A TRIAL OF JAPANESE TEXT INPUT SYSTEM USING SPEECH RECOGNITION

K.Shirai Y.Fukazawa T.Matzui H.Matzuura
Department of Electrical Engineering
Waseda University
3-4-1 Okubo Shinjuku-ku
Tokyo, 160 Japan

Summary

Since written Japanese texts are expressed by many kinds of characters, input technique is most difficult when Japanese information is processed by computers. Therefore a task-independent Japanese text input system which has a speech analyzer as a main input device and a keyboard as an auxiliary device was designed and it has been implemented. The outline and experience of this system is described in this paper.

The system consists of the phoneme discrimination part and the word discrimination part.

Acoustic analysis and phonemic discrimination are effectively performed using an estimation method of the articulatory motion from speech waves in the phoneme discrimination part. And its outputs are lattices of Japanese pseudo-phonemes corresponding to speech inputs.

In the word discrimination part, phonemic strings are corrected using various kinds of a priori knowledge and transformed into suitable kinds of characters. On behalf of it, this part mainly consists of the retrieval of the word dictionary indexed by vowels, and the similarity calculation using dynamic programming.

1. Introduction

Speech recognition systems can be classified according to whether they recognize isolated or connected words. But in either case, word recognition has been thought a basic problem. In the speech understanding system, a method of utilizing various kinds of a priori knowledge, which is obtained from a phonemic, a syntactic and a semantic level on clearly defined tasks, has been developed. And it has been clarified how information at each level contributes to recognition. [1-3]

While many understanding systems set the goal at the understanding of contents of speech, our main object is the task-independent Japanese texts input system. Because a system which utilizes properties of a task too much is not so practical and the phonemic discrimination must be the core of the speech recognition system. This system consists of the phoneme discrimination part and the word discrimination part.

In the phoneme discrimination part, connected spoken words is segmented into discrete phonemic strings. The main difficulties at this step are the large variation of each utterance and the coarticulation, and the difference between talkers. So it is important to select acoustic parameters which aren't relatively affected by these variations.

This system employs the feature extracton method based on the speech production model. The articulatory motion for the utterance is estimated from speech waves, and the phonemic discrimination is performed using the trace of them. The followings are the merits to describe the acoustic feature at an articulatory level:

(1) Processing of coarticulation may be considered most clearly in relation to the physiological factor of speech production.

(2) Adaptation to a speaker is easier than other methods.

Nevertheless it is almost impossible to construct perfect phonemic strings corresponding to spoken words only by the phonemic discrimination. So in the word discrinimation part, the phonemic strings are corrected using various kinds of a priori information. At the same time, it is neccessary to transform the phonemic strings into the suitable kinds of characters in order to print the texts. Because written Japanese texts are expressed by many kinds of characters, which are Kanji (ideo-graphic character), Hira-gana (cursive form of Kana, Japanese syllabary), Kata-kana (square form of Kana) and special symbols.

The confusion matrix between phonemes, the phonemic connective rules, the word dictionary and so on have been already used as a priori information in word discrimination systems, and their validity has been also ascertained. [4-6]

Therefore, the word discrimination part mainly consists of the retrieval of the word dictionary which is indexed by vowels, and the similarity calculation using dynamic programming and the phonemic confusion matrix. Because the phoneme discrimination part can recognize vowels more correctly than consonants and the phonemic confusion matrix reflects the characteristics of the phoneme discrimination part exceedingly.

In order to translate from Kana to Kanji, some systems with the word dictionary for Kana-Kanji translation and with syntax and semantics analyzer, are suggested. But the task domain should be restricted so as to get the high translation rate.

Main difficulties of Kana-kanji translation are as follows:

(1) Segmentation from input character strings to words.

(2) Identification of synonyms

In this system, identifiers inserted from a simple keyboard serve as spaces between words. And the difficulty of the latter is solved by redundant inputs since speech input is much easier than other input techniques.

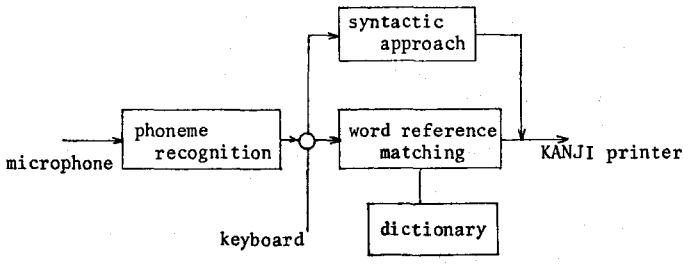

Fig 1-1 Block diagram of the system

The paper is structured as follows. The next section gives a feature extraction method using the articulatory model. And then the segmentation of the continuous speech, and the discrimination technique of vowels and consonants are described. Section 3 gives an outline of the word discrimination method which uses the word dictionary constructed systematically and employs some a priori knowledge. Section 4 concludes with a brief description of the results obtained with the system thus far.

2. Phoneme discrimination

2.1 Segmentation

In the recognition of continuous speech, it is effective to discriminate voiced, unvoiced and silence. Next four parameters are used for voiced-unvoiced-silence decision.[14]

1) Power of signal.

$$E=10.\log\left[\frac{1}{N}\sum_{n=1}^{N} s^2(n)\right] \quad (2\text{-}1)$$

where N is the number of samples in one frame.

