J. Acoust. Soc. Jpn. (*E*) **7**, 3 (1986)

Speaker independent telephone speech recognition and reference pattern generation

Hiroshi Iizuka, Makoto Morito, and Kozo Yamada

Research Laboratory, OKI Electric Ind. Co., Ltd.,
550–5, Higashi-Asakawa-cho, Hachioji, 193 Japan

(*Received 30 November 1984*)

This paper describes the speaker independent isolated word speech recognition method developed for telephone speech response systems. To recognize speech, input utterances are first frequency analyzed by 19 channel BPFs. The frame cycle used is 8 ms. Then the analyzed data undergo logarithmic conversion, normalization of voice chords sound source characteristics by least squares approximation line and time normalization by linear companding to 32 frames. The speech patterns thus obtained undergo pattern matching with multiple reference patterns generated separately for male and female speakers in advance. In applying this recognition method, it is necessary to optimize the reference patterns so that the speech can be correctly recognized in spite of the difference of formant frequencies, the differences in individual speaker's habits, the variations of phonetic positions, non-vocalization, and slight segmentation errors. To evaluate the performance of this recognition method, voices of about 2,000 persons were recorded through long distance telephone lines. A 16 Japanese words vocabulary was used. A total of 256 male and female reference patterns were generated using the training voice data of about 570 persons. The speech recognition accuracy of this method in recognizing non-training voice data was 97.8%.

PACS number: 43.70.Sc, 43.70.Qa, 43.85.Ta

1. INTRODUCTION

The telephone speech response service is spreading widely, forming a part of the office automation system. To use the telephone speech response service from ordinary dial-up type telephone sets, it is necessary to incorporate speaker independent mode telephone speech recognizer in the speech response system. We have been developing a speaker independent isolated word speech recognition system for these years.[1-3]

It is well known that time domain dynamic programming algorithm proves effective for speaker dependent isolated word speech recognition.[4] However, in the speaker independent mode speech recognition, dynamic programming is effective only for the non-linear variations of utterance speed and the slight segmentation errors. For other recognition error factors, i.e. the differences in formant frequencies, non-vocalization of utterances, and an extreme deviation of the phonetic position caused by individual speaker's habits, dynamic programming does not necessarily prove effective.

We have developed a method of learning the voice data of 500 to 1,000 speakers and generating the reference patterns in the optimum form in off-line mode. The optimum reference patterns thus obtained can absorb the above-mentioned error factors, thereby obtaining a high recognition accuracy.

Besides, this system performs only linear matching between voice input patterns and reference patterns so that the structure of a recognizer becomes simple.

This paper first discusses the recognition system in Section 2. The voice input is first frequency-analyzed by 19 channel BPFs. Then the voice data are normalized, and finally normalized data are

matched with 256 male and female reference patterns. At the same time, the discrimination between male and female is performed. After male and female are discriminated, the pattern matching is performed using only the reference patterns corresponding to the discrimination results to reduce the recognition time.

We will discuss the generation method of reference patterns in Section 3. To generate reference patterns, the recognition of training patterns and updating of the reference patterns are performed alternatively. This cycle is repeated several ten of times and finally the reference patterns are optimized.

In Section 4, we will discuss the evaluation of this method. To evaluate, the speech data of 16 words spoken by about 2,000 male and female adults were used. The evaluation items contain the recognition accuracy at the reference patterns generation, recognition accuracy using the length of input voice and power information, and an overall recognition accuracy under an environment equivalent to actual recognizer.

Finally in Section 5, we summarize the conclusion.

2. RECOGNITION METHOD

This Section deals with the recognition system. Figure 1 illustrates recognition flowchart of the system. The input voice is amplified by the gain programmable amplifier. The amplified voice is frequency limited to the telephone voice band (0.3～3.4 kHz) and then undergoes A/D conversion by 8-kHz sampling frequency and 12-bit coding.

Next, it is frequency-analyzed by 19 channel BPFs ($Q=6$, single tuning type) aligned at an equidistance in 1/5 octave. Figure 2 illustrates the frequency characteristics of the BPFs. Then an absolute value of the BPF output is taken for each channel, and the mean value is calculated within every 8 ms frame period. The result is put as an analyzed data $U(i,j)$. Here, i points BPF channels and j points the frames. Next, the data $U(i,j)$ are logarithmically converted into 8 bits by Eq. (1) to obtain $V(i,j)$.

$$V(i,j)=\begin{cases}\dfrac{255}{\log_{10}2047}\cdot\log_{10}U(i,j), & U(i,j)>0\\ 0, & U(i,j)=0\end{cases} \quad (1)$$

To normalize utterance power and vocal cord sound source characteristics, the least squares approximation line values are calculated for every frame period. These values are subtracted from the original data $V(i,j)$ using Eq. (2).[5] Here, Eq. (3) represents the slope of least squares approximation line and Eq. (4) represents the segment of the line.

