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

January 01, 2014s

Michihiro Jinnai

# Abstract

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.

Furthermore, we describe that there are the following three shortcomings with the above geometric distance algorithm. (1) Since standard and input patterns are normalized to have the same area, a pseudo difference in shapes occurs between them. (2) Since “shape variation” is calculated in each combination of the standard and input patterns, the processing overhead increases when the number of standard patterns increases. (3) Since reference patterns are evaluated for each movement position of a normal distribution, the computational memory overhead increases when the number of components of standard and input patterns increases. To counter these shortcomings, a new geometric distance algorithm is proposed. (1) It is derived without normalization of the standard and input patterns, so that the pseudo difference in shapes is removed. (2) It reduces the processing overhead by separating the calculation of “shape variation” into a registration process and a recognition process. (3) It reduces the computational memory overhead by sharing a single reference pattern. Experiments in vowel recognition were carried out using the same voice data as the above experiments. At a mean of 5 dB SNR, the recognition accuracy improved from 78% to 82% over the above algorithm.

Moreover, in the above algorithm, we have performed the optimization of the Geometric Distance using the “clean vowels in the continuous speech” for vowel recognition. However, there is a shortcoming with the above optimization method because only the

clean vowels are used. To improve the shortcoming, we propose a new optimization method using the weighted random numbers generated by the computer and five patterns of long vowels, instead of the “clean vowels in the continuous speech”. By using our proposed method, we have checked the relationship between the variance of the normal distribution and the vowel recognition accuracy, and estimated the optimum variance value. Also, by using the estimated value, we have performed evaluation experiments for the “long vowels with actual noise of 5 dB SNR” and achieved the vowel recognition accuracy of 80.3%. We have verified the effectiveness of the proposed method.

# Acknowledgements

First and foremost, I would like to thank my supervisor Professor Minoru Fukumi. With Professor Fukumi's strong confident support and guidance, I was able to accomplish top-level technical works, and able to publish high-quality academic papers with innovative proposals.

I sincerely express my thanks to Associate Professor Satoru Tsuge with Daido University, Professor Shingo Kuroiwa with Chiba University and Professor Fuji Ren with Tokushima University, for their thoughtful comments and invaluable suggestions.

I am deeply grateful to Neil Boucher, Hollis Taylor, Jeremy Robertson and Sonia Kleindorfer. They have given me many invaluable suggestions for my English papers.

I would like to thank the members of West Nippon Expressway Engineering Shikoku Company Limited, Yukio Akashi, Kazuaki Hashimoto and Shogo Hayashi. They have given me many invaluable sound data to carry out experiments.

Thanks to my family and my friends, they have been supportive for my study. And thanks to many others who have been supportive of the research program.

# Contents

# Chapter 1

# Introduction

Human beings, dogs, cats, and other such animals have "the sense of similarity" in hearing and sight. To realize "the sense of similarity" using an algorithm called "similarity scale" is an important subject for developing computer intelligence. 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, image retrieval and detection of abnormal vibration, many distance functions have been proposed for comparing histograms.[11, 12, 13, 14, 15, 24]

A similarity scale is a concept that should intuitively concur with the human concept of similarity in hearing and sight. Therefore, 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. 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.

Then, we proposed an algorithm based on one-to-many point mapping to realize the mathematical model. Within the algorithm, the difference in shapes between the standard and input patterns is replaced by the shape change of a reference pattern having the initial shape 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.

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. Chapter 2 describes the shortcomings that are found in the conventional similarity scales. Chapter 3 describes the mathematical model and algorithm of the new similarity scale, describes numerical experiments, and describes that the algorithm performs well. Chapter 4 describes the speech recognition tests that have been carried out, and describes the effectiveness of the mathematical model and algorithm.

Furthermore, we describe that there are the following three shortcomings with the above algorithm. (1) Since the standard and input patterns are normalized to have the same area, a pseudo difference in shapes occurs between the standard and input patterns and the recognition performance of geometric distance becomes unpredictable. (2) Since "shape variation" is calculated in each combination of the standard and input patterns if we use multiple standard patterns and a single input pattern, the processing overhead increases when the number of standard patterns increases. (3) Since positive and negative reference patterns are evaluated for each movement position of the normal distribution, the computational memory overhead increases in proportion to the square of the number of components of the standard and input patterns.

In Chapter 6, we propose a new geometric distance algorithm that can realize the above mathematical model and that can also improve the above shortcomings. (1) The new algorithm is derived without normalization of the standard and input patterns, so that the pseudo difference in shapes is removed and the recognition performance of geometric distance becomes stable. (2) The new algorithm reduces the processing overhead during an input pattern recognition process by separating the calculation of "shape variation" into a standard pattern registration process and an input pattern recognition process. (3) The new algorithm reduces the computational memory overhead by sharing a single reference pattern.

Chapter 5 describes the shortcomings that are found in the conventional algorithm. Chapter 6 describes the new algorithm, provides the evaluation results of the processing overhead and the computational memory required for the algorithm, describes numerical experiments, and describes that the algorithm performs well. Chapter 7 describes the speech recognition tests that have been carried out, and describes the stabilized recognition performance.

Moreover, within the above algorithm, the difference in shapes between the standard and input patterns is replaced by the shape change of a reference pattern having the initial shape of a normal distribution, and the magnitude of this shape change is numerically evaluated as a variable of the moment ratio. In such a case, from its principle, it is important to optimize the shape (variance $\sigma^2$) of the normal distribution that covers the standard and input patterns. Until now, we have determined the optimum variance value of the normal distribution using the "clean vowels in the continuous speech" for vowel recognition.

However, there is a shortcoming with the above optimization method. That is, the characteristic <1> of the above mathematical model is ignored because only the clean vowels are used. The optimization needs to be made to maximize the effect of the characteristics <1> and <2> of the mathematical model simultaneously. Besides, since the optimum variance value of the normal distribution needs to be re-calculated each time the speaker changes, a low processing overhead is also required to calculate the optimum value. To improve the shortcoming and to satisfy the requirement, in Chapter 9, we propose a new method to determine the optimum variance value of the normal distribution for vowel recognition, where we consider both characteristics <1> and <2> of the mathematical model and reduce the processing overhead. We perform an experiment to estimate the optimum value by using our proposed method. Also, we perform evaluation experiments of vowel recognition by using the estimated value that we have calculated. Chapter 8 describes the shortcoming that is found in the conventional optimization method of the new similarity scale. Chapter 9 describes the new optimization method of the new similarity scale, and describes the optimization experiment using the weighted random numbers generated by the computer and five patterns of long vowels. Chapter 10 describes the evaluation experiments of vowel recognition that have been carried out by using the calculated optimum value (estimated value), and describes the effectiveness of the proposed method. Chapter 11 describes the conclusions and touches on future work.

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.

# Chapter 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".

Figure 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 $\tau$ in the relationship shown in Figure 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.

Figure 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 $\rho$ in the

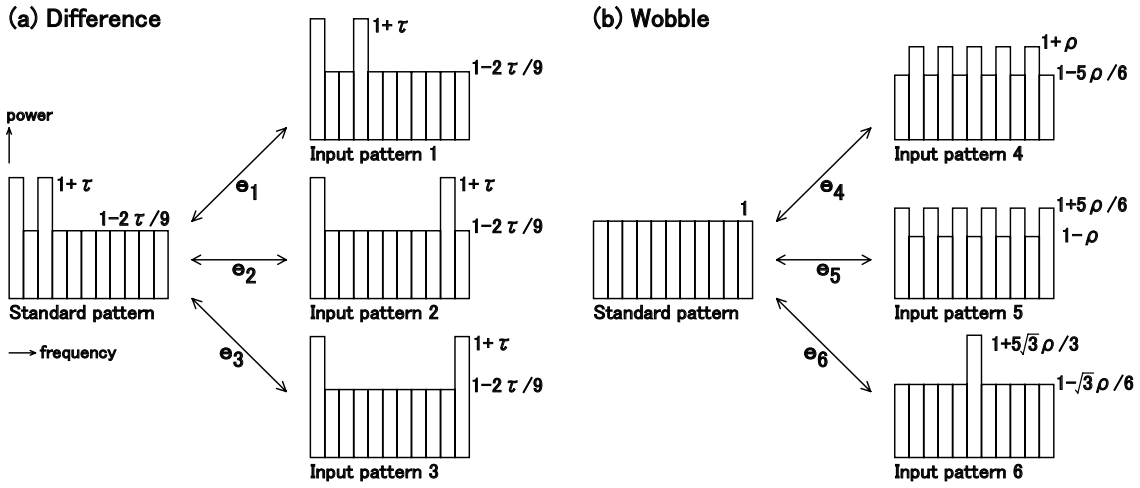


Figure 1. Typical examples of standard and input patterns.

relationship shown in Figure 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.

# Chapter 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, $\mu$ is mean, and $\sigma^2$ 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 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

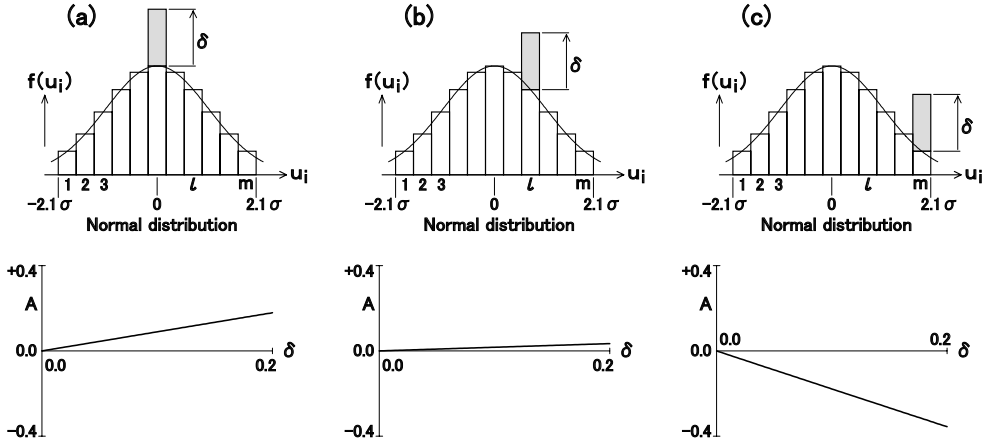


Figure 2. Change of moment ratio A.

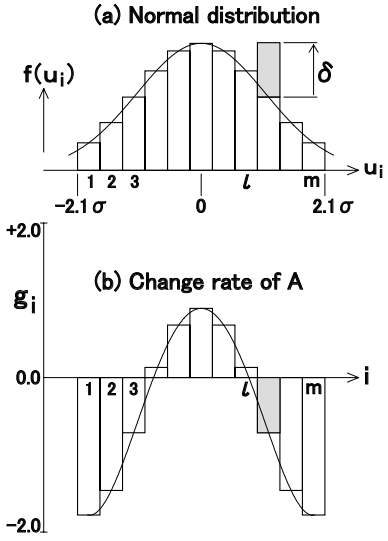


Figure 3. Change rate of moment ratio $A$.

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 $\mu$. 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 Figure 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

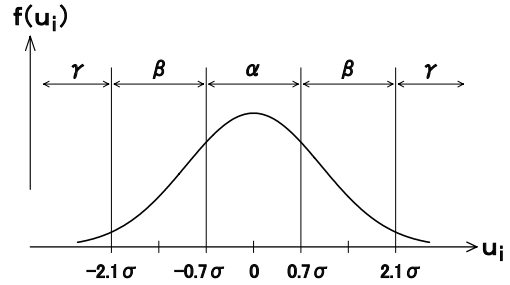

Figure 4. Normal curve.

Table 1. Features of moment ratio $A$.

| Figure 4 | $\alpha$ | Boundary area between $\alpha$ and $\beta$ | $\beta$ | $\gamma$ |
|---|---|---|---|---|
| Increase of $f(u_i)$ | $A > 0$ | $A \approx 0$ | $A < 0$ | — |

ratio $A$ and the increment value $\delta$ of bar graphs seen in Figures 2 and 3. Graphs in the upper side of Figures 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 $-2.1\sigma \leq u_i \leq 2.1\sigma$ ($\sigma = 1$) of the normal distribution. On these bar graphs, only a single bar increases by value $\delta$ 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 $\delta$ is shown in the lower side of Figures 2(a)–(c). For now we only consider positive values of $\delta$. From these graphs, it is discovered that $A = 0.0$ if $\delta = 0.0$. Also, $A$ changes approximately linearly when value $\delta$ increases. Note that if only a single bar increases by value $\delta$ in the graph with $m$ bars, it is the same as when only a single bar with an $1/m$ ratio increases by value $\delta$. If value $m$ changes ($m$ is an odd numbered value), the gradient of moment ratio graphs in the lower side of Figures 2(a)–(c) changes by the same $1/m$ weight. This property holds for all values of $m$ and for any variance $\sigma^2$ of the normal distribution.

Figures 3(a) and (b) show the change rate of $A$ ($\mathrm{g}_i$, where a change of $\delta$ occurs at the $i$-th position) for a normal distribution and a single instance of $\delta$. Change rate $\mathrm{g}_i$ is described by the following equation.

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

The $\mathrm{g}_{(1+m)/2}$, $\mathrm{g}_l$ and $\mathrm{g}_m$ are equal to the gradients of respective graphs shown in the lower side of Figures 2(a)–(c). Next, in Figure 3(a), position $i$ of the bar that has increased by value $\delta$ is scanned from 1 to $m$, and Eq. (4) is calculated. Figure 3(b) shows a bar graph

of the calculated value $g_i$, where $\delta = 0.2$. From Figure 3(b), $g_i > 0$, $g_i \approx 0$ and $g_i < 0$ 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. Figure 4 shows a normal curve $f(u_i)$ with mean $\mu = 0$ and variance $\sigma^2$, and the moment ratio becomes $A = 0$. Also, if the value $f(u_i)$ exceeds the value of the normal curve in area $\alpha$ shown in Figure 4, the moment ratio becomes $A > 0$. If the value $f(u_i)$ increases in area $\beta$, the moment ratio becomes $A < 0$. If the value $f(u_i)$ increases in the boundary area between $\alpha$ and $\beta$ (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 $\gamma$, $A$ 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 Figures 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 Figures 5(a) and (b).

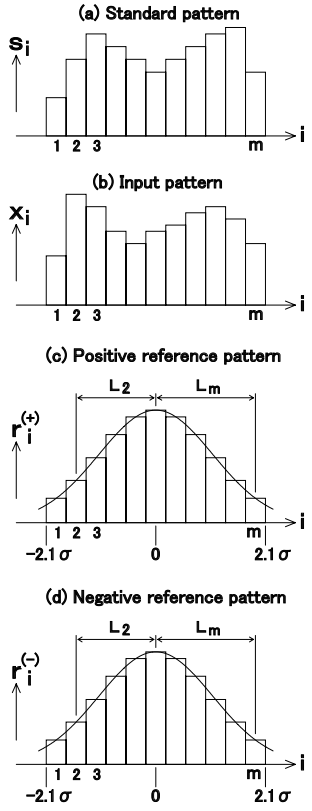


Figure 5. Shape expression of pattern vectors.