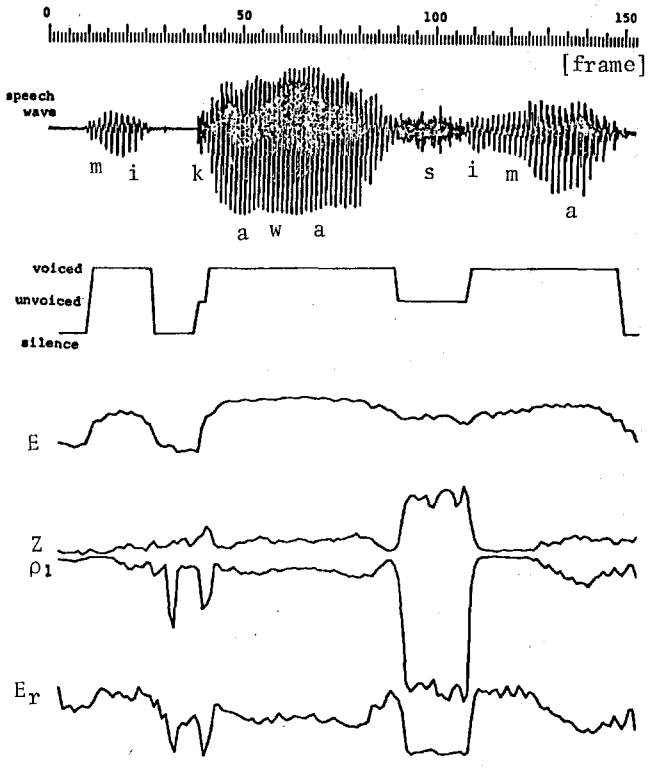

Fig.2-1 Example of analysis /mikawasima/

2) The ratio of the power in low frequency and the total power of the signal.

$$Er=\frac{\text{Power in low frequency}}{\text{Total Power}}=\sum_{n=1}^{N} s_{\ell}^2(n) \Big/ \sum_{n=1}^{N} s^2(n) \quad (2\text{-}2)$$

3) Zero-crossing count in one frame.

4) Normalized autocorrelation coefficient at unit delay, ρ_1 as

$$\rho_1=\sum_{n=1}^{N} s(n)s(n-1) \Big/ \sqrt{\left(\sum_{n=1}^{N} s^2(n)\right)\left(\sum_{n=0}^{N-1} s^2(n)\right)} \quad (2\text{-}3)$$

These parameters are computed for each frame of samples and the example of analysis shown in Fig.2-1. The decision is performed using the next quadratic discrimination function.

$$d_i(x)=(x-\mu_i)^t c_i^{-1}(x-\mu_i)+\ln|c_i| \quad (2\text{-}4)$$

After this decision is made for each frame, smoothing algorithm is used . By this smoothing algorithm, noises and disturbances in the VUS-decision can be modified. The result of this decision is shown in Fig.2-1.

2.2 Articulatory model

Several authors have studied the estimation of the vocal tract shape from the speech wave. If the articulatory motion is successfully estimated, the result may be a good feature of the speech and effective for the recognition.[7-9] In this paper, a new technique is used to estimate the vocal tract shape, which depends on the precise analysis of the real data of the articulatory mechanism.[10-13]

In this section, the articulatory model which constitutes the base of this method is introduced. This model relies on the statistical analysis of the real data and makes it possible to represent the position and the shape of each articulatory organ very precisely using the small number of parameters. And further, the physiological and phonological constraints are included automatically in the model.

The total configuration of the model is shown in Fig.2-2. The jaw is assumed to rotate with the fixed radius FJ about the fixed point F, and the location J is given by the jaw angle X_J. The lip shape is specified by the lip opening height L_h, the width L_w and the protruction L_p relating to the jaw position on the midsagittal plane. The tongue contour is described in terms of a semi-polar coordinate system fixed to the lower

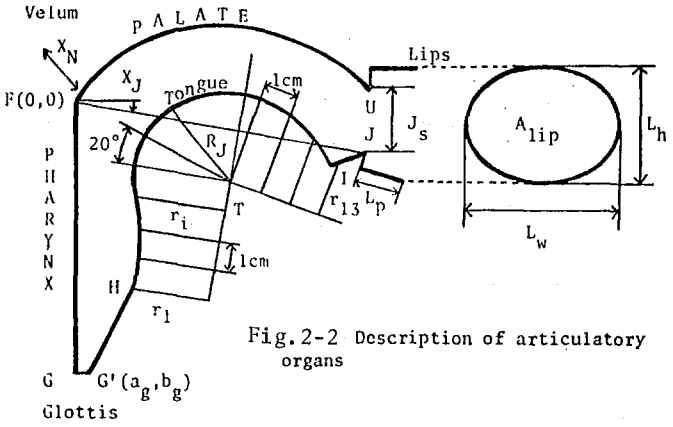

Fig.2-2 Description of articulatory organs

jaw and represented as a 13 dimensional vector $r=(r_1,\ldots,r_{13})$. The variability of the usual articulatory process, there may be some limitations on account of physiological and phonological constraints, which can be expressed by the strong correlation of the position of each segment along the tongue contour. Therefore, it becomes possible to represent the tongue shape effectively using only a few parameters. These parameters can be extracted from the statistical analysis of X-ray data.

A principal component analysis is applied and the tongue contour vector for vowels r_v may be expressed in a linear form as,

$$r_v = \sum_{i=1}^{p} X_{T_i} v_i + \bar{r}_v \qquad (2\text{-}5)$$

where v_i's (i=1,2,...,p) are eigenvectors, and $\bar{r}_v$ is a mean vector for vowels which corresponds roughly to the neutral tongue contour.

