International Journal of Advanced Intelligence
Volume 1, Number 1, pp.59-88, November, 2009.

AIA
International
Advanced
Information
Institute

New Similarity Scale to Measure the Difference in Like Patterns with Noise

Michihiro Jinnai
Department of Electro-Mechanical Systems Engineering,
Kagawa National College of Technology
355 Chokushi-cho, Takamatsu, 761-8058, Japan
jinnai@t.kagawa-nct.ac.jp

Satoru Tsuge
Faculty of Engineering, University of Tokushima
2-1 Minami-josanjima, Tokushima, 770-8506, Japan
tsuge@is.tokushima-u.ac.jp

Shingo Kuroiwa
Department of Information and Image Science, Chiba University
1-33 Yayoi-cho Inage-ku, Chiba, 263-8522, Japan
kuroiwa@faculty.chiba-u.jp

Fuji Ren
Faculty of Engineering, University of Tokushima
2-1 Minami-josanjima, Tokushima, 770-8506, Japan
ren@is.tokushima-u.ac.jp

Minoru Fukumi
Faculty of Engineering, University of Tokushima
2-1 Minami-josanjima, Tokushima, 770-8506, Japan
fukumi@is.tokushima-u.ac.jp

Received (January 2009)
Revised (August 2009)

A new similarity scale called the Geometric Distance, that numerically evaluates the degree of likeness between two patterns is proposed. Traditionally, the similarity scales known as the Euclidean distance and cosine similarity have been widely used to measure likeness. Traditional methods do not perform well in the presence of noise or pattern distortions. In this paper, a new mathematical model for a similarity scale is proposed which overcomes these limitations of the earlier models, while improving the overall recognition accuracy. Experiments in speech vowel recognition were carried out under various SNR levels in a variety of noisy environments. In all cases a significant improvement in recognition accuracy is demonstrated, with the improvement most pronounced in the noisiest conditions. In fact, at a SNR of 5 dB in a subway, the recognition accuracy improved from 65% to 75% and at 20 dB SNR from 98.4% to 99.6% over the MFCC method. Numerical modeling of simple patterns is used to demonstrate the principles behind the Geometric Distance.

Keywords: Similarity measures; Distance functions; Pattern matching; Noise robust.

1. Introduction

In pattern recognition, a known pattern stored in a PC memory is called as the "standard pattern", and a pattern to be compared is called the "input pattern". The degree of likeness between the standard pattern and the input pattern is evaluated using a similarity scale. If the similarity of the standard and input patterns is close, then those two patterns are considered to be in the same category and the input pattern is recognized. The similarity is often measured as a "distance" between the two patterns.

Conventionally, the similarity scales known as the Euclidean distance and cosine similarity have been widely used.[1,2] Conventional similarity scales compare the patterns using a one-to-one mapping. The result of the one-to-one mapping is that, the distance metric is highly sensitive to noise, and the distance metric changes in a staircase pattern when a difference occurs between peaks of the standard and input patterns.

To improve the shortcomings, various techniques have been applied. For example, in speech recognition, the Itakura-Saito distance measure,[2,3,4] LLR,[5] WLR,[6,7] WSM,[8] and projection distance[9] have been proposed for the purpose of comparing the shapes of the power spectra.[10] Besides, in pattern classification or clustering and image retrieval, many distance functions have been proposed for comparing histograms.[11,12,13,14,15]

A similarity scale is a concept that should intuitively concur with the human concept of similarity in hearing and sight. First we need to develop a mathematical model for the similarity scale so that we can perform numerical processing by computer. In this paper, a mathematical model of the similarity scale is proposed to improve the shortcomings that are found in the Euclidean distance, cosine similarity and others, and a new algorithm based on a one-to-many point mapping is proposed to realize the mathematical model. Then, numerical experiments are carried out using some geometric patterns, and the algorithm is confirmed to perform well. Finally, some speech recognition tests are carried out using the proposed algorithm with real voices. The effectiveness of the mathematical model and algorithm is evaluated based on the result of speech recognition.

A mathematical model incorporating the following two characteristics is used.
<1> The distance metric must show good immunity to noise.
<2> The distance metric must increase monotonically when a difference increases between peaks of the standard and input patterns.

The proposed similarity scale can be applied widely to pattern recognition such as pattern classification or clustering and image retrieval using the distance between histograms. This paper explains this technique using power spectrum patterns of voice. The paper consists of the following sections. Section 2 describes the shortcomings that are found in the conventional similarity scales. Section 3 describes the mathematical model and algorithm of the new similarity scale, describes numerical experiments, and describes that the algorithm performs well. Section 4

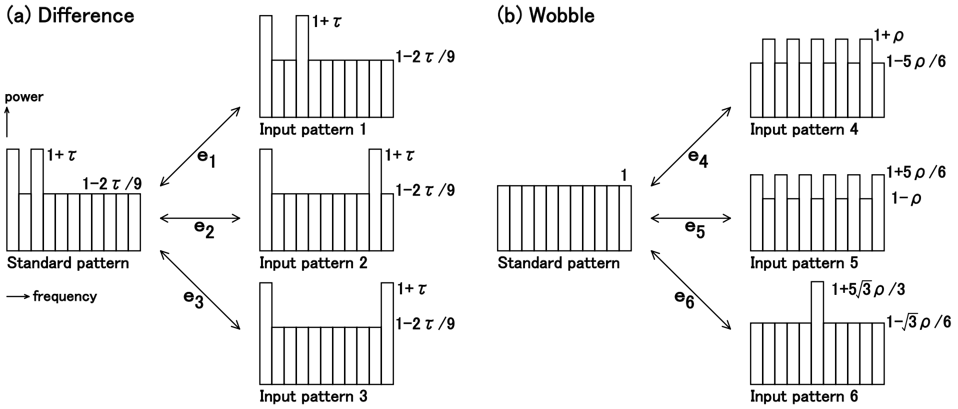


Fig. 1. Typical examples of standard and input patterns.

describes the speech recognition tests that have been carried out, and describes the effectiveness of the mathematical model and algorithm. Section 5 provides the conclusions and touches on future work.

2. Conventional Similarity Scale

In this paper, for example, for the power spectrum of voice, a random variation of power spectrum caused by noise and air turbulence such as fricative sound is defined as the "wobble". Also, the difference between peaks of the power spectra such as formant is defined as the "difference".

Conventional similarity scales Euclidean distance and cosine similarity compare the patterns using a one-to-one mapping. The result of the one-to-one mapping is that, input patterns with different shapes may have the same distance from the standard pattern when the power spectrum patterns have the "difference" and "wobble".

Fig. 1(a) gives an example of the "difference" where the standard pattern has two peaks in the power spectrum, and input patterns 1, 2 and 3 have a different position on the second peak. However, each pattern is assumed to have variable τ in the relationship shown in Fig. 1(a). Therefore, the standard pattern and the input patterns always have the same area. In this case, the Euclidean distance and cosine similarity e_1, e_2 and e_3 have the relationship of $e_1 = e_2 = e_3$ between the standard pattern and each of input patterns 1, 2 and 3. Therefore, input patterns 1, 2 and 3 cannot be distinguished.

Fig. 1(b) gives an example of the "wobble" where the standard pattern has a flat power spectrum, input patterns 4 and 5 have the "wobble" on the flat power spectrum, and input pattern 6 has a single peak. However, each pattern is assumed to have variable ρ in the relationship shown in Fig. 1(b). Therefore, the standard pattern and the input patterns always have the same area. In this case, the Euclidean distance and cosine similarity e_4, e_5 and e_6 have the relationship of

$e_4 = e_5 = e_6$ between the standard pattern and each of input patterns 4, 5 and 6. Therefore, input patterns 4, 5 and 6 cannot be distinguished.

To deal with these shortcomings, the cepstrum is used as the feature parameter in the speech recognition, for example.[16] The cepstrum is a result of taking the Inverse Fourier transform of the logarithmic power spectrum. In particular, the Mel-Frequency Cepstrum Coefficient (MFCC),[17] which is a combination of this cepstrum and Mel filter bank, is used in many speech recognition systems.[18] Although this MFCC is the feature parameter that can absorb a certain level of "difference" and "wobble" of the power spectrum, the remaining "difference" and "wobble" are finally absorbed using statistical models and adaptation techniques.[19,20] Insufficient attention has been paid to date to the role of the similarity scale in both speech and non-speech sound recognition. Therefore, we propose a new similarity scale that we will introduce in the next section.

3. New Similarity Scale

A new algorithm based on a one-to-many point mapping is proposed to realize the mathematical model. The difference in shapes between standard and input patterns is replaced by the shape change of a normal distribution, and the magnitude of this shape change is numerically evaluated as a variable of the moment ratio that is derived from the kurtosis. In this method, when a "difference" occurs between peaks of the standard and input patterns with "wobble" due to noise or similar occurrence, the "wobble" is absorbed and the distance metric increases monotonically according to the increase of the "difference". In the second half of this section, numerical experiments are carried out using some geometric patterns with the "difference" and "wobble", and the proposed algorithm is confirmed to perform well.