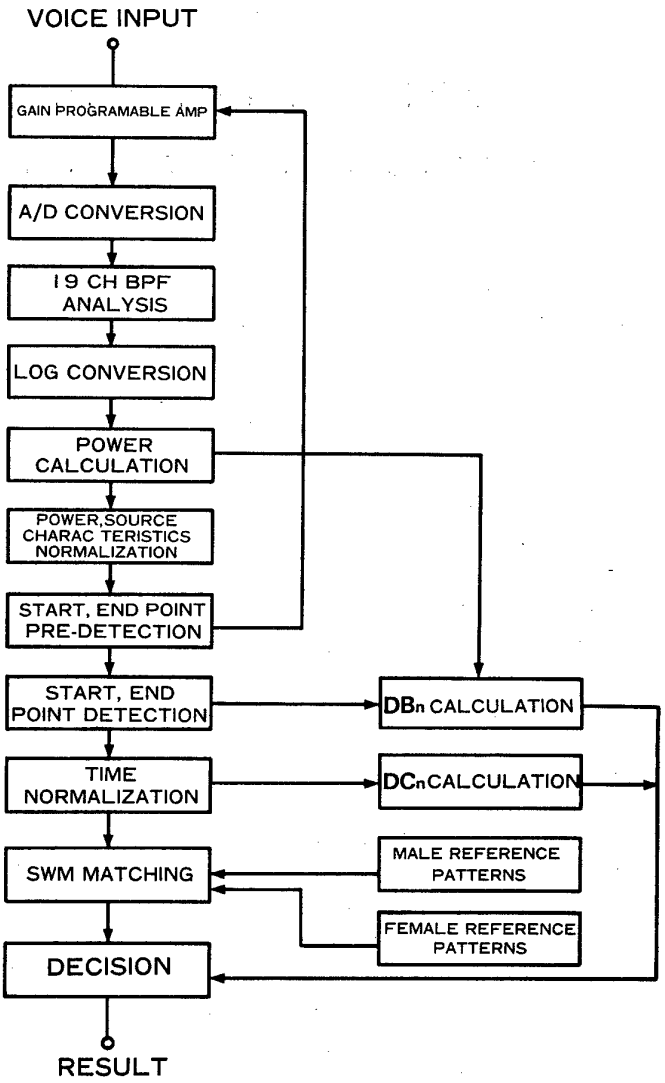

Fig. 1 Recognition flowchart.

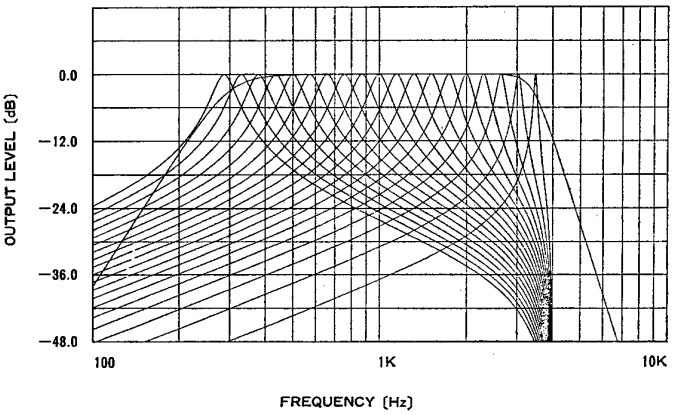

Fig. 2 BPF characteristics. Telephone voice band filter and BPFs characteristics are illustrated.

$$W(i,j)=V(i,j)-(a\cdot i+b) \tag{2}$$

$$a=\frac{1}{c}\left\{19\cdot\left\{\sum_{i=1}^{19} i\right\}\cdot V(i,j)-\left\{\sum_{i=1}^{19} i\right\}\cdot\sum_{i=1}^{19} V(i,j)\right\} \tag{3}$$

$$b=\frac{1}{c}\left\{\left\{\sum_{i=1}^{19} i^2\right\}\cdot\sum_{i=1}^{19} V(i,j)-\left\{\sum_{i=1}^{19} i\right\}\cdot\sum_{i=1}^{19} i\cdot V(i,j)\right\} \tag{4}$$

$$c=19\cdot\sum_{i=1}^{19} i^2-\left\{\sum_{i=1}^{19} i\right\}^2 \tag{5}$$

The voice start/end point pre-detection section performs rough segmentation using the summed value of voice power for every 10 frames obtained from the sound source characteristics normalization section. At the same time, the noise power in the soundless section is obtained as a parameter to detect the voice start/end points.

The voice start/end points are precisely determined by the start/end point detection section. Next, the time normalization section performs time normalization by linear companding to 32 frames using Eq. (6). In Eq. (6), $I(i,j)$ are time normalized data, $STFR$ is voice start frame number, IF is number of input voice frames, and int is a function that returns integer part of the argument.

$$I(i,j)=\frac{1}{4}\sum_{P=P_1}^{P_2} W(i,\mathrm{int}(P/4+STFR))$$
$$\begin{cases}P_1=\mathrm{int}((4\cdot IF-4)\cdot(j-1)/31)\\ P_2=P_1+3\end{cases} \tag{6}$$

The distance D_n for each reference pattern is obtained from summed value of distances DA_n, DB_n, and DC_n. These distances are described below.

