FEATURE EXTRACTION AND SENTENCE RECOGNITION ALGORITHM IN SPEECH INPUT SYSTEM

K. Shirai
Department of Electrical Engineering, Waseda University,
Tokyo, Japan

Abstract

A feature extraction method for speech waves and an algorithm for sentence recognition are studied. The feature extraction is based on an articulatory model constructed from the statistical analysis of X-ray data. The model holds implicitly the physiological constraints and made possible to estimate the state of the articulatory mechanism. The estimated articulatory parameters provide a set of good features for the speech recognition. The sentence recognition problem is mathematically formulated as an optimization problem with constraints by introducing sentence structures from the syntactic and semantic considerations. The algorithm presents an optimal solution in the Bayesian sense.

Introduction

In this paper two major components of the speech understanding system are discussed. One is a feature extraction method for the speech wave and the other is a sentence recognition algorithm. Many speech pattern recognition systems have been made to classify spoken words and a few have been tried to treat spoken sentences. A speech recognition system which employs information from all levels - from the aconstic to the semantic - to understand the meaning of the speech is of current interest.[1~5] It is true that the reseaches at the higher levels such as the syntactic and semantic levels have not been sufficiently applied to speech recognition. However, the balance of each level of the system is Important. The application of the information at the higher levels might improve the total performance of the system but it should be noted that the communication system in which little redundancy is remained is deprived of variability and adaptability. Though the information at the upper level may be used in the manner of human speech understanding, it does not mean that he cannot correctly recognize a single word which is pronounced clearly. Therefore, the more effort should be necessary for the recognition of words or phonemes. If the speech understanding system is objected. In this study, first, a feature extraction method for the speech wave is treated as the most elemental problem.

Recently, several authors have studied the estimation of the vocal tract shape from the speech wave. But till now the results are not necessarily satisfactory, because, as is well known, the spectral characteristics of the speech wave do not correspond one to one to the vocal tract shape as an acoustic tube and because the speech wave carries the effects of the vocal cord oscillation, noise from the turbulence and so on. Then it is desirable to apply the knowledge from physiological and phonological study for the estimation of the vocal tract shape. And such a reseach has potentiality to make possible the estimation of the motor commands that move the articulatory mechanism. It is doubtful that, if the motor commands are precisely estimated, the phoneme recognition becomes easy. However, they may be good features for the speech recognition. In this study, an articulatory model is constructed on the basis of X-ray data and a nonlinear regression method is used to estimate the articulatory parameters. The results of the estimation preserve the typical nature of each phoneme and the method would provide a useful feature extraction technique.

Second, the problem of sentence recognition is mathematically formulated as an optimization problem with the constraint of sentence structures and is solved by a method of dynamic programming.

Here, it might be necessary to define what is recognition of sentence. The fundamental situation of the speech is conversation. In the conversation between A and B, the representation " B understands the sentence spoken by A," has too much content. Then we consider only," B responds to the sentence spoken by A." The response is described by a state transition of the machine. Of course, the machine may exhibit some outputs at the transition.

Each moment of the conversation is accompanied by a scene and usually the speaker and the receiver have a common recognition of the scene. The concept of the scene Is naturally taken Into account by the state of the machine. Probable sentences that may appear under a state are limited an ' the effective number of the sentences that affects the recognition score is reduced.

In a practical application, the purpose and the ability of the machine Is always limited, and the contents of the conversation may be finite. Then, it is allowable to set sentence structures to order words in restricted ways.

In the framework mentioned above, the sentence recognition can be considered on the extension of a classification problem and an effective optimal algorithm for the sentence recognition can be obtained.

Feature Extraction via Estimation of Articulatory Parameters

Construction of Articulatory Model

An articulatory model presents an effective representation for the structures of the articulatory mechanism and the dynamical characteristics of the articulatory motion, and further it relates the articuatory parameters to the acoustic ones? [9]

The configulation of the articulatory model is shown in Fig.I. The mIdsaggital vocal tract outline can be represented by the variables specifying the positions of the movable structures, i.e. jaw, tongue and lips. The maxilla, rear-pharyngeal wall and larynx outlines are fixed and approximated by the sequence of circular arcs and straight lines. The Jaw Is assumed to rotate with the fixed radius F_jJ about the fixed point Fj, and its location J 1s given by the angle θ_j with respect to the reference line which is tangent to the hard palate. The jaw movement executes the passive effect to the position of

the tongue and lips, and Influences not only r^{Ua} mouth opening area but the overall vocal tract shape.

The lip shape is specified by the height L_n and the protrusion Lp relative to the jaw position on the midsagittal plane. Only the lip protrusion parameter may be necessary to specify the lip movement for the vowels, but both are required to explain the different gestures in the usual speech containing labial consonants.

The tongue contour is described in terms of a semi-polar coordinate system defined with reference to the jaw position. The center F_t rotates synchronously with the jaw movement[1.1] Therefore, the tongue contour is measured with the jaw based coordinate system. Though the tongue may be able to form various shape, it has limited freedom to move about in articulatory process on account of the physiological and phonological constraints. Thdse constraints can be expressed by the strong correlation of the position of each segment along the tongue contour and may be extracted from the statistical analysis of the X-ray data.

It ia known that for the tongue articulation of vowels, the extrinsic muscle activity is more significant than that of the intrinsic one. Then, the principal components for the extrinsic actvity are obtained from the tongue contour data for vowels, and the tongue contour vector for vowels X_v can be expressed in the linear form as,

$$X_v = \sum_{j=1}^{p} a_j V_j + \bar{X}_v \quad , \qquad (1)$$

where, V_j $(j=1.2,\ldots,p)$ are eigenvectors and $\bar{X}_v$ is a mean vector for vowels which corresponds to the neutral tongue contour. The eigenvectors are calculated from the next equation.

$$C_{xxv} V = \lambda V \quad , \qquad (2)$$

$$C_{xxv} = E[(X_v-\bar{X}_v)(X_v-\bar{X}_v)^T] \quad ,$$

and A is the corresponding eigenvalue to satisfy the characteristic equation.

$$| C_{xxv} - \lambda I | = 0. \qquad (A)$$

For the consonants, the effect of the intrinsic muscle activity appears particularly in the front part of the tongue but it is difficult to separate precisely the intrinsic muscle activity from the extrinsic one. At first the contribution which comes from the extrinsic components are subtracted by projecting the tongue contour vector X_c to the vowel space which is spanned by the eigenvectors for the vowels. And the remainder $\tilde{X}_c$ is calculated as,

$$\begin{aligned} \tilde{X}_c &= X_c - \hat{X}_c \\ &= X_c - \{ \sum_{j=1}^{p} V_j V_j^T (X_c-\bar{X}_v) + \bar{X}_v \} \end{aligned} \qquad (5)$$

where $\hat{X}_c$ means the projection of X_c to the vowel space. Again tho principal component analysis is performed on $\tilde{X}_c$ and the eigenvectors C_k (k»l,2, ...,q) is calculated In the same manner as V_j. Finally the expression for the consonants can be obtained as,

$$X_c = \sum_{j=1}^{p} a_j V_j + \sum_{k=1}^{q} b_k C_k + \bar{X}_c \quad , \qquad (6)$$

where $\bar{X}_c$ is the sum of the mean vector $\tilde{X}_c$ and $\bar{X}_v$.