3.1. *Normal distribution and kurtosis*

In statistical analysis, the normal distribution shown in the following equation is often used for models exhibiting many phenomena.

$$f(u) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right\} \tag{1}$$

Where, μ is mean, and σ^2 is variance. When the normal distribution is applied to a model exhibiting phenomenon, it is important to check whether the phenomenon meets the normal distribution or not. The kurtosis of a probability distribution is a measure of its relative peakedness or flatness compared to the normal distribution. A positive kurtosis indicates peakedness and a negative one, flatness relative to the normal distribution with the same mean and variance. In Eq. (1), if the continuous value u is replaced by discrete value u_i, kurtosis a can be calculated

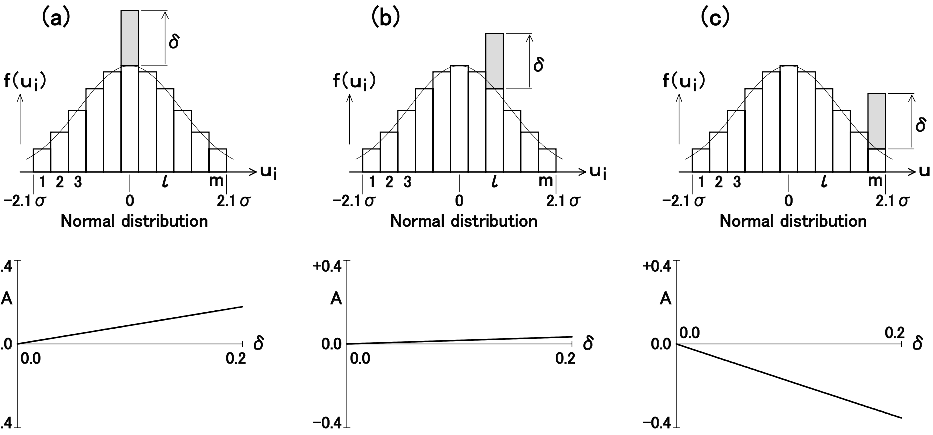


Fig. 2. Change of moment ratio A.

using the following equation.

$$a = \frac{\left\{\sum_i f(u_i)\right\} \cdot \left\{\sum_i (u_i - \mu)^4 \cdot f(u_i)\right\}}{\left\{\sum_i (u_i - \mu)^2 \cdot f(u_i)\right\}^2} - 3 \tag{2}$$

If a probability distribution of the phenomenon follows the normal distribution, then $a = 0$. If it has peakedness relative to the normal distribution, then $a > 0$. Adversely, if it has flatness relative to the normal distribution, then $a < 0$. Eq. (2) shows a ratio of the forth moment to the square of second moment around mean μ. When the proposed method is used, a shape change around the component position needs to be detected based on the center of each component position of the power spectrum as shown in Fig. 7 of Section 3.7. Therefore, we assume $\mu = 0$ and change Eq. (2) as follows.

$$A = \frac{\left\{\sum_i f(u_i)\right\} \cdot \left\{\sum_i (u_i)^4 \cdot f(u_i)\right\}}{\left\{\sum_i (u_i)^2 \cdot f(u_i)\right\}^2} - 3 \tag{3}$$

Eq. (3) shows a ratio of the forth moment to the square of second moment around the origin. In this paper, Eq. (3) is called "Moment ratio A".

Then, numerical experiments are carried out to study the relationship between moment ratio A and the increment value δ of bar graphs seen in Figs. 2 and 3. Graphs in the upper side of Figs. 2(a)–(c) show the bar graphs each having m bars whose height is the same as function value $f(u_i)$ of the normal distribution. Note that $m = 11$ and the bar graphs are created by using the area of $-2.1\sigma \leq$

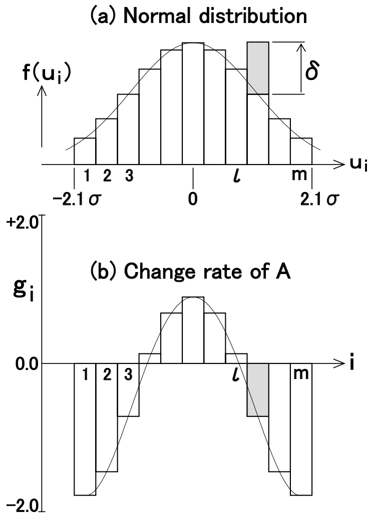


Fig. 3. Change rate of moment ratio A.

$u_i \leq 2.1\sigma$ ($\sigma = 1$) of the normal distribution. On these bar graphs, only a single bar increases by value δ in the center, an intermediate position, and an end of the normal distribution. Here, the moment ratio A is calculated using Eq. (3) for the bar graph whose shape is changed as described above. The obtained relationship between values A and δ is shown in the lower side of Figs. 2(a)–(c). For now we only consider positive values of δ. From these graphs, it is discovered that $A = 0.0$ if $\delta = 0.0$. Also, A changes approximately linearly when value δ increases. Note that if only a single bar increases by value δ in the graph with m bars, it is the same as when only a single bar with an $1/m$ ratio increases by value δ. If value m changes (m is an odd numbered value), the gradient of moment ratio graphs in the lower side of Figs. 2(a)–(c) changes by the same $1/m$ weight. This property holds for all values of m and for any variance σ^2 of the normal distribution.

Figs. 3(a) and (b) show the change rate of A (g_i, where a change of δ occurs at the i-th position) for a normal distribution and a single instance of δ. Change rate g_i is described by the following equation.

$$\mathrm{g}_i = A/\delta \qquad (i = 1, 2, 3, \cdots, m) \tag{4}$$

The $\mathrm{g}_{(1+m)/2}$, g_l and g_m are equal to the gradients of respective graphs shown in the lower side of Figs. 2(a)–(c). Next, in Fig. 3(a), position i of the bar that has increased by value δ is scanned from 1 to m, and Eq. (4) is calculated. Fig. 3(b) shows a bar graph of the calculated value g_i, where $\delta = 0.2$. From Fig. 3(b), $\mathrm{g}_i > 0$, $\mathrm{g}_i \approx 0$ and $\mathrm{g}_i < 0$ are found in the center, an intermediate position, and an end of the normal distribution.

The following summarizes the features of moment ratio A that have been obtained from the above numerical experiments. Fig. 4 shows a normal curve $f(u_i)$

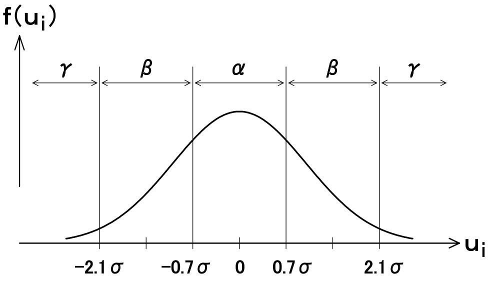


Fig. 4. Normal curve.

Table 1. Features of moment ratio A.

Fig. 4	α	Boundary area between α and β	β	γ
Increase of $f(u_i)$	$A > 0$	$A \approx 0$	$A < 0$	—

with mean $\mu = 0$ and variance σ^2, and the moment ratio becomes $A = 0$. Also, if the value $f(u_i)$ exceeds the value of the normal curve in area α shown in Fig. 4, the moment ratio becomes $A > 0$. If the value $f(u_i)$ increases in area β, the moment ratio becomes $A < 0$. If the value $f(u_i)$ increases in the boundary area between α and β (close to area $u_i = \pm 0.7\sigma$), the change of A is small and it is $A \approx 0$. Meanwhile, if the value $f(u_i)$ increases in area γ, A is unstable as it becomes greater than or less than 0. They have been summarized on Table 1. This paper uses area $-2.1\sigma \leq u_i \leq 2.1\sigma$ to obtain stable value A.

3.2. *Creation of standard and input pattern vectors*

An example of standard and input patterns, that have been created using the power spectrum of standard and input voices, are given in Figs. 5(a) and (b). Note that the power spectrum is generated from the output of filter bank with the m frequency bands (where, m is an odd number). Also, we suppose that the i-th power spectrum values (where, $i = 1, 2, \cdots, m$) of standard and input voices are divided by their total energy and normalized power spectra s_i and x_i have been calculated. At this moment, the standard and input patterns have the same area size. Here, we create a standard pattern vector $\boldsymbol{s}$ having s_i components, and an input pattern vector $\boldsymbol{x}$ having x_i components, and represent them as follows.

$$\begin{aligned} \boldsymbol{s} &= (s_1\,, s_2, \cdots, s_i, \cdots, s_m\,)^T \\ \boldsymbol{x} &= (x_1, x_2, \cdots, x_i, \cdots, x_m)^T \end{aligned} \tag{5}$$

Eq. (5) expresses the shapes of the power spectra of the standard voice and input voice by the m pieces of component values of the pattern vector respectively. Note that in this paper the width of each bar graph is $1/m$ for standard and input patterns shown in Figs. 5(a) and (b).

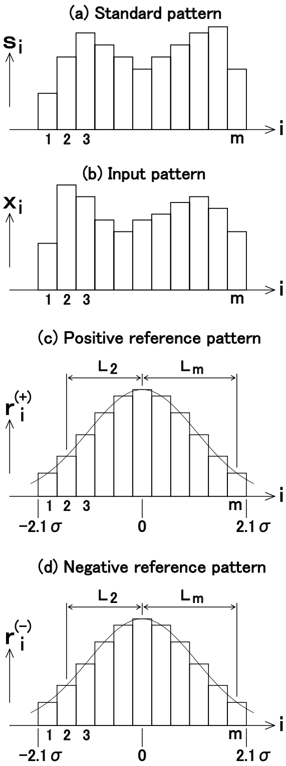


Fig. 5. Shape expression of pattern vectors.