The eigenvectors are calculated from the next equation

$$C_v v_i = \lambda_i v_i \qquad (2\text{-}6)$$

where C_v is the covariance matrix which is defined by

$$C_v = \frac{1}{N} \sum_{k=1}^{N} \{(r_{v_k} - \bar{r}_v)(r_{v_k} - \bar{r}_v)^T\} \qquad (2\text{-}7)$$

and λ_i is the corresponding eigenvalue to satisfy the characteristic equation

$$|C_v - \lambda I| = 0 \qquad (2\text{-}8)$$

The same stastical technique as described above may be used for the expression of lip shape.

$$\begin{pmatrix} L_h \\ L_w \\ L_p \end{pmatrix} = \begin{pmatrix} k_1 J_s + \bar{L}_h \\ k_2 J_s^{k_3} + \bar{L}_w \\ \bar{L}_p \end{pmatrix} + \begin{pmatrix} \ell_h \\ \ell_w \\ \ell_p \end{pmatrix} X_L \qquad (2\text{-}9)$$

where Js signifies the distance between the upper and the lower incisors. The active movement of the lips is reflected in the second term, and X_L is the lip parameter.

The total model with the nasal cavity is shown in Fig.2-3 and the characteristics of the articulatory parameters are summarized in Table 2-1. Details about the model are found in the references.

2.3 Estimation of articulatory motion

In this section, the estimation of the articulatory parameters from the speech wave is considered in the framework of the speech production model. This problem becomes nonlinear optimization of parameters under a certain criterion, and it must be solved by an iterative procedure. Therefore, the uniquness of the solution and the stability of the convergence are significant problems. In this method, these problems are solved by introducing constraints to confine the articulatory parameters within a restricted region and by the selection of the appropriate initial estimate using the continuity property. And also the model adjustment to the speaker brings good effects for the estimation.

Let an m-dimensional vector y be the acoustic feature to represent the vocal tract transfer function of the model. In this study the cepstrum coefficients are adopted as the acoustic parameters. The acoustic parameters are expressed as a nonlinear function h(x) of the articulatory parameters which means the transformation from x to y through the speech production model. On the other hand, let $_sy$ be acoustic parameters measured from the speech wave after glottal and radiation characteristics are removed. Then, the estimate $\hat{x}$ of the articulatory parameters is obtained as to minimize the next cost function.

$$J(x_k) = \| {}_sy_k - h(x_k) \|_R^2 + \| x_k \|_Q^2 + \| x_k - \hat{x}_{k-1} \|_\Gamma^2 \qquad (2\text{-}10)$$

where R,Q and Γ are the matrices of weight, k is frame number and $\hat{x}_{k-1}$ is the estimate at the previous frame. In the cost function, the 1st term is the weighted square errors between the acoustic parameters of the model and the measured values. The 2nd and 3rd terms are used to restrict the deviation from the neutral condition (x=0) and from the estimate of the previous frame, respectively. These terms are also efficient to reduce the compensation of the articulatory parameters.

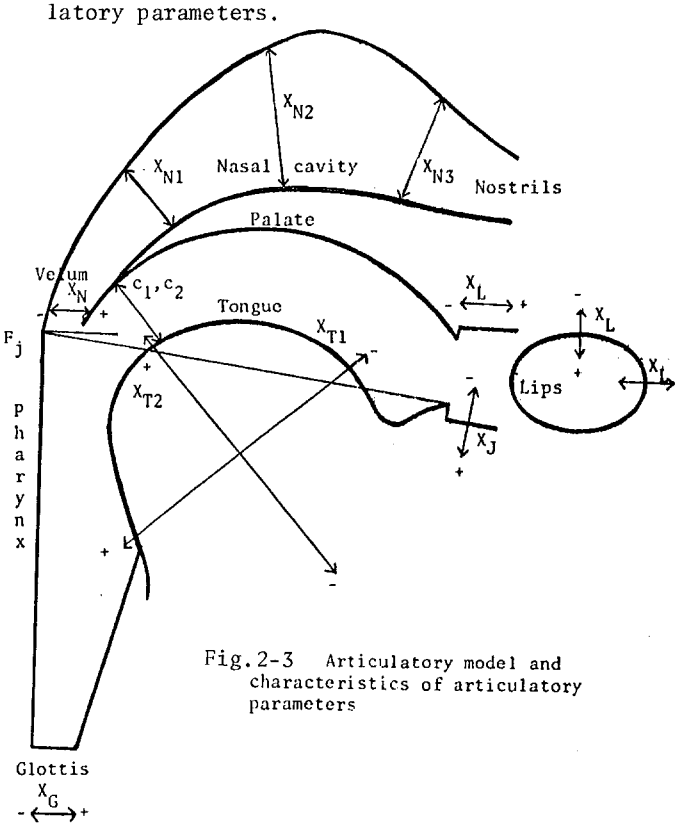

Fig.2-3 Articulatory model and characteristics of articulatory parameters

Table 2-1 Articulatory parameters

Articulatory parameters	Tongue X_{T1}	Tongue X_{T2}	Jaw X_J	Lip X_L	Glottis X_G	Velum X_N
+	back	high	open	round	open	open
-	front	low	close	spread	close	close

If the method is applied to a speaker, first of all, the mean length of the vocal tract should be adjusted for the speaker. The length factor $\bar{S}_{FK}$ which gives a uniform change in the vocal tract length, and the articulatory parameters can be estimated simultaneously. Once the mean vocal tract length is fixed, only the articulatory parameters are estimated. In Fig.2-4 the estimated result for the utterance /aiueo/ is shown.[15-17]

2.4 Normalization of articulatory parameters

The size and shape of each articulatory organ are different between talkers, and further the manner of articulation varies each other. These differences make difficult the talker independent speech recognition problem. Therefore, a method to be able to adapt for the speaker differences is inevitable for speech recognition.