## 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. Figures 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 number of components of Eq. (5),

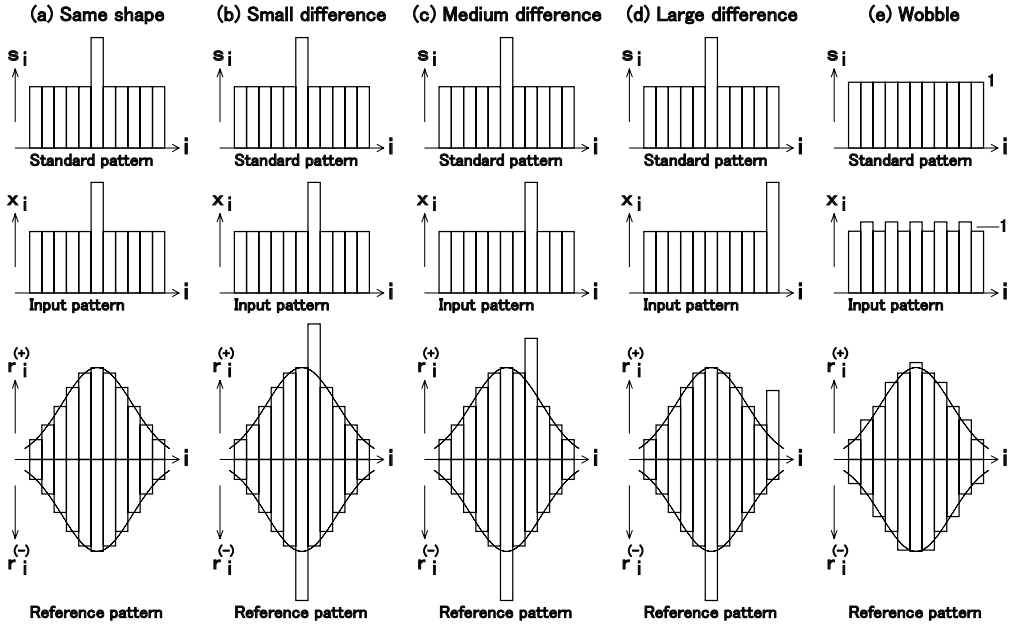


Figure 6. Shape changes of reference patterns.

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

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

and all bar graphs of Figures 5(a)–(d) have the same width. Also, as shown in Figures 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.

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

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). Figure 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 Figure 6.

• Figure 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 Figure 6(a) do not change in their shapes from the normal distribution.

• Figures 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 Figures 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.

• Figure 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 Figure 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).

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

Where, $L_i$ $(i = 1, 2, \cdots, m)$ is a deviation from the center axis of the normal distribution shown in Figures 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 $D$".

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

Figure 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 $\alpha$ shown in Figure 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 $\alpha$, 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 $\alpha$ and $\beta$, 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 $\beta$, 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 Figure 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 Figures 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.

Figures 7(a) and (b) give an example of standard and input patterns. Also, Figures 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 Figures 7(a)–(f) have the same width. Here, as shown in Figures 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 Figure 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 Figures 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 Figures 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.

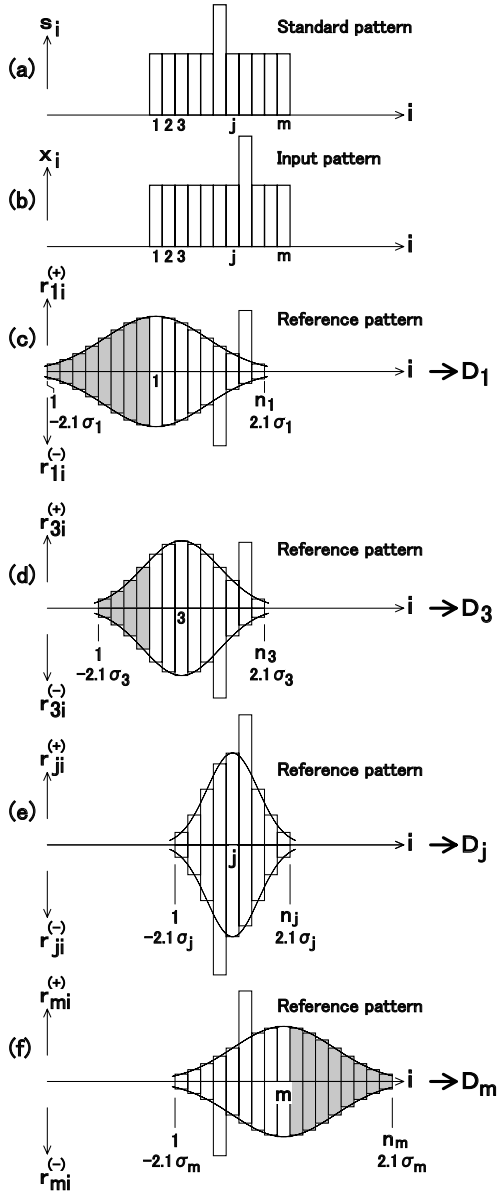


Figure 7. Movement of normal distribution.

In the proposed algorithm, the values $n_j$ and $\sigma_j$ 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.

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

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).

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

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. Figure 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 Figures 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

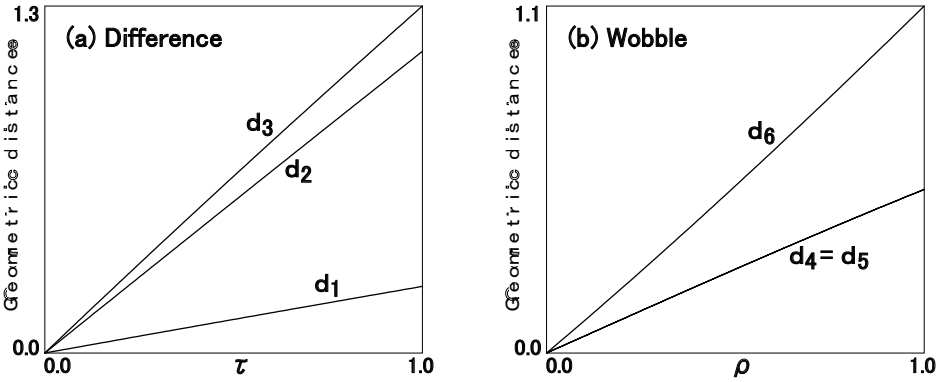


Figure 8. Calculation in geometric distance.

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 $d \approx 0$.

## 3.9 Numerical experiments of geometric distance

To confirm that the geometric distance algorithm matches the mathematical model that we have assumed in Chapter 1, we performed numerical experiments to calculate the geometric distances of the standard and input patterns shown in Figure 1. However, we have developed

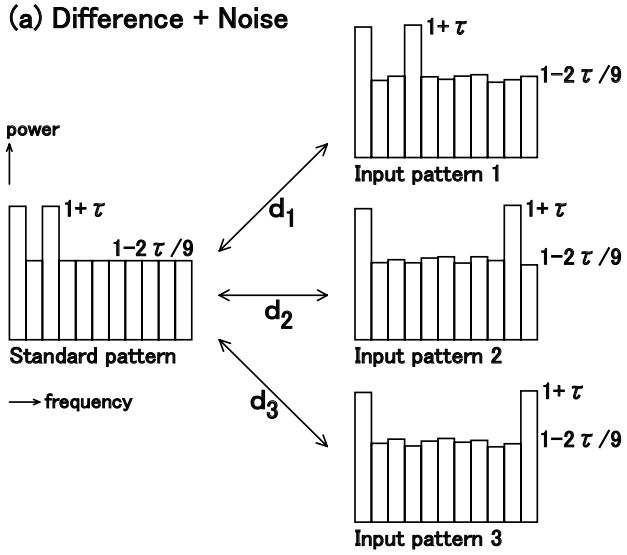


Figure 9. Typical example of standard and input patterns.

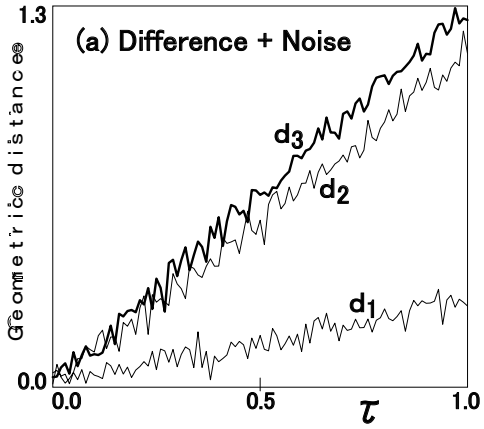


Figure 10. Calculation in geometric distance.

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 Figure 1 as geometric distances $d_1$ to $d_6$ respectively.

Figure 8(a) shows the calculation result of geometric distances $d_1$, $d_2$ and $d_3$ by increasing value $\tau$ from 0.0 to 1.0 in Figure 1(a). From Figure 8(a), if value $\tau$ is fixed, we can determine that the geometric distance increases monotonically according to the increase of the "difference" of the input pattern peak. Figure 8(b) shows the calculation result of geometric distances $d_4$, $d_5$ and $d_6$ by increasing value $\rho$ from 0.0 to 1.0 in Figure 1(b). In Figure 8(b), if value $\rho$ 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, Figure 9 shows input patterns 1, 2 and 3 of Figure 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 $\tau$. Figure 9 is developed by assuming that $\tau = 0.5$. Figure 10 shows that effect on geometric distances $d_1$, $d_2$ and $d_3$ of increasing value $\tau$ from 0.0 to 1.0 in Figure 9. Note that we have set to change the random number if value $\tau$ changes. From Figure 10, if value $\tau$ is fixed within the range of $0.5 \leq \tau$, it is discovered that the geometric distance increases 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 Figures 8(a), (b) and Figure 10, we could verify that the geometric distance algorithm matches the characteristics <1> and <2> of the mathematical model.

## 3.10 Calculation of median

Figure 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.

$$\bar{d}_i = (\sum_j d_{ij})/4 \qquad (15)$$
$$(i = 1, 2, 3, 4, 5;\ j = 1, 2, 3, 4, 5;\ i \neq j)$$

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

In Figure 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

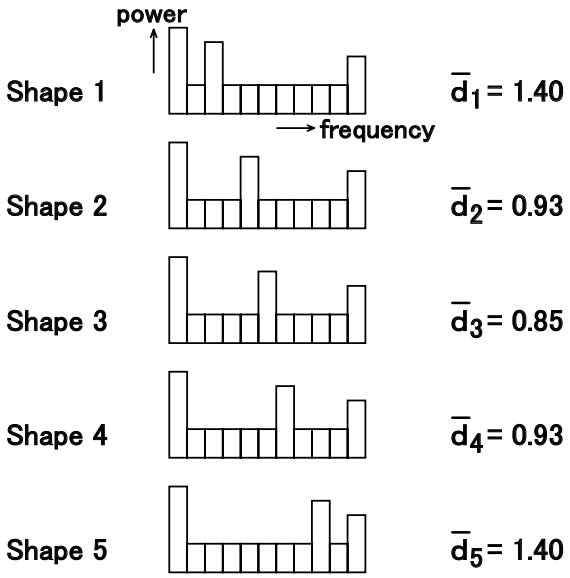


Figure 11. Example of calculation for median.

geometric distance, to be the median. Figure 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 Figure 11 matches the characteristic <2> of the mathematical model.

# Chapter 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 $\times$ 72 sounds $\times$ 4 noises $\times$3 SNRs" has been created. These voice data are also similarly referred to as "/a/01Babble20dB" to "/o/72Subway5dB".

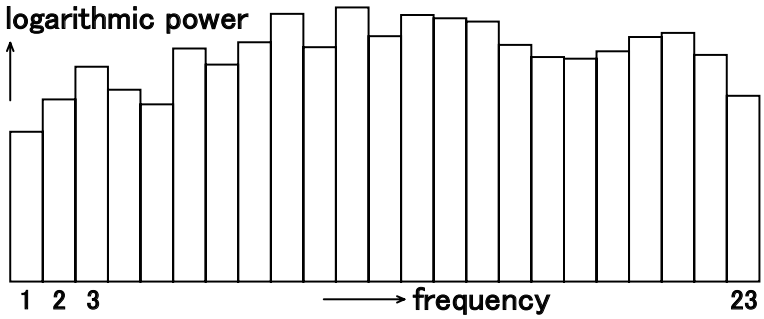


Figure 12. 23-rd dimensional power spectrum of vowel /a/.

## 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".

Figure 12 gives an example of the 23-rd dimensional power spectrum that has been extracted from the "clean vowel /a/". The power spectrum of Figure 12 has $m = 23$ in Eq.

(5), and the standard and input patterns are created based on this value. Note that Figure 12 is referred to as the "1-frame power spectrum" in this paper.

## 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 $\sigma^2$) 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 Figure 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 Figure 7(e), if the value $n_j$ is changed for each 2, the value $\sigma_j$ changes as a discrete value for each $1/(2.1m)$. Thus, if value $m$ is small, the accuracy of optimum value $\sigma_j$ drops. In order to improve the accuracy, we subdivide the bar graph of the positive and negative reference patterns.

Figures 13(a) and (b) show a typical example of bar graph of the standard and input patterns consisting of the 23 bars. Figures 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 Figure 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 Figures 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 Figures 13(c) and (d), for example, the bar graph is structured with 20.2 bars (where, $\omega = 20.2$). This $\omega$ is the number of white bar graphs of the positive and negative reference patterns. In Figure 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 $\sigma_j$ changes as a discrete value for each $0.1/(2.1m)$. The accuracy of the optimum value $\sigma_j$ is improved.

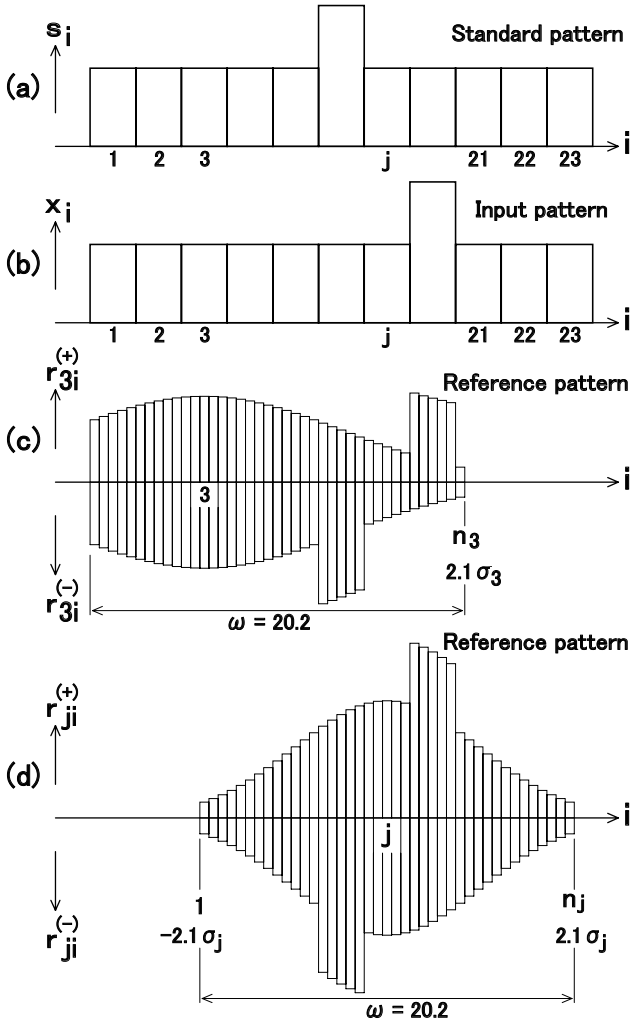


Figure 13. Subdivision of reference patterns.

### 4.3.2 Optimization of $\sigma$

Figure 14, Figure 15 and Table 3 show the processing procedure to determine the optimum value of $\sigma$ (the optimum value of $\omega$) using the "vowel in the continuous speech". Figure 14 is a flowchart used to determine the optimum value by scanning value $\omega$ in the range of 3.0 to 23.0. In Step 1 of Figure 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 Figure 13 are created according to the $\omega$ set value. Then, we explain Steps 3–7 by referring to Figure 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 Figure 15 shows the process required to determine

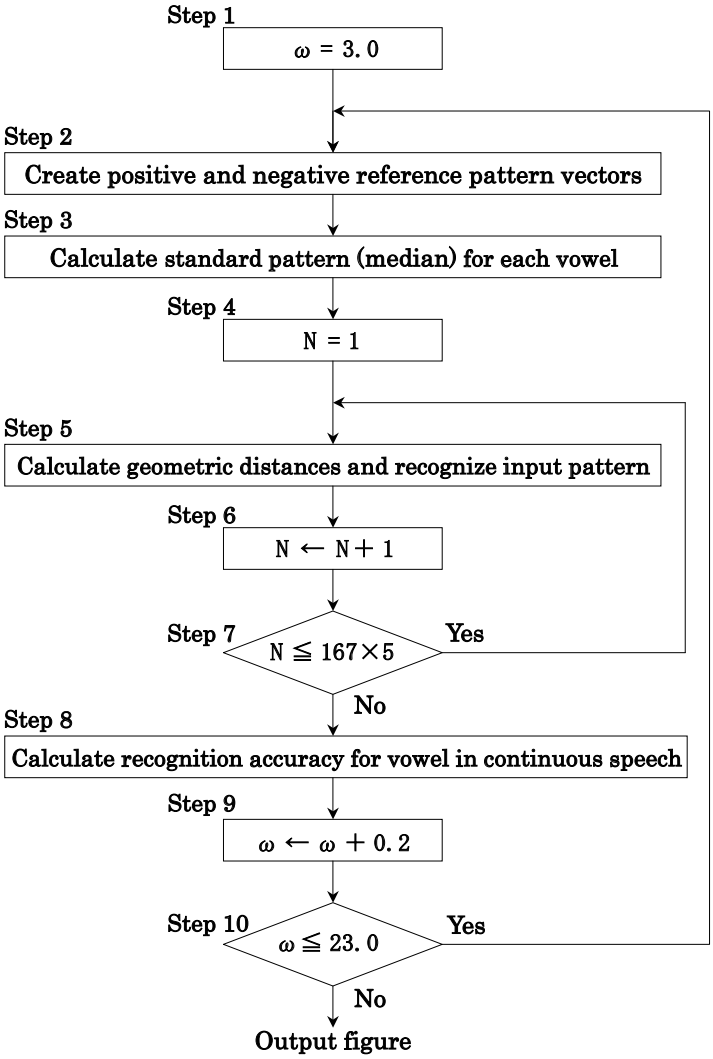


Figure 14. Flowchart for optimizing normal distribution.