3.3. *Creation of reference pattern vectors*

With the proposed algorithm, the difference in shapes between standard and input patterns is replaced by the shape change of the normal distribution, and the magnitude of this shape change is numerically evaluated as a variable of the moment ratio. However, in general, Eq. (3) cannot be defined if the value $f(u_i)$ is negative. Therefore, we create a pair of reference patterns that have the initial shape of a normal distribution so that the change of the value $f(u_i)$ does not decrease. Figs. 5(c) and (d) show the bar graphs, each having the same height as function values $r_i^{(+)}$ and $r_i^{(-)}$ of their normal distribution. Here, we create a positive reference pattern vector $\boldsymbol{r}^{(+)}$ having $r_i^{(+)}$ components, and a negative reference pattern vector $\boldsymbol{r}^{(-)}$ having $r_i^{(-)}$ components, and represent them as follows.

$$\begin{aligned}\boldsymbol{r}^{(+)} &= (r_1^{(+)}, r_2^{(+)}, \cdots, r_i^{(+)}, \cdots, r_m^{(+)})^T \\ \boldsymbol{r}^{(-)} &= (r_1^{(-)}, r_2^{(-)}, \cdots, r_i^{(-)}, \cdots, r_m^{(-)})^T\end{aligned} \tag{6}$$

$\boldsymbol{r}^{(+)}$ and $\boldsymbol{r}^{(-)}$ are equivalent vectors. Eq. (6) expresses the shape of a normal distribution by the m pieces of component values of pattern vector respectively. Note that the number of components of Eq. (6) is supposed to be equal to the

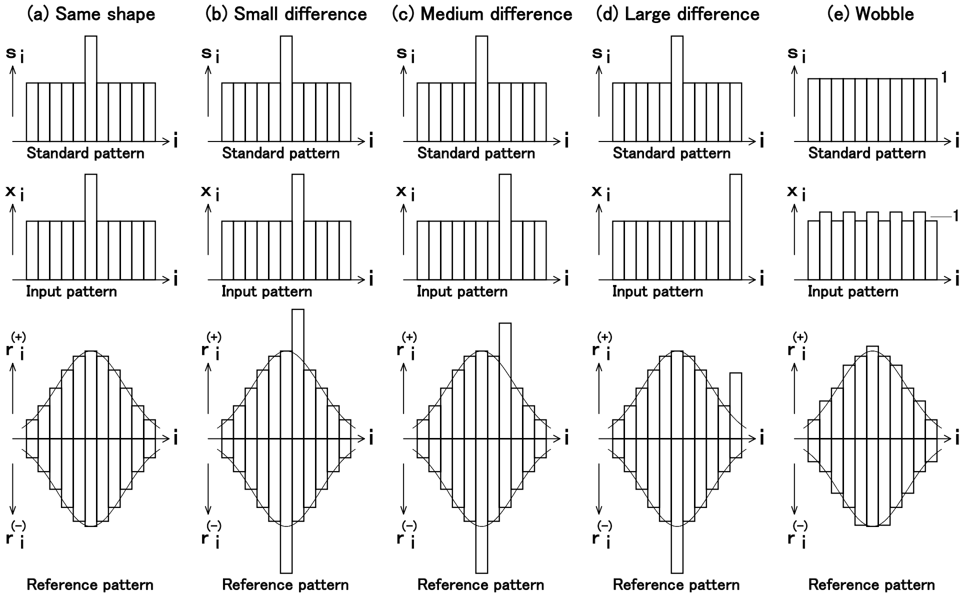


Fig. 6. Shape changes of reference patterns.

Table 2. Relationship between shape variation and shape changes of reference patterns.

Fig. 6	(a)	(b)	(c)	(d)	(e)
Increase of $r_i^{(+)}$	$A^{(+)}=0$	$A^{(+)}>0$	$A^{(+)}\approx 0$	$A^{(+)}<0$	$A^{(+)}\approx 0$
Increase of $r_i^{(-)}$	$A^{(-)}=0$	$A^{(-)}>0$	$A^{(-)}>0$	$A^{(-)}>0$	$A^{(-)}\approx 0$
$A^{(+)}-A^{(-)}$	$D=0$	$D\approx 0$	$D<0$	$D\ll 0$	$D\approx 0$

number of components of Eq. (5), and all bar graphs of Figs. 5(a)–(d) have the same width. Also, as shown in Figs. 5(c) and (d), the center axis of a normal distribution assumes to locate at the center of standard and input patterns, and Eq. (6) is created using area $-2.1\sigma \leq u_i \leq 2.1\sigma$ of the normal distribution. Note that $\sigma = 1/4.2$ as $2.1\sigma \times 2 = (1/m) \times m$.

3.4. *Shape changes of reference pattern vectors*

A difference in shapes between standard pattern vector $\boldsymbol{s}$ and input pattern vector $\boldsymbol{x}$ is replaced by the shape changes of positive reference pattern vector $\boldsymbol{r}^{(+)}$ and negative reference pattern vector $\boldsymbol{r}^{(-)}$ using the following equation.

$$\begin{aligned}&For\ \ i = 1, 2, 3, \cdots, m\,;\\ &\bullet\ \ if\ \ x_i > s_i,\ \ then\ \ r_i^{(+)} \longleftarrow r_i^{(+)} + |x_i - s_i|\\ &\bullet\ \ if\ \ x_i < s_i,\ \ then\ \ r_i^{(-)} \longleftarrow r_i^{(-)} + |x_i - s_i|\end{aligned} \tag{7}$$

In Eq. (7), $r_i^{(+)}$ and $r_i^{(-)}$ on the right side show the component values of positive and negative reference pattern vectors having the shape of the normal distribution, and those on the left side show the components after the shape has changed. In Eq. (7), if component value x_i of the input pattern vector is greater than component value s_i of the standard pattern vector, component value $r_i^{(+)}$ of the positive reference pattern vector increases by $|x_i - s_i|$ from the normal distribution value. Also, if x_i is smaller than s_i, component value $r_i^{(-)}$ of the negative reference pattern vector increases by $|x_i - s_i|$ from the normal distribution value. Thus, the values $r_i^{(+)}$ and $r_i^{(-)}$ do not decrease in Eq. (7). Fig. 6 shows the shape of Eq. (7). However, $\boldsymbol{r}^{(-)}$ is shown upside down in order to compare it with $\boldsymbol{r}^{(+)}$. Next, we explain Eq. (7) using Fig. 6.

- Fig. 6(a) gives an example of the case where standard pattern and input pattern have the same shape. Because values $r_i^{(+)}$ and $r_i^{(-)}$ of Eq. (7) do not change during this time, a pair of the reference patterns shown in Fig. 6(a) do not change in their shapes from the normal distribution.
- Figs. 6(b)–(d) respectively show an example exhibiting a small, medium, and large "difference" of peaks between the standard and input patterns. If Eq. (7) is represented by the shapes, as shown in Figs. 6(b)–(d), value $r_i^{(-)}$ increases at peak position i of each standard pattern. At the same time, value $r_i^{(+)}$ increases at peak position i of each input pattern.
- Fig. 6(e) typically shows the standard pattern having a flat shape and an input pattern where a "wobble" occurs in the flat shape. Because values $r_i^{(+)}$ and $r_i^{(-)}$ increase alternatively in Eq. (7) during this time, a pair of reference patterns shown in Fig. 6(e) have small shape changes from the normal distribution.

3.5. *Moment ratios of reference pattern vectors*

For the positive and negative reference pattern vectors whose shapes have changed by Eq. (7), the magnitude of shape change is numerically evaluated as the variable of moment ratio. The moment ratios of the positive and negative reference pattern vectors can be calculated using the following equation that has been modified from Eq. (3).

$$\begin{aligned} A^{(+)} &= \frac{\left\{\sum_{i=1}^{m} r_i^{(+)}\right\} \cdot \left\{\sum_{i=1}^{m} (L_i)^4 \cdot r_i^{(+)}\right\}}{\left\{\sum_{i=1}^{m} (L_i)^2 \cdot r_i^{(+)}\right\}^2} - 3 \\ A^{(-)} &= \frac{\left\{\sum_{i=1}^{m} r_i^{(-)}\right\} \cdot \left\{\sum_{i=1}^{m} (L_i)^4 \cdot r_i^{(-)}\right\}}{\left\{\sum_{i=1}^{m} (L_i)^2 \cdot r_i^{(-)}\right\}^2} - 3 \end{aligned} \qquad (8)$$

Where, L_i $(i = 1, 2, \cdots, m)$ is a deviation from the center axis of the normal distribution shown in Figs. 5(c) and (d).

3.6. *Calculation of shape variation*

The initial value of the moment ratio of both positive and negative reference pattern vectors is equal to 0. Therefore, the amount of change of moment ratio in positive direction is $A^{(+)}$, and the amount of change in negative direction is $A^{(-)}$. The total amount of change is the difference between them. Thus, the difference in shapes between standard and input patterns is calculated using the following equation, and it is defined as "Shape variation D".

$$D = A^{(+)} - A^{(-)} \tag{9}$$

Fig. 6 and Table 2 show how D varies with $r_i^{(+)}$, $r_i^{(-)}$, $A^{(+)}$ and $A^{(-)}$.

- In (a), values $r_i^{(+)}$ and $r_i^{(-)}$ do not change. The shape variation becomes $D = 0$ as $A^{(+)} = 0$ and $A^{(-)} = 0$.
- In (b)–(d), because peak position i of the standard pattern locates in area α shown in Fig. 4, the moment ratio becomes $A^{(-)} > 0$ when value $r_i^{(-)}$ increases.
- In (b), because peak position i of the input pattern also locates in area α, the moment ratio becomes $A^{(+)} > 0$ when value $r_i^{(+)}$ increases. The entire shape variation becomes $D \approx 0$.
- In (c), because peak position i of the input pattern locates in the boundary area between α and β, the moment ratio becomes $A^{(+)} \approx 0$ even when value $r_i^{(+)}$ increases. The entire shape variation becomes $D < 0$.
- In (d), because peak position i of the input pattern locates in area β, the moment ratio becomes $A^{(+)} < 0$ when value $r_i^{(+)}$ increases. The entire shape variation becomes $D \ll 0$.
- In (e), a pair of reference patterns have small shape changes from the normal distribution, and the shape variation becomes $D \approx 0$ as $A^{(+)} \approx 0$ and $A^{(-)} \approx 0$. Also, if values $r_i^{(+)}$ and $r_i^{(-)}$ increase randomly, the shape variation becomes $D \approx 0$.