Then a_j $(j=1,2,\ldots,p)$ and b_k $(k=1,2,\ldots,q)$ become the articulatory parameters for the tongue.

The above procedure was applied to the X-ray data by J.S.Perkell.[2]

The extracted principal components of the tongue contour and their cumulative distributions are shown in Fig.2. It is found that the four components (p=3, q=1) account for roughly 96% of the tongue data variance.

The first component represents the movement of the tongue body between the rear-pharyngeal wall and the hard palate direction and produces mainly an antisymmetric perturbation of the vocal tract. It Indicates the opposite feature of back and front vowels, i.e. (a] vs [i]. The second component represents the movemerat of the tongue towards the velum and produces a symmetric perturbation that is effective for the rounded vowel [u]. The third component is less clearly explained and may be interpreted as the resulting tongue deformation from the contraction of the posterior fibers of genioglosus and the intrinsic muscle of the tongue tip. The fourth component Is an intrinsic component and represents the tongue tip retroflex.

As an example the loci of the tongue movement on ai~a2 space for three utterances /h3tV/ (V: a,i,u) are illustrated in Fig.3. The points A on the loci indicate the onset and the offset of the tongue tip closure.

From the above discussions, it is seen that the following articulatory parameters are enough to describe the midsagIttal vocal tract outline, i.e. the jaw angle $_J$, the lip protrusion L_p and the weighting coefficients of the tongue components aj, bfc. For the vowels the lip height is dependent on the lip protrusion and can be approximated by,

$$L_h - 0.3 - 0.25(L_p-1.0). \qquad (7)$$

The relation (7) was determined from the analysis of the front and side photographs.

The cross-sectional dimension along the vocal tract is determined from a semi—polar coordinate system fixed with regard to the maxilla and the rear-pharyngeal wall. The relation between the cross dimension d and the cross sectional area S is approximated by power function $S\text{-}2d^{1.fi}$. In the labial region the area is approximated by an ellipse with the width given by

$$L_w = (7.8 - 4L_p)\sqrt{L_s/(L_s+0.2)} \quad . \qquad (8)$$

where L_s is the vertical separation of the lips.

The vocal tract is devided Into 30 uniform cylindrical tubes and the reflection coefficients between the adjoining sections are calculated. Regarding the losses at the glottis, lips and within the tract, a transmission-line model is constructed and the transfer function is expressed using z-transform.

$$|G(z^{-1})| = \frac{K}{|1+a_1 z^{-1} + \cdots + a_{30} z^{-30}|} \qquad (9)$$

where z-exp(j2πfΔT), ΔT=2h/c, h the length of one section and c is the velocity of sound. The formant frequencies are calculated from Eq.(9) by the Fibonacci searching method.

Estimation of Articulatory Parameters

It is well known that the vocal tract shape is not uniquely determined from the spectral characteristics of the speech wave without the additional constraints in terms of the speech production process. Such constraints must be considered from the physiological, phonological and personality points of view. The constraints can be reflected on the articulatory model in two ways One is in the physical dimension of the articulatory organs and In the components vectors V_i and O_k The other is the maimer of the control of the articulatory parameters. The latter is considered in this study only a little. It is clear that the number of the articulatory parameters is much smaller than that of the cylindrical tubes to describe the vocal tract shape and this fact will make the estimation easy. Therefore, the vocal tract shape is determined by estimating these articulatory parameters. The present model is useful for the estimation from the speech wave, because it is constructed not only to describe the articulatory state strictly but also to be directly related to the variation of the vocal tract shape. However, there remains a possibility to bring about a freedom in the articulatory parameters for a certain region of the spectral characteristics of a speech sound. Therefore, it is desirable to apply the cooperative relation between the articulators in static and dynamic senses to avoid such a freedom.

The time constants of the articulatory motion are large compared with the sound propagation phenomena and in the case of the articulation of vowels, the power spectral density of the speech sound is considered to be stationary in a smal1 inLerval. Therefore, the static correspondence between the articulatory parameters and the acoustic ones is very important.

The relationship between the articulatory parameters and the acoustic features is formulated in nonlinear regression as,

$$y_i = P_o + \sum_{j=1}^{m} P_{ij}(f_j-\bar{f}_j) + \sum_{j=1}^{m}\sum_{k=1}^{m} Q_{ijk}(f_j-\bar{f}_j)(f_k-\bar{f}_k) + \cdots , \quad (10)$$

where y_i is an articulatory parameter, f_j is an acoustic feature and P_o, P_{ij} and Q_{ijk} are the regression coefficients. Since Eq.(10) is linear with respect to the regression coefficients, it is rewritten as,

$$Y = \Phi U, \quad (11)$$

where Y is a vector $(y_1\ y_2\ \cdots\ y_r)^T = (O_1\ a_1\ a_2 \cdots a_p\ b_1\ b_2 \cdots b_q\ L_p\ L_h)^T$, Φ is a regression matrix and U is a vector $(1, f_1-\bar{f}_1, f_2-\bar{f}_2, \ldots, (f_1-\bar{f}_1)^2, (f_2-\bar{f}_2)^2, (f_1-\bar{f}_1)(f_2-\bar{f}_2), \ldots)^T$.

Applying the least squares method, the estimates of the regression coefficients are given by,

$$\hat{\Phi} = C_{YU}\ C_{UU}^{-1} \quad (12)$$

where Cy and C are the covariance matrices and $\hat{\Phi}$ means the estimated value of Φ.

Four variables are used as the articulatory parameters, namely the jaw opening O_1, the weighting coefficients of the principal components of the tongue contour a1 and a2, and the lip protrusion L_p, which were introduced in the preceding section. As the acoustic features, first two formant frequencies F_1 and F_2 are used, because they are the most significant features to be closely related to the vocal tract shape for the vowel-like sounds. The data from which the regression coefficients are determined consist of the samples distributed around the five Japanese vowels and real ones obtained from the X-ray pictures, and the total number of the samples is 300. By utilizing the real articulatory data for the estimation, some cooperative relation between the articulators is included in the regression coefficients. The formant frequencies are calculated according to the algorithm presented in the preceding section.

The estimated result for the synthesized continuous speech /a i u e o/ is shown in Fig. 4 in comparison with the original articulatory parameters. Solid lines show the original articulatory motion. The first and the second formant frequencies at every moment are calculated for this articulatory movement and conversely the articulatory parameters are estimated by Eq.(II) for those formant frequencies. The estimated values agree with the original ones except for the slight deviation in the tongue parameters a^ and .IT, However, this result means that if the model is just fitted for the speaker, the estimation would be successful by the nonlinear regression method.