In the estimated articulatory parameters in the preceding section, the average length of the vocal tract is taken into consideration and automatically normalized. While the vocal tract length is the most significant in this problem, other factors still remain.

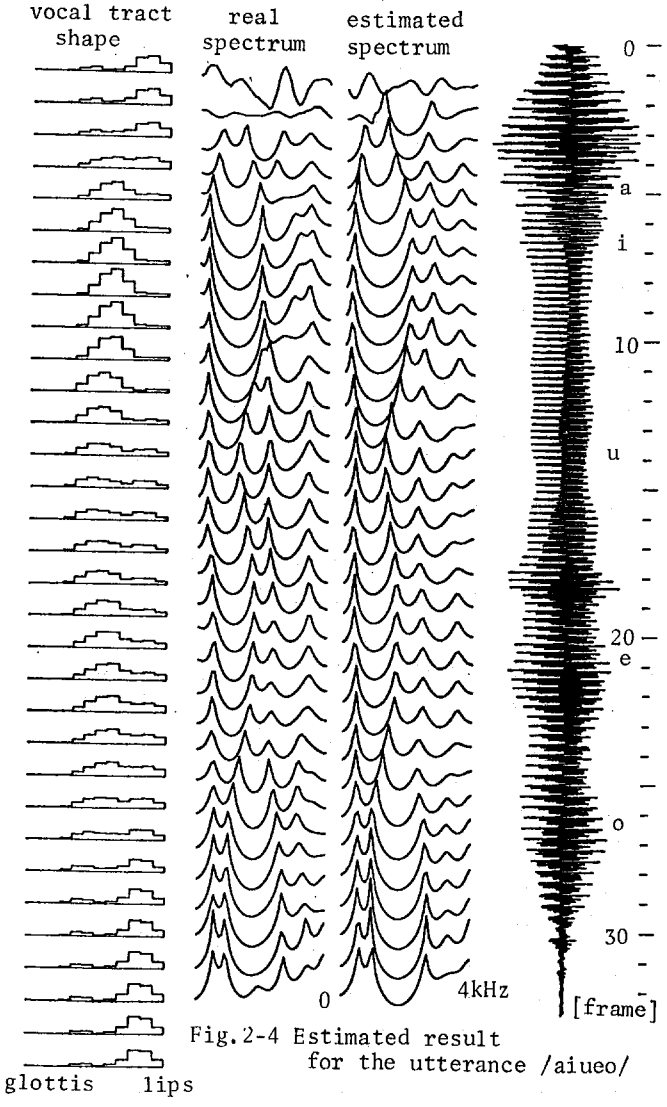

Fig.2-4 Estimated result for the utterance /aiueo/

The differences in Japanese 5 vowels between adult females are shown in Fig.2-6. In order to cancel these differences, a linear transformation which changes the distribution of female articulatory parameters into that of male ones is introduced.

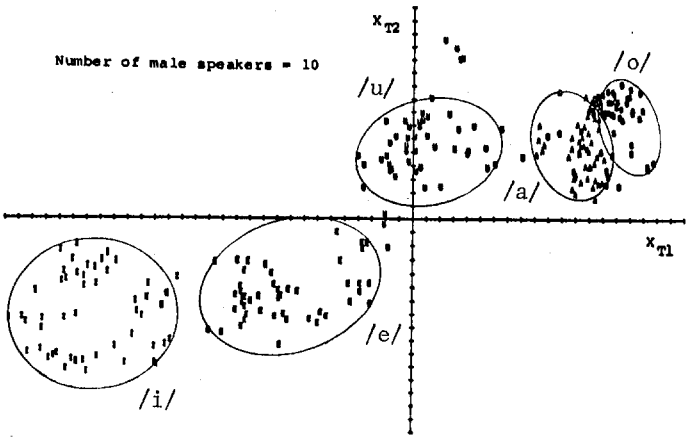

Fig.2-5 Distribution for male speakers (A group) in X_{T1}-X_{T2} space

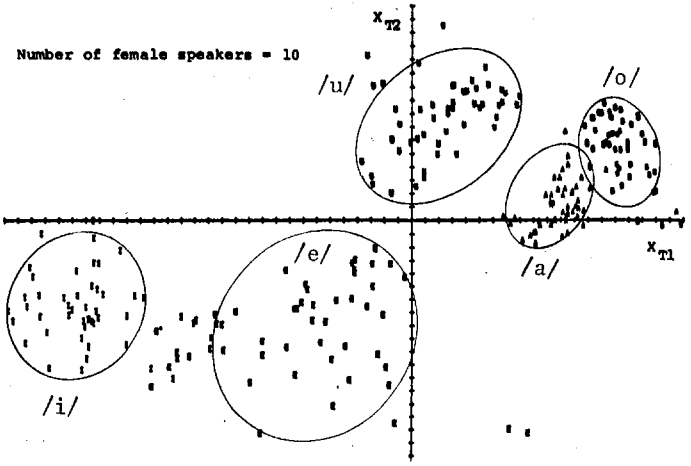

Fig.2-6 Distribution for female speakers (b group) in X_{T1}-X_{T2} space

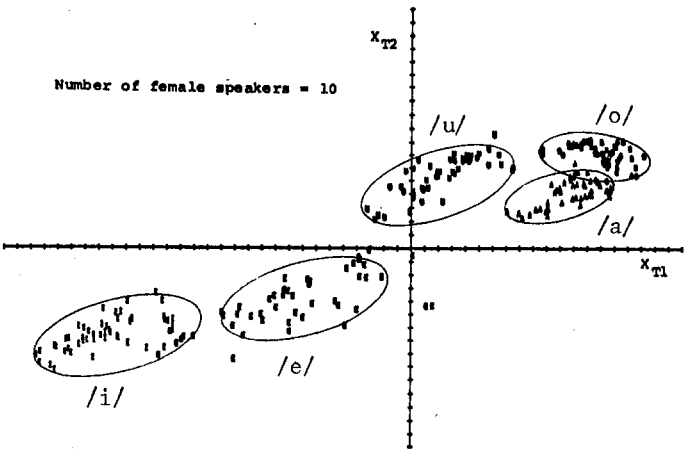

Fig.2-7 Distribution for female speakers (b group) in X_{T1}-X_{T2} space after linear transformation