In Fig. 3 (b), the bar graph of the change rate of moment ratio A decreases monotonically from the center to the outer end. From this result and from above (a)–(d), we can understand that value $|D|$ increases monotonically according to the increase of the "difference" between peaks of the standard and input patterns. Also, from (e), it is clear that $D \approx 0$ for the "wobble".

3.7. *Movement of normal distribution*

In the previous section, we have determined the shape variation D by assuming that the center axis of the normal distribution locates at the center of standard and input patterns as shown in Figs. 5 and 6. In this section, however, we determine the amount of shape variation D_j for each j in the case where the center axis of the normal distribution moves to any component position j (where, $j = 1, 2, \cdots, m$) of the standard and input patterns.

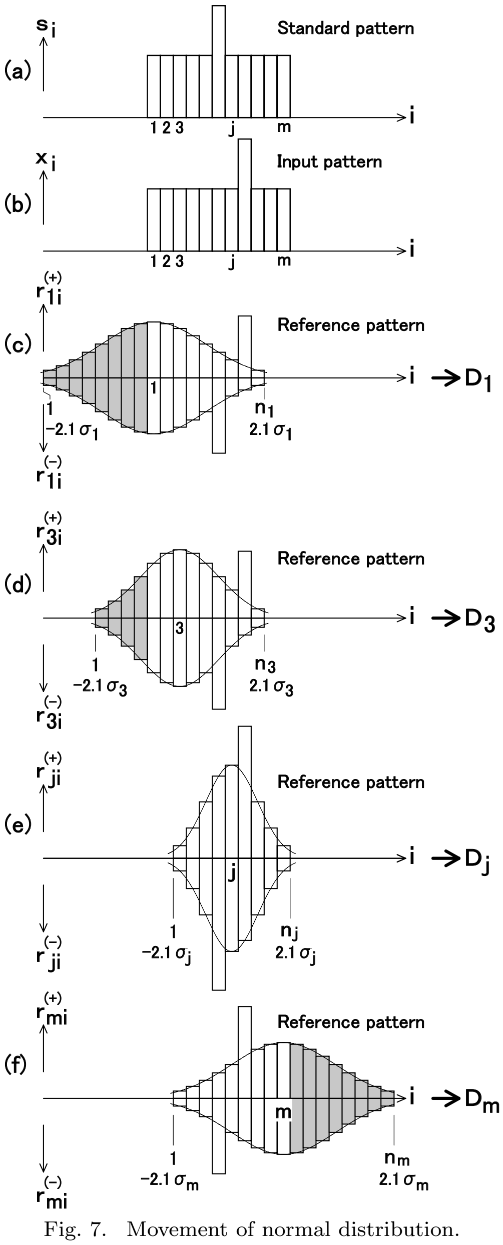


Fig. 7. Movement of normal distribution.

Figs. 7(a) and (b) give an example of standard and input patterns. Also, Figs. 7(c)–(f) show the positive and negative reference patterns when the center axis of the normal distribution moves to positions 1, 3, j and m, respectively. Note that all bar graphs of Figs. 7(a)–(f) have the same width. Here, as shown in Figs. 7(c)–(f), we create positive and negative reference patterns for each j so that bar graphs 1 to n_j of positive and negative reference patterns correspond to area $-2.1\sigma_j \leq u_i \leq 2.1\sigma_j$ of the normal distribution. Where, $\sigma_j = n_j/(4.2m)$ because $2.1\sigma_j \times 2 = (1/m) \times n_j$. As shown in Fig. 7(e), the positive and negative reference patterns do not necessarily cover the entire standard and input patterns.

Then, we process the ends so that the sensitivity to the "wobble" in the positive and negative reference patterns is equated regardless of the movement position of the normal distribution. In the positive and negative reference patterns shown in Figs. 7(c)–(f), the "white" bar graph corresponds to the component numbers i of the input pattern and, therefore, its value changes according to the "wobble" of the input pattern. However, the "gray" bar graph does not correspond to it and its value does not change. Therefore, we set value n_j so that the number of white bar graphs is equated in all the positive and negative reference patterns. In Figs. 7(c)–(f), for an example, each positive and negative reference patterns consists of 9 white bar graphs. By this means, the sensitivity to the "wobble" in the positive and negative reference patterns is equated.

In the proposed algorithm, the values n_j and σ_j must be set appropriately to the pattern recognition application. In Section 4.3, an example method to set these values is given. We can expand Eq. (6) as described above, create positive and negative reference pattern vectors $\boldsymbol{r}_j^{(+)}$ and $\boldsymbol{r}_j^{(-)}$ which have different variance values of the normal distribution for each movement position j, and represent them as follows.

$$\begin{aligned} \boldsymbol{r}_j^{(+)} &= (r_{j1}^{(+)}, r_{j2}^{(+)}, \cdots, r_{jk}^{(+)}, \cdots, r_{jn_j}^{(+)})^T \\ \boldsymbol{r}_j^{(-)} &= (r_{j1}^{(-)}, r_{j2}^{(-)}, \cdots, r_{jk}^{(-)}, \cdots, r_{jn_j}^{(-)})^T \\ & \qquad\qquad\qquad (j = 1, 2, 3, \cdots, m) \end{aligned} \tag{10}$$

Then, we replace the difference in shapes between standard pattern vector $\boldsymbol{s}$ and input pattern vector $\boldsymbol{x}$ into the shape changes of the vectors $\boldsymbol{r}_j^{(+)}$ and $\boldsymbol{r}_j^{(-)}$ by using the following equation instead of Eq. (7).

$$\begin{aligned} & For \;\; i = 1, 2, 3, \cdots, m\,; \\ & \qquad when \;\; k{=}i{-}j{+}(1{+}n_j)/2 \quad (where, 1{\leq}k{\leq}n_j)\,; \\ & \qquad\quad \bullet\, if \; x_i > s_i, \; then \; r_{jk}^{(+)} \longleftarrow r_{jk}^{(+)} + |x_i - s_i| \\ & \qquad\quad \bullet\, if \; x_i < s_i, \; then \; r_{jk}^{(-)} \longleftarrow r_{jk}^{(-)} + |x_i - s_i| \\ & \qquad\qquad\qquad\qquad\qquad (j = 1, 2, 3, \cdots, m) \end{aligned} \tag{11}$$

Note that $(1 + n_j)/2$ is the center component number of $\boldsymbol{r}_j^{(+)}$ and $\boldsymbol{r}_j^{(-)}$, and $i - j$ is a deviation from the center component number. Also, if value k does not satisfy $1 \leq k \leq n_j$, we assume that values $r_{jk}^{(+)}$ and $r_{jk}^{(-)}$ do not change. Fig. 7 represents the shape of Eq. (11), and it shows the example of the increase of values $r_{jk}^{(+)}$ and $r_{jk}^{(-)}$. Then, the magnitude of the shape change of $\boldsymbol{r}_j^{(+)}$ and $\boldsymbol{r}_j^{(-)}$ is numerically evaluated as the variable of moment ratio. The moment ratio of $\boldsymbol{r}_j^{(+)}$ and $\boldsymbol{r}_j^{(-)}$ can be calculated by using the following equation instead of Eq. (8).

$$\begin{aligned} A_j^{(+)} &= \frac{\left\{\sum_{k=1}^{n_j} r_{jk}^{(+)}\right\} \cdot \left\{\sum_{k=1}^{n_j} (L_{jk})^4 \cdot r_{jk}^{(+)}\right\}}{\left\{\sum_{k=1}^{n_j} (L_{jk})^2 \cdot r_{jk}^{(+)}\right\}^2} - 3 \\ A_j^{(-)} &= \frac{\left\{\sum_{k=1}^{n_j} r_{jk}^{(-)}\right\} \cdot \left\{\sum_{k=1}^{n_j} (L_{jk})^4 \cdot r_{jk}^{(-)}\right\}}{\left\{\sum_{k=1}^{n_j} (L_{jk})^2 \cdot r_{jk}^{(-)}\right\}^2} - 3 \\ & \qquad\qquad (j = 1, 2, 3, \cdots, m) \end{aligned} \tag{12}$$

Note that value L_{jk} is a deviation from the center axis of the normal distribution that corresponds to position j. At this time, the shape variation D_j can be calculated by using the following equation instead of Eq. (9).

$$D_j = A_j^{(+)} - A_j^{(-)} \qquad (j = 1, 2, 3, \cdots, m) \tag{13}$$

As shown in Figs. 7(c)–(f), the value D_j is calculated from the respective positive and negative reference patterns for each position j. Thus, if all positive and negative reference patterns cover the peaks of standard and input patterns, all values $|D_j|$ increase monotonically according to the increase of the “difference” between peaks of the standard and input patterns as described in Section 3.6. Also, because the number of white bar graphs has been equated in the positive and negative reference patterns, the shape variation equally becomes $D_j \approx 0$ for the “wobble”.

3.8. *Calculation of geometric distance*

Using the m pieces of the shape variation D_j that we have obtained in Eq. (13), we can calculate the difference in shapes between standard and input patterns by the following equation and we define it as the “Geometric distance d”.

$$d = \sqrt{\sum_{j=1}^{m} (D_j)^2} \tag{14}$$

As described above, the geometric distance can be calculated by using Eqs. (5) and (10)–(14) sequentially. Note that d is the square root of a square sum of the m pieces of values D_j. Thus, as described in the previous section, d also increases monotonically if all values $|D_j|$ increase monotonically according to the increase of the “difference” between peaks of the standard and input patterns. Also, if the shape variation equally becomes $D_j \approx 0$ for the “wobble”, the geometric distance also becomes $d \approx 0$.