The distance DA_n between the time normalized data $I(i,j)$ and the reference pattern $S_n(i,j)$ is obtained from Eq. (7). In Eq. (7), $KW_n(i,j)$ is predetermined weight region for each reference pattern. We call this matching method as Selective Weighted Matching (SWM).

$$DA_n=\sum_{i=1}^{19}\sum_{j=1}^{32}\{S_n(i,j)-I(i,j)\}^2\cdot WE$$
$$WE=\begin{cases}4: & \mathrm{sign}(S_n(i,j))\neq\mathrm{sign}(I(i,j))\\ & \text{and } KW_n(i,j)=1\\ 1: & \text{other}\end{cases} \tag{7}$$

The distance DB_n for category of reference pattern is obtained from the coefficient PDV which represents the magnitude of power dip. PDV is calculated from Eqs. (8)~(11). The $f(j)$ in Eq. (9) represents the straight line tying two peaks of the input voice power, and PA in Eq. (11) represents a slope

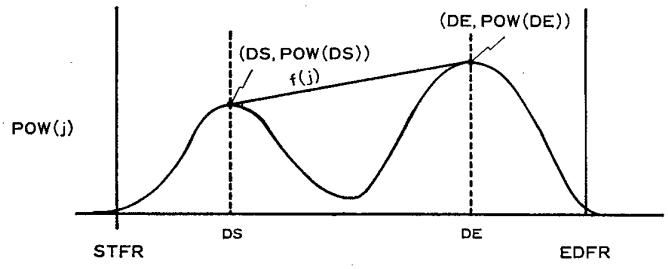

Fig. 3 Power information. *STFR*: voice start point, *EDFR*: voice end point. The speech, for example "ichi," contains power dip. We find 2 power local peaks, and calculate dip magnitude. *PM* is maximum power of input voice. In this case *PM* is *POW*(*DE*).

of the straight line $f(j)$. PM is the maximum value of $POW(j)$. The relationship between these variants is shown in Fig. 3. When a power dip is detected in the input voice, a small DB_n is determined for a category which contains an explosive sound, and a large DB_n for a category which does not contain an explosive sound. If no power dip is detected in the input voice, the above procedure is reversed.

$$PDV=\frac{2\cdot SS}{WW\cdot AA\cdot PM} \tag{8}$$

$$SS=\sum_{j=DS}^{DE}\{f(j)-POW(j)\}^2 \tag{9}$$

$$WW=DE-DS+1 \tag{10}$$

$$AA=PA^2+1 \tag{11}$$

The distance DC_n is converted from the length of input voice using Eq. (12). In Eq. (12), IF represents the number of frames of the input voice, and $FMIN_c$ and $FMAX_c$ represent utterance length variants predetermined for each category. Their values are given in Table 1. Distance DC_n is determined by the category c of reference pattern n. MA and DM are variants for balancing DC_n with DA_n.

$$DC_n=\begin{cases}\min((FMIN_c-IF)\cdot MA, DM), & IF<FMIN_c\\ 0, & FMIN_c\leqq IF\leqq FMAX_c\\ \min((IF-FMAX_c)\cdot MA, DM), & FMAX_c<IF\end{cases} \tag{12}$$

Finally, the total distance D_n is obtained from Eq. (13). $\min(DA)$ in Eq. (13) is the minimum value of DA_n ($n=1\sim N$). DA_n is normalized for combining with DB_n and DC_n. The category of a reference pattern with the minimum D_n is the recognition result.[6]

Table 1 $FMIN_c$, $FMAX_c$.

Category	$FMIN_c$	$FMAX_c$
zero	39	46
ichi	45	53
ni	29	36
san	41	44
yon	34	41
go	28	35
roku	47	55
nana	43	51
hachi	48	57
kyu	39	46
hai	33	40
iie	54	63
dozo	53	61
moichido	83	97
owari	52	60
horyu	53	62

These values represent standard length of utterance for each category.

$$D_n=\frac{DA_n\cdot 1024}{\min(DA)}+DB_n+DC_n \qquad (13)$$

Generally speaking, voice recognition process is done ten-odd times for each user of the speech response service. In this recognition method, both category pre-selection and male/female decision are done in the first N_1 times ($N_1\fallingdotseq 5$) utterances. Once the speaker's sex is determined, only the reference patterns of the corresponding sex are used for recognition process.

Several reports have been presented regarding the validity of category pre-selection and male/female decision.[7-9] The methods of category pre-selection and sex decision used by this system are described below.

A priority is determined for each reference pattern in reference pattern generation process. In category pre-selection, M_1% ($M_1=30\sim 45$) reference patterns with high priorities in each category are pattern-matched. As a result of the pattern matching, the higher M_2 categories ($M_2=4\sim 8$) are selected and pattern-matched with all the remaining reference patterns of these categories. In the final decision stage, the minimum distance obtained from the 1st and 2nd pattern matchings is judged. Male/female decision is identified as follows. Assume that the total number of reference patterns is N. The reference patterns consist of $N/2$ reference patterns generated from the male training data and $N/2$ generated from the female training data. In recognition of the first N_1 words, distances DA_n are summed up seperately for male and female. Thus added values of DA_n for male and female are respectively obtained as DM and DF. If $DM\leqq DF$, the speaker is judged to be male, and if $DF<DM$, the speaker is judged to be female.[10]

3. REFERENCE PATTERN GENERATION

This section describes the reference pattern generation method.