The result for the real speech data /a i u e 0/ is shown in Fig.5. The speaker is different from the person whose data were used to construct the model. Although the estimated values cannot be compared with the true ones because the X-ray data were not taken for this utterance, they convey the typical nature of each vowel. For example, vowel [u] is characterized by the most positive value of a2 and the lip protrusion and clearly distinguished from [o] by the differences of the jaw opening and the tongue parameter a*. Front vowel [i] is characterized by the most negative value of a_1 and distinguished from [e] by the difference of the degree of the jaw opening.

In Fig.4 and Fig.5, the marks which indicate the estimated values of the fonnant frequencies mean the calculated formant values corresponding to the estimated articulatory parameters. It is seen that the correspondence in the fonnant domain is very well. Nevertheless, there are slight deviation In the articulatory parameters in Fig.4. This indicates that the first two fonnant frequencies are not sufficient to decide precisely the articulatory parameters In some region.

In a practical point of view, the outputs of filter bank is more convenient than the fonnant frequencies as a set of acoustic parameters, because the calculation of the formant frequencies is not so easy matter. Then the estimation using the output of the filter bank was tried. The out-

put signals of the filterbank were reduced to 3 or 4 components by the principal components analysis and those main components were used as the acoustic-parameters in the nonlinear regression. 'the result is shown in Fig.6 for the synthesized voice /a i u e o/. Compared with Fig,4, the accuracy is almost the same. The original and the estimated spectral patterns are shown in Fig.7. Formant frequencies from first to fourth are in good agreement.

It may be concluded that the articulatory parameters estimated by the nonlinear regression method can be employed as a feature vector for the speech recognition.

Sentence Recognition Algorithm

Formulation of the Sentence Recognition Problem

in this section the sentence recognition problem will be given a mathematical formulation. Sentence structures mean the categolization of the types of the sentences according to the syntactic and semantic contents. The method to set the sentence structures depends upon the scale and the complexity of the problem. In this study, from the practical view point, it is assumed that the number of the words is not so large and the language can be described by the context free grammar. The construction of the C.F.G. for the given problem becomes important. It may be difficult to find the general procedure. However, if an appropriate restriction for the speaker is settled, the categolization of the sentences is not so hard in a small scale problem.

First, a set of parts of speech $W=W_i$ $(i=1,2,\cdot,m)$ Is Introduced. The concept of a part of speech may be different from the linguistic one. It is defined so as to include its meaning in addition to its role in a sentence. The j-th word of the i-th part of speech is denoted by wj*, so that $W_i=\{w_{ij} | j=1,2,\cdot\cdot,N_i\}$. It may happen that a word is registered in two or more p a of speech The total vocabulary of the system is $\bigcup_{i=1}^{m} W_i$.

Second, sentence structures are described by a context free grammar $G_\alpha(V_n, V_t, P, z_\alpha)$ which gives the arrangement oi W^ in a sentence. The grammar G_α is assumed to be not ambiguous and

V_n : a set of non-terminal symbols,
V_t : a set of terminal symbols,
z_α : initial symbol,
P : a set of production rules.

The set V_t is the set of parts of speech, i.e. $V_t=\{W_i | i=1,2,\cdot\cdot,m\}$. The initial symbol z_a means that the state of the machine is at the a-th state The set of the state $Z=\{z_\alpha | \alpha=1,2,\cdot\cdot,c\}$, where c denotes the total number of the states of the machine. The grammar G_α produces the sentence structures which can appear under the a-th state. It is assumed that G_α is not ambiguous. Therefore one sentence corresponds to only one left most derivation and is described by a sequence of the used production rules. The set of the sentences which Is generated by a sequence of the production rules $P_{k(\alpha\beta,1)}P_{k(\alpha\beta,2)}\cdot\cdot P_{k(\alpha\beta,\ell)}$ denoted by $S_{\alpha\beta}$.

$$S_{\alpha\beta}=\{s_{\alpha\beta\gamma} | \gamma=1,2,\cdot\cdot,N_{\alpha\beta}\} = \{W_{i(\alpha\beta,1)}W_{i(\alpha\beta,2)}\cdot\cdot W_{i(\alpha\beta,L_\beta)}\}. \quad (13)$$

$$z_\alpha \xrightarrow{P_{k(\alpha\beta,1)}P_{k(\alpha\beta,2)}\cdots P_{k(\alpha\beta,\ell)}} W_{i(\alpha\beta,1)}W_{i(\alpha\beta,2)}\cdot\cdot W_{i(\alpha\beta,L_\beta)}, \quad (U)$$

when' LB shows the length of the sentence $S_{\alpha\beta}$. An element $s_{\alpha\beta\gamma}$ means the Y-th sentence with the β-th sentence structure which may be arise under the a-th state, and can be expressed as.

$$s_{\alpha\beta\gamma}= W_{i(\alpha\beta,1)j(\alpha\beta,\gamma,1)}W_{i(\alpha\beta,2)j(\alpha\beta,\gamma,2)}\cdots \cdots W_{j(\alpha\beta,L_\beta)j(\alpha\beta,\gamma,L_\beta)}, \quad (15)$$

where $i(\alpha\beta,h)$ indicates the part of speech o£ the h-th word in the sentence of $S_{\alpha\beta}$ and further $j(\alpha\beta,\gamma,h)$ addresses a word In W_i The fact that one sequence of words has meaning is considered as that the sequence is one possible $s_{\alpha\beta\gamma}$.

The purpose to introduce the sentence structure is that the strong mutual dependence of words in a sentence Is* absorbed in the sentence structure a n $j(\alpha\beta,\cdot,h)$ be e s nearly statistically independent in $S_{\alpha\beta}$. The operation of the machine can be shown as Fig.8 that is an example to give orders for a robot in a dialogue to move forward or turn in an assigned manner.

Method of Sentence Recognition

Let the speech pattern representing one word be Ω, which is a random vectors sequence. A sentence can be represented by a string of Ω as $\mathcal{A}=[\Omega_1,\Omega_2,\ldots,\Omega_L]$. When a sentence $\mathcal{A}$ is observed at the state z_α, the sentence recognition is to determine β and γ. The optimal determination of β and γ in the sense of minimum error is that of maximizing the a posterior probability $f_\alpha(\beta,\gamma|\mathcal{A})$.

$$f_\alpha(\beta,\gamma|\mathcal{A}) = f_\alpha(s_{\alpha\beta\gamma}|\mathcal{A},\beta)\, f_\alpha(\beta|\mathcal{A}). \quad (16)$$

The numbers $j(\alpha\beta,\cdot,h)$ in β are assumed statistically independent. Then, Eq.(16) is rewritten as,

$$f_\alpha(\beta,\gamma|\mathcal{A}) = \left[\prod_{h=1}^{L} f_\alpha(j(\alpha\beta,\gamma,h)|\beta,\Omega_h)\right] f_\alpha(\beta|\mathcal{A}). \quad (17)$$

Using the Bayes theorem and taking the logarithm, the objective function to be maximized is got as,

$$D_\alpha(\beta,\gamma|\mathcal{A}) = \sum_{h=1}^{L} d_{i(\alpha\beta,h)j(\alpha\beta,\gamma,h)}(\Omega_h) + \log f_\alpha(\beta), \quad (18)$$

where the common term was discarded and,

$$d_{i(\alpha\beta,h)j(\alpha\beta,\gamma,h)}(\Omega_h) = \log f_\alpha(\Omega_h | w_{i(\alpha\beta,h)j(\alpha\beta,\gamma,h)}) + \log f_\alpha(w_{i(\alpha\beta,h)j(\alpha\beta,\gamma,h)}|\beta). \quad (19)$$

The first term of Eq.(19) comes from the word recognition and the second terms of Eq.(18) and Eq.(19) carry information on the context and the situation. Then the sentence recognition problem

was given an effective formulation as an optimization of $D_\alpha(\beta,\gamma \mid \mathcal{A})$ with constraints come from the fact that β and γ must be selected from a restricted set.