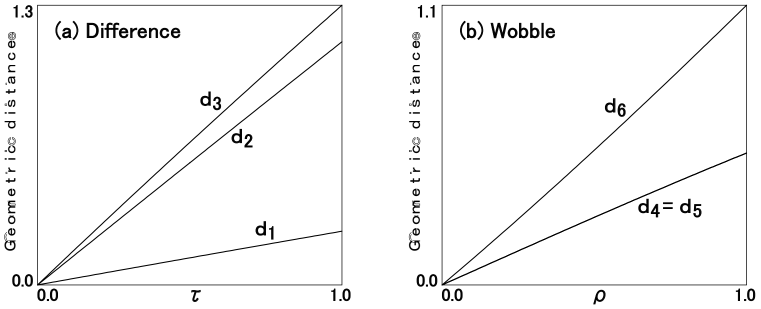


Fig. 8. Calculation in geometric distance.

3.9. *Numerical experiments of geometric distance*

To confirm that the geometric distance algorithm matches the mathematical model that we have assumed in Section 1, we performed numerical experiments to calculate the geometric distances of the standard and input patterns shown in Fig. 1. However, we have developed Eq. (10) by using values $n_j = 27$ ($\sigma_j = n_j/(4.2m) = 0.58$) that are fixed regardless of movement position value j. During this time, the number of white bar graphs of positive and negative reference patterns is 11 for all j values. Note that we read Euclidean distances e_1 to e_6 in Fig. 1 as geometric distances d_1 to d_6 respectively.

Fig. 8(a) shows the calculation result of geometric distances d_1, d_2 and d_3 by increasing value τ from 0.0 to 1.0 in Fig. 1(a). From Fig. 8(a), if value τ is fixed, we can determine that the geometric distance increases monotonically according to the increase of the "difference" of the input pattern peak. Fig. 8(b) shows the calculation result of geometric distances d_4, d_5 and d_6 by increasing value ρ from 0.0 to 1.0 in Fig. 1(b). In Fig. 8(b), if value ρ is fixed, values d_4 and d_5 are smaller than value d_6. That is, if input patterns 4, 5 and 6 have the same area, input patterns 4 and 5 have the energy that is distributed to multiple peaks as the "wobble" when compared with input pattern 6 that has the energy concentrated on a single peak. Thus, the geometric distance of input patterns 4 and 5 is smaller than that of input pattern 6. As a result, it is discovered that the change of geometric distance to the "wobble" is small.

Moreover, Fig. 9 shows input patterns 1, 2 and 3 of Fig. 1(a) where uniformly distributed random numbers are added to the power spectrum of all frequency bands, and they are normalized so that the area of each input pattern becomes equal to the area of standard pattern. However, as uniformly distributed random numbers, the values that uniformly distribute within the range of 0 to 10% of average height 1 of the standard pattern are used regardless of value τ. Fig. 9 is developed by assuming that $\tau = 0.5$. Fig. 10 shows that effect on geometric distances d_1, d_2 and d_3 of increasing value τ from 0.0 to 1.0 in Fig. 9. Note that we have set to change the random number if value τ changes. From Fig. 10, if value τ is fixed within the range of $0.5 \leq \tau$, it is discovered that the geometric distance increases

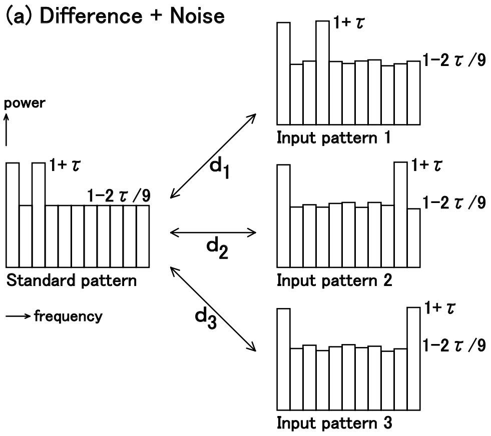


Fig. 9. Typical example of standard and input patterns.

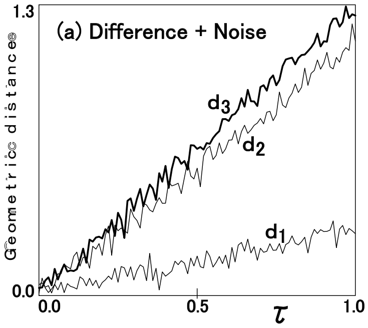


Fig. 10. Calculation in geometric distance.

monotonically according to the increase of the "difference" of the input pattern peak in the "wobble" due to random numbers. From the numerical experiments shown in Figs. 8(a), (b) and Fig. 10, we could verify that the geometric distance algorithm matches the characteristics <1> and <2> of the mathematical model.

3.10. *Calculation of median*

Fig. 11 shows a typical example of 5 shapes, each having a different position on the second peak. We assume that the geometric distance between shape i and shape j is d_{ij}, determine the value d_{ij} between shape i and other 4 shapes j respectively, and calculate mean value $\bar{d}_i$ using the following equation.

$$\bar{d}_i = \Big(\sum_j d_{ij}\Big)/4 \tag{15}$$

$$(i = 1, 2, 3, 4, 5;\ j = 1, 2, 3, 4, 5;\ i \neq j)$$

Note that we have developed Eq. (10) under the conditions described in Section 3.9.

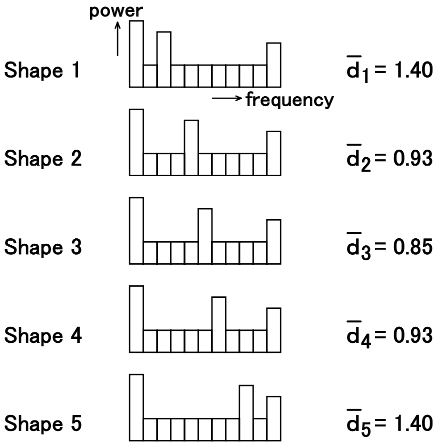


Fig. 11. Example of calculation for median.

In Fig. 11, because shape 1 has a larger "difference" of the second peak when compared with shapes 4 and 5, we can estimate that mean value $\bar{d}_1$ of the geometric distance becomes a large value. Meanwhile, because shape 3 has a smaller "difference" of the second peak when compared with the other 4 shapes, we can estimate that mean value $\bar{d}_3$ becomes a small value. Therefore, we determined shape i, that has the minimum mean value $\bar{d}_i$ of the geometric distance, to be the median. Fig. 11 shows the values $\bar{d}_1$ to $\bar{d}_5$ that have been calculated by numerical experiments. From these values, it is discovered that values $\bar{d}_1$ and $\bar{d}_5$ are large, but value $\bar{d}_3$ is minimal. Therefore, we have determined shape 3 to be the median. The result of numerical experiment of Fig. 11 matches the characteristic <2> of the mathematical model.

4. Experiments of Vowel Recognition

To check the effectiveness of mathematical model and geometric distance algorithm described in the previous section, we have performed the speech recognition experiments using the geometric distance algorithm and actual voices. We used Japanese speech produced by one female speaker in the experiments. We performed the experiments in the following two stages.
(Stage 1) First, we optimized the variance of the normal distribution using the "vowel in the continuous speech" that is different from the voice data for the evaluation experiments.
(Stage 2) Next, we performed the evaluation experiments for the "clean vowel" and the "vowel with noise" by using the optimized normal distribution.
Note that, in this section, a vowel without noise is called the "clean vowel". Also, Stage 1 and Stage 2 are, respectively, divided into Substages Stage 1A, Stage 1B and Stage 1C and Substages Stage 2A, Stage 2B and Stage 2C which are described in the following sections.

4.1. *Voice data*

(Stage 1A) First, we recorded the continuous speech (phonetically-balanced sentences) of the subject female in a soundproof room and created speech data.
(Stage 2A) Next, we recorded each vowel (/a/, /i/, /u/, /e/, /o/) produced by the same speaker in the soundproof room for a period of 2 seconds for each vowel. We repeated this recording 6 times on one day each week over a period of 12 weeks, and we created voice data of the 72 resultant sounds for each vowel (the vowels produced 6 times over 12 weeks). These 5 vowels in 72 voice data sounds are called "/a/01Clean", "/i/01Clean", "/u/01Clean",$\cdots$,"/e/72Clean","/o/72Clean" for each sound, according to the time sequence of the sounds. Then, Babble, Car, Exhibition, and Subway noises[21] have been added with the 20 dB, 10 dB and 5 dB SNR, and the voice data of "5 vowels × 72 sounds × 4 noises ×3 SNRs" has been created. These voice data are also similarly referred to as "/a/01Babble20dB" to "/o/72Subway5dB".

4.2. *Feature parameters*

We have set the voice analysis conditions with the 8kHz sampling frequency, 16bit quantization, 25msec frame width (Hamming window), 10msec frame period, 0.97 pre-emphasis coefficient, 64Hz start frequency of the first filter bank, and 4000Hz end frequency of the 23-rd filter bank.
(Stage 1B) First, we sampled the vowel zone from the continuous speech data of Stage 1A, and extracted the logarithmic power spectrum array of the 23-rd dimensional Mel filter bank output (abbreviated as "power spectrum" hereafter).[22] We repeated them and finally extracted the power spectra of a total of 168 frames for each vowel. The power spectra of these "5 vowels × 168 frames" are the feature parameters that have been extracted from the "vowel in the continuous speech".
(Stage 2B) Next, we sampled the central 100 frames from "5 vowels × 72 sounds" for "/a/01Clean" to "/o/72Clean" voice data and from "5 vowels × 72 sounds × 4 noises × 3 SNRs" for "/a/01Babble20dB" to "/o/72Subway5dB" voice data, and extracted their power spectrum. The power spectra of these "5 vowels × 72 sounds × 13 types × 100 frames" are the feature parameters that have been extracted from the "clean vowel" and the "vowel with noise".

At the same time, we extracted the 12-th dimensional MFCC[22] under the same conditions as those for Stage 2B in order to compare our proposed technique with the conventional technique. The MFCCs of these "5 vowels × 72 sounds × 13 types × 100 frames" are the feature parameters that have been extracted from the "clean vowel" and the "vowel with noise".