This recognition method has multiple reference patterns to overcome recognition error factors such as the difference of formant frequencies, non-vocalization caused by individual speaker's habits, non-linear variations of utterance speed, and slight segmentation errors. Remarkable advance in semiconductor technology enabled us to adopt the linear matching system with multiple reference patterns. We judged by primary experiment that it was more profitable to adopt the linear matching with increased reference patterns than non-linear matching.[11,12] The size of one reference pattern is 608 bytes. Today, it is possible to build a system having as many as 1,000 reference patterns.

To generate reference patterns, it is necessary to use as many training data as possible and to optimize the number of reference patterns per category and the alignment of reference patterns. To achieve this, we have developed a following reference pattern automatic generation method. Reference patterns will be generated by repeating the recognition of the training data by a certain reference pattern set and updating it based on the recognition result. One updating process is not decisive, but it improves the recognition accuracy at the least. Therefore, an optimum reference pattern can be obtained from 20$\sim$30 times of updating processes.[13] A reference pattern generation flowchart is shown in Fig. 4. Each processing item is discussed below.

The training data used for reference pattern generation are the voice data with normalized vocal chords speech sound characteristics and normalized utterance time as explained in Section 2. The voice data of 500$\sim$1,000 speakers are used to generate reference patterns.

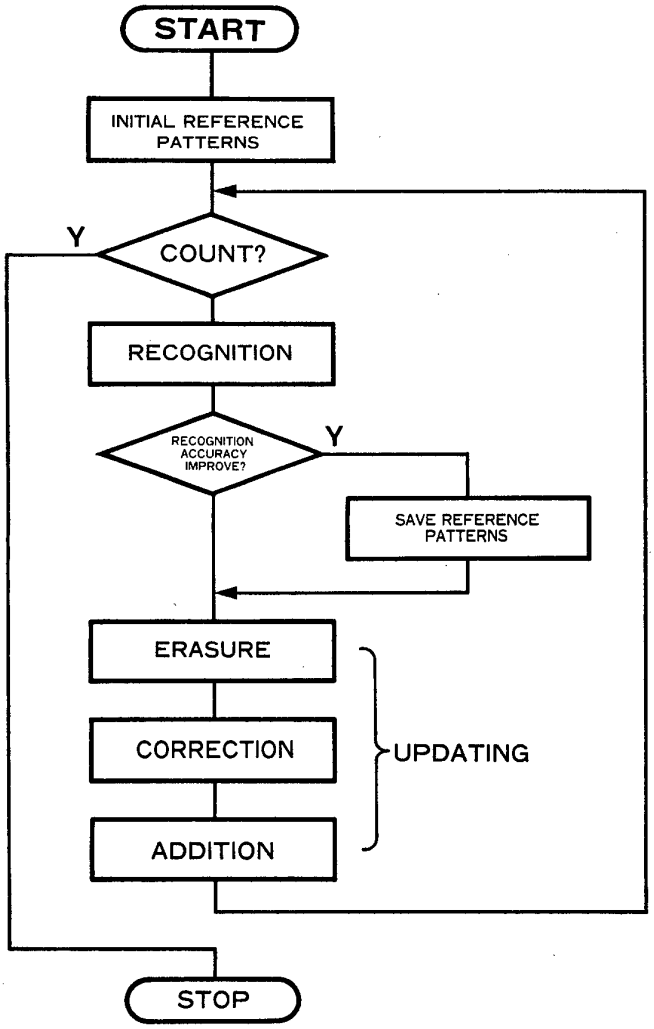

Fig. 4 Reference patterns generation flowchart.

3.1 Reference Pattern Initialization

Let the number of reference patterns to be generated be N. We select total of $N/2$ patterns from the training patterns equally from each category. The selected patterns are the initial reference patterns. It has been already confirmed that the change in the initial reference pattern selection does not affect the finally obtained reference patterns.

3.2 Recognition

All training patterns are recognized. And at the same time, the data required for reference pattern updating are saved. If the result of recognition accuracy has been improved, the relevant reference patterns are saved. The reference pattern set giving the highest recognition accuracy among several tens times of recognition processes is stored as a final result.

3.3 Reference Pattern Updating

In this section, the positional relationship between the reference patterns and training patterns is shown by a two dimensional model (Figs. 5~7). In the figures, ○ and × respectively represent the voice patterns of a different category. ◎ and ※ represent the reference patterns of respective category. In Fig. 5, the dot line is located equidistant from reference patterns "a" and "c" of two categories. Two × marked voice patterns located of the left of the dot line are nearer to ◎ "a", so they may be incorrectly recognized.

The reference patterns are updated through the following 3 steps.

(1) Erasure of the reference pattern
(2) Correction of the reference pattern
(3) Addition of the reference pattern

(1) to (3) are briefed below.

3.3.1 Erasure of the reference pattern

Reference patterns not useful for recognition are erased. Figure 5 shows the concept of pattern erasure. The following patterns are considered useless for recognition.

(a) A pattern which fetches voice patterns of a different category to a greater degree.
(b) A pattern which does not come under (a), but does not fetch many patterns of the same category.