If the constraint about γ is neglected, the optimization can be performed by using dynamic programming. Let the maximum value of $D_\alpha(\beta, j(\alpha\beta,\cdot,h) \mid \mathcal{A})$ at the h-th stage be,

$$\Lambda_h(\beta) = \max D_\alpha(\beta, j(\alpha\beta,\cdot,h) \mid \mathcal{A}). \qquad (20)$$

The following recursive formula is derived.

$$\Lambda_h(\beta) = \max \{ d_{i(\alpha\beta,h)j(\alpha\beta,\cdot,h)}(\Omega_h) + \Lambda_{h-1}(\beta) \}, \qquad (21)$$

where

$$\Lambda_0(\beta) = \log f_\alpha(\beta).$$

From this formula one optimal sequence of words for each β that satisfies $L_\beta = L$ can be determined. Therefore, if a sequence is found to fit for any β^*, γ^* the optimal solution is decided as $\beta = \beta^*$ and $\gamma = \gamma^*$. In this method, there is a possibility to appear combinations of $j(\alpha\beta,\cdot,h)$ which correspond to not allowable or meaningless sentences. In such case the calculated β^* and γ^* should be rejected and the second optimal solution should be tested.

An important problem arises in the above formulation. That Is the aberration of the scene recognition between the speaker and the machine. The speaker does not know the state of the machine or the speaker utters a sentence by mistake that should not be permitted in that situation. These phenomena often occur in the actual conversation. Particularly in the case that the machine made a misrecognltion and went to a state unexpected by the speaker, It is difficult for the speaker to do a suitable action. In such aberration condition the above algorithm cannot work satisfactorily. This phenomenon always appear when the relation between the speaker and the receiver is made tight to improve the recognition score.

Conclusion

In this study two important parts of the speech understanding system were considered. The feature extraction method that utilizes the physiological and the phonological constraints was proposed. It will be effective for the speech recognition by improving the word or phoneme recognition score.

Recently, several attempts have been made to estimate the cross-sectional area function of the vocal tract using the state space expression of the acoustic vave in the vocal tract. However, in those framework, it will make the problem too much complicated one to consider the various constraints of the artlculatory motion. If the dynamics is taken into account in any sense, the dynamic character of the artlculatory motion should be considered first and the state space expression of the acoustic level may be ignored because of the difference in their time constants.

The sentence recognition algorithm is well formulated and very compact. Then, it makes easy the real time operation of the speech understanding without using special hard wares.

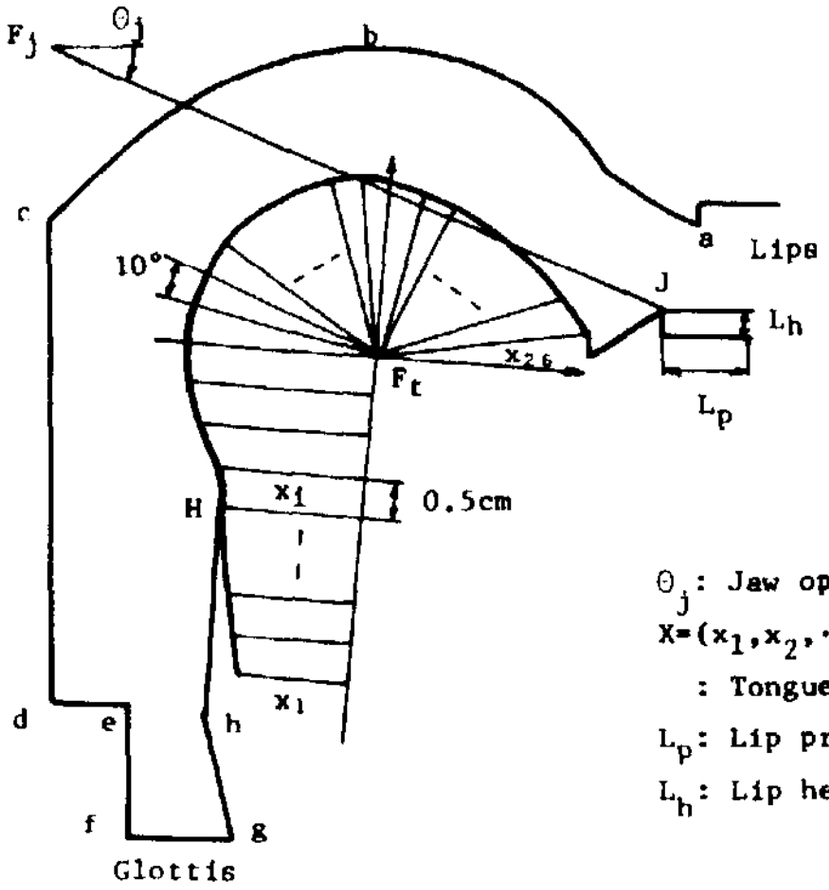
The two methods were discussed separately in this paper. But, if the suitable algorithm to decide phonemes from the extracted feature is added, the system will be completed.

The author thanks Dr. H. Fujisawa and Mr, M. Honda for their cooperation.

This research was partly supported by Kawakami Memorial Foundation.