$$\bar{X}=AX+B \qquad (2\text{-}11)$$

where

$\bar{X}=(\bar{X}_{T1},\bar{X}_{T2},\bar{X}_J,\bar{X}_L)^T$:modified vectors coming near the male distribution

$X=(X_{T1},X_{T2},X_J,X_L)^T$:estimation value for the female

The matrices A and B are calculated by the least squares method. In Fig.2-5 ,2-6 and 2-7 distribution for 10 males and females are shown respectively. The distribution of Fig.2-6 is transformed by the above linear transformation into Fig.2-7. It should be noticed that the matrices A and B were determined by data of 10 female who are not contained in the group shown in Fig.2-6. The results for the vowel recognition experiment are shown in Table 2-2.

In this experiment, the number of subjects is 50, that is 30 males and 20 females. Each subject uttered every vowel five times. They are grouped every ten subjects, then there are three male groups A, B, C and two female groups a, b.

Experiment I ; Recognition of male voices after learning by male voices

Experiment II ; Recognition of female voices after learning by female voices

Experiment III ; Recognition of female voices after leaning by male voices

Experiment IV ; Recognition of female voices using normalization of linear transformation, where the reference is made by male voices

These results show the effectiveness of the articulatory parameters for the talker independent speech recognition and the normalization procedure is useful for the male-female transformation.

2.5 Discrimination of continuous vowels by using articulatory parameters

The locus of estimated articulatory parameters about /aiueo/ is shown in Fig.2-8.

Initially, continuous vowels are discriminated frame by frame using a quadratic discrimination function of articulatory parameters. As this results involve transient parts, it is necessary to estimate them. Then S(k) is computed as the parameter of discriminating them.

$$S(k)=|X_{T1}(k+1)-X_{T1}(k)|+|X_{T2}(k+1)-X_{T2}(k)| +|X_J(k+1)-X_J(k)|+|X_L(k+1)-X_L(k)| \qquad (2\text{-}12)$$

It can be regarded the valley of S(k) as the stable part and the high part of S(k) than the threshold t as transient part. But, it is difficult to distiguish between stable part and transient part enough. So stable part is decided with changing threshold t.

The speakers in this experiment are 2 male adults. Each speaker spoke 23 kinds of continuous 5 vowels 2 times. This continuous vowels involve all 60 kinds of connected vowel $V_1V_2V_3$(V_1,V_2,V_3: a,i,u,e,o) such as aie, oei.

Table 2-3 shows the errors occurred by coarticulation.

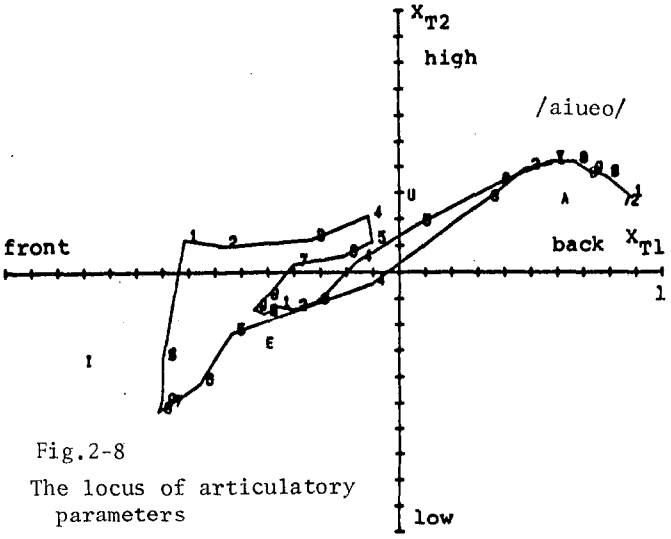

Fig.2-8
The locus of articulatory parameters

Table 2-2 Confusion matrix for the result of vowel recognition

Experiment I 98.8%

I \ O	/a/	/i/	/u/	/e/	/o/
/a/	298	-	-	-	2
/i/	-	296	-	4	-
/u/	-	-	300	-	-
/e/	-	2	-	298	-
/o/	9	-	1	-	290

Experiment II 98.0%

I \ O	/a/	/i/	/u/	/e/	/o/
/a/	99	-	-	-	1
/i/	-	96	-	4	-
/u/	-	-	99	-	1
/e/	-	-	-	100	-
/o/	2	-	2	-	96

Experiment III 94.9%

I \ O	/a/	/i/	/u/	/e/	/o/
/a/	295	-	-	1	4
/i/	-	286	-	14	-
/u/	1	-	298	1	-
/e/	2	1	9	288	-
/o/	37	-	6	-	257

Experiment IV 97.3%

I \ O	/a/	/i/	/u/	/e/	/o/
/a/	289	-	-	-	11
/i/	-	298	-	2	-
/u/	-	-	299	1	-
/e/	-	15	-	285	-
/o/	1	-	11	-	288

