STUDIA PHONOLOGICA XIV (1980)

# Differences in Feature Parameters of Japanese Vowels with Sex and Age

Sei-ichi NAKAGAWA, Hironori SHIRAKATA,
Masatoshi YAMAO and Toshiyuki SAKAI

## Summary

In this paper, we consider the problem of speaker differences in Japanese vowels which are the most important phonemes to recognize Japanese.

The speaker differences are divided into two kinds. One is inter-group differences—speaker differences in age and sex. The other is intra-group differences. The former is the physical differences of the apparatus, and the latter is caused by the minute differences of articulators and the differences of linguistic environments.

First of all, we investigate the speaker differences in sex and age by using many materials spoken by 120 persons, and test the existence of the inter-group differences by using a variance analysis technique. Next, we experiment on Japanese vowel recognition on the basis of the result of analyses. From these results, we obtain in conclusion that the classification of speaker classes in terms of sex and age is effective for speaker independent vowel recognition and that the 3 classes of male, female and children are the best kind as speaker-grouping ways.

## I. Introduction

Speech is the most natural communication media for man. Therefore, it is useful for man-machine communication. However, a machine must recognize conversation or continuous speech to realize the natural communication between man and machine. Automatic recognition of continuous speech must solve very difficult problems such as segmentation, coarticulation, speaker differences, word juncture, prosody and so on. Until now, many researchers have studied such problems, but they are still open problems. In this paper, we consider the problem of speaker differences in Japanese vowels which are the most important phonemes to recognize Japanese speech.

The speaker differences are divided into two kinds. One is inter-group differences—speaker differences in age and sex. The other is intra-group differences. The former is the physical differences of the apparatus (most of hardware differences)

Sei-ichi NAKAGAWA (中川聖一) Assistant Professor, School of Information Engineering, Toyohashi University of Technology
Hironori SHIRAKATA (白方博教) Shikoku Electric Power Company
Masatoshi YAMAO (山尾雅利) Toshiba Ltd.
Toshiyuki SAKAI (坂井利之) Professor, Department of Information Science, Kyoto University

and the latter is mainly the minute differences of articulators (part of hardware differences) and the differences of linguistic environments (software differences).

Some studies of speaker normalization in acoustic feature parameters revealed the insufficiency of a simple uniform normalization method, for example, usage of relative values between formant frequencies[1],[2],[3] or vocal tract length[4],[5]. Some analyses of feature parameters showed the existence of speaker differences in sex and age[6],[14]. Such differences are not a simple linear relationship. G. Fant reported that the direction of distribution of Italian /a/ in the F1–F2 plane is different on sex[7]. R. D. Kent and L. L. Forner reported that the scale factor between adult male and children is different on vowel or formant frequency[8]. Furthermore, S. F. Disner showed that the optimal speaker normalization procedure depends on language[9].

G. Fant proposed a non-linear normalization method on formant frequencies which did not assume the homology of vocal tract shape[7]. However, his method did not consider explicitly the difference on sex or vowel. We propose a method which eliminates the inter-group differences by grouping speakers.

First of all, we investigate the speaker differences in sex and age by using many materials spoken by 120 persons, and test the existence of the inter-group differences by using a variance analysis technique. Next, we experiment on Japanese vowel recognition on the basis of the result of analyses.

## II. SPEECH MATERIALS AND FEATURE PARAMETERS

(i) Speaker

Speakers are predolesent schoolboys and schoolgirls (fifth grade, 10 or 11 years old), males and females of about 20 years old, and males and females over 40 years old. Each group consists of 20 persons (total of 120 persons), and everybody is normal speakers with Kansai dialect.

(ii) Speech Material

Each speaker in the above uttered three times Japanese five vowels (/a/, /i/, /u/, /e/, /o/). Each utterance was sampled by 10 KHz and digitized at 10 bit/sample. The 256 samples in the most stationary were extracted by manual.

(iii) Feature Parameter

The extracted samples were analyzed by the procedure shown in Fig. 1.

(a) pitch frequency (F0)—fundamental frequency of glottal source. This was extracted by the cepstrum technique.

(b) first three formant frequencies (F1, F2, F3)—resonance frequencies of vocal tract. They were extracted by an operator on the basis of visual observation of cepstrum-spectrum[10], the results of peak picking method and pole frequencies of LPC model[11].

(c) inclination of spectrum (INCL)—characteristics of glottal source. This was extracted from the least squares fit line of speech spectrum.

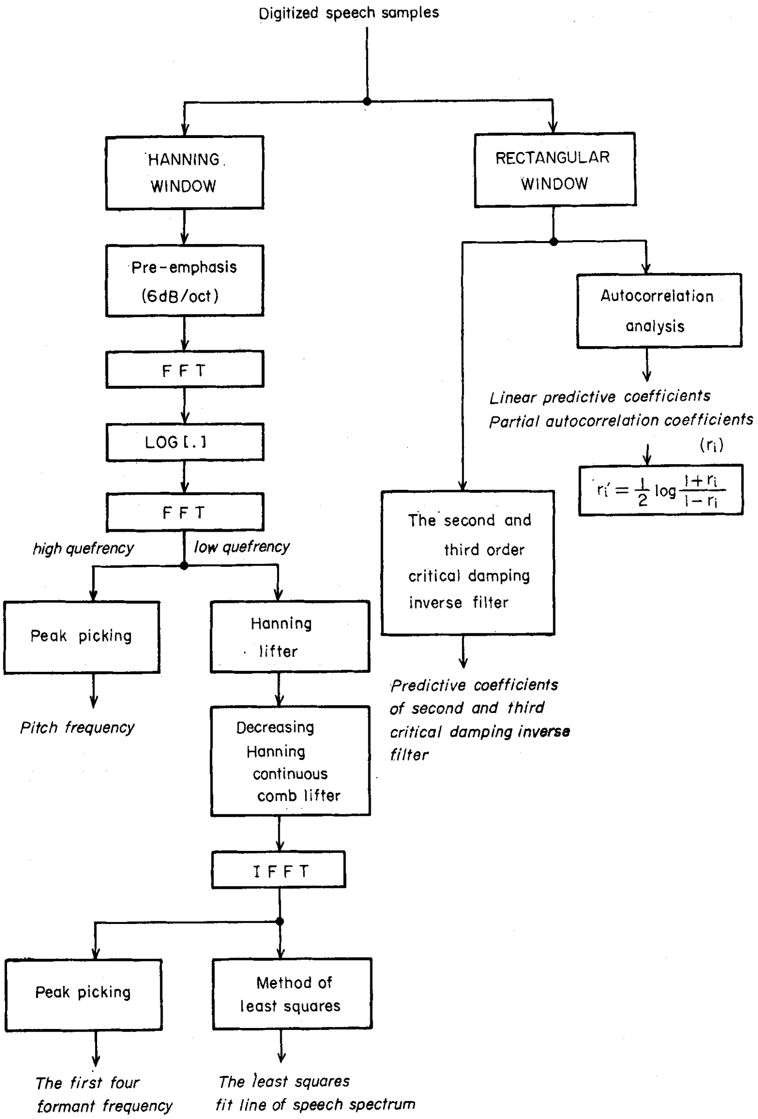
Digitized speech samples
HANNING WINDOW
RECTANGULAR WINDOW
Pre-emphasis (6dB/oct)
Autocorrelation analysis
F F T
Linear predictive coefficients
Partial autocorrelation coefficients (ri)
LOG [.]
$r_i' = \frac{1}{2} \log \frac{1+r_i}{1-r_i}$
F F T
The second and third order critical damping inverse filter
high quefrency
low quefrency
Peak picking
Hanning lifter
Predictive coefficients of second and third critical damping inverse filter
Pitch frequency
Decreasing Hanning continuous comb lifter
I F F T
Peak picking
Method of least squares
The first four formant frequency
The least squares fit line of speech spectrum

Fig. 1. Procedure of feature parameters extraction.

(d) partial autocorrelation coefficients (PARCOR, $r_1$, ..., $r_{14}$)—reflection coefficients of vocal tract. They were calculated by an autocorrelation method[11].

(e) log area ratio of vocal tract (LAR, $r_1'$, ..., $r_{14}'$) —$r_i'=\frac{1}{2}\cdot\log\frac{1+r_i}{1-r_i}$ This transformation makes a normal distribution of r[12].

(f) predictive coefficient of second and third critical damping inverse filter (CD2, CD3)—slope and flection of frequency characteristic including both glottal source and radiation[13].