Fig. 12 gives an example of the 23-rd dimensional power spectrum that has been extracted from the "clean vowel /a/". The power spectrum of Fig. 12 has $m = 23$ in Eq. (5), and the standard and input patterns are created based on this value. Note that Fig. 12 is referred to as the "1-frame power spectrum" in this paper.

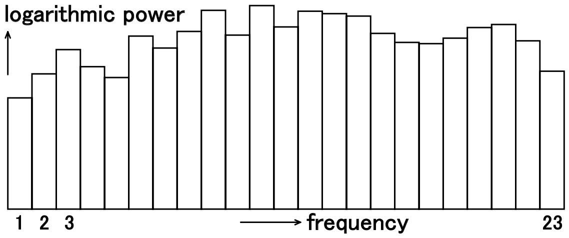


Fig. 12. 23-rd dimensional power spectrum of vowel /a/.

4.3. *Variance optimization of normal distribution*

(Stage 1C) For vowel recognition, it is important to be able to accurately detect a "difference" between the formants of the standard and input patterns. The proposed technique replaces the amount of "difference" between the formants by the shape change of normal distribution and detects it. In such a case, it is important to optimize the shape (variance σ^2) of normal distribution that covers the standard and input patterns. Therefore, we show the optimization procedure in Subsections 4.3.1 and 4.3.2.

4.3.1. *Subdivision of reference pattern*

In the previous section, as shown in Fig. 7, we have determined the geometric distance by assuming that all bar graphs of the standard and input patterns and those of the positive and negative reference patterns have the same width. In this case, because $\sigma_j = n_j/(4.2m)$ in Fig. 7(e), if the value n_j is changed for each 2, the value σ_j changes as a discrete value for each $1/(2.1m)$. Thus, if value m is small, the accuracy of optimum value σ_j drops. In order to improve the accuracy, we subdivide the bar graph of the positive and negative reference patterns.

Figs. 13(a) and (b) show a typical example of bar graph of the standard and input patterns consisting of the 23 bars. Figs. 13(c) and (d) show the positive and negative reference patterns when the center axis of normal distribution moves to positions 3 and j, respectively. Here, as shown in Fig. 13, for example, we use a single-bar graph of the standard and input patterns and we subdivide the positive and negative reference patterns into the 10-bar graph. Then, as shown in Figs. 13(c) and (d), each of the positive and negative reference pattern (where, $j = 1, 2, \cdots, 23$) is configured by the same number of bars of the white bar graph. In Figs. 13(c) and (d), for example, the bar graph is structured with 20.2 bars (where, $\omega = 20.2$). This ω is the number of white bar graphs of the positive and negative reference patterns. In Fig. 13 (d), the relationship of $\omega = n_j/10$ and $\sigma_j = \omega/(4.2m) = n_j/10/(4.2m)$ (where, $m = 23$) is established. Thus, if the value n_j is changed for each 2, the value σ_j changes as a discrete value for each $0.1/(2.1m)$. The accuracy of the optimum value σ_j is improved.

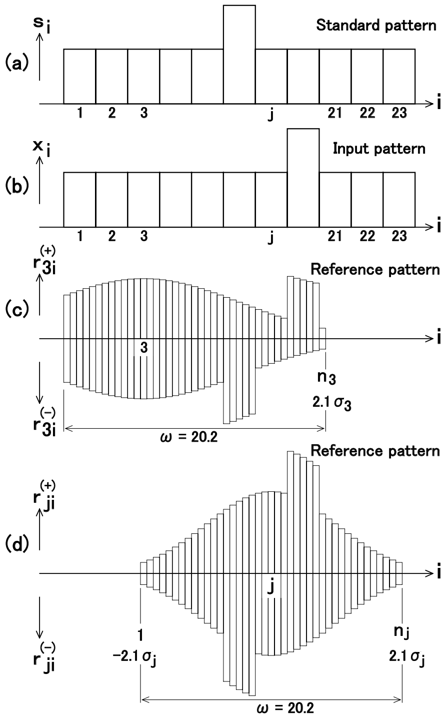


Fig. 13. Subdivision of reference patterns.

4.3.2. *Optimization of σ*

Fig. 14, Fig. 15 and Table 3 show the processing procedure to determine the optimum value of σ (the optimum value of ω) using the "vowel in the continuous speech". Fig. 14 is a flowchart used to determine the optimum value by scanning value ω in the range of 3.0 to 23.0. In Step 1 of Fig. 14, $\omega = 3.0$ is set as the initial value. In Step 2, the positive and negative reference pattern vectors that are equivalent to those of Fig. 13 are created according to the ω set value. Then, we explain Steps 3–7 by referring to Fig. 15 and Table 3. Table 3 shows the type and the number of the 23-rd dimensional power spectrum that has been used for the standard and input patterns. The power spectra, each consisting of 168 frames shown on the first row of Table 3, have been extracted from the "vowel in the continuous speech" in Stage 1B of Section 4.2. In Step 3, a single standard pattern is calculated for each vowel. Step 3 of Fig. 15 shows the process required to determine the median from the above 168 frames using the technique of Section 3.10, and to set the standard pattern of each vowel. The power spectra, each consisting of one frame shown on the second row of Table 3, are the standard patterns that have been determined for each vowel. The power spectra, each consisting of 167 frames shown on the third row of Table 3, are the patterns of the above 168 frames from which each standard pattern has been removed. These "167×5" frames are

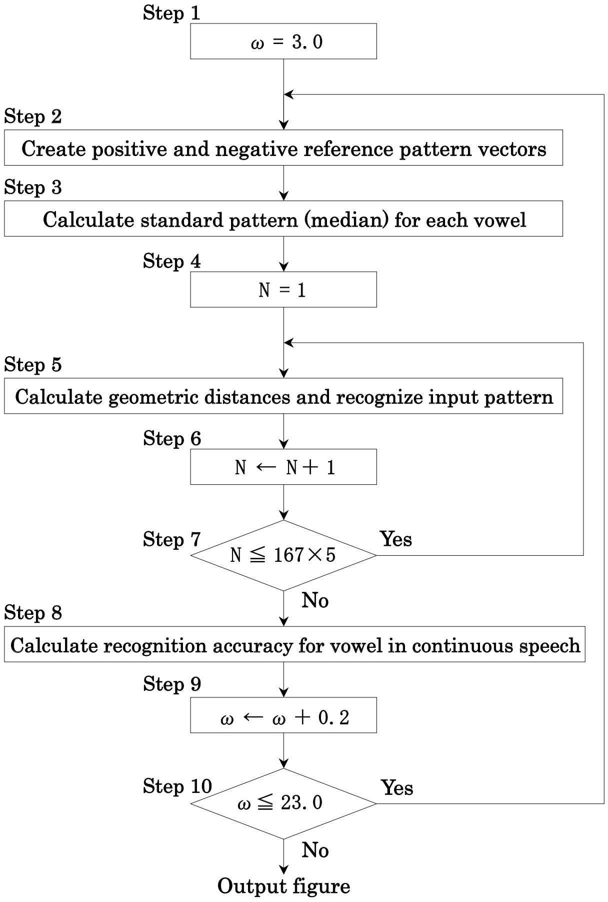

Step 1
ω = 3.0
Step 2
Create positive and negative reference pattern vectors
Step 3
Calculate standard pattern (median) for each vowel
Step 4
N = 1
Step 5
Calculate geometric distances and recognize input pattern
Step 6
N ← N+1
Step 7
N ≦ 167×5
Yes
No
Step 8
Calculate recognition accuracy for vowel in continuous speech
Step 9
ω ← ω + 0.2
Step 10
ω ≦ 23.0
Yes
No
Output figure

Fig. 14. Flowchart for optimizing normal distribution.

the input patterns. In Step 4, $N = 1$ is set as the initial value and, as shown in Step 4 of Fig. 15, the first input pattern is specified from the "167×5" frames. In Step 5, the geometric distance is calculated and the input pattern is recognized. As shown in Step 5 of Fig. 15, the geometric distance between the standard and input patterns is calculated for each of the 5 vowels, and the minimum value is determined among the 5 geometric distance values obtained. Then, the category to which the standard pattern having the minimum value belongs is selected as the recognition result of the input pattern. In Steps 6 and 7, value N is incremented by 1, the N-th input pattern is specified among the "167×5" frames, and Step 5 is repeated. After the recognition result of all input patterns has been obtained, in Step 8, the recognition accuracy is calculated by setting the total "167×5" frames as the denominator and by setting the number of correctly recognized input patterns as the numerator. In Steps 9 and 10, value ω is incremented by 0.2 until it reaches 23.0, and the process of Steps 2–8 is repeated.

Table 3. Power spectra for optimizing normal distribution.

	/a/	/i/	/u/	/e/	/o/
Vowel in continuous speech	168	168	168	168	168
Standard pattern	1	1	1	1	1
Input pattern	167	167	167	167	167

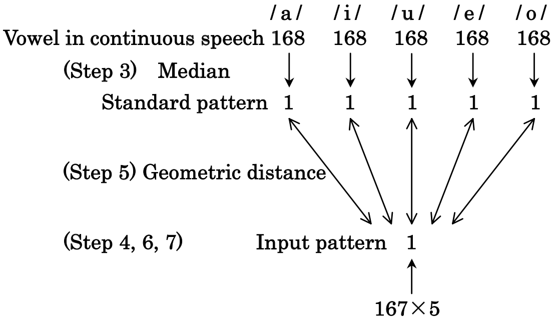

/a/
/i/
/u/
/e/
/o/
Vowel in continuous speech 168 168 168 168 168
(Step 3) Median
Standard pattern 1 1 1 1 1
(Step 5) Geometric distance
(Step 4, 6, 7)
Input pattern 1
167×5

Fig. 15. Diagram for optimizing normal distribution.