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 the input patterns. In Step 4, $N = 1$ is set as the initial value and, as shown in Step 4 of Figure 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 Figure 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.

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 |

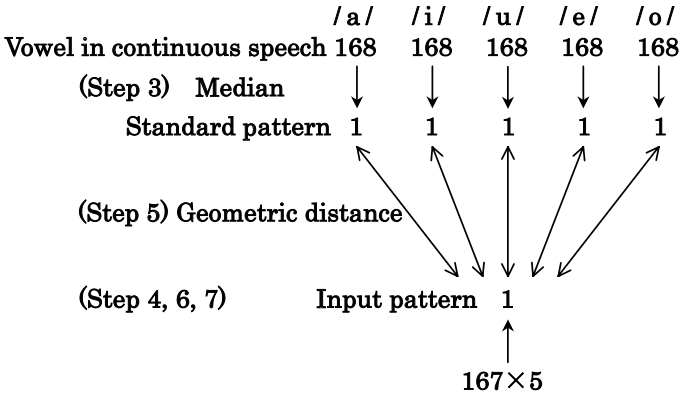


Figure 15. Diagram for optimizing normal distribution.

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 $\omega$ is incremented by 0.2 until it reaches 23.0, and the process of Steps 2–8 is repeated.

Figure 18 shows the relationship between the value $\omega$ and the recognition accuracy obtained by the above process. From Figure 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. Figure 16, Figure 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

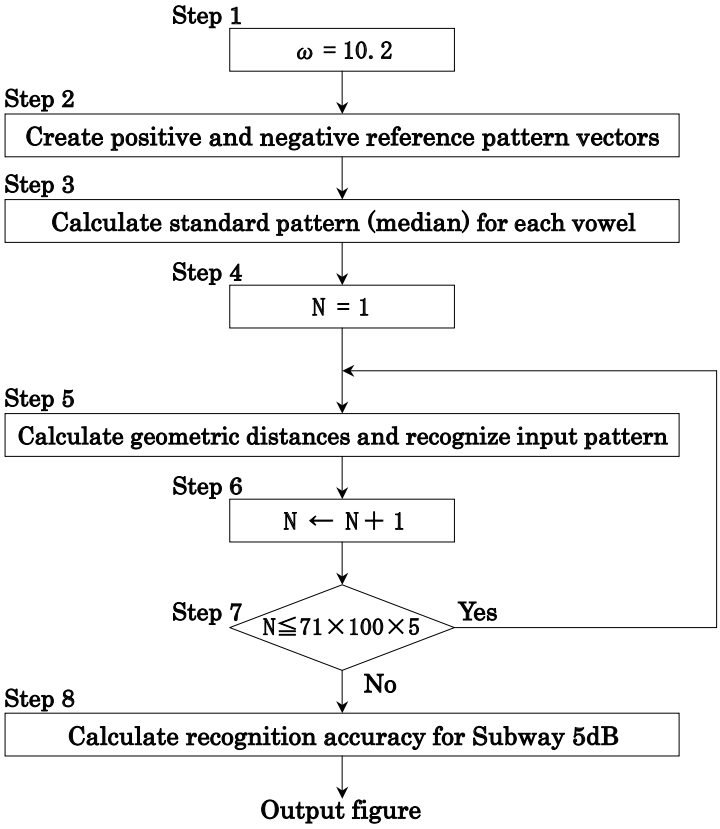


Figure 16. Flowchart for vowel recognition.

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 Figure 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. Figures 16 and 17 show the procedure for evaluation, by using both 5 standard patterns obtained from 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 Figure 16 and Steps 3–7 of Figure 17, the same process is executed as those of Figures 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.

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 |

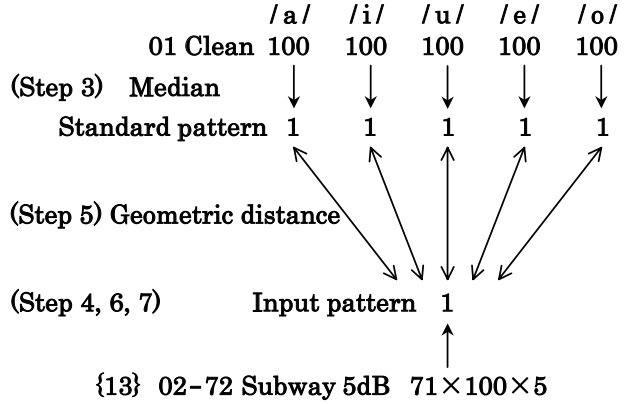


Figure 17. Diagram for vowel recognition.

### 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 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% |

## 4.5 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.

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

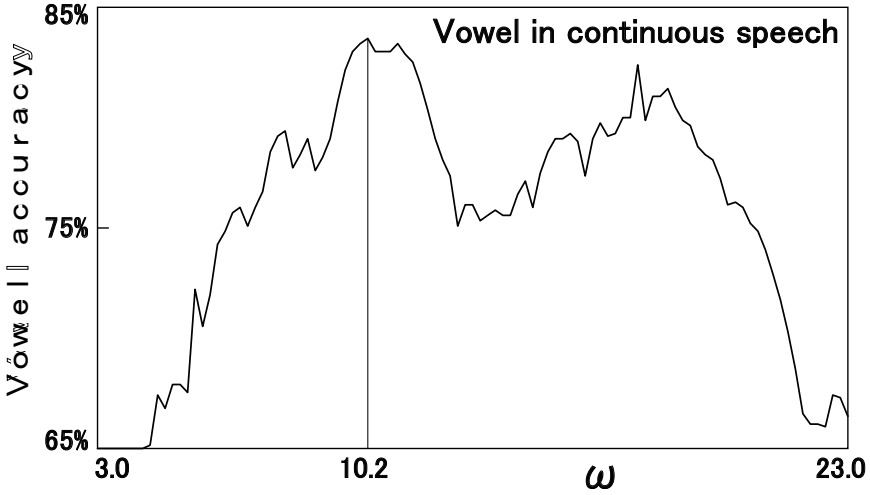


Figure 18. Vowel recognition accuracy and optimum value $\omega$.

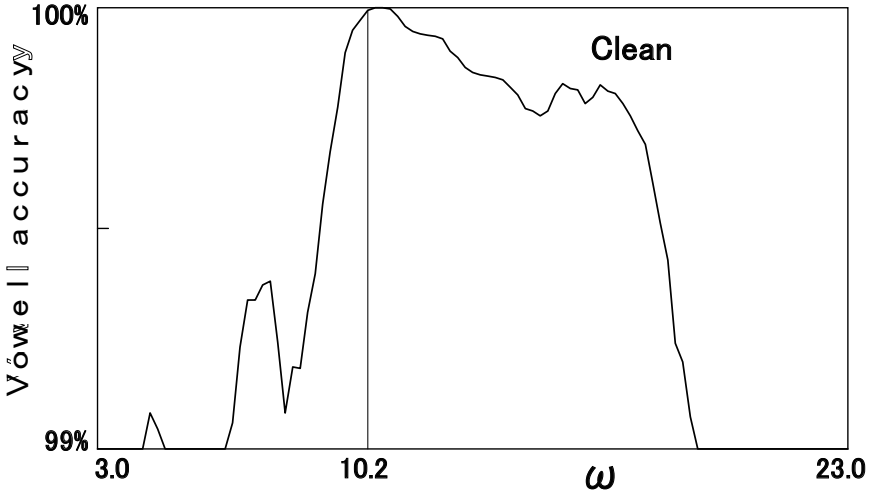


Figure 19. Vowel recognition accuracy with geometric distance.

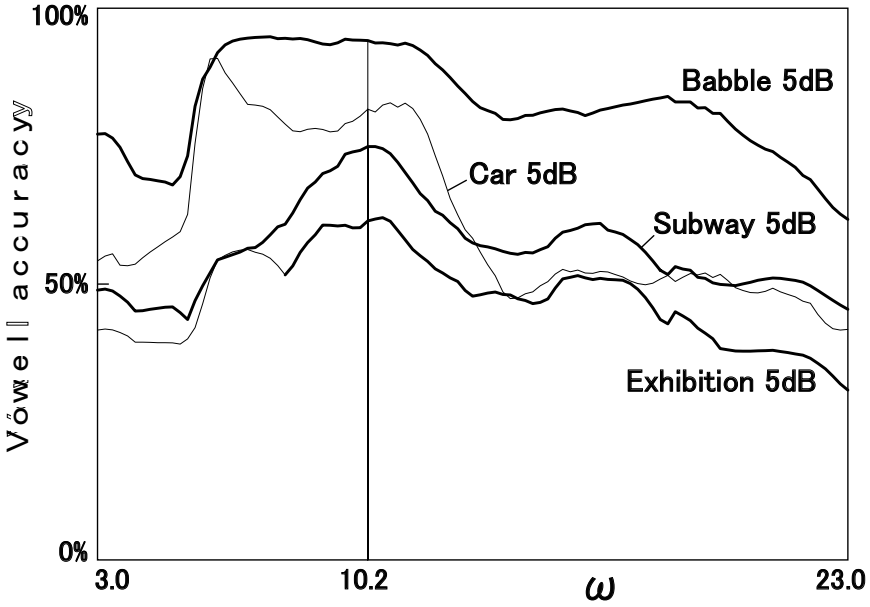


Figure 20. Vowel recognition accuracy with geometric distance.

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

In Subsection 4.3.2, we determine the optimum value $\omega$ 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".

Figure 19 shows the relationship between the value $\omega$ and the recognition accuracy obtained from the "Clean" input patterns. These voice data have the variability with time of 12 weeks. In Figure 19, the recognition accuracy is 100% in part of the $\omega$ value range. From the results of vowel recognition experiments, we have found that the recognition accuracy reaches 100% in the relatively wide $\omega$ 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 $\omega$ 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 are produced in a short period. Therefore, the maximum position of recognition accuracy is most obvious as shown in Figure 18. Thus we use the "vowel in the continuous speech" to determine the optimum value of $\omega$.

# Chapter 5

# Conventional Geometric Distance Algorithm

With the conventional algorithm described in Chapter 3, the standard and input patterns are normalized to have the same area. Then, a difference in shapes between standard and input patterns is replaced by a shape change of a normal distribution. If this method is used, a pseudo difference in shapes may occur between standard and input patterns due to normalization of power spectrum. As an example, Figures 21(a) and (b) show the standard pattern, input patterns 1 and 2 having the same shape in the power spectrum. In the input pattern 2, however, noise has been added to the power spectrum in frequency band $f_N$, and the input pattern 2 has been normalized to have the same area as the standard pattern. As a result, a pseudo difference $\delta$ in shapes occurs at the peaks of the standard pattern and the input pattern 2 as shown in Figure 21(b). Figures 21(c) and (d) show another example of this. However, the input pattern 4 has been normalized to have the same maximum value as the standard pattern. After normalization, a pseudo difference $\delta$ in shapes occurs again at the peaks of the standard pattern and the input pattern 4 as shown in Figure 21(d). Because the pseudo difference in shapes always occurs regardless of the use of any normalization method, it results in an actual shape change of the normal distribution and the recognition performance of geometric distance becomes unpredictable.

Moreover, with the conventional algorithm described in Chapter 3, we need to calculate the moment ratios (shape variation) in each combination of standard and input patterns if

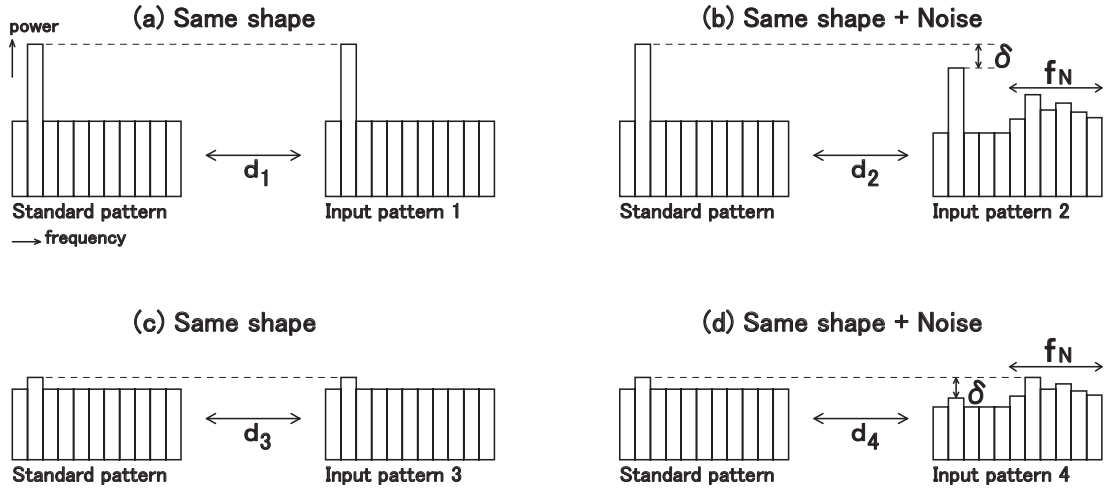


Figure 21. Pseudo difference in shapes.

we use multiple standard patterns and a single input pattern. Hence the processing overhead increases when the number of standard patterns increases. If the calculation of pattern recognition is separated into a standard pattern registration process and an input pattern recognition process, then the moment ratios (shape variation) are calculated during the input pattern recognition process. Therefore, the calculation time of the input pattern recognition process increases in proportion to the number of standard patterns.

However, with the conventional algorithm described in Chapter 3, we need to evaluate positive and negative reference patterns for each movement position of the normal distribution. Therefore, the computational memory overhead increases in proportion to the square of the number of components of the standard and input patterns.

Because of these shortcomings, we propose a new algorithm that we will introduce in the next section.

# Chapter 6

# New Geometric Distance Algorithm

In this section, we use the same mathematical model as the conventional algorithm. We propose a new algorithm that can realize the mathematical model and that can also improve the above shortcomings. Specifically, we use a weighting vector that consists of a rate of change of the moment ratio, and create two weighted pattern vectors by performing the product-sum operation using the weighting vector and the standard pattern vector and the product-sum operation using the weighting vector and the input pattern vector. Then, we use the angle between these weighted pattern vectors as a new geometric distance. As a result, we can remove the pseudo difference in shapes and stabilize the recognition performance of the geometric distance. Also, we can reduce the processing overhead during the input pattern recognition process and reduce the computational memory overhead for the positive and negative reference pattern vectors. 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.