## III. DIFFERENCES OF PITCH AND FORMANT FREQUENCIES IN SEX AND AGE

The analyses of pitch frequency and formant frequencies have been studied by many researchers[6),14),15),16)]. These were extracted by the visual observation of sonagraph. Therefore the accuracy was not satisfactory. The many data have not been obtained by an automatic extraction. In particular, the automatic extraction of formant frequency is very difficult. Therefore, we extracted pitch frequency

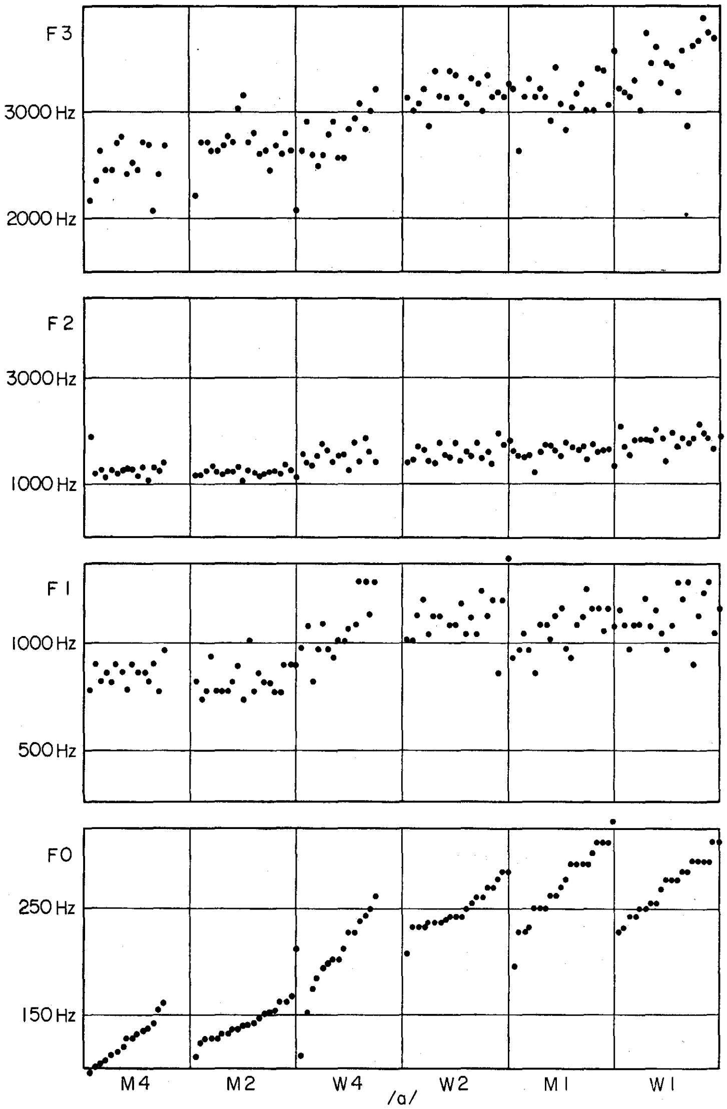

Fig. 2. Pitch and formant frequencies.
F0: pitch frequency. Fl: first formant frequency. F2: second formant frequency. F3: third formant frequency. Ml: schoolboy. Wl: schoolgirl. M2: male about 20 years old. W2: female about 20 years old. M4: male over 40 years old. W4: female over 40 years old.

and formant frequencies by a semi-automatic procedure from many materials of 120 speakers. We divided 120 speakers into six groups as follows. Each group consists of 20 speakers.

| | |
|---|---|
| M1: | schoolboy |
| W1: | schoolgirl |
| M2: | adult male about 20 years old |
| W2: | adult female about 20 years old |
| M4: | adult male over 40 years old |
| W4: | adult female over 40 years old |

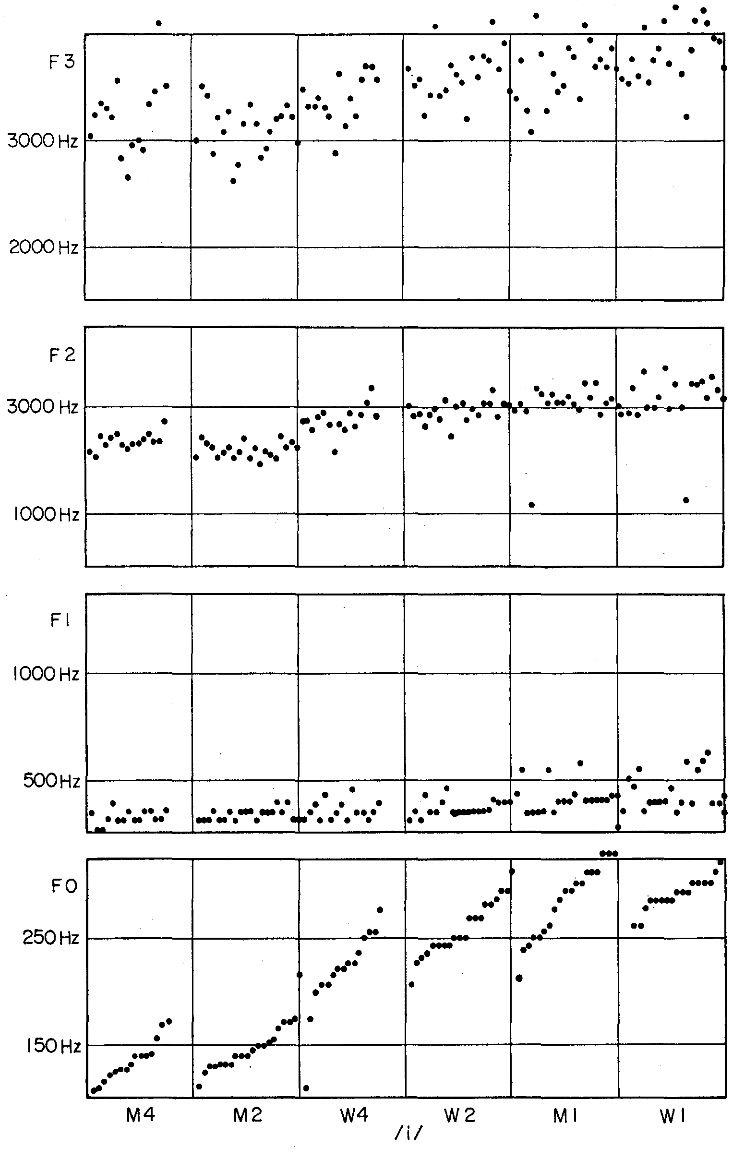

Fig. 2.

Fig. 2 shows the extracted pitch frequency and formant frequencies. They were arranged by the increasing order of pitch frequency in each speaker group. Fig. 3 and Table 1 show the average values of every parameters.

The pitch frequency of vowel /u/ is high and that of /e/ is low for every speakers. The differences is about 20%. We cannot still decide whether these phenomena are inherent[21] or were caused by the order of utterance (/a, i, u, e, o/). The increasing order of pitch frequency is M4, M2, W4, W2, M1, W1. This is nearly the same order as formant frequencies. However, there is no correlation between pitch and formant frequencies in intra-speaker groups. The speaker groups are merged into three classes: (M4, M2), (W4), (W2, M1, W1) with according to pitch frequency, (M4, M2), (W4), (M1, W1) with according to formant frequencies, respectively.

