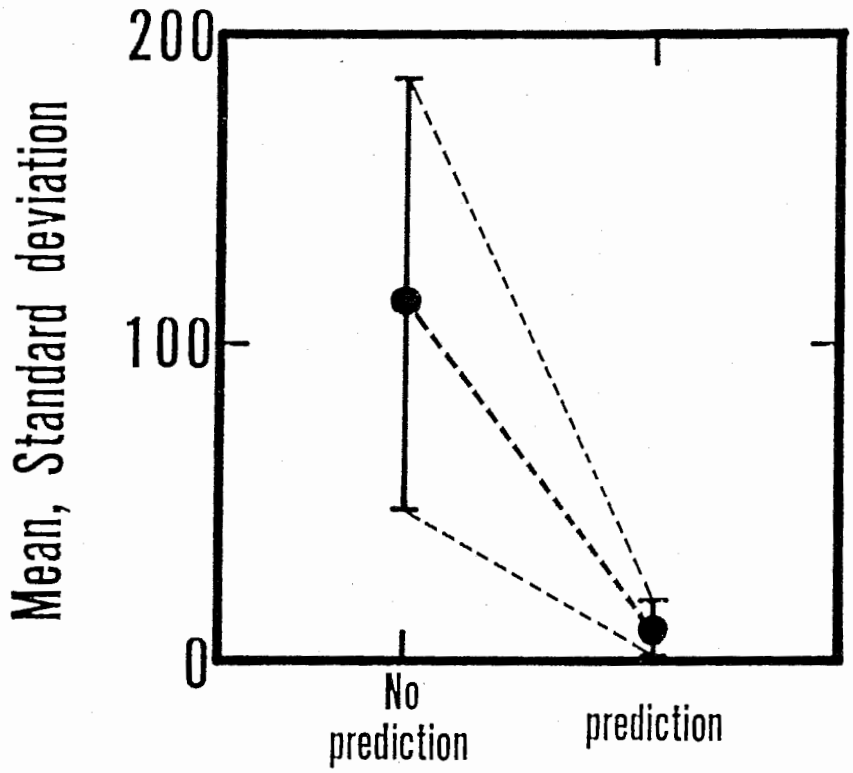


Fig.11 Mean (μ) and standard deviation (σ) of the sum total of the eigen values, S. The value S is the measure of variations.