The pattern mentioned in (a) directly causes degradation of the recognition accuracy. For the pattern in (b), the erasure of such reference patterns does not always cause increase in recognition errors, and there is the possibility that such pattern is replaced by other more effective reference patterns.

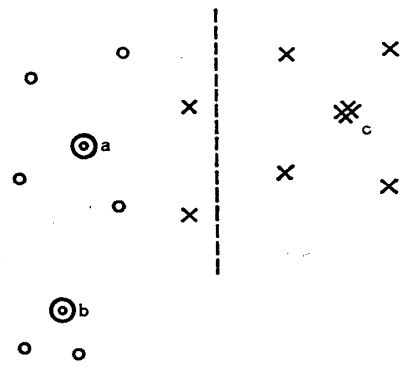

Fig. 5 Erasure. Dot line represents equidistance from 2 opposite reference patterns "a" and "c." These reference patterns are useless for recognition.

Actually the problem lies in the balancing methods of patterns (a) and (b).

3.3.2 Correction of the reference pattern

One reference pattern is generated by averaging out K voice patterns $I_k(i,j)$ using Eq. (14).

$$S_n(i,j)=\frac{1}{K}\sum_{k=1}^{K} I_k(i,j) \tag{14}$$

The positions of the reference pattern can be shifted depending on what patterns are added in averaging process. The reference pattern is shifted based on the following 3 principles.

(a) The reference pattern is not shifted too far away from the previous reference pattern.
(b) The reference pattern is shifted nearly to that of a different category.
(c) The reference pattern is shifted far away from that of the same category.

(a) means that an extreme shift operation should be avoided. (b) is preferable because the reference pattern of different categories should be located closer to each other in order to correctly recognize a voice pattern located on the border line. (c) is done because it is considered effective to distribute reference patterns widely in the pattern group of the same category.

Figure 6 shows the concept of reference pattern correction. In Fig. 6, it is assumed that the reference pattern "d" was composed of 8 peripheral voice patterns. It should be shifted closer to the reference pattern "e" of a difference category to correct the reference pattern. At the same time, it is desirable to shift the reference pattern "d" further away from the reference patterns "f" and "g" of the same category. If five marked patterns of the 8 voice patterns are averaged out, the above principle will be satisfied. The number K of voice patterns to be averaged out is 5 to 32.

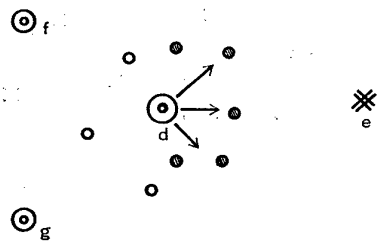

Fig. 6 Correction. Reference pattern "d" is shifted closer to other category's reference patterns "e" and further away from same category's reference patterns "f," "g."

3.3.3 Addition of the reference pattern

Of erroneously recognized patterns, patterns which are far from the reference patterns of the same category but near to that of a different category are added as new reference patterns. In this case, a voice pattern which belongs to a category of low recognition accuracy and is erroneously recognized by a category liable to erroneous recognition is selected with high priority. For example, if the recognition accuracy of category "go" is low and the pattern belonging to that category is liable to be erroneously recognized as "yon," the voice pattern which primarily belongs to category "go" but is erroneously recognized as "yon" is added with a high priority.

Figure 7 shows the concept of reference pattern addition. As voice pattern "h" is closer to the reference pattern "j" of the different category than the reference pattern "i" of the same category, it is erroneously recognized. In this case, voice pattern "h" is added to the reference patterns. If this reference pattern remains valid at the next updating process, averaging process among this pattern and neighbouring voice patterns is carried out and makes a new reference pattern.

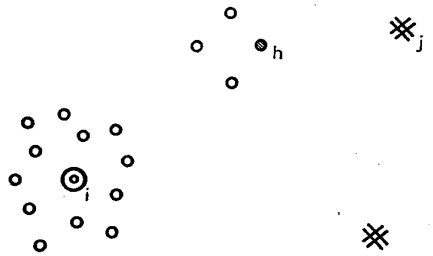

Fig. 7 Addition. Pattern "h" is erroneously recognized. "h" is added as new reference pattern without averaging.

A newly added pattern, which has not undergone the averaging process, is erased with a high possibility at the next updating process. Once the number of reference patterns reaches its maximum, a pattern which was erased immediately after its addition will be excluded at the next adding process.

4. EVALUATION

This section deals with the evaluation of the recognition method using a large amount of voice data.

For evaluation purpose, we used the voice data spoken by about 2,000 male and female adults who each spoke once, recorded through long distance

telephone lines. The recognition vocabulary consists of 16 Japanese words containing 10 digit words from 0 to 9 and 6 control words such as "Hai" (Yes) and "Iie" (No).

The recorded voice data are frequency-analyzed by the BPFs and then their classification and segmentation are checked by the voice data processing system VRDS1 implemented on the computer.[14] If required, some of them are corrected or deleted manually. No errors are permitted in classification and segmentation for reference pattern generation. To evaluate the recognition performance correctly, the work was proceeded based on the principle that voice patterns should be deleted only when an obvious utterance error occurred or a large noise was mixed by a telephone line.