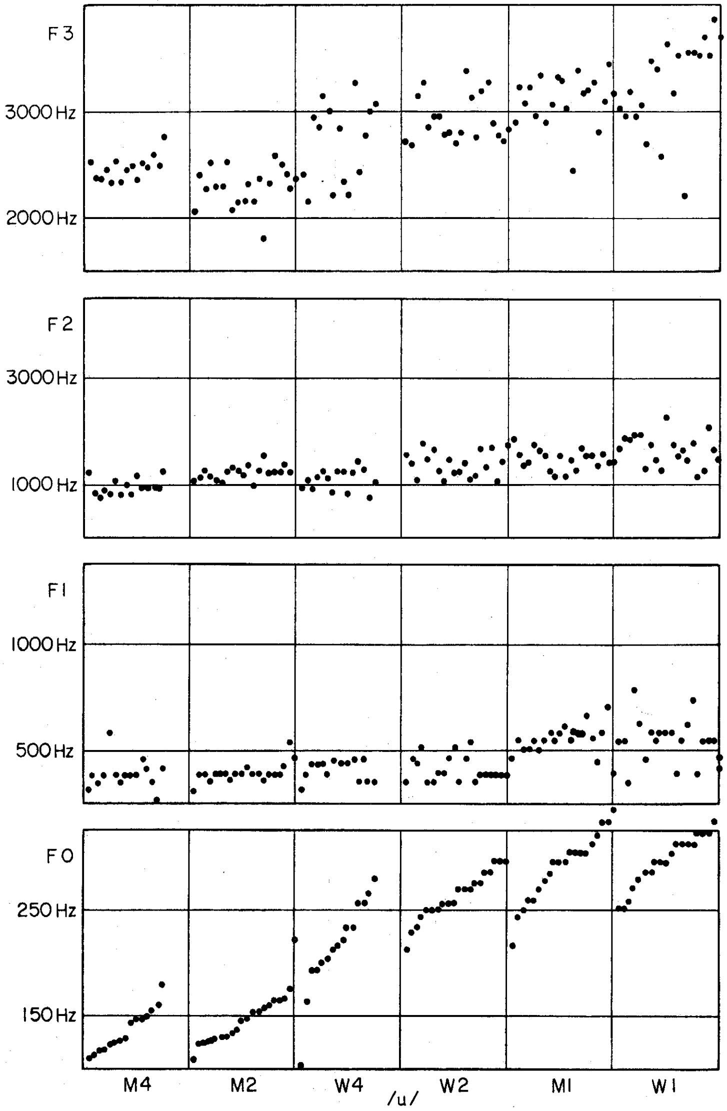

Fig. 2.

In the latter case, the speaker group (W2) belongs to either (W4) or (M1, W1). The characteristic of (W2) is similar to that of (W4, M1) more than that of (W1). There is also the same relationship on spectra of vowels and spoken words[20].

The differences of formant frequencies with sex and age are not a simple relationship. They depend on vowels and formants. For example, the second formant frequencies (F2) of /o/ are almost constant. F1, F2 of /a/ and F2 of /i/ classify (M4, M2) and (W4, W2, M1, W1). F3 of /a/ classifies (M4, M2, W4) and (W2, M1, W1).

Although we cannot classify /a/ of (M2, M4) and /o/ of (M1, W1), and /e/ of

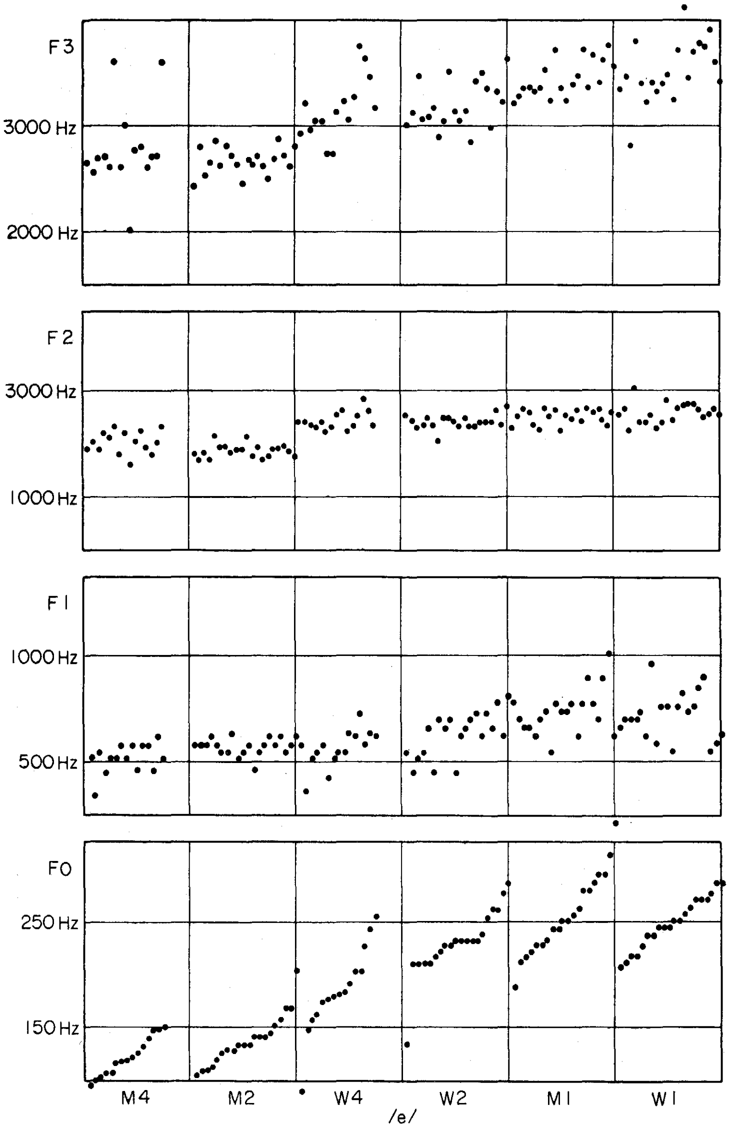

Fig. 2.

(M2, M4) and /u/ of (M1, W1) by only F1 and F2[6], if we take F3 into consideration, we can classify them. The first three formant frequencies have high performance for vowel recognition. As described above, however, their automatic extraction is very difficult. Therefore we think they are not useful as feature parameters in automatic speech recognition systems.

## IV. STATISTICAL AANLYSES OF FEATURE PARAMETERS IN VOWEL, SEX AND AGE

### IV-1 F Ratio of Feature Parameters

Each parameter is evaluated in terms of this ability to classify speaker groups.

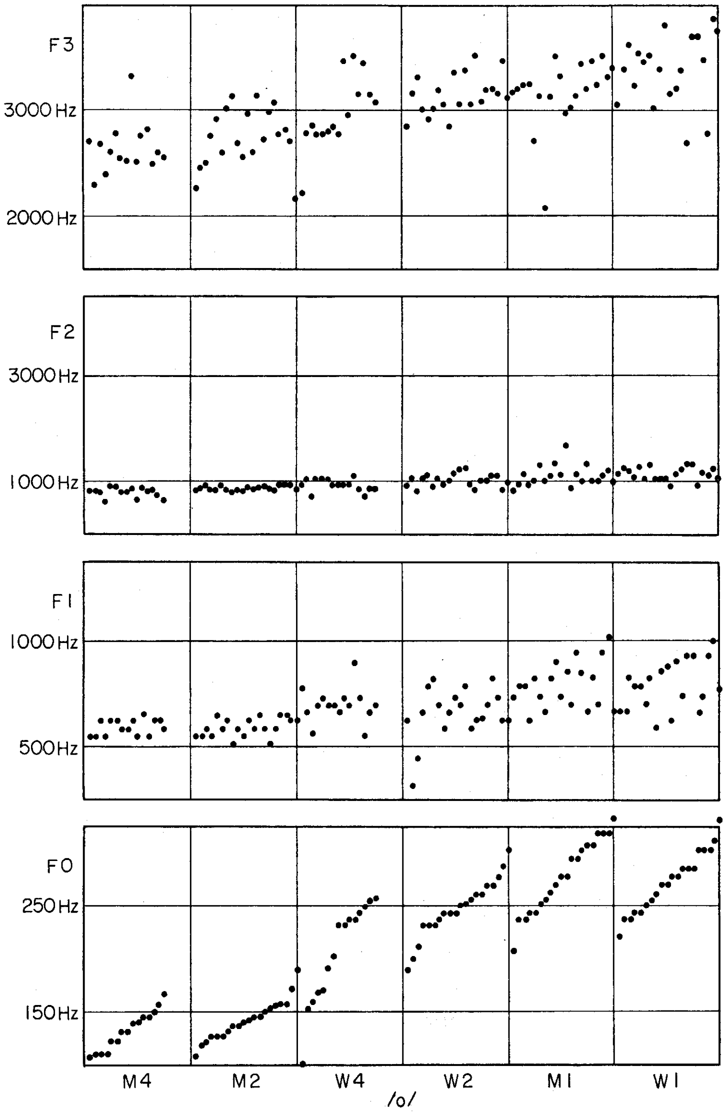

Fig. 2.