## 6.1 Properties of moment ratio

With the conventional 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. If variable $u_i$ is a discrete value, moment ratio $A$ of function $f(u_i)$ can be calculated using the following equation.

$$A = \frac{\left\{\sum_i f(u_i)\right\} \cdot \left\{\sum_i (u_i)^4 \cdot f(u_i)\right\}}{\left\{\sum_i (u_i)^2 \cdot f(u_i)\right\}^2} - 3 \tag{16}$$

Then, numerical experiments are carried out to study the relationship between moment ratio $A$ and the increment value $\delta$ of bar graphs seen in Figures 22–24. The upper side of graphs (a)–(c) of Figures 22–24 shows the bar graphs each having $m$ bars whose height is the same as function value $f(u_i)$ of the normal distribution. Note that, as described in Section 3.1, $m = 11$ and the bar graphs are created by using the area of $-2.1\sigma \leq u_i \leq 2.1\sigma$ ($\sigma = 1$) of the normal distribution. On bar graphs of Figures 22(a)–(c), only a single bar increases by value

$\delta$ in the center, an intermediate position, and an end of the normal distribution. In Figures 23(a)–(c), two bars of the graph increase by the same value $\delta$. Also, in Figures 24(a)–(c), only one bar increases by value $\delta$ and another bar increases by value 0.2 at the same time. Here, the moment ratio $A$ is calculated using Eq. (16) for the bar graph whose shape is changed as described above. The obtained relationship between values $A$ and $\delta$ is shown by graphs (i) to (ix) in the lower side of graphs (a)–(c) of Figures 22–24.

From graphs (i) to (iii) shown in Figures 22(a)–(c), it is discovered that $A = 0.0$ if $\delta = 0.0$. Also, the value of $A$ changes approximately linearly when value of $\delta$ increases. In Figures 23(a)–(c), graphs (i)+(ii), (ii)+(iii), and (i)+(iii) are the results obtained by addition of graphs (i), (ii) and (iii) respectively. From these graphs, it is discovered that graphs (iv), (v) and (vi) are approximated to respective graphs (i)+(ii), (ii)+(iii), and (i)+(iii). Also, from Figures 24(a)–(c), it is discovered that the gradients of graphs (vii), (viii) and (ix) are equal to those of graphs (i), (ii) and (iii) respectively, and that the intercepts on the vertical axis are equal to the change amounts of moment ratio $A$ if $\delta = 0.2$ on graphs (ii), (iii) and (i) respectively.

From the above description, it is discovered that we can plot approximate graphs (iv) to (ix) using graphs (i) to (iii) if we have already plotted graphs (i) to (iii) using Eq. (16) in advance. In other words, if the rate of change $\mathrm{g}_i$ $(i = 1, 2, \cdots, m)$ of moment ratio $A$ is calculated in advance based on the gradients of graphs (i) to (iii), we can determine the product of $\mathrm{g}_i$ multiplied by $\delta_i$ for each bar graph even when multiple bar graphs change by different values $\delta_i$. Also, we can calculate an approximate value of moment ratio $A$ by summing $\mathrm{g}_i \cdot \delta_i$ for all $i$. This property holds for all values of $m$ and for any variance $\sigma^2$ of the normal distribution.

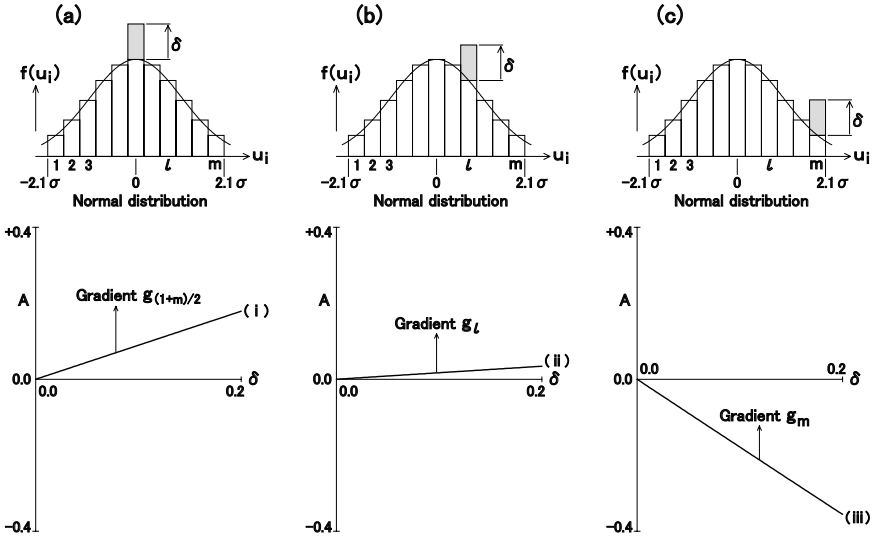


Figure 22. Change of moment ratio $A$.

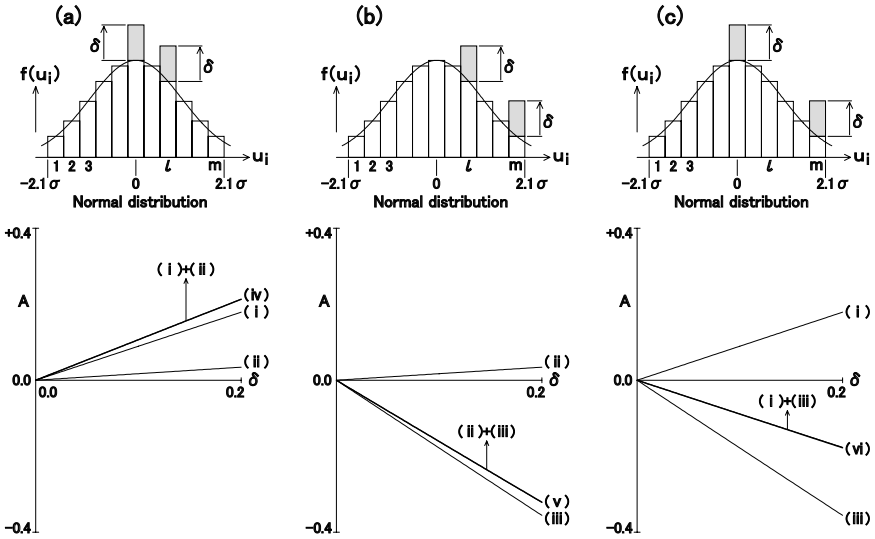


Figure 23. Change of moment ratio $A$.

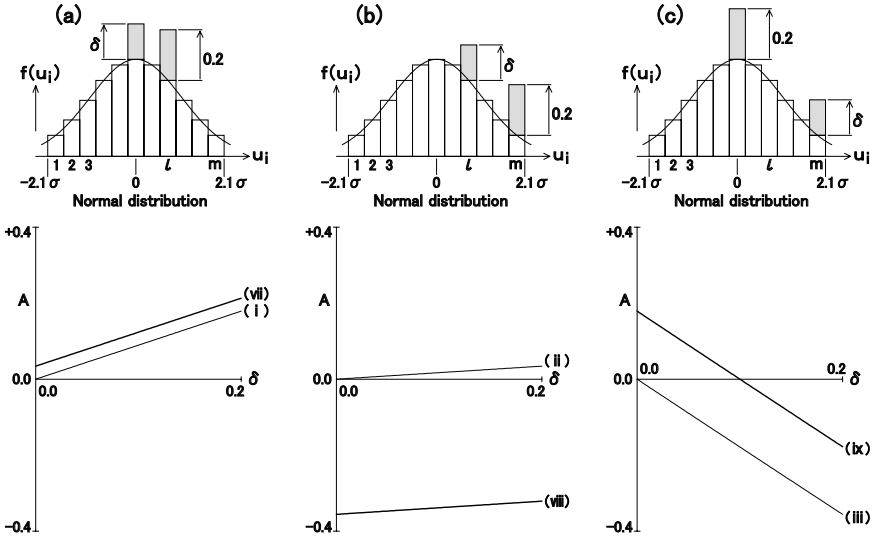


Figure 24. Change of moment ratio $A$.

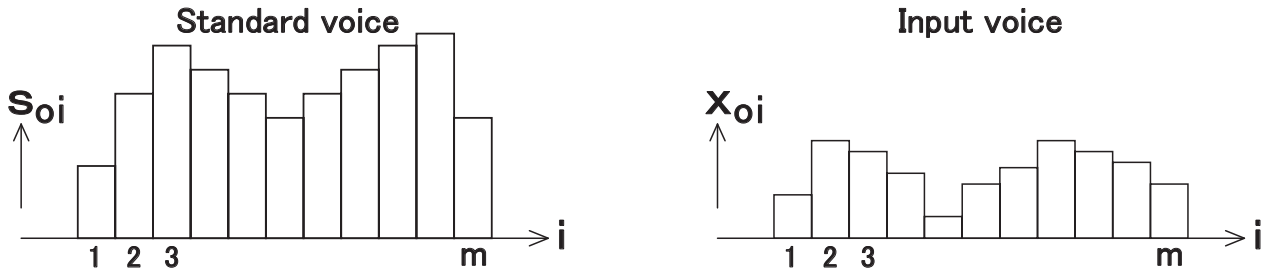


Figure 25. Power spectra of standard voice and input voice.

## 6.2 Creation of pattern vectors

Figure 25 gives an example of the power spectrum of standard and input voices. Note that the power spectrum is generated from the output of filter bank with the $m$ frequency bands (where, $m$ is an odd number). If the $i$-th power spectrum values (where, $i = 1, 2, \cdots, m$) of standard and input voices are $s_{oi}$ and $x_{oi}$ respectively, we create an original standard pattern vector $\boldsymbol{s_o}$ having $s_{oi}$ components, and an original input pattern vector $\boldsymbol{x_o}$ having $x_{oi}$ components, and represent them as follows. In Eq. (17), the function of "T" means a transposed matrix.

$$\begin{aligned} \boldsymbol{s_o} &= (s_{o1}\,, s_{o2}, \cdots, s_{oi}, \cdots, s_{om}\,)^T \\ \boldsymbol{x_o} &= (x_{o1}, x_{o2}, \cdots, x_{oi}, \cdots, x_{om})^T \end{aligned} \tag{17}$$

Moreover, the component values $s_{oi}$ and $x_{oi}$ are divided by the summation of $s_{oi}$ and the summation of $x_{oi}$ respectively, and normalized power spectra $s_i$ and $x_i$ have been calculated. Then, 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.

$$\begin{aligned} \boldsymbol{s} &= (s_1\,, s_2, \cdots, s_i, \cdots, s_m\,)^T \\ \boldsymbol{x} &= (x_1, x_2, \cdots, x_i, \cdots, x_m)^T \end{aligned} \tag{18}$$

If we assign constants $c_s$ and $c_x$ to the summation of $s_{oi}$ and the summation of $x_{oi}$ respectively in Eq. (17), we can show the relationship between component values of Eqs. (17) and (18) as follows.

$$\begin{aligned} s_i &= s_{oi}/c_s \\ x_i &= x_{oi}/c_x \qquad\qquad (i = 1, 2, 3, \cdots, m) \end{aligned} \tag{19}$$

Also, the component values $s_{oi}$ and $x_{oi}$ are divided by the maximum value of $s_{oi}$ and the maximum value of $x_{oi}$ respectively, and normalized power spectra $s'_i$ and $x'_i$ have been calculated.

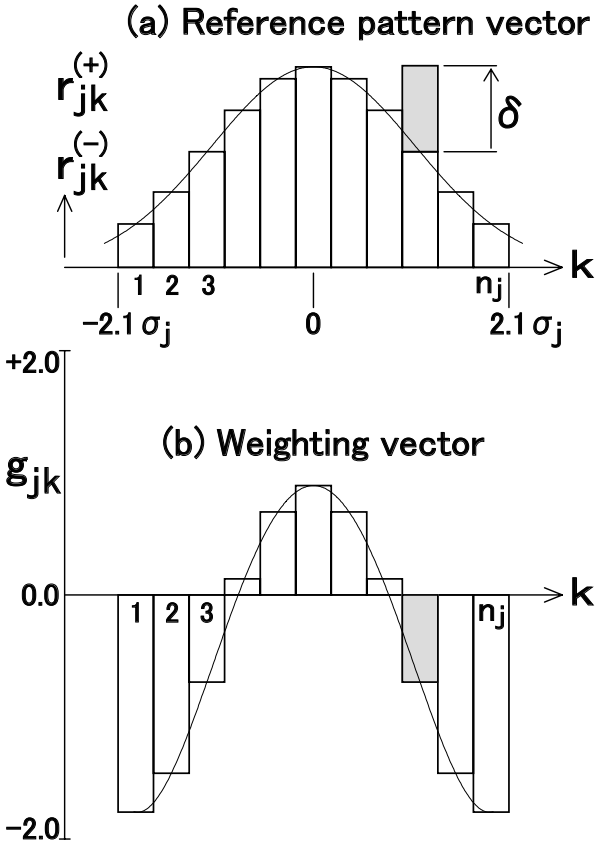


Figure 26. Creating weighting vector.

Then, 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.

$$\begin{aligned} \boldsymbol{s'} &= (s'_1, s'_2, \cdots, s'_i, \cdots, s'_m)^T \\ \boldsymbol{x'} &= (x'_1, x'_2, \cdots, x'_i, \cdots, x'_m)^T \end{aligned} \tag{20}$$

If we assign constants $c'_s$ and $c'_x$ to the maximum value of $s_{oi}$ and the maximum value of $x_{oi}$ respectively in Eq. (17), we can show the relationship between component values of Eqs. (17) and (20) as follows.

$$\begin{aligned} s'_i &= s_{oi}/c'_s \\ x'_i &= x_{oi}/c'_x \qquad\qquad (i = 1, 2, 3, \cdots, m) \end{aligned} \tag{21}$$

Eqs. (17),(18) and (20) express 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 power spectrum shown in Figure 25. The area and the maximum values usually differ between $\boldsymbol{s_o}$ and $\boldsymbol{x_o}$ shown in Figure 25. Meanwhile, the area of $\boldsymbol{s}$ and $\boldsymbol{x}$ are the same and the maximum values of $\boldsymbol{s'}$ and $\boldsymbol{x'}$ are the same.

## 6.3 Creation of weighting vector

From the conventional algorithm, as shown in Figure 26(a), we created positive and negative reference pattern vectors $\boldsymbol{r}_j^{(+)}$ and $\boldsymbol{r}_j^{(-)}$ having function values $r_{jk}^{(+)}$ and $r_{jk}^{(-)}$ of the normal distribution as components for each movement position $j$, and represented them as follows.

$$\begin{aligned} \boldsymbol{r}_j^{(+)} &= (r_{j1}^{(+)}, r_{j2}^{(+)}, \cdots, r_{jk}^{(+)}, \cdots, r_{jn_j}^{(+)})^T \\ \boldsymbol{r}_j^{(-)} &= (r_{j1}^{(-)}, r_{j2}^{(-)}, \cdots, r_{jk}^{(-)}, \cdots, r_{jn_j}^{(-)})^T \\ & \qquad\qquad (j = 1, 2, 3, \cdots, m) \end{aligned} \tag{22}$$

Figures 26(a) and (b) show the rate of change of $A$ ($\mathrm{g}_{jk}$, where a change of $\delta$ occurs at the $k$-th position, $k = 1, 2, \cdots, n_j$ ) for a normal distribution and a single instance of $\delta$. Note that each bar graph has $n_j$ bars. The rate of change $\mathrm{g}_{jk}$ is described by the following equation.

$$\begin{aligned} \mathrm{g}_{jk} = A/\delta \qquad & (k = 1, 2, 3, \cdots, n_j\,) \\ & (j = 1, 2, 3, \cdots, m\,) \end{aligned} \tag{23}$$

The $\mathrm{g}_{j(1+n_j)/2}$, $\mathrm{g}_{jl}$ and $\mathrm{g}_{jn_j}$ correspond to the gradients of respective graphs shown in the lower side of Figures 22(a)–(c). Next, in Figure 26(a), position $k$ of the bar that has increased by value $\delta$ is scanned from 1 to $n_j$, and Eq. (23) is calculated. Figure 26(b) shows a bar graph of the calculated value $\mathrm{g}_{jk}$, where $\delta = 0.2$. Here, we create a weighting vector $\mathbf{g}_j$ having $\mathrm{g}_{jk}$ components, and represent it as follows.

$$\begin{aligned} \mathbf{g}_j &= (\mathrm{g}_{j1}, \mathrm{g}_{j2}, \cdots, \mathrm{g}_{jk}, \cdots, \mathrm{g}_{jn_j})^T \\ & \qquad\qquad (j = 1, 2, 3, \cdots, m) \end{aligned} \tag{24}$$

Eq. (24) expresses the rate of change of moment ratio $A$ by the $m$ pieces of component values of the vector. As $\boldsymbol{r}_j^{(+)}$ and $\boldsymbol{r}_j^{(-)}$ are equivalent vectors in the initial state, the weighting vector calculated from $\boldsymbol{r}_j^{(+)}$ and the weighting vector calculated from $\boldsymbol{r}_j^{(-)}$ are equal to each other. Thus, symbols (+) and (–) are omitted in Eq. (24). Also, the curve shown in Figure 26(b) is the envelope curve of the $\mathrm{g}_{jk}$ bar graph, that has been calculated assuming the value $n_j$ is sufficiently large, and it is called "Weighting curve" in this paper. As shown in Figures 26(a) and (b), the normal curve corresponds to the weighting curve, and the positive and negative reference pattern vectors correspond to the weighting vector.