4.1 Reference Pattern Generation

As an example of generating reference patterns from the voice data, the voice of about 480 male speakers is used. It was designated that the number of reference patterns to be adopted should be 192, the number of reference patterns to be deleted at one updating process should be 6, and the maximum number of newly added reference patterns should be 12. Figure 8 shows transition of the recognition accuracy during this process. Once the number of reference patterns reaches the maximum, a few ripples in recognition accuracy are observed. In this example, the maximum recognition accuracy reached 97.64% at the 18th recognition process.

Table 2 shows the confusion matrix at the time of

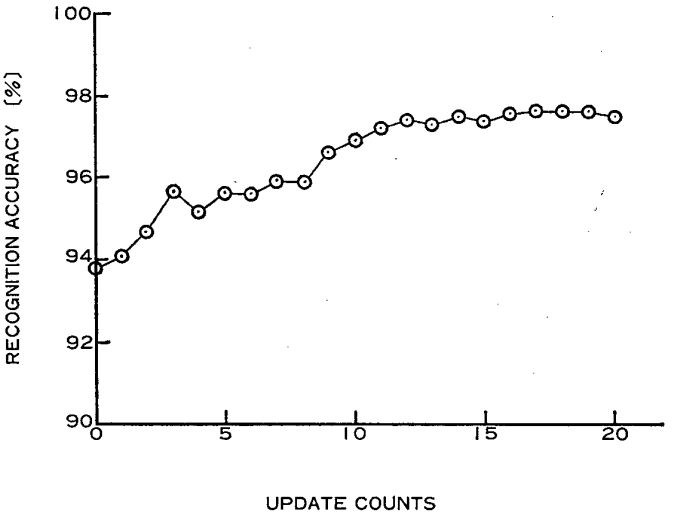

Fig. 8 Recognition accuracy and update procedure.

Table 2 Confusion matrix.

No.	Input category	Recognition results																Accuracy (%)
		0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	
0	zero	361			1	2		1	1		1			1				98.1
1	ichi		418	2	1	1			1	2	2		3					97.2
2	ni		3	438		3	3			1	2		3	1	1			96.3
3	san	3		1	444	1			1	1		1						98.2
4	yon	3		1	15	433								1				95.6
5	go	9				6	435		1					3		1		95.6
6	roku			1	1	3		433	5					3		1	1	96.7
7	nana						1	1	447		1							99.3
8	hachi			1	3					419		10						96.8
9	kyu	3	1	1		1	1				440		1				3	97.6
A	hai		1		5				1	5		444						97.4
B	iie	7	1								1		445					98.0
C	dozo	1										1		452				99.6
D	moichido			1		1		1							444	2	1	98.7
E	owari				1							3				453		99.1
F	horyu			1			1			2	3				1		443	98.2
Total																		97.6

At the time of generating reference patterns using 480 male speakers. Number of reference patterns is 192.

the above reference pattern generation. There are few variances in recognition accuracy among categories. Table 3 shows the number of reference patterns by category for this reference pattern generation. As is clear from the table, the number of reference patterns varies considerably from category to category. The number of reference patterns for each category may sometimes differs in the number shown in this example depending on the training voice data used. As the number of training voice patterns increases, the difference between the recognition accuracy at training and non-training data becomes smaller, the reason being that if there are few training data, reference patterns covering the voice of many speakers are not generated.

Table 3 Number of reference patterns.

Category	Number of patterns
zero	12
ichi	13
ni	18
san	12
yon	27
go	10
roku	18
nana	6
hachi	14
kyu	10
hai	8
iie	6
dozo	14
moichido	8
owari	7
horyu	9

The number of reference patterns varies for each category.

4.2 Pre-selection

We investigated to what degree the pre-selection processing could decrease the times of pattern matching and would degrade the recognition accuracy. The 192 reference patterns were generated from the voice data of 480 male speakers. In this case, the priority is given to the reference patterns which show relatively so many correct recognition results at reference pattern generation.

Table 4 gives the example of pre-selection experiment for training data. This table shows the times of matching and the recognition accuracy when the rate M_1 of the primary matching is set at $30\sim45\%$ and the number of categories M_2 for the secondary matching is set at $4\sim8$.

Table 4 Recognition accuracy with pre-selection.

Primary		Secondary categories (M_2)	Results			
M_1 (%)	Number of patterns		Matching times	Rate with 192 (%)	Accuracy (%)	Recognized patterns
30	50	4	86	45	96.16	876
		6	103	54	96.81	882
		8	121	63	97.47	888
35	60	4	93	48	96.60	880
		6	110	57	97.47	888
		8	126	66	97.69	890
40	69	4	100	52	96.49	879
		6	115	60	97.58	884
		8	131	68	97.80	891
45	78	4	107	56	97.07	884
		6	121	63	97.47	888
		8	135	70	97.80	891
Without pre-selection					98.02	893

M_1 is primary matching rate, and M_2 is secondary matching categories.

If the pre-selection processing is done and the times of matching are decreased to about 60%, the recognition accuracy is degraded approximately from 98% to 97.5%. In the actual recognizer, M_1 is set to 40% and M_2 to 6.