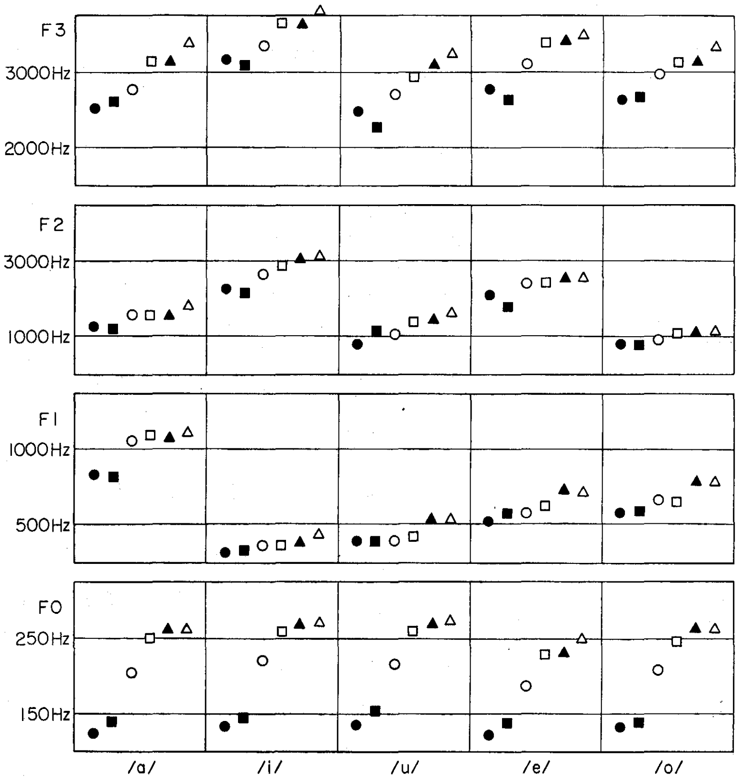

Fig. 3. Average values of pitch and formant frequencies for each group. ▲: M1 (schoolboy), △: W1 (schoolgirl), ■: M2 (male about 20 years old), □: W2 (female about 20 years old) ●: M4 (male over 40 years old), ○: W4 (female over 40 years old)

For this purpose, the F ratio of the analysis of variance is used. A good parameter is one for which the individual group distributions are narrow and as widely separated as possible. The F ratio is given by

$$F_k = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\overline{X}_{ik}-\overline{\overline{X}}_k)^2}{\frac{1}{(m-1)n}\sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ijk}-\overline{X}_{ik})^2}$$

where $X_{ijk}$ is the k-th parameter value on the j-th repetition by the i-th group, $j=1, 2, \ldots, m, i=1, 2, \ldots, n$;

$$\overline{X}_{ik} = \frac{1}{m}\sum_{j=1}^{m} X_{ijk}$$

$$\overline{\overline{X}}_k = \frac{1}{n}\sum_{i=1}^{n} \overline{X}_{ik}$$

Thus F is proportional to the ratio of the mean of the intra-group variance to the mean of the inter-group variance. The latger is the value of F, the more suitable is the parameter for speaker-group recognition, that is, the larger is the speaker-group differences.

Table 1. Average of pitch and formant frequencies of isolated vowels.

| Group | Sex | Age (years) | Height (cm) | Pitch formant | Average frequencies (in Hz) of five vowels | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | a | i | u | e | o | average |
| M1 | male | 10. 3 | 137. 4 | pitch | 274 | 289 | 291 | 257 | 279 | 278 |
| | | | | $F_1$ | 1068 | 413 | 550 | 744 | 787 | 712 |
| | | | | $F_2$ | 1609 | 3077 | 1474 | 2521 | 1123 | 1961 |
| | | | | $F_3$ | 3182 | 3655 | 3125 | 3437 | 3170 | 3314 |
| W1 | female | 10. 5 | 140. 2 | pitch | 272 | 292 | 300 | 249 | 274 | 277 |
| | | | | $F_1$ | 1132 | 447 | 550 | 722 | 789 | 728 |
| | | | | $F_2$ | 1859 | 3144 | 1648 | 2593 | 1160 | 2081 |
| | | | | $F_3$ | 3442 | 3867 | 3269 | 3529 | 3393 | 3500 |
| M2 | male | 22. 6 | 169. 8 | pitch | 145 | 148 | 154 | 139 | 142 | 146 |
| | | | | $F_1$ | 832 | 338 | 396 | 578 | 593 | 547 |
| | | | | $F_2$ | 1234 | 2207 | 1215 | 1894 | 894 | 1489 |
| | | | | $F_3$ | 2668 | 3118 | 2302 | 2675 | 2730 | 2699 |
| W2 | female | 19. 0 | 156. 8 | pitch | 251 | 259 | 263 | 231 | 249 | 251 |
| | | | | $F_1$ | 1124 | 367 | 413 | 636 | 661 | 667 |
| | | | | $F_2$ | 1619 | 2929 | 1406 | 2451 | 1043 | 1890 |
| | | | | $F_3$ | 3192 | 3634 | 2939 | 3197 | 3149 | 3222 |
| M4 | male | 53. 0 | 165. 6 | pitch | 125 | 135 | 136 | 124 | 133 | 131 |
| | | | | $F_1$ | 854 | 325 | 393 | 520 | 594 | 582 |
| | | | | $F_2$ | 1268 | 2369 | 981 | 2082 | 831 | 1506 |
| | | | | $F_3$ | 2515 | 3244 | 2502 | 2593 | 2676 | 2746 |
| W4 | female | 51. 1 | 153. 6 | pitch | 206 | 220 | 215 | 186 | 207 | 207 |
| | | | | $F_1$ | 1071 | 359 | 408 | 539 | 701 | 616 |
| | | | | $F_2$ | 1552 | 2774 | 1101 | 2471 | 957 | 1771 |
| | | | | $F_3$ | 2807 | 3398 | 2739 | 3159 | 2978 | 3016 |

Table 2 shows the F-ratio of each feature parameter. The F-ratio of pitch frequency, F3, inclination of spectrum and the 10-th PARCOR is large. We should note that F ratio is smaller than 1 except for pitch frequency, because the variance for vowels is not separated from the variance for speakers. If F ratio is calculated for every vowel, it would become larger than 1. In the next section, we separate the variance for vowel-factor and the variance for speaker-factor.

## IV–2 Variance Analysis of Feature Parameters

The F ratio described in the previous section showed the ratio of the speaker differences of intra-group to that of inter-group, and the number of groups was fixed to six. In this section, we investigate the relationship in each feature parameter among the speaker differences of intra-group, the speaker differences of inter-group, the differences among vowels and so on. K. Tabata et al[17]. andl S. Furui[12] investigated the speaker differences in vowels of male adults by a variance analysis. We test statistically how it is effective to eliminate speaker differences in the case that

Table 2. F ratio.

F0: pitch frequency
Fi: the i-th formant frequency
INCL: inclination of spectrum
CDi: predictive coefficient of the i-th critical damping inverse filter
$\gamma_i$: the i-th PARCOR ceefficient

| parameter | F ratio |
|---|---|
| F0 | 4.10 |
| F1 | 0.11 |
| F2 | 0.11 |
| F3 | 0.74 |
| INCL | 0.28 |
| CD 2 | 0.06 |
| CD 3 | 0.04 |
| $\gamma_1$ | 0.08 |
| $\gamma_2$ | 0.03 |
| $\gamma_3$ | 0.003 |
| $\gamma_4$ | 0.01 |
| $\gamma_5$ | 0.13 |
| $\gamma_6$ | 0.09 |
| $\gamma_7$ | 0.04 |
| $\gamma_8$ | 0.14 |
| $\gamma_9$ | 0.04 |
| $\gamma_{10}$ | 0.25 |
| $\gamma_{11}$ | 0.10 |
| $\gamma_{12}$ | 0.09 |
| $\gamma_{13}$ | 0.16 |
| $\gamma_{14}$ | 0.27 |

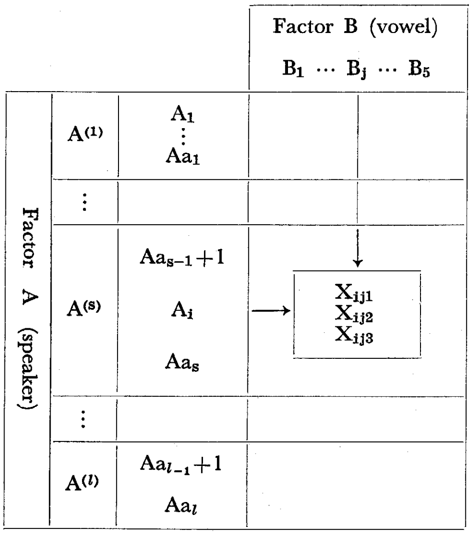

Fig. 4. Multi-divided type of variance analysis for two-factor design with repeated measurements.

the speaker group (class) of an unknown speaker is given. For this purpose, we extend Tabata's two divided type[18)] of model of analysis of variance to multi divided type as follows (see Fig. 4).