Fig. 18 shows the relationship between the value ω and the recognition accuracy obtained by the above process. From Fig. 18, it is discovered that the recognition accuracy becomes maximum if $\omega = 10.2$. Thus, we determine $\omega = 10.2$ as the optimum value and use it in the following evaluation experiments.

4.4. *Evaluation experiments*

4.4.1. *Vowel recognition with geometric distance*

(Stage 2C) We have performed the evaluation experiments for the "clean vowel" and the "vowel with noise" by using the value $\omega = 10.2$ determined in the previous section. Fig. 16, Fig. 17 and Table 4 show the procedure. Table 4 shows the type and the number of the 23-rd dimensional power spectrum that has been used for the standard and input patterns. The power spectra, each consisting of 100 frames shown on the first row of Table 4, have been extracted from "01Clean" of each vowel in Stage 2B of Section 4.2. "01Clean" is the first "clean vowel" that was produced among 72 sounds in 12 weeks. Then, as shown in Step 3 of Fig. 17, the median was determined from the above 100 frames and it was used as the standard pattern of each vowel. The power spectra, each consisting of one frame shown on the second row of Table 4, are the standard patterns that have been determined for each vowel. Also, the power spectra, each consisting of 100 frames shown in {1} to {13} of Table 4, have been extracted from the "clean vowel" and the "vowel with noise" in Stage 2B of Section 4.2. Then, the power spectra of these "13×71×100×5" frames were used as the input patterns. Figs. 16 and 17 show the procedure for evaluation, by using both 5 standard patterns obtained from

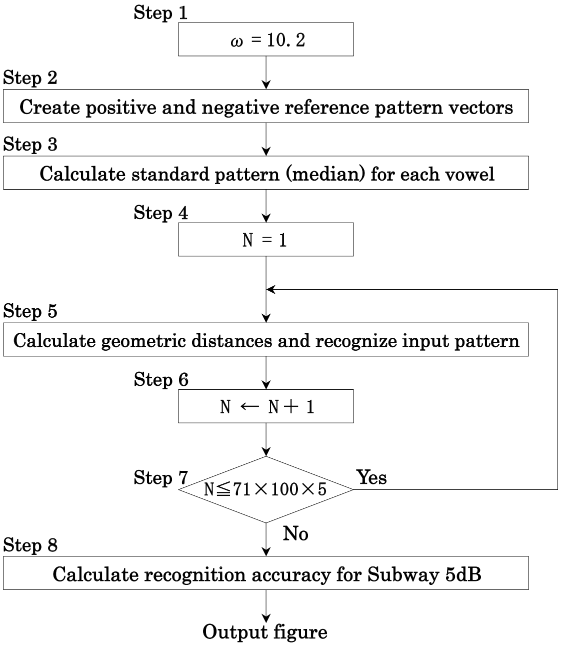

Step 1
ω = 10. 2
Step 2
Create positive and negative reference pattern vectors
Step 3
Calculate standard pattern (median) for each vowel
Step 4
N = 1
Step 5
Calculate geometric distances and recognize input pattern
Step 6
N ← N + 1
Step 7
N≦71×100×5
Yes
No
Step 8
Calculate recognition accuracy for Subway 5dB
Output figure

Fig. 16. Flowchart for vowel recognition.

the "01Clean" and 71×100×5-frame input patterns shown in {13} of Table 4. A similar process is also carried out if the 71×100×5-frame input patterns shown in {1} to {12} are used. In Steps 2–8 of Fig. 16 and Steps 3–7 of Fig. 17, the same process is executed as those of Figs. 14 and 15. Then, the recognition accuracy is calculated by setting the total "71×100×5" frames as the denominator and by setting the number of correctly recognized input patterns as the numerator.

4.4.2. *Vowel recognition with MFCC*

To compare the proposed technique with the conventional techniques, we performed the evaluation experiments of vowel recognition using the 12-th dimensional MFCC. The MFCC was extracted from the "clean vowel" and the "vowel with noise" in Section 4.2, and its type and number are the same as those shown on Table 4. First, we determined the mean and variance in each dimension using the 12-th dimensional MFCCs of 100 frames in "01Clean", and created the 12-th dimensional normal distribution. We created this 12-th dimensional normal distribution for each vowel, and used it as the standard pattern of each vowel. Then, we used the 12-th dimensional MFCCs of 13×71×100×5 frames shown in {1} to {13} as the input patterns. We calculated the likelihood between the input pattern and the standard pattern of each vowel, and determined that the category of the input pattern is equal to the category of the standard pattern having the maximum likelihood among 5 standard patterns.

Table 4. Power spectra for vowel recognition.

		/a/	/i/	/u/	/e/	/o/
	01 Clean	100	100	100	100	100
	Standard pattern	1	1	1	1	1
	02 Clean	100	100	100	100	100
{1}	: Input pattern	:	:	:	:	:
	72 Clean	100	100	100	100	100
	02 Babble 20dB	100	100	100	100	100
{2}	: Input pattern	:	:	:	:	:
	72 Babble 20dB	100	100	100	100	100
:		...	...	...	...	...
:		...	...	...	...	...
	02 Subway 5dB	100	100	100	100	100
{13}	: Input pattern	:	:	:	:	:
	72 Subway 5dB	100	100	100	100	100

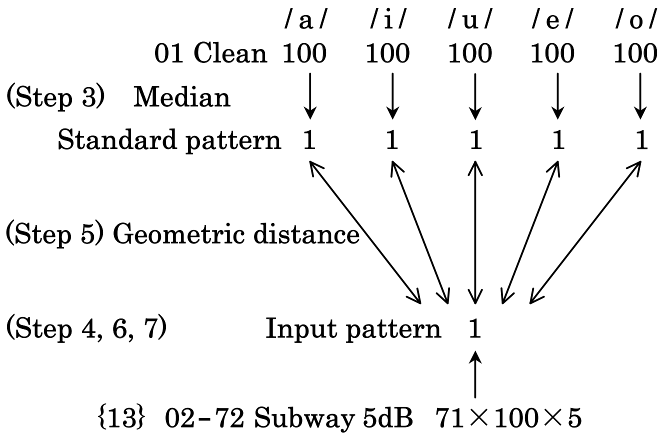

/a/ /i/ /u/ /e/ /o/
01 Clean 100 100 100 100 100
(Step 3) Median
Standard pattern 1 1 1 1 1
(Step 5) Geometric distance
(Step 4, 6, 7) Input pattern 1
{13} 02-72 Subway 5dB 71×100×5

Fig. 17. Diagram for vowel recognition.

4.5. *Results of evaluation experiments*

Tables 5 and 6 show the results of vowel recognition using the geometric distance and MFCC, respectively. From these tables, it is learned that the recognition accuracy with the geometric distance is higher than that with the MFCC in all cases. In particular, "mean" of 10 dB and 5 dB SNR has improved approximately by 10%. For both Tables 5 and 6, the recognition accuracy of "Exhibition5dB" is low. This reason may be the insertion of a background male voice in the "Exhibition". Thus we confirm the effectiveness of the mathematical model and the geometric distance algorithm.

Table 5. Vowel recognition accuracy with geometric distance. ($\omega = 10.2$)

	Babble	Car	Exhibition	Subway	Mean
Clean					99.99%
SNR 20 dB	99.90%	99.82%	99.00%	99.56%	99.57%
SNR 10 dB	99.26%	97.72%	83.80%	90.66%	92.86%
SNR 5 dB	94.14%	81.69%	61.42%	74.89%	78.04%

Table 6. Vowel recognition accuracy with MFCC.

	Babble	Car	Exhibition	Subway	Mean
Clean					99.54%
SNR 20 dB	98.83%	97.55%	96.57%	98.43%	97.84%
SNR 10 dB	91.05%	80.92%	78.23%	83.57%	83.44%
SNR 5 dB	78.62%	68.10%	60.84%	64.67%	68.06%

4.6. *Verification of optimum value*

Table 5 shows the result of recognition accuracy using the optimum value $\omega = 10.2$ that we have determined from Fig. 18. Here, in order to verify that the value $\omega = 10.2$ is truly the optimum value, we have scanned the value ω from 3.0 to 23.0 in Fig. 16 and calculated the recognition accuracy. Figs. 19 and 20 show the calculated relationship between the value ω and the recognition accuracy for the input patterns of the "clean vowel" and the "vowel with 5 dB noise", respectively. From Figs. 19 and 20, we can find that the recognition accuracy is almost maximum in the value $\omega = 10.2$.

4.7. *The reason why "vowel in continuous speech" was used for optimization*

In Subsection 4.3.2, we determine the optimum value ω using 168 frames of each "vowel in the continuous speech" shown on the first row of Table 3. While in Subsection 4.4.1, we determine the standard pattern using 100 frames of each vowel of "01Clean" shown on the first row of Table 4. This section describes the reason why we have used the "vowel in the continuous speech".

Fig. 19 shows the relationship between the value ω and the recognition accuracy obtained from the "Clean" input patterns. These voice data have the variability with time of 12 weeks. In Fig. 19, the recognition accuracy is 100% in part of the ω value range. From the results of vowel recognition experiments, we have found that the recognition accuracy reaches 100% in the relatively wide ω value range in the variability with time below 4 weeks. In such a case, we have a problem determining the maximum position of recognition accuracy. This means that we will find it difficult to determine the optimum value of ω by using the voices with few variations produced in a short period. Meanwhile, if the "vowel in the continuous speech" is used, the power spectrum of the vowel changes appropriately even if the voices

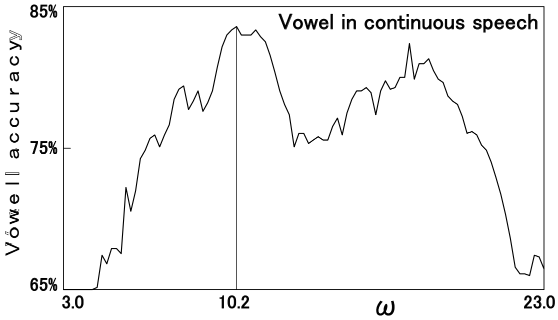


Fig. 18. Vowel recognition accuracy and optimum value ω.