Table 2-3 Vowel discrimination errors

speaker	S.M.		T.S.	
	No.1	No.2	No.3	No.4
1 /aiueo/				
2 /aeiou/				
3 /aeuoi/				
4 /iaoiu/		5oiu→ouu	16ao→o	26 ao→a
5 /iuoea/		6uo→u		
6 /ieaou/				aou→auu 27
7 /uaieo/				28
8 /uaoei/	1ei→ii	7ao→o 8ei→ii	17uao→uoo 18ei→i	uao→uou 29eai→eei
9 /ueaio/			19eai→eei	
10 /uoaeo/				
11 /eiuae/		9a→o		
12 /euioa/	2e→i		20uio→ueo	
13 /eoaui/				
14 /eouei/		10ei→ii	21ei→i	
15 /oiaue/				
16 /oieua/		ie→ii 11	22iae→iee	
17 /ouiae/				
18 /auoua/		12ou→oo	23auo→ao	
19 /uieia/		13uau→uou	24uie→ue	
20 /eauau/		14e→i		30 eau→eeu
21 /eoioe/	3ioe→iue			ioe→iue 31
22 /oaiai/	4iai→iei	15ueu→uiu	25iai→iei	
23 /oeueu/				

2.6 Consonant recognition using articulatory parameters

By the data of the gliding section between consonant and vowel, consonants are discriminated. In this discrimination experiment, a method of DP matching applied for the loci of the articulatory parameters which are estimated for the gliding section. Simple acoustic parameters are used together.

Speech data in this experiment is uttered by 2 male adults, each speaker spoke continuous V_1CV_2. (V_1 is fixed at /a/, V_2 is vowel /a/,/i/, /u/,/e/,/o/, C is consonant /g/,/z/,/d/,/b/,/h/, /s/,/p/,/t/,/k/ where /adi/,/adu/ were omitted.) One data is for reference pattern, another is for the test.

Since the articulatory organs show relatively speedy motion in the interval from the consonant to the vowel, the frame length is selected as 12 ms with Hamming Window and the pitch syncronous method is used for the analysis. For the estimation initial value is obtained by the piecewise linear estimation method at stable vowel section. And articulatory parameters are estimated backward to consonant section, using the parameters at the previous frame as the initial value. The process of the discrimination experiment is roughly shown in Fig.2-9.

Initially, we estimate the articulatory parameters of every data by the method mentioned above and perform the end-free DP matching for the reference pattern of every kind of data for each speaker. The total power and the power in high frequency are also used for the discrimination. For the recognition of the voiced consonants, the results of the DP matching of articulatory parameters are reliable. For that of the unvoiced consonants, the acoustic parameters are mainly used. These results are shown in Table 2-4.

The attempt to extract the feature of consonants from the gliding part of the articulatory motion in the following vowel was successful to some extent. But there are some problems in the accuracy in the estimation of the articulatory parameters. And the individual difference has influence on the parameters.

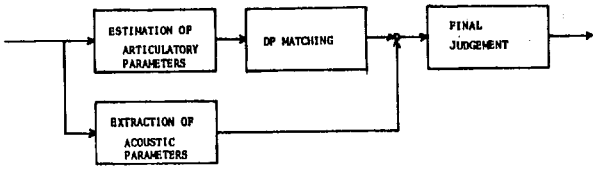

Fig.2-8 Block diagram of the recognition experiment

Table 2-4 Results of the recognition experiment

in \ out	/ga/	/za/	/da/	/ba/
/ga/	20	0	0	0
/za/	2	16	2	0
/da/	0	0	20	0
/ba/	0	0	0	20

in \ out	/ka/	/sa/	/ta/	/pa/	/ha/
/ka/	20	0	0	0	0
/sa/	1	17	2	0	0
/ta/	6	0	13	1	0
/pa/	3	0	1	15	1
/ha/	0	1	0	0	19

in \ out	/gi/	/zi/	/bi/
/gi/	18	1	1
/zi/	0	20	0
/bi/	0	1	19

in \ out	/ki/	/shi/	/chi/	/pi/	/hi/
/ki/	13	0	2	0	5
/shi/	0	20	0	0	0
/chi/	0	0	20	0	0
/pi/	0	0	0	20	0
/hi/	5	0	0	0	15

in \ out	/gu/	/zu/	/bu/
/gu/	19	1	0
/zu/	2	18	0
/bu/	0	0	20

in \ out	/ku/	/su/	/tsu/	/pu/	/fu/
/ku/	19	0	0	1	0
/su/	0	18	1	1	0
/tsu/	1	9	10	0	0
/pu/	1	0	0	18	1
/fu/	0	0	0	1	19

in \ out	/ge/	/ze/	/de/	/be/
/ge/	20	0	0	0
/ze/	0	18	2	0
/de/	0	1	19	0
/be/	0	0	1	19

in \ out	/ke/	/se/	/te/	/pe/	/he/
/ke/	19	0	0	0	1
/se/	0	20	0	0	0
/te/	0	2	14	4	0
/pe/	0	0	1	16	3
/he/	0	0	0	2	18

in \ out	/go/	/zo/	/do/	/bo/
/go/	15	0	3	2
/zo/	0	17	3	0
/do/	0	1	16	3
/bo/	0	1	4	15

voiced : 91.4 %

in \ out	/ko/	/so/	/to/	/po/	/ho/
/ko/	17	0	3	0	0
/so/	0	20	0	0	0
/to/	0	1	17	2	0
/po/	1	0	4	14	1
/ho/	0	0	0	4	16

unvoiced : 85.4 %

3. Word discrimination

3.1 How to input Japanese texts

Ideally speaking, texts which are arbitrarily pronounced should be recognized perfectly and suitable Kana-Kanji translation should be performed. But the phonemic discrimination, the processing of synonyms and so on are actually very difficult. So the system uses a simple keyboard as an auxiliary input device. it has only 7 keys and it is enough manageable by one hand.