*The linear model of l-divided type of variance analysis for two-factor design with repeated measurements.*

$$X_{ijk}=\mu+\mu^{(s)}+\alpha_i^{(s)}+\beta_j+\tau_j^{(s)}+\gamma_{ij}^{(s)}+\varepsilon_{ijk},$$

where,

$$s=\begin{cases} 1, & a_0+1\leqq i\leqq a_1 \ ; \ a_0=0 \\ 2, & a_1+1\leqq i\leqq a_2 \\ \vdots & \\ l, & a_{l-1}+1\leqq i\leqq a_l \ ; \ a_l=120 \end{cases}$$

$$d_s=a_s-a_{s-1}$$

$$\sum_{s=1}^{l} d_s\cdot\mu^{(s)}=0, \quad \sum_{j=1}^{5}\beta_j=0, \quad \varepsilon_{ijk}\sim N(0,\sigma^2)$$

$$\sum_{i=a_0+1}^{a_1}\alpha_i^{(1)}=\sum_{i=a_1+1}^{a_2}\alpha_i^{(2)}=\cdots=\sum_{i=a_{l-1}+1}^{a_l}\alpha_i^{(l)}=0$$

$$\sum_{j=1}^{5}\tau_j^{(1)}=\cdots=\sum_{j=1}^{5}\tau_j^{(l)}=0, \quad \sum_{s=1}^{l} d_s\cdot\tau_j^{(s)}=0$$

$$\sum_{j=1}^{a_1} \gamma_{ij}{}^{(1)} = \cdots = \sum_{i=a_{l-1}+1}^{a_l} \gamma_{ij}{}^{(l)} = 0$$

$$\sum_{j=1}^{5} \gamma_{ij}{}^{(1)} = \cdots = \sum_{j=1}^{5} \gamma_{ij}{}^{(l)} = 0$$

$X_{ijk}$: the k-th observation of the j-th vowel of the i-th speaker. $1 \leqq k \leqq 3$, $1 \leqq j \leqq 5$ (1=/a/, 2=/i/, 3=/u/, 4=/e/, 5=/o/) and $1 \leqq i \leqq 120$.

s: class number of the i-th speaker.

$d_s$: number of speakers in the s-th class.

$\mu$: general level.

$\mu^{(s)}$: main effect of the s-th class-factor.

$\alpha_i{}^{(s)}$: main effect of speaker-factor for the s-th class.

$\beta_j$: main effect of vowel-factor.

$\tau_j{}^{(s)}$: interaction effect of vowel-factor and the s-th class-factor.

$\gamma_{ij}{}^{(s)}$: interaction effect of vowel-factor and speaker-factor for the s-th class.

$\varepsilon_{ijk}$: residual.

*The breakdown of total variance Q of $X_{ijk}$ is as follows.*

$$Q = \sum_{i=1}^{120} \sum_{j=1}^{5} \sum_{k=1}^{3} (X_{ijk} - X_{\cdots})^2$$

$$= Q_1 + Q_2{}^{(1)} + \cdots + Q_2{}^{(l)} + Q_3 + Q_4 + Q_5{}^{(1)} + \cdots + Q_5{}^{(l)} + R,$$

where

$$Q_1 = 5 \times 3 \times \sum_{s=1}^{l} ds \cdot (X^{(s)}{}_{\cdots} - X_{\cdots})^2$$

$$Q_2{}^{(s)} = 5 \times 3 \times \sum_{i=a_{s-1}+1}^{a_s} (X_{i\cdot\cdot} - X_{\cdots})^2$$

$$Q_3 = 120 \times 3 \times \sum_{j=1}^{5} (X_{\cdot j \cdot} - X_{\cdots})^2$$

$$Q_4 = 3 \times \sum_{j=1}^{5} \sum_{s=1}^{l} ds \cdot (X_{\cdot j \cdot}{}^{(s)} - X^{(s)}_{\cdots} - X_{\cdot j \cdot} + X_{\ldots})^2$$

$$Q_5{}^{(s)} = 3 \times \sum_{i=a_{s-1}+1}^{a_s} \sum_{j=1}^{5} (X_{ij\cdot} - X_{i\cdot\cdot} - X_{\cdot j \cdot}{}^{(s)} + X^{(s)}_{\cdots})^2$$

$$R = \sum_{i=1}^{120} \sum_{j=1}^{5} \sum_{k=1}^{3} (X_{ijk} - X_{ij\cdot})^2$$

$$X_{\cdots} = \frac{1}{120 \times 5 \times 3} \sum_{i=1}^{120} \sum_{j=1}^{5} \sum_{k=1}^{3} X_{ijk}$$

$$X_{ij\cdot} = \frac{1}{3} \sum_{k=1}^{3} X_{ijk}$$

$$X^{(s)}_{\cdots} = \frac{1}{ds \times 5 \times 3} \sum_{i=a_{s-1}+1}^{a_s} \sum_{j=1}^{5} \sum_{k=1}^{3} X_{ijk}$$

$$X_{i\cdot\cdot} = \frac{1}{5 \times 3} \sum_{j=1}^{5} \sum_{k=1}^{3} X_{ijk}$$

$$X_{\cdot j\cdot}=\frac{1}{120\times 3}\sum_{i=1}^{120}\sum_{k=1}^{3}X_{ijk}$$

$$X_{\cdot j\cdot}^{(s)}=\frac{1}{\mathrm{ds}\times 3}\sum_{i=a_{s-1}+1}^{a_s}\sum_{k=1}^{3}X_{ijk}$$

These $Q_i$, $Q_i^{(1)}\cdots Q_i^{(l)}$ and R correspond to the above $\mu, \alpha, \beta, \tau, \gamma, \varepsilon$ as follows. $Q_1$: $\mu^{(s)}$, $Q_2^{(1)}$: $\alpha_i^{(1)}\cdots Q_2^{(l)}$: $\alpha_i^{(l)}$, $Q_3$: $\beta_j$, $Q_4$: $\tau_j^{(s)}$, $Q_5^{(1)}$: $\gamma_{ij}^{(l)}\cdots Q_5^{(l)}$:$\gamma_{ij}^{(l)}$, R: $\varepsilon_{ijk}$

*The likelihood ratio test for null hypothesis*

Let us consider a test of the hypothesis for each effect, for example,

hypothesis $H_0$: $\mu^{(1)}=\mu^{(2)}=\cdots=\mu^{(l)}=0$

that is, all the effects of each class are equal (there is no effect of speaker-class). In this case, we can test the hypothesis since it is possible to prove that the likelihood ratio criterion

$$\nu=\{n-l_2-1-l_1/2\}\cdot\log(Q_1+R)/R$$

is distributed asymptotically according to $\chi^2$-distribution with $l_1$ degrees of freedom under the conditions, $n=120\times 5\times 3=1800$, $l_1+l_2=120\times 5=600$, and $l_1=l-1$[18),19)]. The hypotheses concerned with the other factors or interactions may be tested in the similar way with the each corresponding $Q_i$, $Q_i^{(m)}$, $l_1$ and $l_2$ shown in Table 3.

We normalize $\nu$ by the value of significant level as the following equation, because the degrees of freedom corresponding to main effects and interactions are different from each other and so are the values of 1% significant level $\chi^2$ test.

$$\nu'=\frac{\nu}{\text{value of 1\% significant level of }\chi^2\text{ test corresponding to the degrees of freedom of }\nu}$$

We tried the variance analysis about four kinds as speaker-grouping ways, that is, 2 classes (male, female: M1 & M2 & M4/W1 & W2 & W4), 2′ classes (children & female, adult male: M1 & W1 & W2 & W4/M2 & M4), 3 classes (children, adult male, adult female: M1 & W1/M2 & M4/W2 & W4) and 6 classes (schoolboy, schoolgirl, male about 20 years old, female about 20 years old, male over 40 years old, female over 40 years old: M1/W1/M2/W2/M4/W4). These correspond to $l$=2, 2, 3 and 6, respectively.