4.3 Male and Female Discrimination

The following experiments confirm the validity of male and female discrimination. The 96 reference patterns were generated from about 240 male speakers. Also 96 reference patterns were generated from about 200 female speakers. Thus a total of 192 reference patterns were generated. Each speaker spoke 16 isolated words. The first N_1 words were used for male and female discrimination. It is assumed that the above N_1 words correspond to a subscriber number or a password of the speech response system. Words from (N_1+1)th to 16th are recognized with either male or female reference pattern based on the discrimination. In this method, the male and female discrimination result is not applied to the N_1 words which are primarily used to discriminate male and female, and therefore, the recognition accuracy tends to be a little lower than the normal counterpart. Table 5 shows the recognition accuracy and the male and female discrimination. The recognition accuracy has been slightly improved. The male and female discrimination rate is almost 100%. One male in the FILE 1 who was mistaken for female can be corrected by putting $N_1=7$.

When 192 reference patterns were generated from the training voice data of 440 male and female speakers, the recognition accuracy is 95.82%. The recognition accuracy of non-training data is lower than the above value. This indicates that it is considerably effective to generate separate reference patterns and discriminate male and female.

Figure 9 shows the summed values *DM* and *DF* (in Section 2) respectively for male and female when N_1 is put as 5 in the MALE FILE 1. Each line segment connects *DM* and *DF* of each speaker.

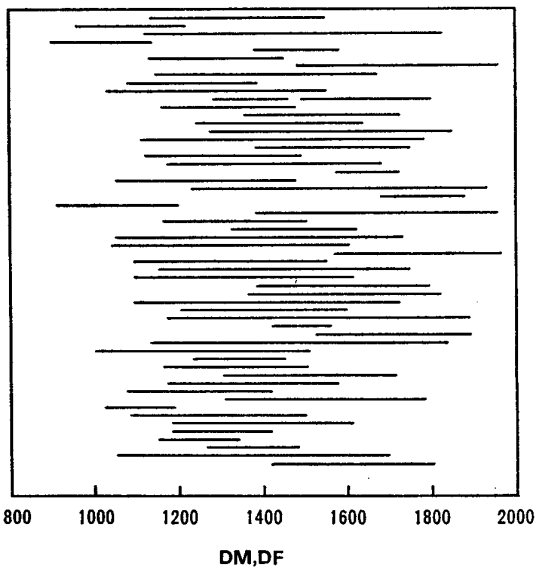

Fig. 9 *DM*, *DF* summation of male/female distances. Each line represents a speaker. The left point of each line is *DM* and the right point is *DF*. It is always $DM \leqq DF$, then male/female discrimination is correct.

Table 5 Evaluation for male/female discrimination.

FILE	N_1	Accuracy (%)	Recognized patterns	Decided as male
	—	97.23	878	
MALE	3	97.34	879	57
1	5	97.34	879	57
	7	97.56	881	58
	—	96.60	880	
MALE	3	96.49	879	58
2	5	96.49	879	58
	7	96.49	879	58
	—	96.32	707	
FEMALE	2	96.87	711	0
1	3	97.00	712	0
	5	97.00	712	0

Number of patterns is 903. Numbers of speakers are 58 for MALE FILE 1 and 2, 48 for FEMALE FILE 1.

Table 6 Recognition with power and length information.

Mode	FILE 4	FILE 5	FILE 6	Average
A	96.95%	96.99%	95.92%	96.61%
B	97.85%	98.10%	96.58%	97.50%
C	98.08%	98.10%	97.24%	97.80%

A: Without power and length information. B: With power information. C: With power and length information. Mode C improves 1.2% than mode A.

ment in the figure represents a speaker. The left end value of the line represents *DM* and the right end value *DF*. As shown in the figure, *DM* values range from 1,000 to 1,400, while *DF* values from 1,100 to 2,000. This assures a stable male and female discrimination.

4.4 Utterance Length and Power Dip Information

The following experiments confirm to what degree the utterance length information and power dip information prove effective for speech recognition. Reference patterns are generated from about 340 male speakers. Then the non-training voice data of about 160 male speakers (FILE 4～6) are provided for evaluation.

The following 3 kinds of recognition modes are provided.

A: Neither utterance length information nor power dip information are applied.
B: Only power dip information is applied.
C: Both utterance length information and power dip information are applied.

In the above processing, the pre-selection and male and female discrimination processes are skipped. Table 6 gives the result. Compare mode A with B, it improves the recognition accuracy about 1% in average. And mode C improves the recognition accuracy by 1.2% compared with mode A. The power dip information is lost by the process of vocal chords characteristics normalization, and utterance length information is lost by time normalization. For example, "ichi" is sometimes mistaken for "ni" and "zero" for "moichido." Mode C proves effective for the improvement of recognition performance because it makes up the lost information again.

4.5 Overall Evaluation

The final evaluation was made in the same manner as when a recognizer is actually operated. The 128 male reference patterns were generated from nearly 340 male speakers and 128 female reference patterns were generated from nearly 230 female speakers. A total reference patterns are 256. Then the non-training voice data of about 170 male speakers (MALE FILE 4～6) and about 60 female speakers (FEMALE FILE 2) were used for overall evaluation. The recognition method same as in Section 2 is applied, but the male and female discrimination result is not applied to the first N_1 words of a speaker as in the case of paragraph 4.3 and N_1 was put as 5. The result is shown in Table 7. As shown in the table, an average recognition accuracy was 97.9% in manual segmentation, which assured the effectiveness of this recognition method. When automatic segmentation is made, the recognition accuracy drops by about 2%. It is considered necessary to improve the above value in future.