## 6.4 Approximate calculation of moment ratio

With the conventional algorithm, a difference in shapes between standard pattern vector $\boldsymbol{s}$ and input pattern vector $\boldsymbol{x}$ has been replaced by the shape changes of positive and negative reference pattern vectors $\boldsymbol{r}_j^{(+)}$ and $\boldsymbol{r}_j^{(-)}$ using the following equation.

$$
\begin{aligned}
&For\ \ i=1,2,3,\cdots,m\,;\\
&when\ \ k{=}i{-}j{+}(1{+}n_j)/2\quad (where, 1{\leq}k{\leq}n_j)\,;\\
&\quad \bullet\ if\ x_i > s_i,\ then\ r_{jk}^{(+)} \longleftarrow r_{jk}^{(+)} + |x_i - s_i|\\
&\quad \bullet\ if\ x_i < s_i,\ then\ r_{jk}^{(-)} \longleftarrow r_{jk}^{(-)} + |x_i - s_i|\\
&\qquad\qquad\qquad\qquad (j=1,2,3,\cdots,m)
\end{aligned}
\tag{25}
$$

With the conventional algorithm, moment ratios of $\boldsymbol{r}_j^{(+)}$ and $\boldsymbol{r}_j^{(-)}$, whose shapes have changed according to Eq. (25), have been calculated using the following equation.

$$
\begin{aligned}
A_j^{(+)} &= \frac{\left\{\sum_{k=1}^{n_j} r_{jk}^{(+)}\right\}\cdot\left\{\sum_{k=1}^{n_j}(L_{jk})^4\cdot r_{jk}^{(+)}\right\}}{\left\{\sum_{k=1}^{n_j}(L_{jk})^2\cdot r_{jk}^{(+)}\right\}^2} - 3\\
A_j^{(-)} &= \frac{\left\{\sum_{k=1}^{n_j} r_{jk}^{(-)}\right\}\cdot\left\{\sum_{k=1}^{n_j}(L_{jk})^4\cdot r_{jk}^{(-)}\right\}}{\left\{\sum_{k=1}^{n_j}(L_{jk})^2\cdot r_{jk}^{(-)}\right\}^2} - 3\\
&\qquad\qquad (j=1,2,3,\cdots,m)
\end{aligned}
\tag{26}
$$

In Section 6.1, we determined the product value $\mathrm{g}_{jk}\cdot|x_i - s_i|$ using the rate of change $\mathrm{g}_{jk}$ of moment ratio $A$ and increment $|x_i - s_i|$, and demonstrated that we can calculate the approximate value of the moment ratio $A$ by summing $\mathrm{g}_{jk}\cdot|x_i - s_i|$ for all $i$. Thus, the values $A_j^{(+)}$ and $A_j^{(-)}$ of Eq. (26) can be calculated approximately using the following equation.

$$
\begin{aligned}
&When\ \ k{=}i{-}j{+}(1{+}n_j)/2\quad (where, 1{\leq}k{\leq}n_j)\,;\\
&\quad \bullet\ for\ \ all\ \ i\ \ where\ \ x_i > s_i\\
&\qquad A_j^{(+)} \approx \sum_{i=1}^{m} \mathrm{g}_{jk}\cdot|x_i - s_i|\\
&\quad \bullet\ for\ \ all\ \ i\ \ where\ \ x_i < s_i\\
&\qquad A_j^{(-)} \approx \sum_{i=1}^{m} \mathrm{g}_{jk}\cdot|x_i - s_i|\\
&\qquad\qquad (j=1,2,3,\cdots,m)
\end{aligned}
\tag{27}
$$

If value of $k$ does not satisfy $1 \leq k \leq n_j$, we assume $\mathrm{g}_{jk} = 0$. Next, we consider the signs and replace $|x_i - s_i|$ by $(x_i - s_i)$, and rewrite Eq. (27) as follows.

$$
\begin{aligned}
&When\ \ k{=}i{-}j{+}(1{+}n_j)/2 \quad (where, 1{\leq}k{\leq}n_j)\,; \\
&\quad \bullet\ for\ \ all\ \ i\ \ where\ \ x_i > s_i \\
&\qquad A_j^{(+)} \approx +\sum_{i=1}^{m} \mathrm{g}_{jk} \cdot (x_i - s_i) \\
&\quad \bullet\ for\ \ all\ \ i\ \ where\ \ x_i < s_i \\
&\qquad A_j^{(-)} \approx -\sum_{i=1}^{m} \mathrm{g}_{jk} \cdot (x_i - s_i) \qquad (28) \\
&\qquad\qquad\qquad (j = 1, 2, 3, \cdots, m)
\end{aligned}
$$

The approximate value of moment ratio can be calculated by product-sum operation using Eq. (28), instead of calculating the moment ratio directly using Eq. (26).

## 6.5 Approximate calculation of shape variation

From the conventional algorithm, the difference in shapes between standard and input patterns has been calculated using the following equation, and it has been defined as "Shape variation $D_j$".

$$
D_j = A_j^{(+)} - A_j^{(-)} \qquad (j = 1, 2, 3, \cdots, m) \qquad (29)
$$

Thus, the value $D_j$ of Eq. (29) can be calculated approximately by substituting Eq. (28) into Eq. (29) as follows.

$$
\begin{aligned}
&When\ \ k{=}i{-}j{+}(1{+}n_j)/2 \quad (where, 1{\leq}k{\leq}n_j)\,; \\
&\qquad D_j \approx \sum_{i=1}^{m} \mathrm{g}_{jk} \cdot (x_i - s_i) \\
&\qquad = \sum_{i=1}^{m} \mathrm{g}_{jk} \cdot x_i - \sum_{i=1}^{m} \mathrm{g}_{jk} \cdot s_i \qquad (30) \\
&\qquad\qquad\qquad (j = 1, 2, 3, \cdots, m)
\end{aligned}
$$

From Eq. (30), it is discovered that the value $D_j$ can be separated into the product-sum operation using the component value $\mathrm{g}_{jk}$ of weighting vector and the component value $x_i$ of input pattern vector, and the product-sum operation using the component value $\mathrm{g}_{jk}$ and the component value $s_i$ of standard pattern vector.

## 6.6 Creation of weighted pattern vectors

We assign $s_{g(j)}$ and $x_{g(j)}$ to the two product-sum operations given by Eq. (30) respectively, and represent them as follows.

$$\begin{aligned} When\ \ k=i-j+(1+n_j)/2 \quad (where, 1 \leq k \leq n_j)\,; \\ s_{g(j)} = \sum_{i=1}^{m} g_{jk} \cdot s_i \\ x_{g(j)} = \sum_{i=1}^{m} g_{jk} \cdot x_i \\ (j = 1, 2, 3, \cdots, m) \end{aligned} \tag{31}$$

Then, we create a weighted standard pattern vector $\boldsymbol{s_g}$ having $s_{g(j)}$ components, and a weighted input pattern vector $\boldsymbol{x_g}$ having $x_{g(j)}$ components, and represent them as follows.

$$\begin{aligned} \boldsymbol{s_g} &= (s_{g(1)}, s_{g(2)}, \cdots, s_{g(j)}, \cdots, s_{g(m)})^T \\ \boldsymbol{x_g} &= (x_{g(1)}, x_{g(2)}, \cdots, x_{g(j)}, \cdots, x_{g(m)})^T \end{aligned} \tag{32}$$

From Eqs. (30) and (31), the value $D_j$ can be represented approximately as follows.

$$D_j \approx x_{g(j)} - s_{g(j)} \qquad (j = 1, 2, 3, \cdots, m) \tag{33}$$

From Eq. (33), it is discovered that the value $D_j$ can be obtained by subtracting the component value $s_{g(j)}$ of weighted standard pattern vector from the component value $x_{g(j)}$ of weighted input pattern vector.

## 6.7 Approximate calculation of geometric distance

Using the conventional algorithm, we have calculated the difference in shapes between standard and input patterns using the following equation and we have defined it as the "Geometric distance $d$".

$$d = \sqrt{\sum_{j=1}^{m} (D_j)^2} \tag{34}$$

Thus, the value $d$ of Eq. (34) can be calculated approximately by substituting Eq. (33) into Eq. (34) as follows. Note that $\widetilde{d}$ is an approximate value of the geometric distance $d$.

$$d \approx \sqrt{\sum_{j=1}^{m} (x_{g(j)} - s_{g(j)})^2} = \widetilde{d} \tag{35}$$

As described above, the value $\widetilde{d}$ can be calculated by using Eqs. (18), (24), (31), and (35) sequentially. From Eqs. (31) and (35), we can find that the value $\widetilde{d}$ can be separated into

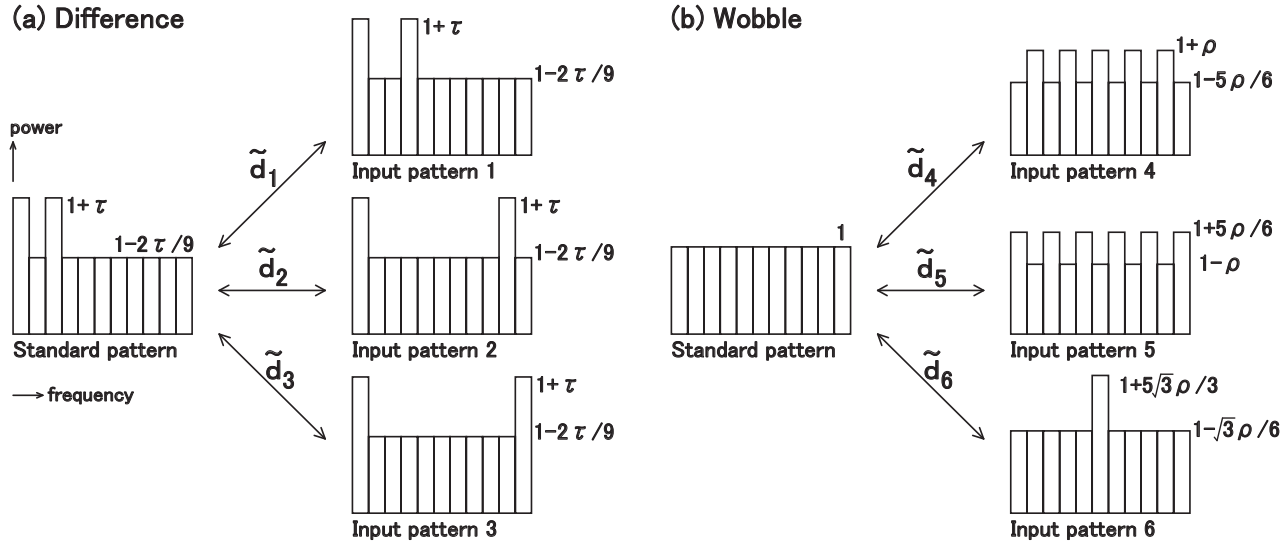


Figure 27. Typical examples of standard and input patterns.

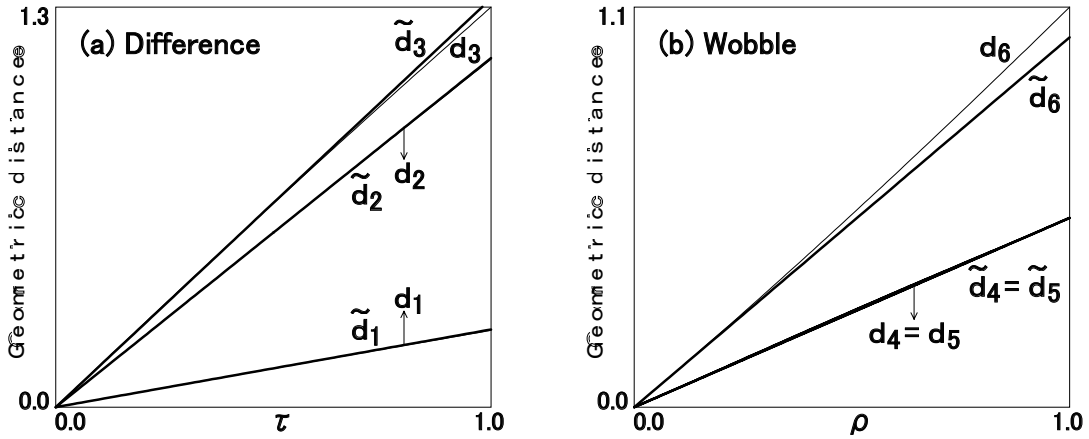


Figure 28. Calculations in geometric distances $d$ and $\widetilde{d}$.

the product-sum operation using the standard pattern vector and the product-sum operation using the input pattern vector.

## 6.8 Numerical experiments of geometric distance $\widetilde{d}$

To confirm the approximation accuracy of $\widetilde{d}$ shown in Eq. (35), we performed numerical experiments to calculate the geometric distances $d_1$ to $d_6$ defined by Eq. (14) and the approximate values $\widetilde{d}_1$ to $\widetilde{d}_6$ of the standard and input patterns shown in Figure 27. However, as described in Section 3.9, we have developed Eqs. (22) and (24) by using values $n_j = 27$ ($\sigma_j = n_j/(4.2m) = 0.58$) that are fixed regardless of movement position $j$ of the normal distribution. Figures 28(a) and (b) show the results of experiments. We can find that values $d_1$ to $d_6$ and $\widetilde{d}_1$ to $\widetilde{d}_6$ are almost identical.

## 6.9 Creation of original and weighted pattern vectors

We assign $s_{og(j)}$ to the product-sum operation using the component value $g_{jk}$ of weighting vector and the component value $s_{oi}$ of original standard pattern vector given by Eq. (17), and assign $x_{og(j)}$ to the product-sum operation using the component value $g_{jk}$ and the component value $x_{oi}$ of original input pattern vector, and represent them as follows.

$$\begin{aligned} &When\ \ k=i-j+(1+n_j)/2 \quad (where, 1\leq k\leq n_j)\,; \\ &\qquad s_{og(j)} = \sum_{i=1}^{m} g_{jk} \cdot s_{oi} \\ &\qquad x_{og(j)} = \sum_{i=1}^{m} g_{jk} \cdot x_{oi} \\ &\qquad\qquad (j = 1, 2, 3, \cdots, m) \end{aligned} \tag{36}$$

Then, we create an original and weighted standard pattern vector $\boldsymbol{s_{og}}$ having $s_{og(j)}$ components, and an original and weighted input pattern vector $\boldsymbol{x_{og}}$ having $x_{og(j)}$ components, and represent them as follows:

$$\begin{aligned} \boldsymbol{s_{og}} &= (s_{og(1)}\,, s_{og(2)}, \cdots, s_{og(j)}, \cdots, s_{og(m)}\,)^T \\ \boldsymbol{x_{og}} &= (x_{og(1)}, x_{og(2)}, \cdots, x_{og(j)}, \cdots, x_{og(m)})^T \end{aligned} \tag{37}$$

Eq. (37) shows the original and weighted pattern vectors that are created without normalization of the power spectrum. Also, we assign $s'_{g(j)}$ to the product-sum operation using $g_{jk}$ and $s'_i$ given by Eq. (20), and assign $x'_{g(j)}$ to the product-sum operation using $g_{jk}$ and $x'_i$, and represent them as follows.

$$\begin{aligned} &When\ \ k=i-j+(1+n_j)/2 \quad (where, 1\leq k\leq n_j)\,; \\ &\qquad s'_{g(j)} = \sum_{i=1}^{m} g_{jk} \cdot s'_i \\ &\qquad x'_{g(j)} = \sum_{i=1}^{m} g_{jk} \cdot x'_i \\ &\qquad\qquad (j = 1, 2, 3, \cdots, m) \end{aligned} \tag{38}$$

Then, we create a weighted standard pattern vector $\boldsymbol{s'_g}$ having $s'_{g(j)}$ components, and a weighted input pattern vector $\boldsymbol{x'_g}$ having $x'_{g(j)}$ components, and represent them as follows:

$$\begin{aligned} \boldsymbol{s'_g} &= (s'_{g(1)}\,, s'_{g(2)}, \cdots, s'_{g(j)}, \cdots, s'_{g(m)}\,)^T \\ \boldsymbol{x'_g} &= (x'_{g(1)}, x'_{g(2)}, \cdots, x'_{g(j)}, \cdots, x'_{g(m)})^T \end{aligned} \tag{39}$$

Eq. (39) shows the weighted pattern vectors that are created with normalization of power spectrum using their maximum values.