Table 3. Degree of freedom of $Q_i$ and $Q_i^{(m)}$.

| Factor | Effective vector | Variance | Degree of freedom |
|---|---|---|---|
| Speaker in inter-class | $\mu^{(s)}$ | $Q_1$ | $l$-1 |
| Speaker in intra-class | $\alpha_i^{(m)}$ | $Q_2^{(m)}$ | dm-1 |
| Vowel | $\beta_j$ | $Q_3$ | 4 |
| Speaker in inter-class and vowel | $\tau_j^{(s)}$ | $Q_4$ | 4•($l$-1) |
| Speaker in intra-class and vowel | $\gamma_{ij}^{(m)}$ | $Q_5^{(m)}$ | 4•($d_m$-1) |

Table 4(a)~(d) show the values of $\nu'$ for each feature parameters. K. Tabata et al. said that the larger was the $\nu'$ of a factor, the easier was the classification on the factor. It is clear from the model that the main-effect of vowel is unchangeability for the every kinds of speaker-grouping ways. It should be noted that the magnitude order of $\nu'$ for $Q_1$ in 6 classes is nearly equal to that of F ratio in Table 2. In general, the interaction-effect of speaker-factor in intra-class and vowel-factor is not larger than the main-effect of vowel-factor or speaker-factor in intra-class. And also, the interaction-effect of speaker-factor in inter-class and vowel-factor is not larger than the main-effect of vowel-factor. However there is no such a relationship between the interaction-effect of speaker-factor in inter-class and vowel-factor, and the main-effect of speaker-factor in inter-class. There are summalized as follows. In general,

$$\nu'(Q_2{}^{(m)}),\ \nu'(Q_3)>\nu'(Q_5{}^{(m)});\ \ \nu'(Q_3)>\nu'(Q_4)$$

In almost all cases,

$$\nu'(Q_3)>\nu'(Q_1);\ \ \nu'(Q_4)>\nu'(Q_2{}^{(m)}),\ \nu'(Q_5{}^{(m)})$$

From these results of variance analyses as described above, if the speaker differences

Table 4. Results of variance analyses.

(a) 6 classes

$\nu'$ : normalized likelihood ratio criterion

$Q_1$ : main effect of class-factor

$Q_2{}^{(m)}$: main effect of speaker-factor for the m-th class

$Q_3$ : main effect of vowel factor

$Q_4$ : interaction effect of vowel-factor and class-factor

$Q_5{}^{(m)}$: interaction effect of vowel-factor and speaker-factor for the m-th class

class1 : M1, class 2: W1, class 3: M2, class 4: W2, class 5: M4, class 6: W4

| $\nu'$ / Parameter | $Q_1$ | $Q_2$ | | | | | | $Q_3$ | $Q_4$ | $Q_5$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | | | 1 | 2 | 3 | 4 | 5 | 6 |
| F0 | 211 | 19 | 14 | 9.3 | 12 | 5.1 | 22 | 27 | 5.0 | 0.8 | 1.2 | 0.5 | 0.8 | 2.2 | 0.9 |
| INCL | 92 | 7.9 | 4.5 | 6.6 | 7.3 | 7.8 | 13 | 169 | 4.9 | 1.5 | 1.2 | 1.6 | 2.0 | 2.0 | 1.6 |
| CD 2 | 16 | 5.6 | 2.7 | 5.0 | 5.1 | 5.1 | 4.0 | 89 | 4.5 | 1.0 | 0.9 | 1.3 | 1.7 | 2.2 | 2.9 |
| CD 3 | 20 | 5.3 | 4.0 | 6.4 | 4.6 | 7.2 | 5.4 | 157 | 9.9 | 2.2 | 2.3 | 1.9 | 2.5 | 2.0 | 1.7 |
| $\gamma_1$ | 32 | 9.8 | 4.6 | 8.0 | 8.1 | 7.0 | 6.9 | 142 | 9.3 | 2.0 | 1.3 | 2.9 | 2.0 | 1.6 | 2.3 |
| $\gamma_5$ | 40 | 2.8 | 2.7 | 4.0 | 2.7 | 4.1 | 2.7 | 115 | 11 | 2.3 | 2.7 | 1.5 | 2.7 | 1.6 | 2.2 |
| $\gamma_1'$ | 60 | 10 | 6.7 | 10 | 13 | 18 | 12 | 187 | 10 | 1.8 | 1.8 | 2.9 | 2.4 | 2.9 | 2.3 |
| $\gamma_2'$ | 19 | 5.5 | 3.1 | 5.8 | 6.6 | 6.4 | 4.7 | 184 | 5.9 | 2.1 | 1.8 | 2.0 | 2.8 | 2.8 | 2.7 |
| $\gamma_3'$ | 1.6 | 3.6 | 1.3 | 5.4 | 3.5 | 2.3 | 2.8 | 63 | 15 | 2.8 | 1.8 | 2.1 | 3.3 | 3.3 | 2.9 |
| $\gamma_4'$ | 6.4 | 1.5 | 2.3 | 6.8 | 2.2 | 7.8 | 3.8 | 176 | 14 | 2.4 | 2.1 | 3.2 | 2.1 | 2.1 | 2.6 |
| $\gamma_5'$ | 40 | 2.8 | 2.6 | 3.9 | 2.6 | 4.0 | 2.6 | 114 | 11 | 2.4 | 2.5 | 1.4 | 2.7 | 1.6 | 2.4 |
| $\gamma_6'$ | 22 | 5.7 | 3.7 | 3.3 | 4.1 | 3.8 | 1.9 | 7.8 | 25 | 1.9 | 1.6 | 2.1 | 2.4 | 2.9 | 1.9 |
| $\gamma_7'$ | 9.5 | 2.0 | 2.8 | 3.4 | 1.5 | 2.5 | 2.0 | 59 | 13 | 2.3 | 1.9 | 1.9 | 2.4 | 2.0 | 1.8 |
| $\gamma_8'$ | 35 | 2.3 | 3.1 | 8.2 | 5.0 | 4.7 | 1.1 | 57 | 13 | 2.1 | 2.6 | 2.5 | 2.6 | 2.4 | 2.5 |
| $\gamma_9'$ | 11 | 4.7 | 3.8 | 3.6 | 6.4 | 2.7 | 3.0 | 18 | 24 | 2.0 | 1.6 | 1.7 | 3.6 | 1.7 | 2.1 |
| $\gamma_{10}'$ | 49 | 2.0 | 2.9 | 2.7 | 2.3 | 3.4 | 2.8 | 1.4 | 21 | 2.6 | 2.0 | 2.0 | 2.4 | 2.4 | 2.9 |

on inter-speaker-groups are eliminated, almost all speaker differences are eliminated and the classification of vowels becomes easy. This elimination may be realized by an automatic speaker-grouping method.

Table 4 (b) 3 classes
class 1: M1•W1, class 2: M2•M4, class 3: W2·W4

| $\nu'$ / Parameter | $Q_1$ | $Q_2$ | | | $Q_3$ | $Q_4$ | $Q_5$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | | | 1 | 2 | 3 |
| F0 | 338 | 16 | 21 | 8.0 | 27 | 8.2 | 1.1 | 0.9 | 1.6 |
| INCL | 113 | 9.7 | 15 | 9.4 | 169 | 6.3 | 1.5 | 2.0 | 2.0 |
| CD 2 | 3.3 | 5.7 | 5.8 | 6.6 | 89 | 2.4 | 1.2 | 2.5 | 2.0 |
| CD 3 | 3.4 | 5.8 | 7.3 | 8.3 | 157 | 14 | 2.4 | 2.4 | 2.1 |
| $\gamma_1$ | 30 | 9.0 | 8.4 | 9.3 | 142 | 11 | 1.9 | 2.4 | 2.5 |
| $\gamma_5$ | 60 | 3.3 | 3.2 | 3.6 | 115 | 10 | 3.2 | 2.7 | 1.7 |
| $\gamma_1'$ | 64 | 10 | 14 | 17 | 187 | 11 | 2.0 | 2.6 | 3.2 |
| $\gamma_2'$ | 3.3 | 5.5 | 7.9 | 7.4 | 184 | 3.1 | 2.3 | 2.9 | 2.7 |
| $\gamma_3'$ | 0.7 | 3.1 | 3.8 | 4.3 | 63 | 19 | 2.8 | 3.3 | 3.1 |
| $\gamma_4'$ | 4.5 | 2.5 | 3.3 | 8.1 | 176 | 20 | 2.4 | 2.7 | 2.8 |
| $\gamma_5'$ | 59 | 3.3 | 3.1 | 5.1 | 114 | 9.6 | 3.2 | 2.7 | 1.6 |
| $\gamma_6'$ | 15 | 5.4 | 4.8 | 5.4 | 7.8 | 44 | 2.0 | 2.3 | 2.7 |
| $\gamma_7'$ | 2.8 | 3.6 | 2.2 | 4.0 | 59 | 20 | 2.5 | 2.2 | 2.0 |
| $\gamma_8'$ | 47 | 3.1 | 3.7 | 8.4 | 57 | 20 | 2.7 | 2.7 | 2.5 |
| $\gamma_9'$ | 8.8 | 6.5 | 5.3 | 3.5 | 18 | 37 | 2.7 | 3.0 | 1.9 |
| $\gamma_{10}'$ | 78 | 3.1 | 3.2 | 3.4 | 1.4 | 36 | 2.4 | 3.0 | 2.3 |