References

1) D.R.Reddy et al: A Model and System for Machine Recognition of Speech, IEEE, Trans.AU-21, June, 1973.
2) M.Kohda, K.Shikano: Speech Recognition of Arithmetic Statements Utilizing Syntactic Information, IECE Japan, Report EA 73-54, March, 1974.
3) W.A.Woods: Motivation and Overview of SPEECHLIS : An Experimental Prototype Speech Understanding Research, IEEE, Trans.ASSP-23, Feb., 1975.
4) V.R.Lesser et al: Organization of Hearsay It Speech Understanding System, IEEE, Trans. ASSP-23, Feb., 1975.
5) J.K.Baker: The DRAGON System-An Overview, IEEE, Trans. ASSP-23, Feb., 1975.
6) T.Nakajima et al: Estimation of Vocal Tract Area Functions by Adaptive Inverse Filtering Methods, Bull, of Electrotechnical Lab. Japan, Vol.37, No.4, 1973.
7) H.Wakita: Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms, IEEE Trans. Vol.AU-21, No.5, Oct. 1973.
8) C.Celter: Speech Synthsis with a Parametric Artlculatory Model, Speech Sympo., Kyoto, 1968.
9) P.Mermelstein: Artlculatory Mode] for the Study of Speech Production, J.A.S.A., No.53, 1973.
10) S.Hiki, K.Niyata: Artlculatory Model for Vowel Production, Speech Data Processing, Tokyo Univ. Press, 1973.
11) B.Lindblom, J.Sundberg: Acoustic Consequences of Lip, Tongue, Jaw and Larynx Movement, J.A.S.A., No.50, 1971.
12) J.S.Perk-ell: Physiology of Speech Production, Monograph, 53, MIT Press, 1969.
13) R.Houde: A Study of Tongue Body Motion during Selected Speech Sound, SCRL Mon., 2, 1968.
14) K.Shirai, H.Fujisawa.Y.Sakai: Ear and Voice of the Wabot, Bull. Sci. & Eng. Research Lab. Waseda Univ., No.62, 1973.
15) K.Shirai, H.Fujisawa: An Algorithm for Spoken Sentence Recognition and Its Application to the Speech Input-Output System, IEEE Trans,, Vol.SMC-4, No.5, Sept. 1974.
16) A.Newell et al: Speech Understanding Systems, North-Holland, 1973.



θ_j: Jaw opening

$X=(x_1,x_2,\cdots,x_{26})^T$
: Tongue contour vector

L_p: Lip protrusion

L_h: Lip height

Fig.1 Articulatory model

Tongue contour components a_j and b_k can be calculated by the next formulae from the tongue contour vector X.

For vowels ($X = X_v$)

$$a_j = V_j^T(X_v - \bar{X}_v).$$

For consonants ($X = X_c$)

$$\begin{cases} a_j = V_j^T(X_c - \bar{X}_v), \\ b_k = C_k^T(X_c - \sum_{j=1}^{p} a_j V_j - \bar{X}_c). \end{cases}$$

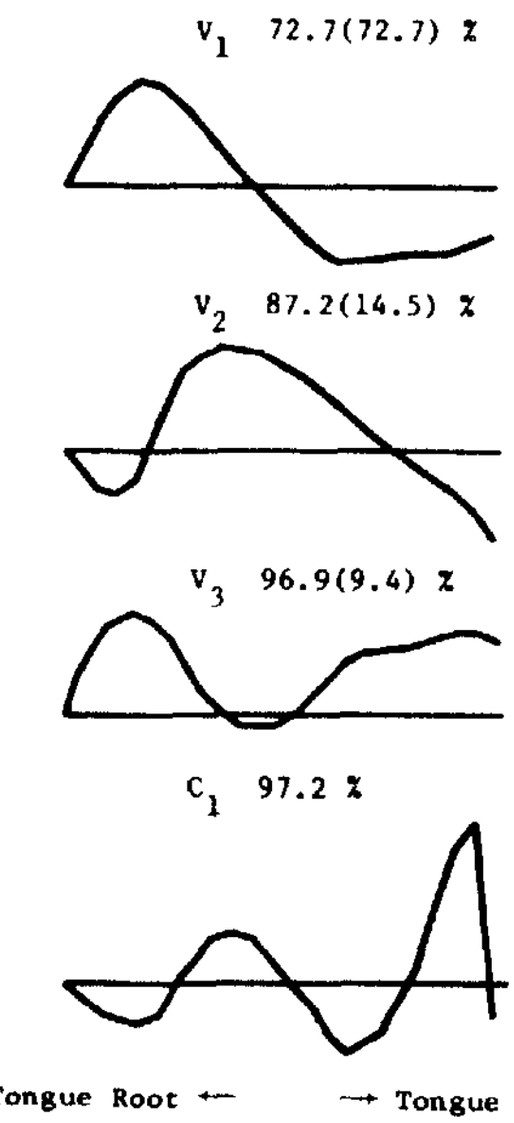


Fig.2 Principal components of tongue contours - v_1,v_2,v_3: Eigen vectors for vowels, c_1: Eigenvector for residuals of consonants after subtraction of vowel components - and cumulative distributions