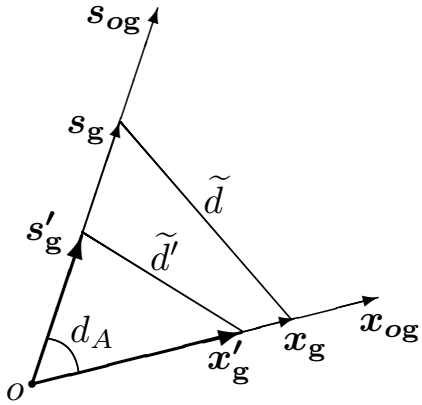


Figure 29. Relationship among weighted pattern vectors.

## 6.10 Relationship among weighted pattern vectors

Eq. (19) is substituted into Eq. (31), and the following equation is obtained using Eq. (36).

$$
\begin{aligned}
&When\ \ k{=}i{-}j{+}(1{+}n_j)/2 \quad (where, 1{\leq}k{\leq}n_j)\,; \\
&\qquad s_{\mathrm{g}(j)} = \sum_{i=1}^{m} \mathrm{g}_{jk} \cdot (s_{oi}/c_s) \\
&\qquad\qquad = s_{o\mathrm{g}(j)}/c_s \\
&\qquad x_{\mathrm{g}(j)} = \sum_{i=1}^{m} \mathrm{g}_{jk} \cdot (x_{oi}/c_x) \\
&\qquad\qquad = x_{o\mathrm{g}(j)}/c_x \qquad\qquad (40) \\
&\qquad\qquad\quad (j = 1, 2, 3, \cdots, m)
\end{aligned}
$$

Similarly, Eq. (21) is substituted into Eq. (38), and the following equation is obtained using Eq. (36).

$$
\begin{aligned}
&s'_{\mathrm{g}(j)} = s_{o\mathrm{g}(j)}/c'_s \\
&x'_{\mathrm{g}(j)} = x_{o\mathrm{g}(j)}/c'_x \qquad (j = 1, 2, 3, \cdots, m) \qquad (41)
\end{aligned}
$$

Figure 29 is a schematic diagram of the $m$-th dimensional pattern space, and it shows six vectors, those are $\boldsymbol{s_{og}}$ and $\boldsymbol{x_{og}}$ given by Eq. (37), $\boldsymbol{s_g}$ and $\boldsymbol{x_g}$ given by Eq. (32), and $\boldsymbol{s'_g}$ and $\boldsymbol{x'_g}$ given by Eq. (39). Note that all vectors begin at origin $o$. From Eq. (40), we can understand that $s_{\mathrm{g}(j)}$ and $s_{o\mathrm{g}(j)}$ are proportional with constant $1/c_s$, and that $x_{\mathrm{g}(j)}$ and $x_{o\mathrm{g}(j)}$ are proportional with constant $1/c_x$. Also, from Eq. (41), we can understand that $s'_{\mathrm{g}(j)}$ and $s_{o\mathrm{g}(j)}$ are proportional with constant $1/c'_s$, and that $x'_{\mathrm{g}(j)}$ and $x_{o\mathrm{g}(j)}$ are proportional with constant $1/c'_x$. Therefore, as shown in Figure 29, vectors $\boldsymbol{s'_g}$, $\boldsymbol{s_g}$ and $\boldsymbol{s_{og}}$ have the same direction. Also, vectors $\boldsymbol{x'_g}$, $\boldsymbol{x_g}$ and $\boldsymbol{x_{og}}$ have the same direction.

## 6.11 Derivation of new geometric distance

From Eq. (35), it is clear that the geometric distance $\widetilde{d}$ can be calculated as the Euclidean distance between the weighted standard pattern vector $\boldsymbol{s}_{\mathbf{g}}$ and the weighted input pattern vector $\boldsymbol{x}_{\mathbf{g}}$. Thus, in Figure 29, we determine the distance between end points of $\boldsymbol{s}_{\mathbf{g}}$ and $\boldsymbol{x}_{\mathbf{g}}$ as value $\widetilde{d}$. Also, if we use Eq. (20) instead of Eq. (18) to determine the standard and input pattern vectors, the geometric distance $\widetilde{d}'$ can be calculated as the Euclidean distance between $\boldsymbol{s}'_{\mathbf{g}}$ and $\boldsymbol{x}'_{\mathbf{g}}$. Thus, in Figure 29, we determine the distance between end points of $\boldsymbol{s}'_{\mathbf{g}}$ and $\boldsymbol{x}'_{\mathbf{g}}$ as value $\widetilde{d}'$. From Figure 29, it is clear that values $\widetilde{d}$ and $\widetilde{d}'$ are changed according to the normalizing method used. To improve on this, we can calculate an angle $d_A$ between $\boldsymbol{s}_{\boldsymbol{og}}$ and $\boldsymbol{x}_{\boldsymbol{og}}$ shown in Figure 29 by the following equation and we define it as the new "Geometric distance $d_A$".

$$\cos(d_A) = \frac{\sum_{j=1}^{m} s_{og(j)} \cdot x_{og(j)}}{\sqrt{\sum_{j=1}^{m} \left(s_{og(j)}\right)^2} \sqrt{\sum_{j=1}^{m} \left(x_{og(j)}\right)^2}} \tag{42}$$

The geometric distance $d_A$ is not affected by the normalizing method used. If $d_A$ is used, we can expect that the shortcoming of the pseudo difference in shapes between the standard and input patterns due to normalization of power spectrum is improved and the recognition performance becomes stable. Therefore, in order to confirm that $d_A$ matches the mathematical model, we perform numerical experiments in Section 6.14. Also, to confirm the stabilized recognition performance of $d_A$, we carry out the speech recognition tests in Chapter 7.

## 6.12 Sharing weighting vector

In Eq. (22), we have created the $m$ pieces of positive and negative reference pattern vectors (normal curves). Figure 30(a) gives an example of three normal curves among these curves. Note that the center axis of the normal curve is drawn in component position $j$. In Eq. (24), we have created the $m$ weighting vectors (weighting curves) from Eq. (22) as shown in Figure 26. The weighting curves created from every normal curve in Figure 30(a) are shown in Figure 30(b). This paper uses a fixed bar width of each graph for both standard and input patterns even when the variance value of the normal distribution has changed. In which case, as shown in Figure 30(b), the maximum and minimum values of those weighting curves are the same respectively, and those weighting curves match when expanded or compressed in the direction of the horizontal axis. Thus, we thought to reduce the computational memory

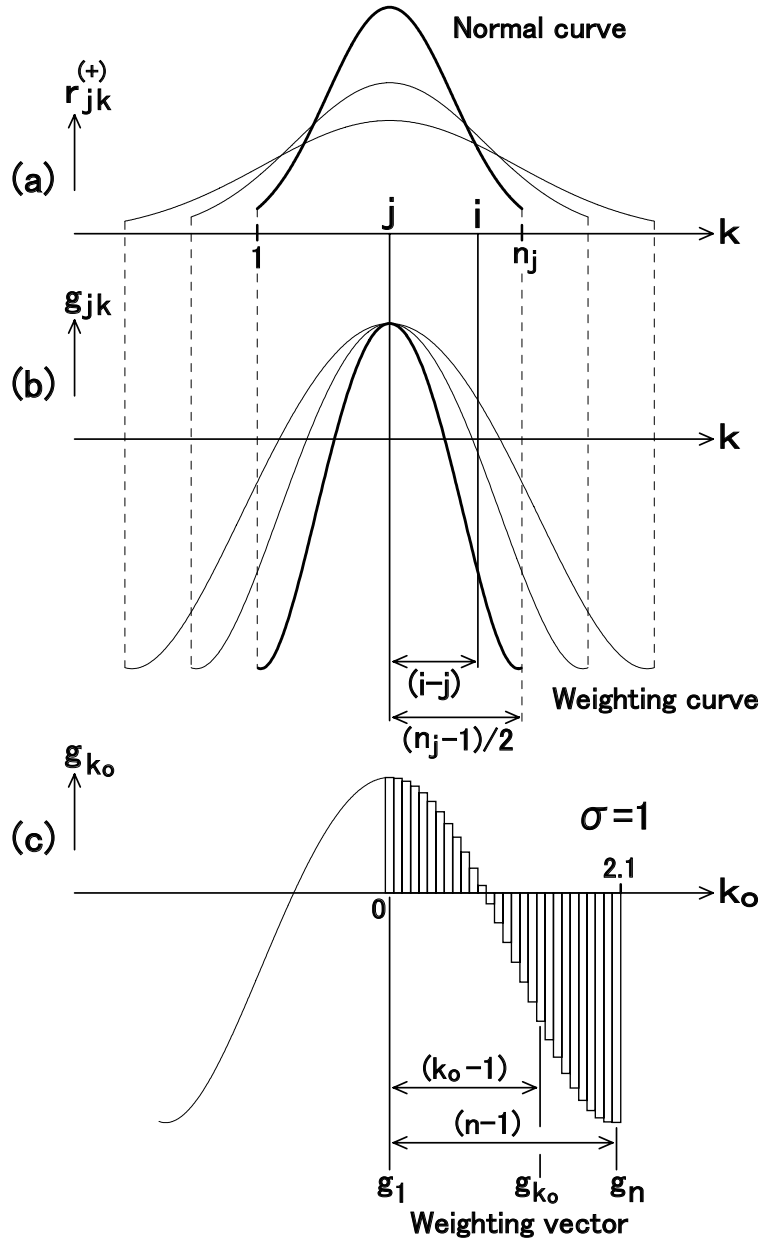


Figure 30. Sharing weighting vector.

overhead by sharing a single weighting vector instead of $m$ vectors. Figure 30(c) shows the weighting curve that has been created from the normal curve of variance $\sigma^2 = 1$. Figure 30(c) also shows a bar graph having the same height as the function value of weighting curve. Here, the right half of the weighting curve is used to create a bar graph for reducing the computational memory overhead. And we create a weighting vector $\mathbf{g}$ having $\mathrm{g}_{k_0}$ (where, $k_0 = 1, 2, \cdots, n$) components whose values are the same as the height of bar graph, and represent it as follows.

$$\mathbf{g} = (\mathrm{g}_1, \mathrm{g}_2, \cdots, \mathrm{g}_{k_0}, \cdots, \mathrm{g}_n, 0, \cdots, 0)^T \tag{43}$$

However, we assume that value $n$ is sufficiently large when compared with the number of

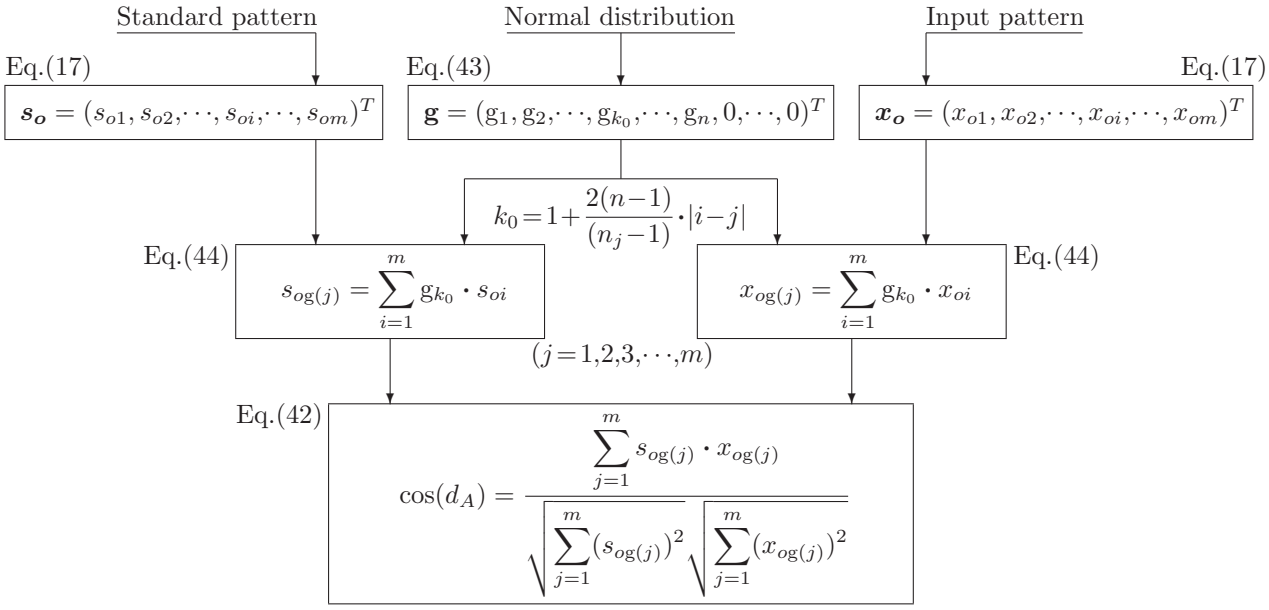


Figure 31. Flowchart for calculating geometric distance $d_A$.

components $n_j$ of Eq. (24). Also, if $n < k_0$, we insert an appropriate number of values $\mathrm{g}_{k_0} = 0$. Eq. (43) is the weighting vector that represents Eq. (24), and Eq. (43) consists of both $n$ components expressing the shape of weighting curve and an appropriate number of component values 0.

As shown by the thick-line weighting curve of Figure 30(b), the difference between component numbers $i$ and $j$ is $(i-j)$ for the weighting vector $\mathbf{g}_j$ given by Eq. (24). The difference between the component number at the center and the component number at the rightmost end position is $(n_j - 1)/2$. On the other hand, as shown in Figure 30(c), the difference between component numbers $k_0$ and 1 is $(k_0 - 1)$ and the difference between component numbers $n$ and 1 is $(n-1)$ for the weighting vector $\mathbf{g}$ given by Eq. (43). As described above, each weighting curve of Figure 30(b) can be obtained by expanding or compressing the weighting curve of Figure 30(c) in the direction of the horizontal axis. Therefore, if the component number $i$ of Figures 30(a) and (b) corresponds to $k_0$ of Figure 30(c), the ratio of $(i-j)$ to $(n_j - 1)/2$ is equal to the ratio of $(k_0 - 1)$ to $(n-1)$. $2(i-j)/(n_j - 1) = (k_0 - 1)/(n-1)$ is satisfied. If we consider that the weighting curve is bilaterally symmetric, we can calculate value $k_0$ using equation $k_0 = 1 + 2|i-j| \cdot (n-1)/(n_j - 1)$. Note that $k_0$ is rounded to an integer value. If value $n$ is sufficiently large, we can reduce the rounding error. In this way, the values $s_{og(j)}$ and $x_{og(j)}$ can be calculated by using the following equation instead of Eq. (36).