In case of input, first of all, a key indicating the kind of the character must be pushed. For example, 'H' indicates Hira-gana. For Kanji, the reading of the compound word, On-yomi (the phonetical reading of Kanji) and Kun-yomi (the Japanese rendering of Kanji) constructing the word are pronounced. Since the acoustic input is much easier than the other input techniques, some redundant data can be inputted with less burden. In the current version, a key indicating the kind of the reading pronounced is pushed for the sake of the easy processing.

For Hira-gana, a speaker may pronounce a word which contains its reading as the ending. Otherwise a user just pronounces Japanese texts as a way of speaking with the control information from the keyboard.

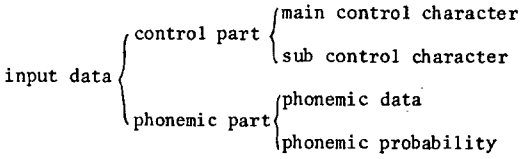

Fig 3-1 the component of input data to the word discrimination part

3.2 Input data

The component of input data to the word discrimination part is indicated in Fig.3-1.

The phonemic part is an acoustic data which is recognized in the phoneme discrimination part. If it cannot decide only one phonemic candidate, the phoneme and its probability for each candidate are passed as a lattice. That is to say, input data is a lattice of Japanese pseudo-phonemes corresponding to the acoustic input. Fig.3-2 is an example of input data.

The control part is inputted from the keyboard in order to correct the phonemic data and perform effective Kana-Kanji translation.

```
letter     天才は９９パーセントの汗である
reading    tensai wa kyuju-kyu paasento no ase dearu
meaning    Genius is ninety-nine per cent perspiration.

data       C J (tensai) T TENSAI
             Y (ten) T TEN , (ama) T AMA
             Y (sai) T SAI , (toshi) T TOSHI
           HH Y (wa) T WA
            S Y (ku) T KU
              Y (ku) T KU
            K Y (paasento) T PAASENTO
            H Y (no) T NO
            C Y (ase) T ASE , (kan) T KAN
            H Y (dearu) T DEARU

  * (ten) --- t8p6k3 e7a1 n7m7
```

Fig 3-2 Example of an input data

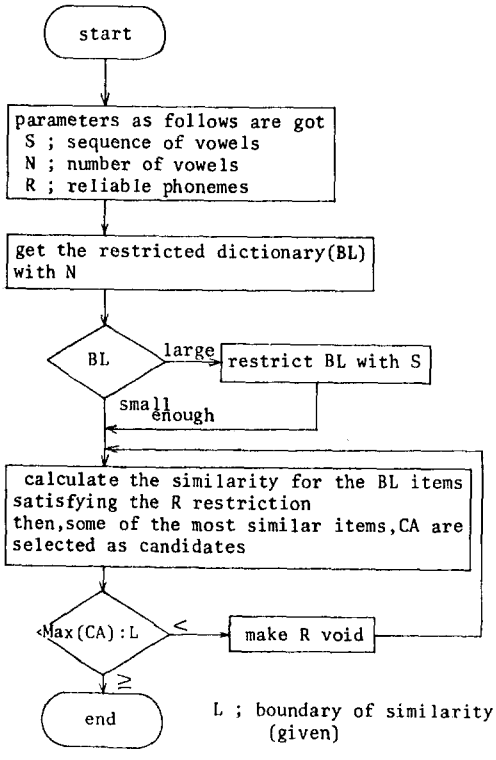

Fig 3-3 Algorithm to decide some candidates

3.3 Word discrimination algorithm

For Kanji, redundant data, which consists of a reading of a compound word, On-yomi, and Kun-yomi, are given. Some candidates of Kanji corresponding to each reading are sought using the algorithm of Fig.3-3. This algorithm depends on the fact that vowels can be discriminated more precisely than consonants, and the tendency which one phoneme is apt to be misdiscriminated from others are statistically known previously.

Kanji dictionary used contains about 1000 readings of Kanji and corresponding characters. This dictionary is divided by the number of vowels and further indexed by the sequence of them. A part of it is shown in Fig.3-4.

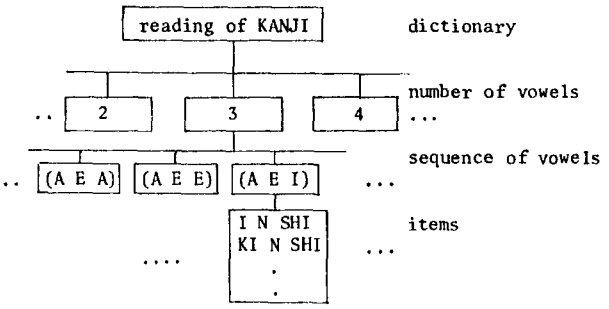

Fig 3-4 Example of a dictionary

Similarity between a discriminated phonemic string and a word contained in Kanji dictionary is computed by dynamic programming using the confusion matrix. This matrix is made using the results of the recognition experiments and contains the misdiscrimination probability to other phonemes, the probability of omission and that of addition of other phonemes. Among these, the probability of addition depends upon the tendency that the addition is influenced by the following phoneme as the nature of the phoneme discrimination part.