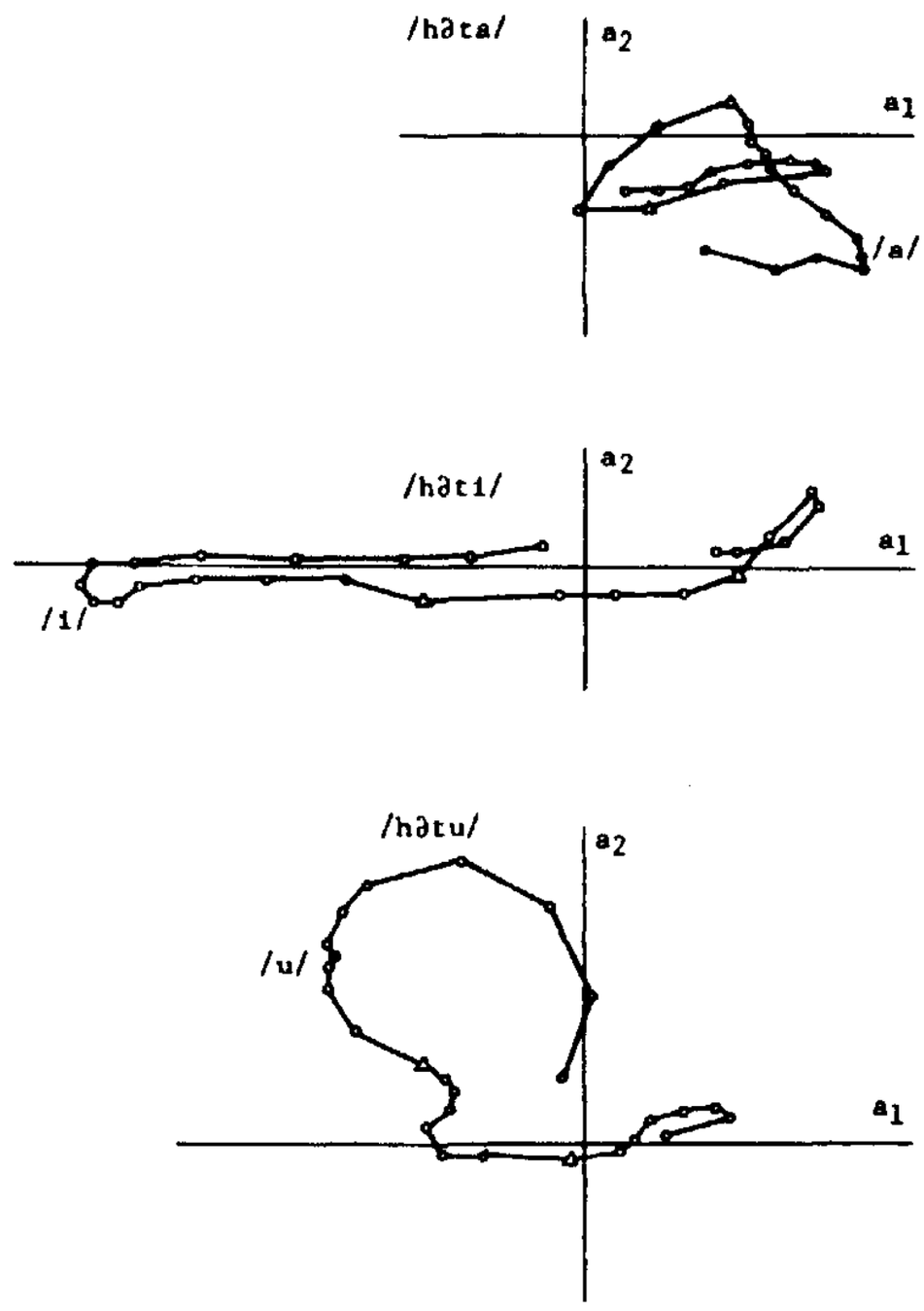


Fig.3 Component loci on a_1-a_2 space for three utterances /hɔtV/ (V: a,i,u)
Δ——Δ :closure period