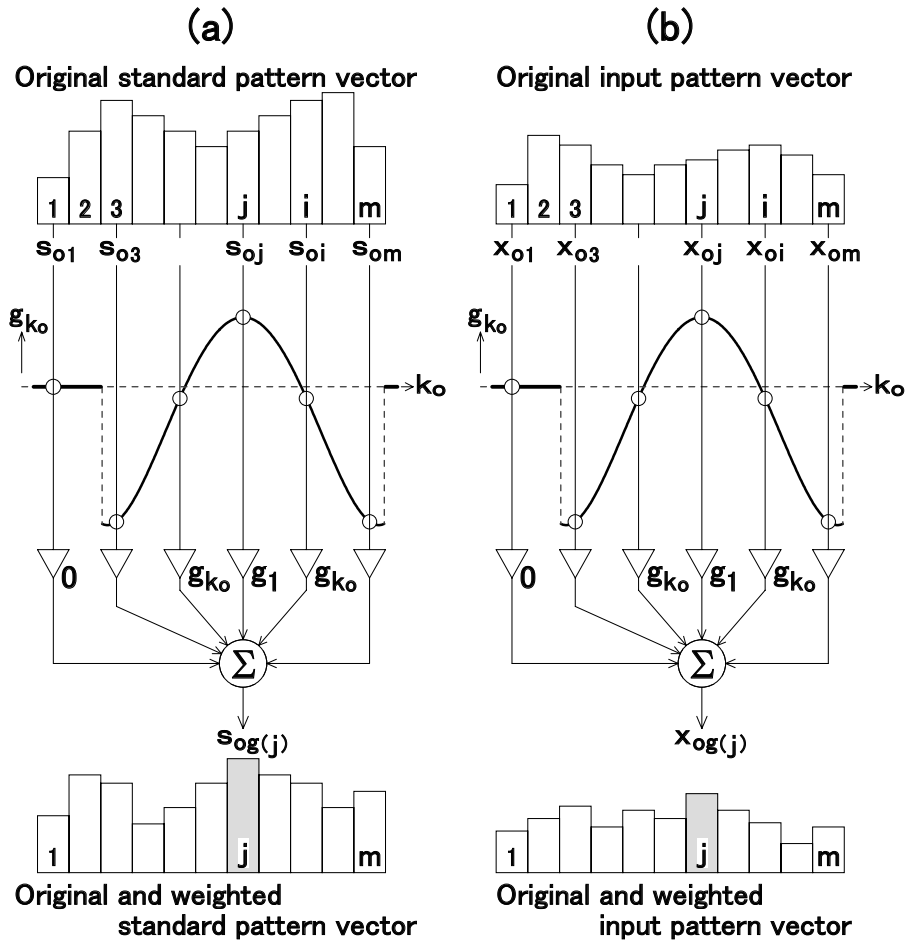


Figure 32. Diagram for calculating product-sum value.

$$
\begin{aligned}
When\ \ k_0 = 1 + \frac{2(n-1)}{(n_j-1)} \cdot |i-j|\,; \\
s_{og(j)} &= \sum_{i=1}^{m} \mathrm{g}_{k_0} \cdot s_{oi} \\
x_{og(j)} &= \sum_{i=1}^{m} \mathrm{g}_{k_0} \cdot x_{oi} \qquad (44) \\
&\qquad (j = 1, 2, 3, \cdots, m)
\end{aligned}
$$

Note that the component number $k$ of Eq. (36) corresponds to $k_0$ of Figure 30(c) or Eq. (44). Using Eq. (44), we can calculate both $s_{og(j)}$ and $x_{og(j)}$ by simply creating a single $\mathbf{g}$ instead of creating $\mathbf{g}_j$ for each movement position $j$ of the normal distribution. In this manner, the computational memory of $\mathbf{g}$ is fixed to the value $n$ in Eq. (43). While in Eq. (24), the memory of $\mathbf{g}_j$ increased in proportion to the square of the value $m$ (in proportion to the value $n_j \times m$). This paper assumes that $n = 2101$. As described above, we can reduce the computational memory overhead by sharing a single weighting vector.

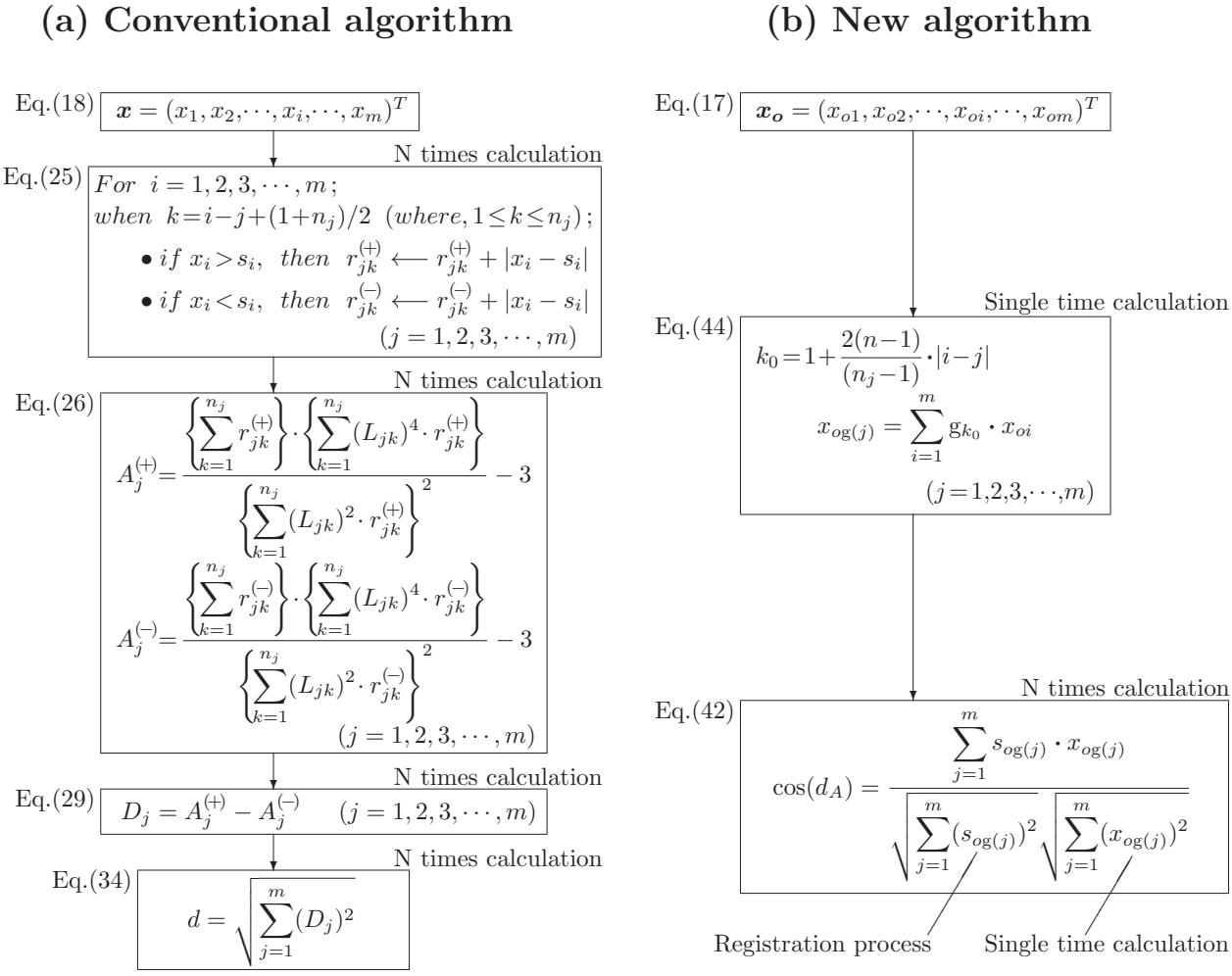


Figure 33. Flowcharts for conventional algorithm and new algorithm during input pattern recognition process.

## 6.13 Procedure for calculating geometric distance $d_A$

Figure 31 shows a flowchart for calculating the new geometric distance $d_A$. From Figure 31, it is clear that we can calculate the value $s_{og(j)}$ in advance during the standard pattern registration process. Moreover, Figures 32(a) and (b) show the flow of product-sum operations given by Eq. (44). Note that the curve in the figure is the weighting curve shown in Figure 30(c), and symbol $\bigtriangledown$ is a multiplier and symbol $\Sigma$ is an adder. In Figure 32(a), by using multiplier $\bigtriangledown$, we calculate the product $g_{k_0} \cdot s_{oi}$ using the component value $g_{k_0}$ of weighting vector and the component value $s_{oi}$ of original standard pattern vector. By using adder $\Sigma$, we calculate the product-sum by addition of the product $g_{k_0} \cdot s_{oi}$ for $i$ (where, $i = 1, 2, \cdots, m$), and use it as the component value $s_{og(j)}$ of original and weighted standard pattern vector. Similarly, in Figure 32(b), we calculate the original and weighted input pattern vector by the product-sum operation using the original input pattern vector and the weighting vector.

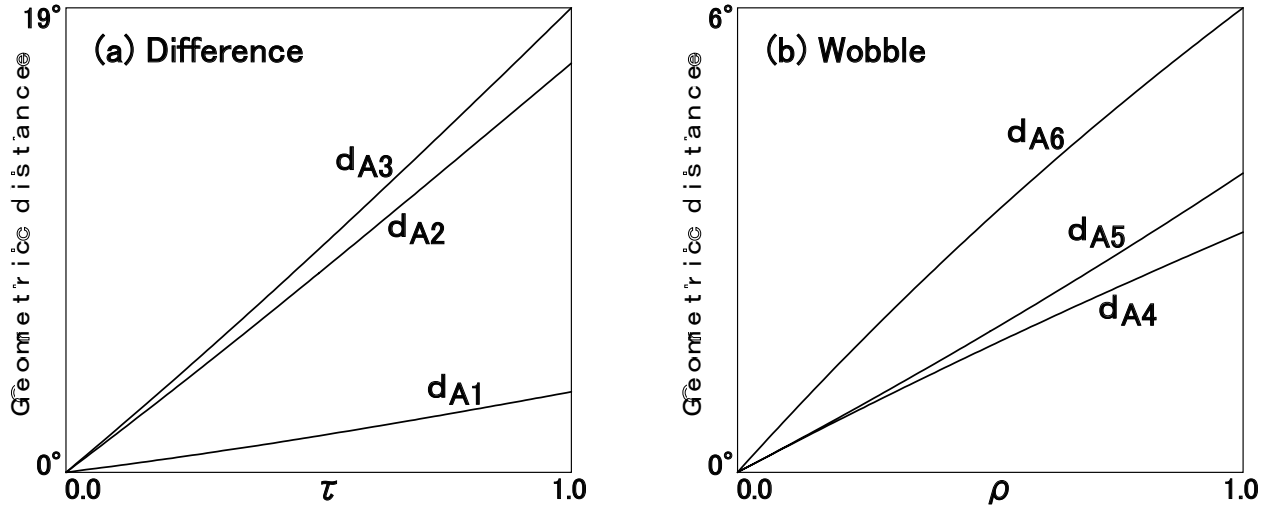


Figure 34. Calculation in geometric distance $d_A$.

From Figures 32(a) and (b), it is discovered that the values $s_{og(j)}$ and $x_{og(j)}$ are calculated from $s_{oi}$ and $x_{oi}$, respectively, by weighting of the weighting curve.

Figures 33(a) and (b) show a comparison between calculation amounts of the conventional algorithm and the new algorithm during the input pattern recognition process. From the conventional algorithm, if we calculate the geometric distances $d$ between $N$ standard patterns and a single input pattern, we need to calculate Eqs. (25), (26), (29) and (34) sequentially in each combination of standard and input patterns during the input pattern recognition process. With the new algorithm, we can obtain the $N$ parts of $d_A$ values by performing a single time calculation of $x_{og(j)}$ and an $N$ times of cosine similarity calculation during the input pattern recognition process. From Figures 33(a) and (b), it is discovered that we can reduce the processing overhead during the input pattern recognition process.

## 6.14 Numerical experiments of geometric distance $d_A$

To confirm that the algorithm of geometric distance $d_A$ matches the mathematical model that we have assumed in Chapter 1, we performed numerical experiments to calculate the geometric distances $d_A$ of the standard and input patterns shown in Figure 27. Note that we used the same $n_j$ value as Section 6.8. Also, note that we read geometric distances $\widetilde{d}_1$ to $\widetilde{d}_6$ in Figure 27 as new geometric distances $d_{A1}$ to $d_{A6}$ respectively.

Figures 34(a) and (b) show the results of experiments. From the figures, we can find that $d_{A4} < d_{A5}$ in Figure 34(b) although $\widetilde{d}_4 = \widetilde{d}_5$ in Figure 28(b). Here, $m = 11$ for the standard and input patterns shown in Figure 27. From the experiments, we found that the larger value was switched between $d_{A4}$ and $d_{A5}$ when value $m$ increased. Also, the two graphs became

close to position $d_{A4}$ shown in Figure 34(b). However, the difference between $d_{A4}$ and $d_{A5}$ is small because we use $m = 23$ in the experiments of vowel recognition performed in the next section. From the numerical experiments described above, we can verify that the algorithm of geometric distance $d_A$ matches the characteristics $< 1 >$ and $< 2 >$ of the mathematical model.

# Chapter 7

# Experiments of Vowel Recognition

To confirm that the geometric distance $d_A$ removes the pseudo difference in shapes and the recognition performance becomes stable, we have performed the speech recognition experiments using the geometric distance $d_A$ and actual voices. We used the same Japanese speech and feature parameters as those used in the experiments with the conventional geometric distance algorithm described in Section 4.2. Similar to the speech recognition experiments of the conventional algorithm, 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.

## 7.1 Variance optimization of normal distribution

Similar to the speech recognition experiments of the conventional algorithm, we determine the optimum value of the variance $\sigma^2$ of the normal distribution (the optimum value of $\omega$) using the "vowel in the continuous speech". This is equivalent to determining the optimum value of the positive and negative reference pattern vectors given by Eq. (22) and to determining the optimum value of the weighting vector $\mathbf{g}_j$ given by Eq. (24). And we convert $\mathbf{g}_j$ into $\mathbf{g}$ as shown in Figures 30(b) and (c) and reduce the computational memory overhead. The value $\omega$ is incremented by 0.2 from 3.0 to 23.0, and the recognition accuracy of the "vowel in the continuous speech" is calculated. Figure 35 shows the relationship between the value $\omega$ and the recognition accuracy obtained by the above process. From Figure 35, it is discovered that the recognition accuracy becomes maximum if $\omega = 11.0$. Thus, we determine $\omega = 11.0$ as the optimum value and use it in the following evaluation experiments.

## 7.2 Evaluation experiments and their results

We have performed the evaluation experiments for the "clean vowel" and the "vowel with noise" using the value $\omega = 11.0$ determined in the previous section. Table 7 shows the result

Table 7. Vowel recognition accuracy with new geometric distance $d_A$. ($\omega = 11.0$)

| | Babble | Car | Exhibition | Subway | Mean |
|---|---|---|---|---|---|
| Clean | | | | | 99.97% |
| SNR 20 dB | 99.93% | 99.88% | 99.22% | 99.49% | 99.63% |
| SNR 10 dB | 98.80% | 98.80% | 88.77% | 93.36% | 94.93% |
| SNR 5 dB | 92.34% | 88.10% | 67.96% | 80.03% | 82.11% |

of vowel recognition using the new geometric distance $d_A$. From Table 7, it is learned that the recognition accuracy with $d_A$ is equalized regardless of noise type[21] when compared with the conventional geometric distance $d$. In particular, the recognition accuracy of "Exhibition5dB" has improved from 61.42% to 67.96%. Also, "mean" of 5 dB SNR has improved from 78.04% to 82.11%. Thus we confirm that the geometric distance $d_A$ removes the pseudo difference in shapes and the recognition performance becomes stable.

## 7.3 Verification of optimum value

Table 7 shows the result of recognition accuracy using the optimum value $\omega = 11.0$ that we have determined from Figure 35. Here, in order to verify that the value $\omega = 11.0$ is truly the optimum value, the value $\omega$ is incremented by 0.2 from 3.0 to 23.0 and the recognition accuracy of the "clean vowel" and the "vowel with noise" is calculated. Figures 36 and 37 show the calculated relationship between the value $\omega$ and the recognition accuracy for the input patterns of the "clean vowel" and the "vowel with 5 dB noise", respectively. From Figures 36 and 37, we can find that the recognition accuracy is almost maximum in the value $\omega = 11.0$.