Table 4 (c) 2′ classes
class 1: M1•M2•W2•W4, class 2: M2•M4

| $\nu'$ / Parameter | $Q_1$ | $Q_2$ | | $Q_3$ | $Q_4$ | $Q_5$ | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | | | 1 | 2 |
| F0 | 450 | 19 | 8.0 | 27 | 12 | 0.2 | 0.2 |
| INCL | 155 | 12 | 9.4 | 169 | 4.0 | 0.4 | 0.6 |
| CD 2 | 4.1 | 5.9 | 6.6 | 89 | 2.8 | 0.2 | 0.2 |
| CD 3 | 2.2 | 6.7 | 8.3 | 157 | 19 | 0.5 | 0.7 |
| $\gamma_1$ | 28 | 8.9 | 9.3 | 142 | 16 | 0.3 | 0.4 |
| $\gamma_5$ | 80 | 5.2 | 3.6 | 115 | 8.9 | 0.5 | 0.4 |
| $\gamma_1'$ | 56 | 12 | 17 | 187 | 13 | 0.4 | 0.5 |
| $\gamma_2'$ | 2.6 | 6.8 | 7.4 | 184 | 2.2 | 0.4 | 0.5 |
| $\gamma_3'$ | 0.4 | 3.6 | 4.3 | 63 | 28 | 0.8 | 0.5 |
| $\gamma_4'$ | 5.4 | 3.1 | 8.1 | 176 | 25 | 0.4 | 0.8 |
| $\gamma_5'$ | 80 | 3.6 | 5.1 | 114 | 8.6 | 0.9 | 0.4 |
| $\gamma_6'$ | 20 | 5.2 | 5.4 | 7.8 | 60 | 0.7 | 0.6 |
| $\gamma_7'$ | 2.9 | 3.1 | 4.0 | 59 | 24 | 0.5 | 0.6 |
| $\gamma_8'$ | 55 | 4.2 | 8.4 | 57 | 25 | 0.6 | 0.6 |
| $\gamma_9,$ | 0.4 | 6.1 | 3.5 | 18 | 45 | 0.8 | 0.3 |
| $\gamma_{10}'$ | 108 | 3.4 | 3.4 | 1.4 | 50 | 0.5 | 0.3 |

Table 4 (d) 2 classes
class 1: M1•W1•W2•W4, class 2: M2•M4

| $\nu'$ / Parameter | $Q_1$ | $Q_2$ | | $Q_3$ | $Q_4$ | $Q_5$ | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | | | 1 | 2 |
| F0 | 267 | 21 | 33 | 27 | 6.0 | 1.7 | 2.5 |
| INCL | 37 | 12 | 17 | 169 | 2.5 | 2.9 | 3.1 |
| CD 2 | 1.0 | 5.2 | 7.2 | 89 | 4.3 | 3.2 | 2.7 |
| CD 3 | 4.3 | 6.9 | 7.5 | 157 | 12 | 3.8 | 3.7 |
| $\gamma_1$ | 0.04 | 7.4 | 12 | 142 | 12 | 3.3 | 3.9 |
| $\gamma_5$ | 27 | 3.3 | 3.7 | 115 | 7.2 | 4.8 | 3.2 |
| $\gamma_1'$ | 0.004 | 12 | 17 | 187 | 13 | 3.8 | 4.4 |
| $\gamma_2'$ | 5.3 | 7.0 | 7.0 | 184 | 1.1 | 4.1 | 4.0 |
| $\gamma_3'$ | 0.4 | 3.4 | 4.4 | 63 | 26 | 4.5 | 4.8 |
| $\gamma_4'$ | 0.9 | 3.4 | 6.5 | 176 | 11 | 4.1 | 4.9 |
| $\gamma_5'$ | 25 | 3.2 | 8.5 | 114 | 6.9 | 4.8 | 3.3 |
| $\gamma_6'$ | 13 | 4.8 | 6.2 | 7.8 | 35 | 3.8 | 5.5 |
| $\gamma_7'$ | 6.4 | 3.1 | 3.6 | 59 | 14 | 3.8 | 4.1 |
| $\gamma_8'$ | 21 | 4.5 | 3.7 | 57 | 9.8 | 4.5 | 4.6 |
| $\gamma_9'$ | 2.2 | 6.0 | 4.6 | 18 | 15 | 5.0 | 5.2 |
| $\gamma_{10}'$ | 43 | 3.3 | 9.5 | 1.4 | 26 | 4.3 | 5.3 |

Next, let us consider about the kind of speaker-grouping way. If an input utterance is given and the group which this speaker belongs to is known, then the residual of speaker differences is the main-effect of speaker-factor in intra-class and the interaction of this and vowel-factor. The residual on 2′, 3 and 6 classes is nearly equal, except for 2 classes. Therefore, the larger is the main-effect of speaker-factor in inter-class, the easier becomes the classification of vowels. Furthermore, we can say approximately that a significant feature parameter for the classification of speaker-class is a parameter on which the difference between the main-effect of speaker-factor in inter-class and intra-class is large. Such parameters are pitch frequency, inclination of spectrum, $\gamma_{10}'$ and $\gamma_1'$. Let us put the kinds of speaker-grouping ways in decreasing order of the above difference for these parameters.

(1) pitch frequency
2′ classes, 3 classes, 6 classes, 2 classes

(2) inclination of spectrum
2′ classes, 3 classes, 6 classes, 2 classes

(3) $\gamma_{10}'$
2′ classes, 3 classes, 6 classes, 2 classes

(4) $\gamma_1'$
3 classes, 6 classes, 2′ classes, 2 classes

Thus, we can say that 2′ classes or 3 classes is better as a speaker-grouping way for speaker-independent vowel classification. The significant feature parameters for vowel recognition are INCL, CD3, $\gamma_1'$, $\gamma_2'$, $\gamma_4'$, $\gamma_5'$ and so on.

## V. Experilent of Vowel Recognition on the Basis of Speraker-Grouping

In this sections, we propose a two-step method for speaker-independent recognition. This method is based on the results of variance analyses.

*⟨First step⟩*

For an input utterance, the speaker is classified into one of $l$ speaker-classes. This operation is to eliminate the speaker differences in inter-speaker class caused by the difference of sex and age.

*⟨Second step⟩*

The input utterance is classified into one of five vowels by using the classification procedure which corresponds to the classified speaker-class.

In this section, we test this two-step method by vowel recognition experiments and consider the best way of speaker-grouping. We divide speech samples into training samples and test samples as follows.

*All data*: speech data of every speakers for six groups (total of 120 speakers).

*Training data*: speech data of 15 speakers for each group (total of 90 speakers).

*Test data*: speech data of 5 speakers for each group (total of 30 speakers).

First of all, the mean vector ($X_l$) and covariance matrix ($\Sigma_l$) of feature vector $X_{ijk}$ are calculated for each speaker class ($l$) by using all data or training data. And also, the mean vector ($X_{li}$) and covariance matrix ($\Sigma_{li}$) are calculated for each vowel (i) of each speaker class ($l$). The recognition is performed on the basis of Mahalanobis' generalized distance, that is, for a given sample X, X is decided as belonging to speaker-class $l$ which minimizes $(X-X_l)\Sigma_l^{-1}(X-X_l)'$, $l=1, 2, \cdots l_1$, and then X is decided as vowel i which minimizes $(X-X_{li})\Sigma_{li}^{-1}(X-X_{li})'$, i=/a/, /i/, /u/, /e/, /o/. We experimented on the following vowel recognition.

*Experiment* **1**: Pitch frequency, inclination of spectrum and $\gamma_1'\sim\gamma_{10}'$ are used as feature parameters for the classification of speaker classes and only $\gamma_1'\sim\gamma_{10}'$ are used for the classification of vowels. The mean vector and covariance matrix are calculated by all data and applied to all the data.