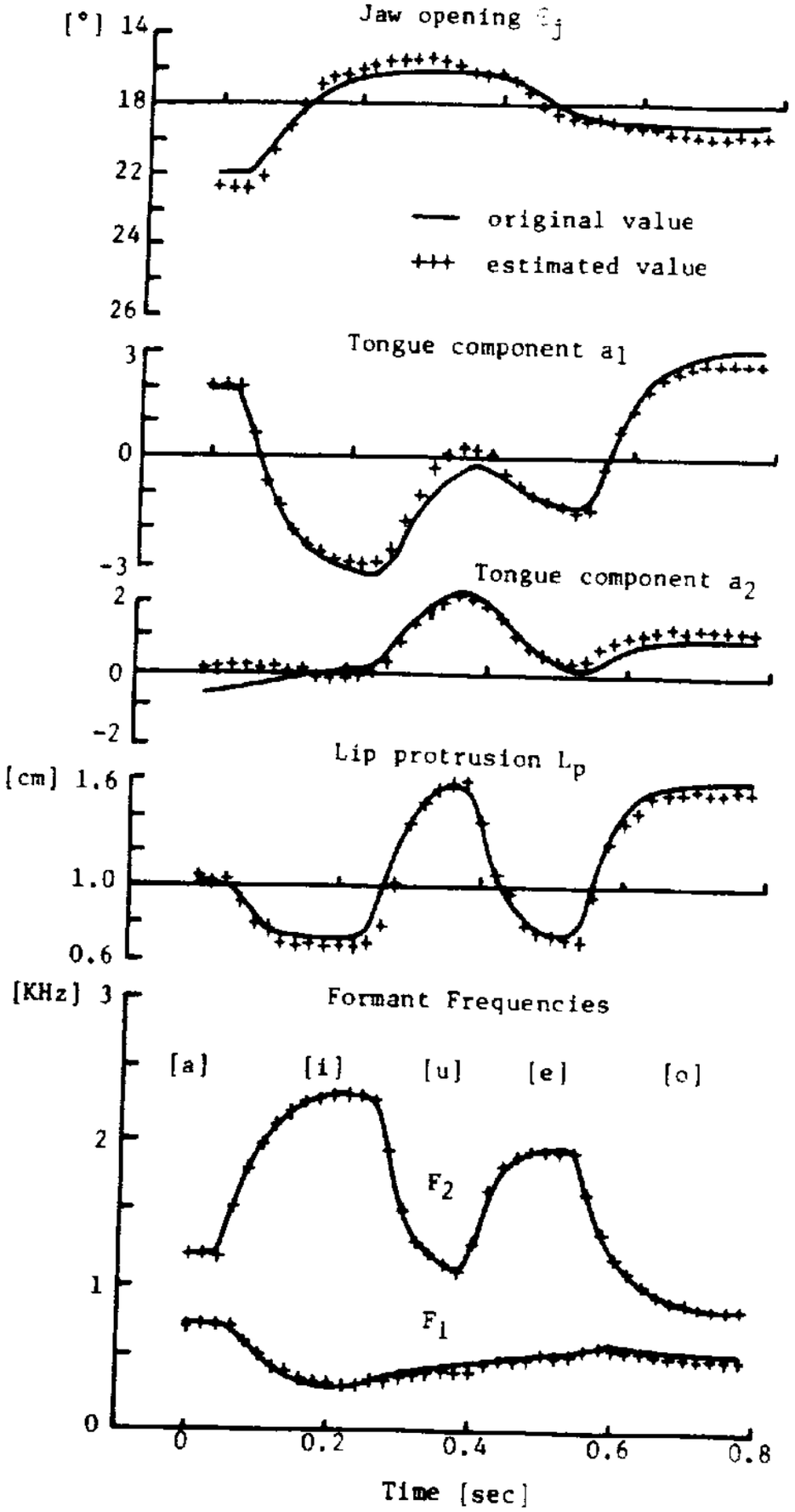


Fig.4 Estimation of articulatory parameters for synthesized voice from F_1, F_2

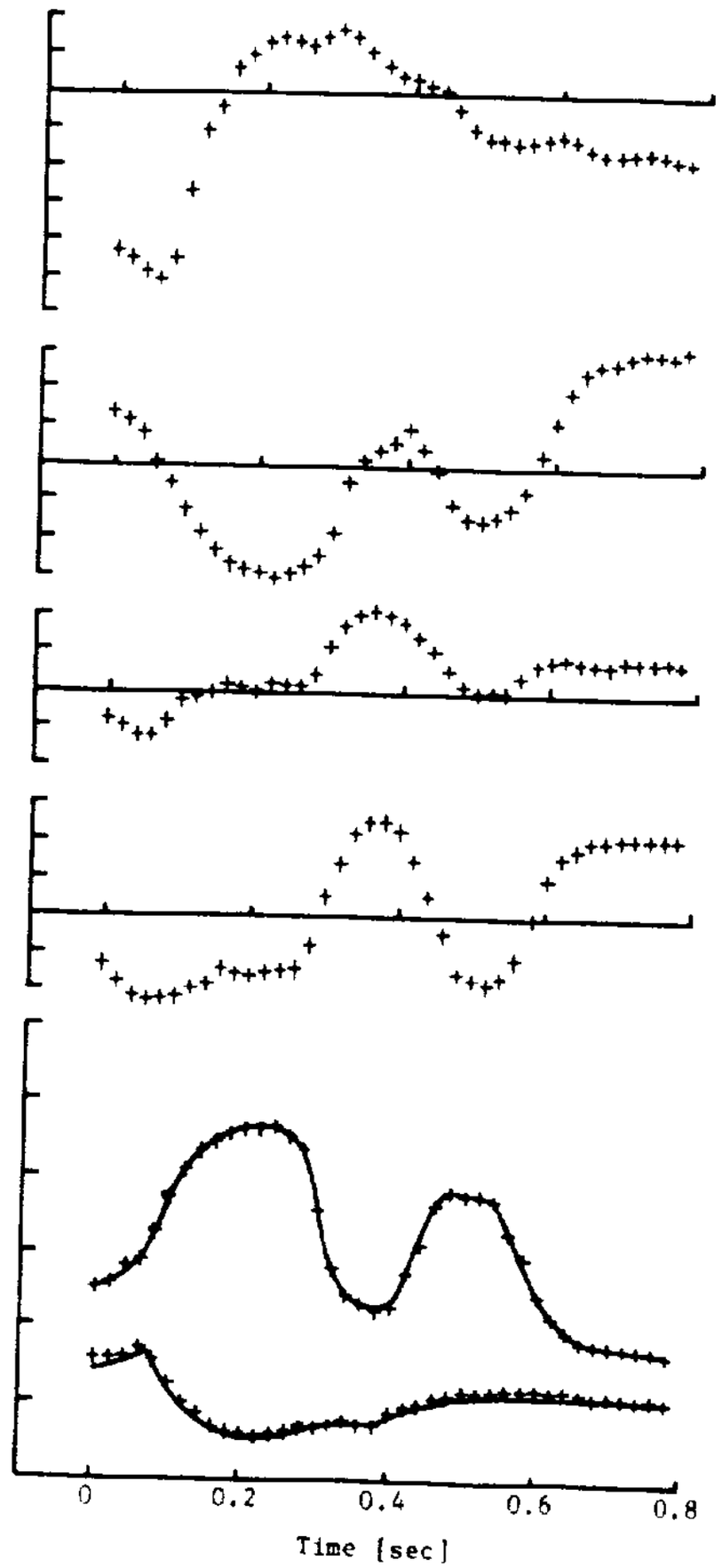


Fig.5 Estimation of articulatory parameters for real speech from F_1, F_2

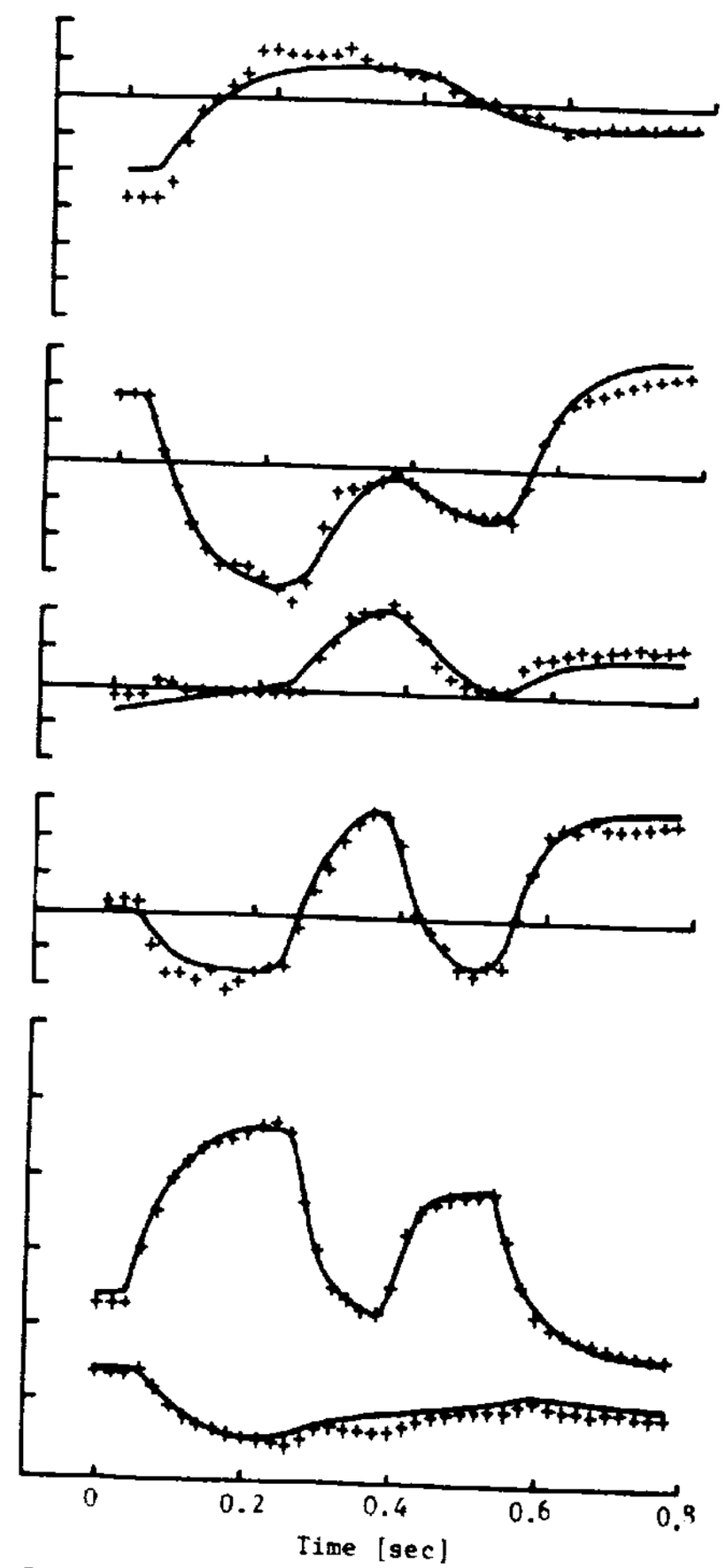


Fig.6 Estimation of articulatory parameters for synthesized voice from the outputs of filter bank