In dynamic programming, it is assumed that the number of the continuous omission or addition of phonemes is less than two. The similarity S(D,W) between the phonemic string D whose length is I, ans W whose length is J is calculated as follows:

$$S(D,W) = g(1,1)/I$$
$$g(i,j) = \max\{L_1(i,j), L_a(i,j), L_o(i,j)\}$$
$$L_1(i,j) = sim(i,j)+g(i+1,j+1)$$
$$L_a(i,j) = sim(i,j+1)+g(i+1,j+2)+adp(i,j)$$
$$L_o(i,j) = sim(i+1,j)+g(i+2,j+1)+omp(j)$$
$$g(I,j) = sup\ (J-j)$$
$$g(i,J) = sup\ (I-i)$$

where sim(i,j) is the similarity between phoneme i and j, and given from the confusion matrix, adp(i,j) is the probability indicating that phoneme i is added before phoneme j, omp(i) is the probability omitting i, and sup(i) is the penalty value of unmatched length.

This algorithm selects some candidates of Kanji which have a high matching score. Since some redundant data is given for each kanji, one most suitable character is selected and printed.

For Hira-gana, the system doesn't have redundant readings originally in contrast with Kanji As it is especially difficult to discriminate too short phonemic strings, phonemic corrections are neccessary. So the system utilizes the nature of Japanese that almost all short Hira-gana strings are found as the inflectional part or Joshi (auxiliary word in Japanese). For the former, a user may pronounce a word which contains the inflection and the system uses them as a redundant data using the inflection dictionary. For the latter, the dictionary containing Joshi and its connection information is used as a priori information.

In case of Kata-kana, as the usage is restricted to the words of foreign origin, the phonemic correction is performed using the loanword dictionary. At this time, the algorithm of Fig.3-2 is available.

Besides, Japanese text input system must process special characters, for example numbers, alphabets, punctuation marks and so on. But they are a few, so the above algorithm can processed them using special character table.

4. Conclusion

At present, the system uses YHP-21MX and the special purpose hardware for the phonemic data extraction, and the phoneme and the word discrimination part are built on HITAC 8800/8700 at the Computing Centre of the University of Tokyo. Though the performance, the facility and the bottle-neck as the total system has not been clarified as yet, partially some results and problems have been found.

The phoneme discrimination part aims at recognizing connected speech uttered by an unspecified speaker. On behalf of it, the feature extraction method at the articulatory level is tried. As a result, stable articulatory motion is estimated with the satisfactory precision for vowels, and the validity of it is confirmed. And the adaptation for speakers is effectively performed. For consonants, the feature extraction method using the transient part of the articulatory motion is tried, but it has a little drawpoint at the precision and the stability of articulatory parameters. It is also influenced by a speaker. A combination to the other sound parameters will perhaps make the system improve.

For the word discrimination, the syntax has not been used except for Hira-gana at this version. Hereafter it is planned to incorporate the syntactic information as much as possible for the improvement of the performance. But the trade-off between the facility and the processing speed is an important problem.

Acknowledgment

The authors wish to thank J.Kubota, T.Kobayashi and M.Ohashi for their contributions to designing and developing this system.

References

(1) Lea,W.A. Ed.'Trends in Speech Recognition' Prentice-Hall (1980)

(2) Ready,D.R. Ed.'Speech Recognition' Academic Press (1975)

(3) Newell,A. et al.'Speech Understanding System: Final Report of a Study Group' North-Holland (1973)

(4) Denes,P.'The Design and Operation of the Mechanical Speech Recognition at University College London' Jour. Brit. I.R.E. Vol.19 No.4 pp.219-234 (1959)

(5) Woods,W.A.'Motivation and Overview of SPEECHLIS: An experimental Prototype for Speech Understanding Research' IEEE Trans. A.S.S.P. Vol.ASSP-23 pp.2 (1975)

(6) Sakai,T.,Nakagawa,S.'A Classification Method of Spoken Words in Continuous Speech for Many Speakers' Jour. Inf. Proc. Soci. of Japan Vol.17 No.7 pp.650-658 (1976)

(7) Shirai,K.'Feature extraction and sentence recognition algorithm in speech input system' 4th Int. J. Conf. on Artificial Intelligence, 506-513, 1975.

(8) Wakita,H.'Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveform, IEEE Trans., ASSP-23, 574-580, 1975.

(9) Nakajima,T.,Omura,T.,Tanaka,H.,Ishizaki,S. 'Estimation of vocal tract area functions by adaptive inverse filtering methods' Bull. Electrotech. Lab., 37, 467-481, 1973.

(10) Stevens,K.N., House,A.S., 'Development of a quantitative description of vowel articulation',JASA, 27, 484-493, 1955.

(11) Coker,C.H.,Fujimura,O.,'Model for specification of the vocal-tract area function',JASA, 40, 1271, 1966.

(12) Lindblom,B.,Sunberg,J.'Acoustical consequences of lip, tongue, jaw and larynx movement' JASA, 50, 1166-1179,1971.

(13) Shirai,K.,Honda,M.,'An articulatory model and the estimation of articulatory parameters by nonlinear regression method' Trans. IECE, Vol. J59-A, No.8, 668-674, 1976.

(14) Atal,B.S.,Rabiner,L.R.,'A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition' IEEE Trans., ASSP-24, 201-212, 1976.

(15) Shirai,K.,Honda,M.,'Estimation of articulatory parameters from speech waves' Trans. IECE, Vol.61-A, No.5, 409-416,1978.

(16) Shirai,K.,Honda,M.,'Feature extraction for speech recognition based on articulatory model' Proc. of 4th Int. J. Conf. on Pattern Recognition 1978.

(17) Shirai,K.,Matzui,T.,'Estimation of articulatory states from nasal sounds' Trans. IECE, Vol. J63-A, No.2, 1980.