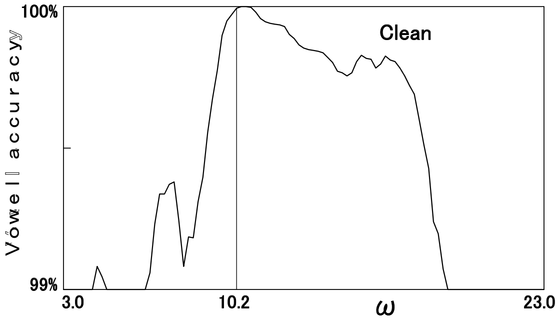


Fig. 19. Vowel recognition accuracy with geometric distance.

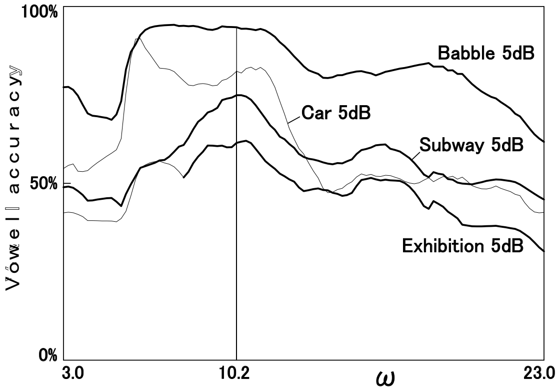


Fig. 20. Vowel recognition accuracy with geometric distance.

are produced in a short period. Therefore, the maximum position of recognition accuracy is most obvious as shown in Fig. 18. Thus we use the "vowel in the continuous speech" to determine the optimum value of ω.

5. Conclusions and Future Work

We have proposed a new similarity scale that replaces the difference in shapes between the standard and input patterns by the shape change of a normal distribution, and that numerically evaluates the magnitude of the shape change as a variable of the moment ratio. At this time, if the number of bar graphs of the standard and input patterns is limited in the actual application of pattern recognition, we have shown that we can avoid the reduced accuracy by subdivision of bar graphs of positive and negative reference patterns. We have performed the vowel recognition experiments and verified the effectiveness of the mathematical model and the geometric distance algorithm.

Finally, we describe future work. This paper describes the vowel recognition experiments that we have carried out using only the vowels produced by one female speaker. We will continue the vowel recognition experiments using various types of voice data and will verify their effectiveness by evaluating the applicable range of mathematical model and algorithm.

We need to calculate Eqs. (11)–(14) in each combination of standard and input patterns if we use multiple standard patterns and a single input pattern. Hence the processing overhead increases when the number of standard patterns increases. Additionally we need to evaluate Eq. (10) for each position j of the normal distribution. Therefore, memory increases in proportion to the square of the m components of the standard and input patterns. These overheads will be addressed in future studies.

Acknowledgments

This research has been partially supported by New Energy and Industrial Technology Development Organization (NEDO) of the Japanese Government under Grant No. 10HC7011, by Queensland Parks and Wildlife Service of the Australian Government under Coxen's Fig-parrot Recovery Plan, and funded by Mitsubishi Heavy Industries, Ltd. of Japan and Tokyo Gas Co., Ltd. of Japan, by West Nippon Expressway Engineering Shikoku Company Limited of Japan.

References

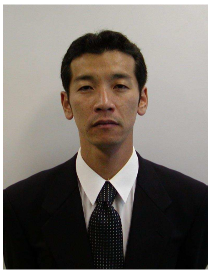
1. R.O. Duda, P.E. Hart and D.G. Stork. *Pattern Classification, second ed.*, Wiley, NewYork, 2000.
2. K.K. Paliwal. *Effect of preemphasis on vowel recognition performance*, Speech Communication, **3**, pp. 101-106, 1984.
3. L.R. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey, 1993.
4. F. Itakura and S. Saito. *An analysis-synthesis telephony based on maximum likelihood method*, Proc. 6th Int. Congr. Acoustics, C-5-5, 1968.
5. F. Itakura. *Minimum prediction residual principle applied to speech recognition*, IEEE Trans. Acoust., Speech and Signal Processing, **23**, pp. 67-72, 1975.
6. S. Furui. *Digital Speech Processing, Synthesis, and Recognition (Electrical and Computer Engineering)*, Marcel Dekker, Inc., NewYork, 1989.

7. K. Shikano and M. Sugiyama. *Evaluation of LPC spectral matching measures for spoken word recognition*, Trans. IECE, 565-D, **5**, pp. 535-541 1982.
8. D. Klatt. *Prediction of perceived phonetic distance from critical band spectra: A first step*, Proc. ICASSP 82, **2**, pp. 1278-1281, 1982.
9. D. Mansour and B.H. Juang. *A family of distortion measures based upon projection operation for robust speech recognition*, IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-37, **11**, pp. 1659-1671, 1989.
10. N. Nocerino, F.K. Soong, L.R. Rabiner and D.H. Klatt. *Comparative study of several distortion measures for speech recognition*, Speech Communication, **4**, pp. 317-331, 1985.
11. S.-H. Cha and S.N. Srihari. *On measuring the distance between histograms*, Pattern Recognition, **35**, pp. 1355-1370, 2002.
12. J.-K. Kamarainen, V. Kyrki, J. Ilonen and H. Kälviäinen. *Improving similarity measures of histograms using smoothing projections*, Pattern Recognition Lett., **24**, pp. 2009-2019, 2003.
13. F.-D. Jou, K.-C. Fan and Y.-L. Chang. *Efficient matching of large-size histograms*, Pattern Recognition Lett., **25**, pp. 277-286, 2004.
14. F. Serratosa and A. Sanfeliu. *Signatures versus histograms: Definitions, distances and algorithms*, Pattern Recognition, **39**, pp. 921-934, 2006.
15. V.V. Strelkov. *A new similarity measure for histogram comparison and its application in time series analysis*, Pattern Recognition Lett., **29**, pp. 1768-1774, 2008.
16. B. Gold and N. Morgan. *Speech and Audio Signal Processing*, John Wiley & Sons, Inc., New Jersey, 2000.
17. S. Davis and P. Mermelstein. *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-28, **4**, pp. 357-366, 1980.
18. S. Nakagawa, M. Okada and T. Kawahara. *Spoken Dialogue Systems*, IOS Press, 2005.
19. F. Jelinek. *Statistical Methods for Speech Recognition*, MIT Press, 1998.
20. S.E. Levinson. *Mathematical Models for Speech Technology*, John Wiley & Sons, Inc., New Jersey, 2003.
21. H.G. Hirsch and D. Pearce. *The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions*, ISCA ITRW ASR2000, 2000.
22. HTK Team in Cambridge University Engineering Department. *HTK Speech Recognition Toolkit (The Hidden Markov Model Toolkit)*, http://htk.eng.cam.ac.uk/

Michihiro Jinnai (Member)

He received the B.S. degree in seismology from Kyoto University, Japan, the M.E. and Ph.D. degrees in speech recognition from Kobe University, Japan, in 1976, 1980, and 1983, respectively. He is currently a professor with the Department of Electro-Mechanical Systems Engineering, Kagawa National College of Technology, Japan. His research interests include similarity scale and pattern matching. He has been developing the application software with geometric distance. It is used for detecting bird call, bat call, and whale call in Australia.

Satoru Tsuge



Satoru Tsuge received his B.E., M.E., and Dr. Eng. degrees from the University of Tokushima, Tokushima in 1996, 1998, and 2001, respectively. From 1997 to 1999, he was an intern researcher at ATR Interpreting Telecommunications Research Laboratories, Kyoto. Since 2000, he has been with the Faculty of Engineering, the University of Tokushima, Tokushima, where he is currently a lecturer. His current research interests include speech recognition, speaker recognition, and information retrieval. He is a member of IPSJ and ASJ.

Shingo Kuroiwa

He received the B.E., M.E. and D.E. degrees in electrocommunications from the University of Electro Communications, Tokyo, Japan, in 1986, 1988, and 2000, respectively. From 1988 to 2001 he was a researcher at the KDD R & D Laboratories. From 2001 to 2007, he was an Associate Professor of Institute of Technology and Science at the University of Tokushima, Japan. Since 2007, he has been with Chiba University, Japan, where he is currently a Professor of Graduate School of Advanced Integration Science. His current research interests include speech recognition, speaker recognition, natural language processing, and information retrieval. He is a member of the IEICE, IPSJ, and ASJ.

Fuji Ren (Member)

He received the Ph.D. degree in 1991 from Faculty of Engineering, Hokkaido University, Japan. He worked at CSK, Japan, where he was a chief researcher of NLP. From 1994 to 2000, he was an associate professor in the Faculty of Information Sciences, Hiroshima City University. From 2001 he joined the faculty of engineering, the University of Tokushima as a professor. His research interests include Natural Language Processing, Artificial Intelligence, Language Understanding and Communication. He is a member of the IEICE, CAAI, IEEJ, IPSJ, JSAI, AAMT and a senior member of IEEE.

Minoru Fukumi

Minoru Fukumi received the B.E. and M.E. degrees from the University of Tokushima, in 1984 and 1987, respectively, and the doctor degree from Kyoto University in 1996. Since 1987, he has been with the Department of Information Science and Intelligent Systems, University of Tokushima. In 2005, he became a Professor in the same department. He received the best paper award from the SICE in 1995 and best paper awards from some international conferences. His research interests include neural networks, evolutionary algorithms, image processing and human sensing. He is a member of the IEEE, SICE, IEEJ, IPSJ and IEICE.