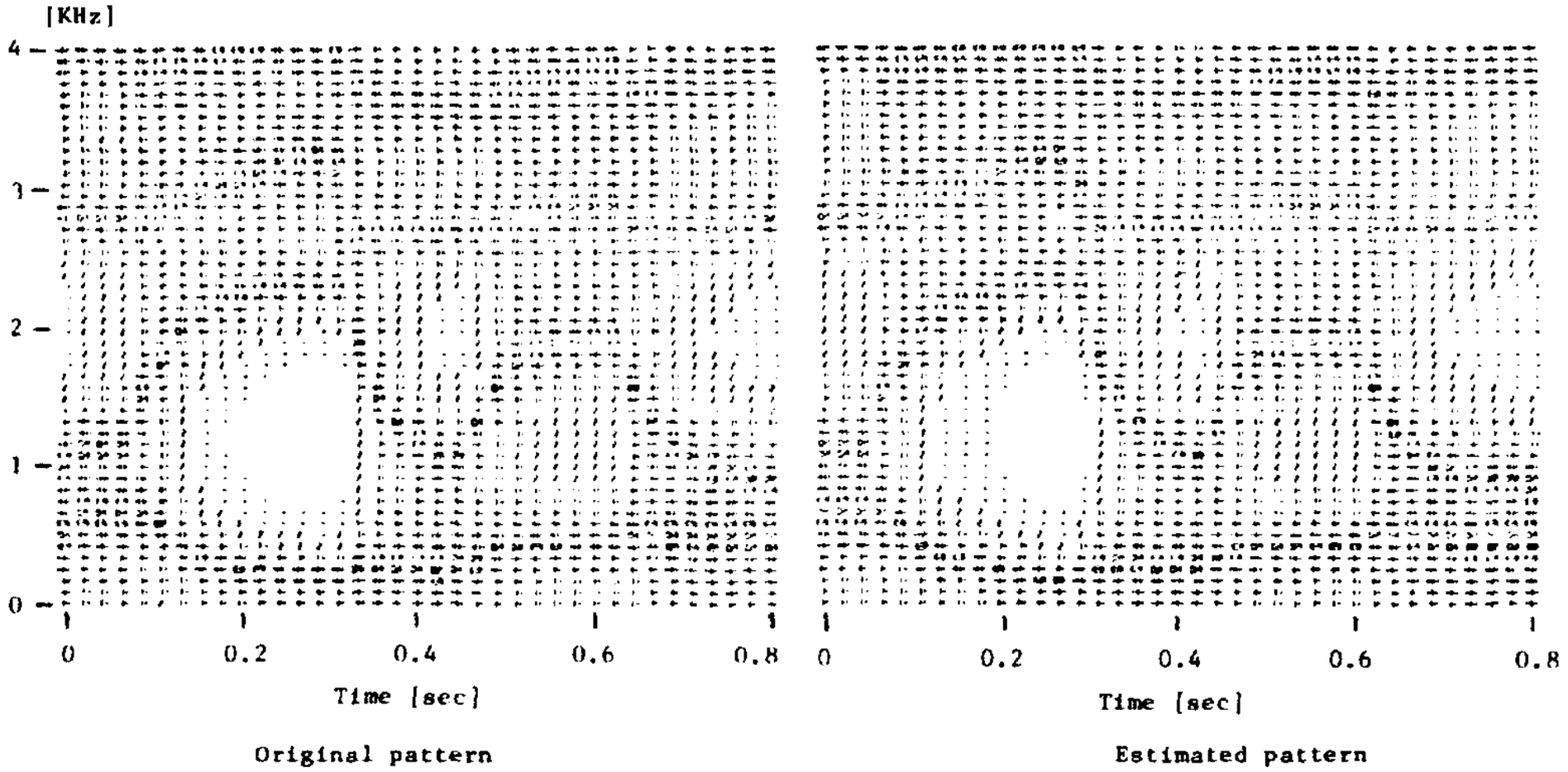


Fig.7 Spectral pattern for synthesized voice and reconstructed voice after estimation Output of filter bank was used for the estimation.

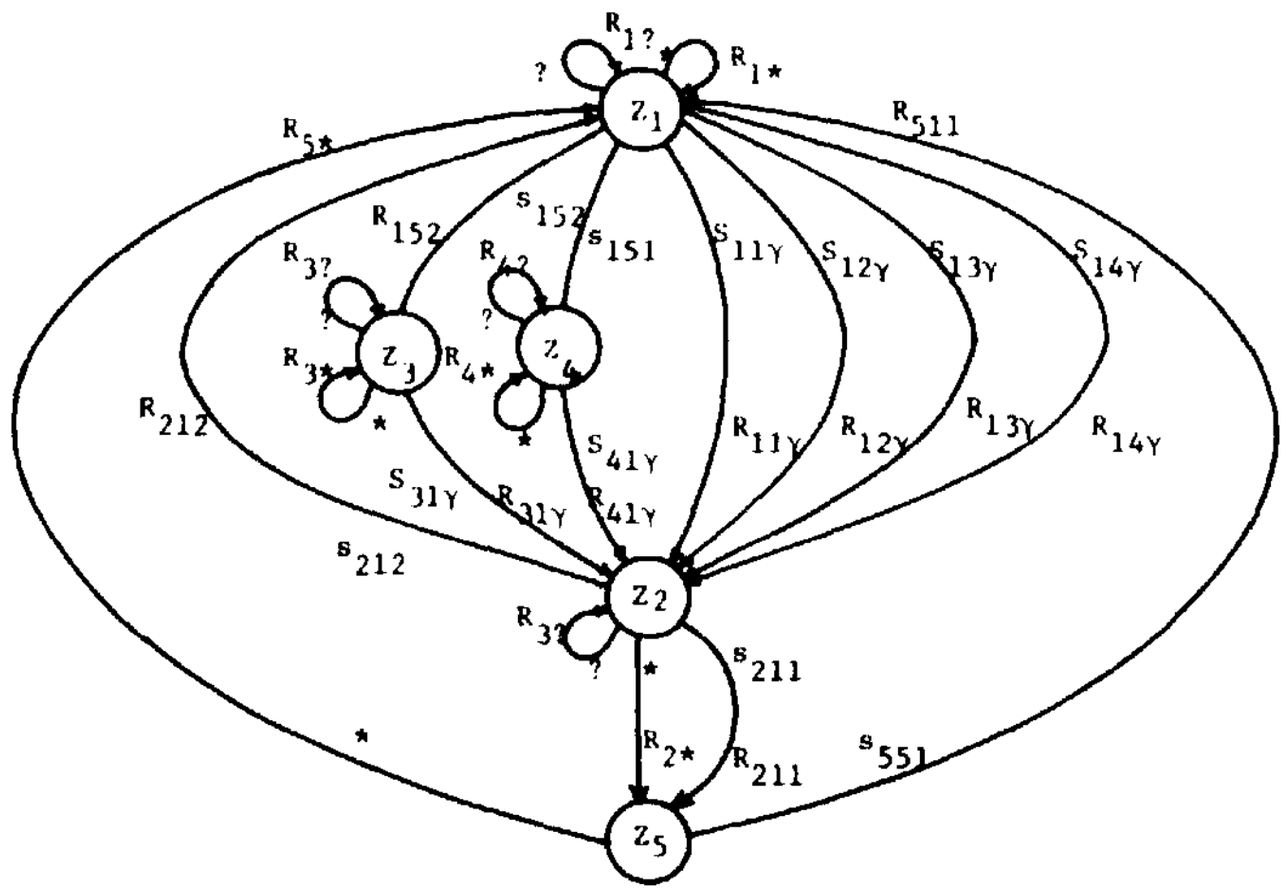


Fig.8 An example of diagram of state transition
z_1: Start waiting for an order
z_2: After the recognition of the order, the machine feedbacks the result to the speaker and waits the permission to work.
z_3: The content of the order is incomplete and the machine asks
z_4: a question.
z_5: State of working
$R_{\alpha\beta\gamma}$: Response of the machine
* : Long silence
? : Rejection