*Experiment* **2**: The three or six reference patterns (mean vector, covariance matrix) are prepared for each vowel. These patterns correspond to the reference patterns for 3 classes or 6 classes in Experiment 1. In this experiment, the classification of speaker classes is not performed explicitly, that is, the number of speaker classes is regarded as one class.

*Experiment* **3**: The same experiment as Experiment 1 except for the use of training data and applying to test data.

*Experiment* **4**: The same experiment as Experiment 2 except for the use of training data and applying to test data.

Following four kinds as speaker classes were tested.

1 class: no classification of speaker-classes.

2 classes: M1 & M2 & M4/W1 & W2 & W4

2′ classes: M1 & W1 & W2 & W4/M2 & M4

Table 5. Experimental results of vowel recognition.

(a) all data

| | Speaker class | | | |
|---|---|---|---|---|
| | Unknown | | Known | |
| | Recognition rate (%) | | | |
| | class | vowel | class | vowel |
| 1 class | — | — | — | 92.6 |
| 2 classes | 72.4 | 92.5 | 100 | 94.7 |
| 2′ classes | 94.6 | 94.6 | 100 | 95.7 |
| 3 classes | 80.7 | 94.4 | 100 | 96.4 |
| 6 classes | 63.8 | 94.7 | 100 | 98.1 |
| 1 class | 3 reference patterns for each vowel | | | 94.1 |
| 1 class | 6 reference patterns for each vowel | | | 95.6 |

(b) test data

| | speaker class | | | |
|---|---|---|---|---|
| | Unknown | | Known | |
| | Recognition rate (%) | | | |
| | class | vowel | class | vowel |
| 1 class | — | — | — | 88.9 |
| 2 classes | 70.7 | 90.0 | 100 | 89.8 |
| 2′ classes | 93.8 | 90.7 | 100 | 90.4 |
| 3 classes | 79.3 | 91.1 | 100 | 92.2 |
| 6 classes | 56.4 | *88.0 | 100 | *89.6 |
| 1 class | 3 reference patterns for each vowel | | | 88.2 |
| 1 class | 6 reference patterns for each vowel | | | *88.0 |

(c) confusion matrix of classification of six speaker classes

all data (63.8%)

| in \ out | M1 | W1 | M2 | W2 | M4 | W4 |
|---|---|---|---|---|---|---|
| M1 | 211 | 23 | 1 | 30 | 0 | 35 |
| W1 | 72 | 163 | 0 | 35 | 1 | 29 |
| M2 | 0 | 0 | 206 | 3 | 77 | 14 |
| W2 | 61 | 14 | 0 | 174 | 2 | 49 |
| M4 | 0 | 1 | 59 | 3 | 201 | 36 |
| W4 | 27 | 8 | 10 | 37 | 25 | 193 |

test data (56.4%)

| in \ out | M1 | W1 | M2 | W2 | M4 | W4 |
|---|---|---|---|---|---|---|
| M1 | 36 | 19 | 1 | 15 | 0 | 4 |
| W1 | 11 | 46 | 0 | 13 | 0 | 5 |
| M2 | 0 | 0 | 34 | 0 | 41 | 0 |
| W2 | 11 | 8 | 0 | 38 | 0 | 18 |
| M4 | 0 | 0 | 9 | 1 | 54 | 11 |
| W4 | 3 | 0 | 1 | 10 | 15 | 46 |

3 classes: M1 & W1/M2 & M4/W2 & W4
6 classes: M1/W1/M2/W2/M4/W4

Table 5 shows the recognition results. We can conclude from these experimental results that the two-step recognition procedure is better than one step, that is, the classification of speaker classes in terms of sex and age is effective for speaker-independent vowel recognition and that the 3 classes are the best kind as speaker-grouping ways. (However, we should not that the number of training samples for reference patterns is not sufficient on six classes, in particular, * in Table 5.)

## VI. Conclusion

In this paper, we investigated the speaker differences in feature parameters of Japanese vowels with sex and age. First of all, we analyzed pitch and first three formant frequencies, and found that there was a relationship between pitch and formant in inter-speaker group but no relationship in intra-speaker group. The relationship among formant frequencies depends on sex, age, vowel and formant, and it is not simple but complex. This shows that there does not exist a simple speaker normalization method by using pitch and formant frequencies. We found there were three speaker groups: (males), (old females), (young females, children) with according to pitch frequency; (males), (old females), (children) with according to formant frequency.

Next, we tested statistically how it was effective to eliminate speaker differences in the case that the speaker group of an unknown speaker was given. For this purpose, we extended Tabata's two divided type of model for analysis of variance to multi divided type, and we showed statistically the existence of inter-speaker group differences.

Finally, we experimented on Japanese vowel recognition on the basis of the results of analyses, and concluded that the classification of speaker classes in terms of sex and age was effective for speaker-independent vowel recognition and the 3 classes (male, female, children) were the best kind as speaker-grouping ways.

In this paper, the 3′ classes (male, old female, young female & children; M2 & M4/W4/W2 & M1 & W1) were not analyzed and investigated. The remaining works are the variance analyses of formant frequencies, other feature parameters (cepstrum, vocal tract length etc.) and 3′ classes, and correlation analyses between feature parameters. We must find out speaker independent feature parameters or speaker normalization procedures on the basis of these analyses.

## References

1) L. J. Gertman: Classification of Self-Normalized Vowels, IEEE Trans. Vol. AU-16, No. 1 (1968).

2) R. M. Schwartz: Automatic Normalization for Recognition of Vowels of All Speakers, B. S. Thesis, MIT (1971).

3) H. Fujisaki, Y. Katagiri and Y. Sato: Feature Extraction and Automatic Recognition of Sustained Vowels Uttered by a Number of Unknown Speakers, ASJ, S77–08 (1977, in Japanese)

4) H. Wakita: Normalization of Vowels by Vocal-Tract Length and Its Application to Vowel Identification, IEEE Trans. Vol. ASSP-25, No. 2 (1977).

5) S. Ishizaki: Vocal Tract Length Estimation by Use of Statistical Method, IECEJ Trans. Vol. 61-E, No. 5 (1978).

6) H. Kasuya, H. Suzuki and K. Kido: Changes in Pitch and First Three Formant Frequencies of Five Japanese Vowels with Age and Sex of Speakers, JASJ, Vol. 24, No. 6 (1968, in Japanese)

7) G. Fant: Non-Uniform Vowel Normalization, Speech Transmission Lab. QPSR, Vol. 2–3 (1975).

8) R. D. Kent and L. L. Former: Development Study of Vowel Formant Frequencies in an Imitation Task, JASA, Vol. 65, No. 1 (1979).

9) S. F. Disner: Evaluation of Vowel Normalization Procedures, JASA, Vol. 67, No. 1 (1980).

10) G. Ooyama, S. Katagiri and K. Kido: Cepstrum Analysis Suppressing Pitch Component with Comb Lifter, ASJ, S77–17 (1977, in Japanese).

11) J. D. Markel and A. H. Gray: Linear Predictive of Speech, Springer-Verlag (1976).

12) S. Furui: Studies on Speaker Characteristics in Speech Waves, Docter thesis, University of Tokyo (1978, in Japanese).

13) T. Nakajima et al.: Estimation of Vocal Tract Area Function by Adaptive Inverse Filtering Methods, ASJ, S72 (1973, in Japanese).

14) R. K. Potter and J. C. Steinberg: Toward the Specification of Speech, JASA, Vol. 22 (1950).

15) G. E. Peterson and H. L. Barney: Control Method Used in a Study of the Vowels, JASA, Vol. 24 (1952).

16) G. Fant: Acoustic Analysis and Synthesis of Speech with Application to Swedish, Ericsson Tech., Vol. 15, No. 1 (1959).

17) T. Sakai and K. Tabata: Multivariate Statistical Analysis of VCV Syllables, IECEJ Trans. Vol. 56–D, No. 1 (1973, in Japanese).

18) K. Tabata: A Divided Type of Model for Multivariate Analysis of Variance and Speech Sounds, Koudou Keiryogaku, Vol. 1, No. 1 (1974, in Japanese).

19) M. Yamao: Considerations of Speaker Clustering Method for Recognition Speech Uttered by Unlimited Speakers, Master thesis, Kyoto University (1980, in Japanese).

20) S. Nakagawa et al.: Considerations on Speaker Grouping by Sex and Age for Automatic Speech Recognition, IECEJ Trans. (in Japanese, to appear)

21) W. A. Lea: Prosodic Aids to Speech Recognition, in Trends in Speech Recognition, edited by W.A. Lea, Prentice-Holl (1980).

(Aug. 31, 1980, received)