Table 7 Results for overall evaluation.

FILE		Segmentation	
		Manual	Automatic
FEMALE	2	97.73%	94.77%
MALE	4	97.47%	95.05%
MALE	5	98.61%	96.81%
MALE	6	97.86%	96.66%
Average		97.92%	95.82%

It is in the same manner as a recognizer is actually operated. The speakers are different from who is used for reference pattern generation.

5. CONCLUSION

We have developed a speaker independent telephone speech recognition method using multiple reference patterns. Differences in formant frequencies, non-linear variation of utterance speed and non-vocalization caused by individual speaker's habits are absorbed by optimized reference patterns.

The structure of the recognizer is very simple by adoption of linear pattern matching. The reference patterns are generated from the voice data of 500 to 600 speakers in the off-line mode. It is required to optimize the reference patterns in terms of the number of reference patterns for every category and their alignment. We have solved the problem of reference pattern generation by new method of repeating recognition and updating. Pre-selection and male and female discrimination are performed in recognition process, and therefore, matching is required only on 60% of the total number of reference patterns. By using the utterance length and power information, the recognition accuracy could be raised from 96.6% to 97.8%. This value is equal to the recognition accuracy obtained at the time of reference pattern generation. The problems to be left for solution are to increase the number of recognition words, and to update the reference patterns generation methods when the required number of reference patterns is increased to the degree of 1,000.

ACKNOWLEDGEMENTS

The authors would like to thank valuable discussion and instruction from Prof. K. Kido of Tohoku University. The authors wish to thank many persons of OKI Electric Ind. Co., Ltd. who were concerned about this work. Especially, we would like to acknowledge Dr. S. Nakaya who encouraged us and gave valuable suggestion, Mr. Y. Tabei and Miss A. Hirota who contributed greatly to this work.

REFERENCES

1) H. Iizuka, M. Morito, and K. Yamada, "Speaker independent spoken word recognition," Trans. Comm. Speech Res., Acoust. Soc. Jpn. S83-54 (1983) (in Japanese).
2) I. Nose, K. Mizuno, and K. Yamada, "Speaker-independent isolated word recognition using a new matching method," J. Acoust. Soc. Am. Suppl. 1, Vol. 71, S7, Spring 1982 (C8. 103rd MASA).
3) K. Mizuno, H. Iizuka, and K. Yamada, "Some experiments of a speaker-independent isolated word recognition using the Selective Weighted Matching," J. Acoust. Soc. Am. Suppl. 1, Vol. 72, S31, Fall 1982 (R9. 104th MASA).
4) H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. **ASSP-26**, 43–49 (1978).
5) J. Miwa, M. Obara, S. Makino, and K. Kido, "A method of spoken word recognition using nonlinear spectral matching," IECE Trans. **J62-D**, No. 1, 46–53 (1981) (in Japanese).
6) A. Hirota, H. Iizuka, M. Morito, and K. Yamada, "Evaluation of speaker-independent telephone speech recognition method," Rec. Fall Meet. Acoust. Soc. Jpn. 1-9-18, 35–36 (1984) (in Japanese).
7) J. Miwa and K. Kido, "Speaker-independent word recognition with large vocabulary using pre-selection and vocal-tract normalization," Rec. Spring Meet. Acoust. Soc. Jpn. 2-2-13, 71–72 (1983) (in Japanese).
8) J. Miwa and K. Kido, "A method of pre-selection for large vocabulary spoken word recognition," Tech. Rep. Electr. Acoust., IECE Jpn. EA82-31 (1982) (in Japanese).
9) N. Hataoka and A. Ichikawa, "Consideration on speaker clustering," Rec. Spring Meet. Acoust. Soc. Jpn. 2-2-15, 75–76 (1983) (in Japanese).
10) H. Iizuka, M. Morito, A. Hirota, and K. Yamada, "Speaker independent telephone speech recognition using male/female reference patterns," Rec. Spring Meet. IECE Jpn. 1671, 6–224 (1984) (in Japanese).
11) L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker independent recognition of isolated words using clustering techniques," IEEE Trans. **ASSP-27**, 336–349 (1979).
12) N. Sugamura, K. Shikano, and M. Kohda, "Speaker-independent isolated word recognition based on multiple word templates using SPLIT method," IECE Trans. **J67-D**, 1210–1217 (1984) (in Japanese).
13) H. Iizuka, I. Nose, K. Mizuno, and K. Yamada, "Reference pattern generation method for speaker independent spoken word recognition," Rec. Spring Meet. IECE Jpn. 1392, 5–377 (1982) (in Japanese).
14) H. Iizuka, K. Mizuno, and K. Yamada, "Data processing system for speaker independent spoken word recognizer development," Rec. Spring Meet. IECE Jpn. 1444, 5–315 (1983) (in Japanese).