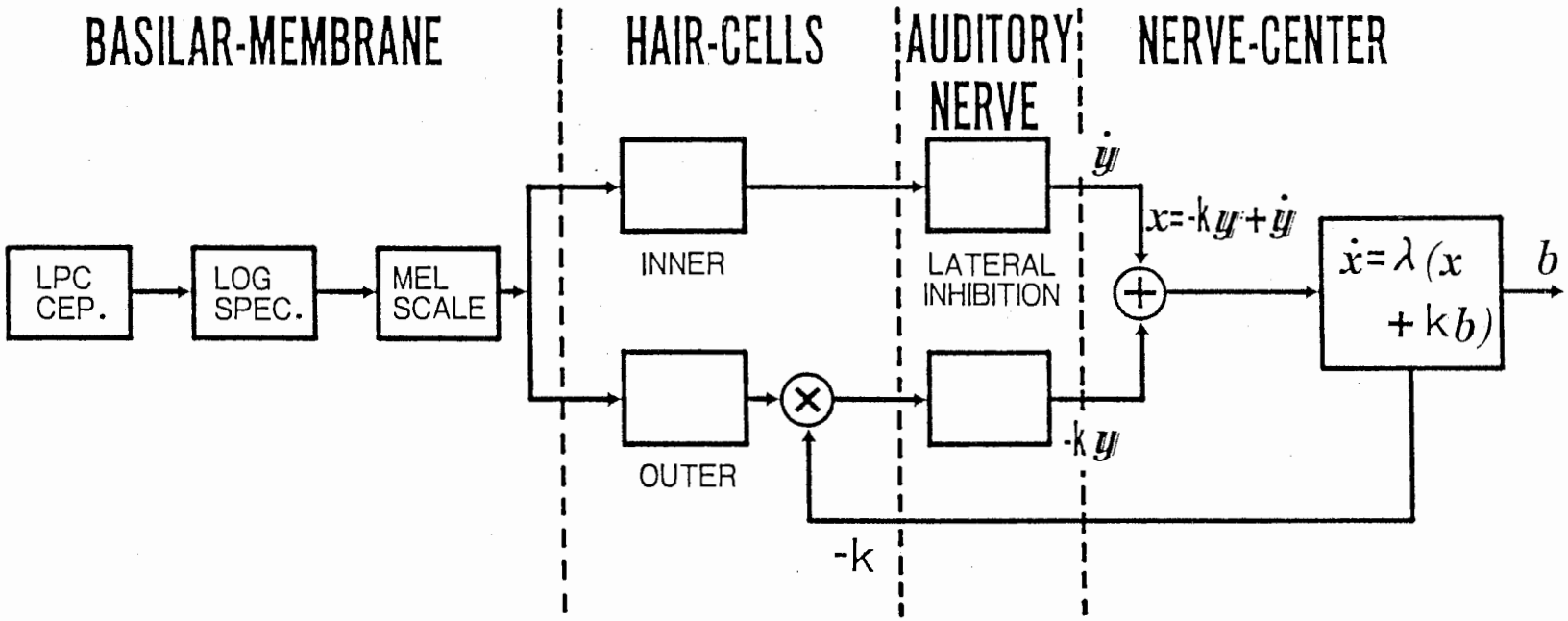


Fig.A-1 A peripheral auditory model showing a phoneme target prediction mechanism.

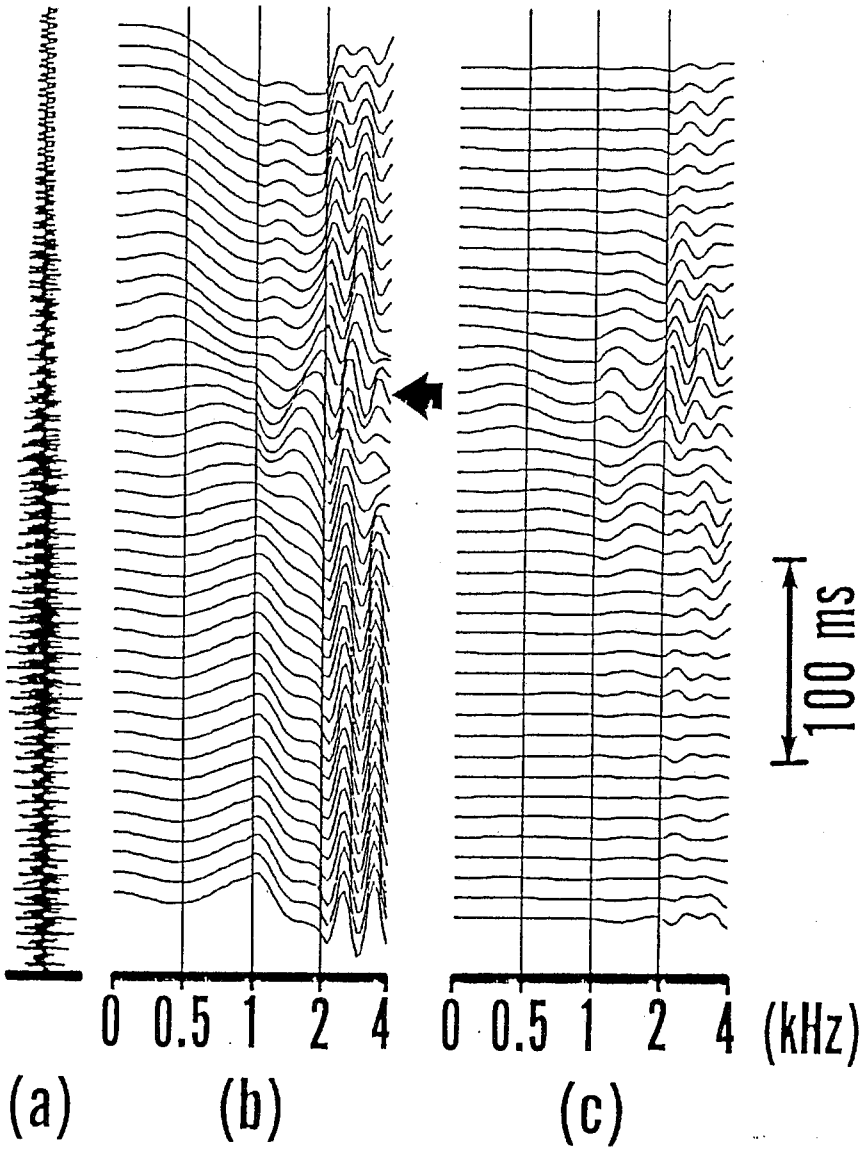


Fig.-A-2 Output spectrum sequences associated with the peripheral auditory model for a diphthong /ai/. (a) wave form, (b) spectrum sequence of the outer hair-cells model, $\boldsymbol{y}$, (c) differential spectrum sequence of the inner hair-cells model, $\dot{\boldsymbol{y}}$. Both $\boldsymbol{y}$ and $\dot{\boldsymbol{y}}$ are obtained through the lateral inhibition model.