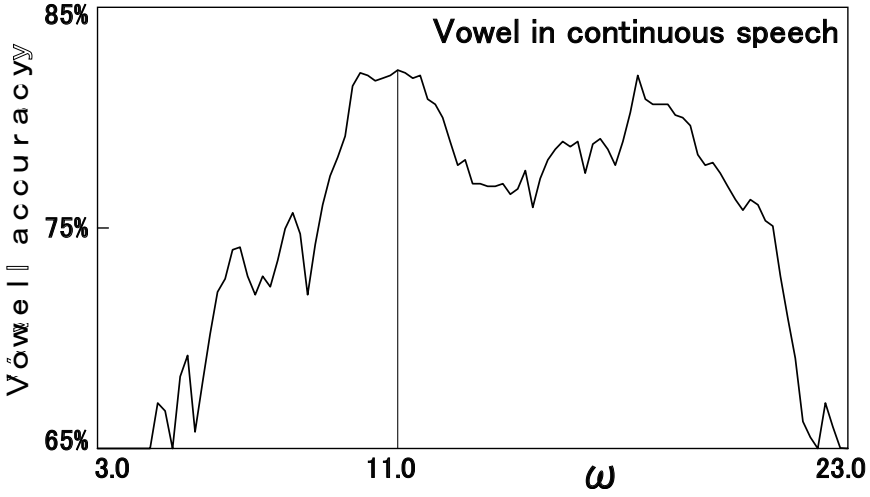


Figure 35. Vowel recognition accuracy and optimum value $\omega$.

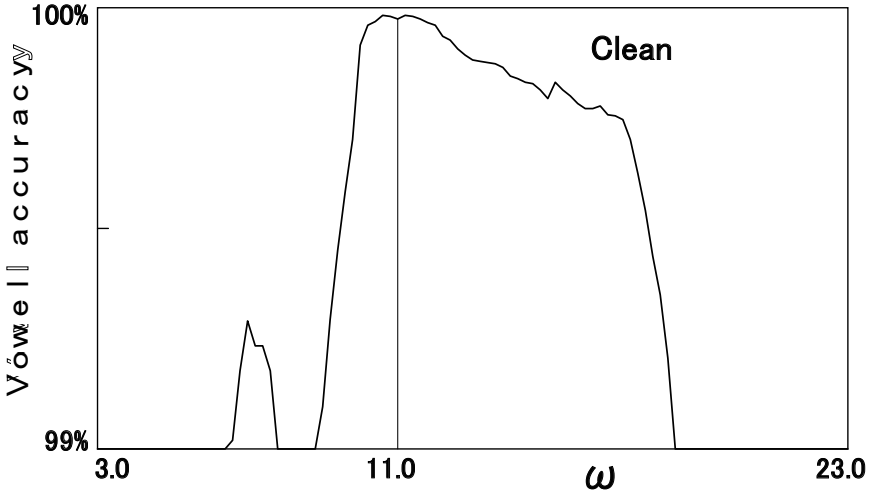


Figure 36. Vowel recognition accuracy with new geometric distance $d_A$.

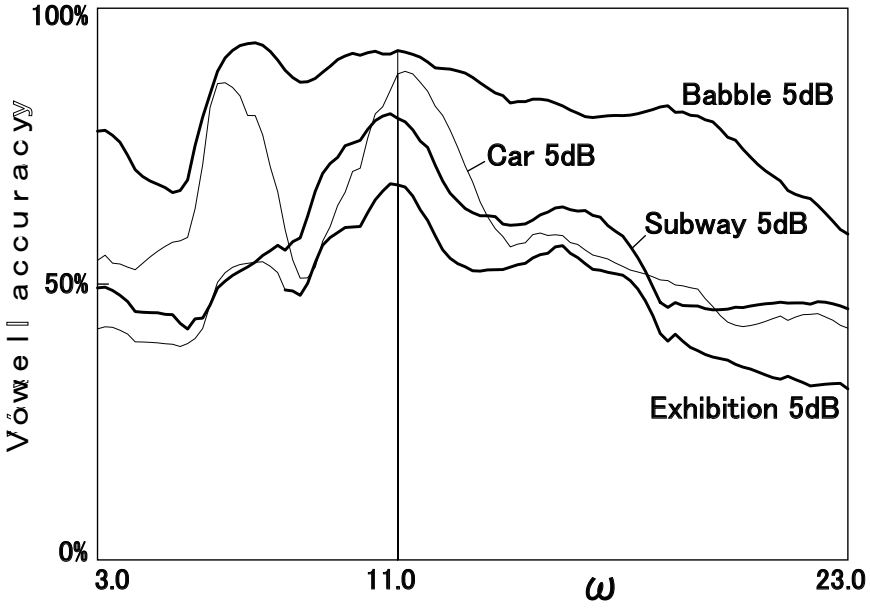


Figure 37. Vowel recognition accuracy with new geometric distance $d_A$.

# Chapter 8

# Conventional Optimization Method

Up to this stage, as shown in Figures 18, 19, 20, 35, 36 and 37, we have checked the relationship between the variance of the normal distribution and the vowel recognition accuracy, using the "clean long vowels having the variability with time of 12 weeks" and the "clean vowels in the continuous speech". From the results of vowel recognition experiments, we have found that the recognition accuracy reaches 100% in a wide variance value range of the normal distribution in the variability with time below 4 weeks if the "clean long vowels having the variability with time" are used. In such a case, we have a problem determining the location of the maximum recognition accuracy. This means that we will find it difficult to determine the optimum variance value of the normal distribution by using the "clean long vowels produced in a short period". Meanwhile, if the "clean vowels in the continuous speech" are used, the power spectrum of the vowel changes minimally even if the voices are produced in a short period. Therefore, the location of the maximum recognition accuracy is most obvious. Owing to the above reason, the conventional optimization method estimates the optimum variance value of the normal distribution using the "clean vowels in the continuous speech". And the evaluation experiments of vowel recognition are performed for the "clean long vowels" and the "long vowels with actual noise" using the estimated value.

However, there is the shortcoming in the above optimization method where the characteristic <1> of the above mathematical model is ignored because only the clean vowels are used. The optimization needs to be made to maximize the effect of the characteristics <1> and <2> of the mathematical model simultaneously. In this case, the shortcoming seemed to be able to be solved by optimization using the "long vowels with actual noise". In other words, optimization is achieved under conditions where the "wobble" caused by the actual noise corresponds to the characteristic <1> of the mathematical model, and the "difference" between the formants of the standard and input patterns corresponds to the characteristic <2>. In this method, however, it is necessary to record all of actual noise in the daily life, create the voice data of long vowels including the actual noise each time the speaker changes, and calculate the optimum value using such voice data. This requires a huge processing overhead, and practical problems remain. As an improvement, we propose a new method that can determine the optimum value with a low processing overhead in the next section.

This method simulates the actual noise in the daily life with a small amount of synthetic noise generated by the computer. Note that the "long vowel" is abbreviated as the "vowel" hereafter.

# Chapter 9

# New Optimization Method

In this section, we have adopted a method to add "wobble" directly to the pattern (the logarithmic power spectrum) whose shape is compared in order to apply the geometric distance to the general pattern recognition. Generally, in the study of speech recognition, the microphone output signal of the actual noise equivalent to the SNR is added to the microphone output signal of the clean vowel, and the voice data is created. Then, this voice data is multiplied by the window function (the "Hamming window" in this research) to calculate the logarithmic power spectrum. If the effect of the window function is considered, this is approximately equivalent to the calculation of the logarithmic power spectrum after adding the power spectrum of the actual noise equivalent to the SNR to the power spectrum of the clean vowel. It is replaced by the direct addition of "wobble" caused by the actual noise to the logarithmic power spectrum of the clean vowel. The proposed method uses weighted random numbers generated by the computer instead of the "wobble" caused by the actual noise. This means that the weighted random numbers generated by the computer are added to the logarithmic power spectrum of the clean vowel and it is used as the input pattern. Also, the logarithmic power spectrum of the clean vowel is used as the standard pattern. In this case, both the characteristics <1> and <2> of the mathematical model are well considered. In this section, we check the relationship between the variance of the normal distribution and the vowel recognition accuracy, using both the standard and input patterns as created above and the algorithm of the geometric distance $d_A$ defined in Eq. (42). Then, we determine the optimum variance value of the normal distribution. In this section, we carry out the optimization experiment using the same voice data as described in Section 4.1.

## 9.1 Difference pattern of actual noise

In order to determine the best weighted random numbers to be added instead of the "wobble" caused by the actual noise, we check the "wobble" of the logarithmic power spectrum caused by the actual noise. An example is shown at the left and center of Figure 38. They are the logarithmic power spectrum arrays of the 23rd dimensional Mel filter bank output (abbreviated as "logarithmic power spectrum" hereafter).[22] Note that the bar graph at the left of Figure 38 shows the logarithmic power spectrum that is extracted from the voice data

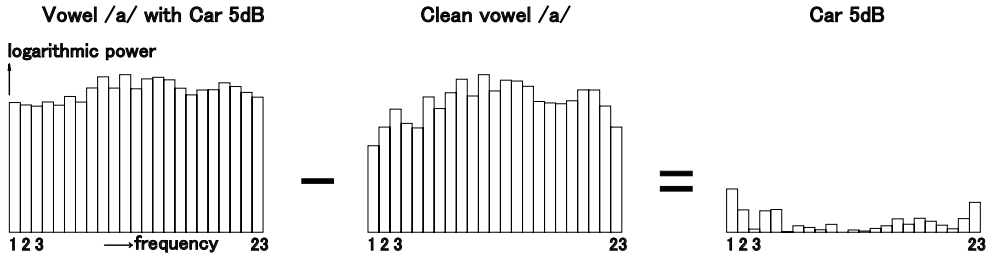


Figure 38. Subtraction of clean vowel from vowel with car 5dB.

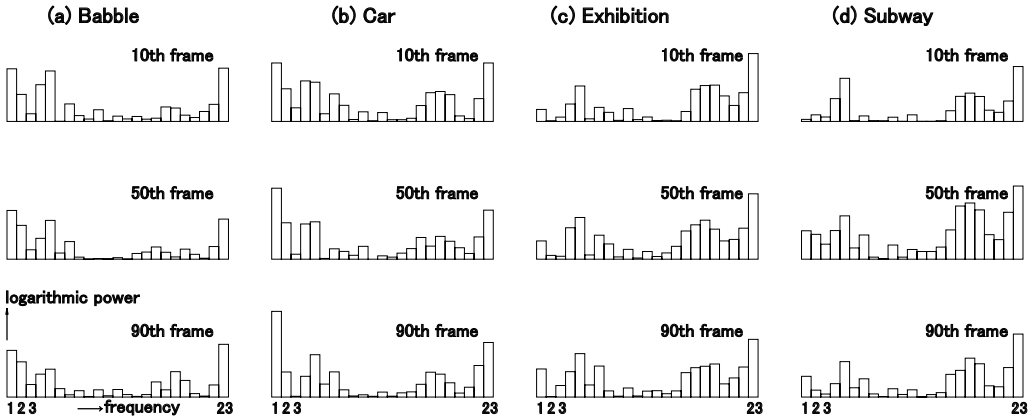


Figure 39. Difference patterns of actual noises.

created by adding the microphone output signal of Car noise equivalent to the SNR of 5 dB to the microphone output signal of the clean vowel /a/. Also, the bar graph at the center of Figure 38 shows the logarithmic power spectrum that is extracted from the clean vowel /a/. Then, the bar graph at the right of Figure 38 shows a difference pattern that is created by subtracting the latter logarithmic power spectrum from the former logarithmic power spectrum. This difference pattern shows the "wobble" of the logarithmic power spectrum caused by the actual noise. Furthermore, Figures 39(a)–(d) show the difference patterns which have been calculated by the above method, using the 10th, 50th and 90th frames of the central 100 frames of the clean vowel /a/ produced for a period of 2 seconds, and using the actual noises of Babble, Car, Exhibition and Subway. From Figure 39, we can understand that the difference pattern of the actual noise changes randomly with time while maintaining a constant shape.

## 9.2 Addition of weighted random numbers

The $m$-th dimensional logarithmic power spectrum of the clean vowel /a/ is shown at the center of Figure 38, where $m = 23$. If the $i$-th logarithmic power spectrum values (where, $i = 1, 2, \cdots, m$) of a clean standard vowel and a clean input vowel are $s_i$ and $x_i$, respectively,

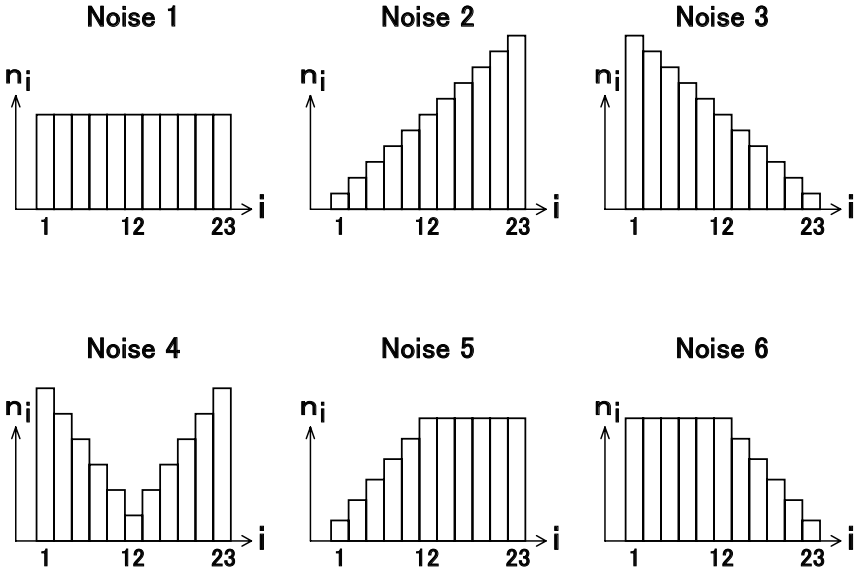


Figure 40. 23rd dimensional noise patterns.

Table 8. Function $n_i$ of Noise 1 to Noise 6.

| | | |
|---|---|---|
| Noise 1 | $n_i = \alpha_1$ | $(\ 1 \leq i \leq 23)$ |
| Noise 2 | $n_i = \alpha_2\, i$ | $(\ 1 \leq i \leq 23)$ |
| Noise 3 | $n_i = \alpha_3\,(24 - i)$ | $(\ 1 \leq i \leq 23)$ |
| Noise 4 | $n_i = \alpha_4\,(13 - i)$<br>$n_i = \alpha_4$<br>$n_i = \alpha_4\,(i - 11)$ | $(\ 1 \leq i \leq 11)$<br>$(i = 12)$<br>$(13 \leq i \leq 23)$ |
| Noise 5 | $n_i = \alpha_5\, i$<br>$n_i = \alpha_5 \times 12$ | $(\ 1 \leq i \leq 11)$<br>$(12 \leq i \leq 23)$ |
| Noise 6 | $n_i = \alpha_6 \times 12$<br>$n_i = \alpha_6\,(24 - i)$ | $(\ 1 \leq i \leq 12)$<br>$(13 \leq i \leq 23)$ |

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. In Eq.(45), the function of "$T$" means a transposed matrix.

$$\begin{aligned} \boldsymbol{s} &= (\ s_1, s_2\ , \cdots, s_i, \cdots, s_m\ )^T \\ \boldsymbol{x} &= (x_1, x_2, \cdots, x_i, \cdots, x_m)^T \end{aligned} \tag{45}$$

Figure 40 shows six types of $m$-th dimensional noise patterns as Noise 1 to Noise 6. They have been generated as a typical example of difference patterns of the actual noise as explained in Figures 39(a)–(d). Also, if the $i$-th value $(i = 1, 2, \cdots, m)$ of the noise pattern shown in Figure 40 is $n_i$, Table 8 shows $n_i$ as the function of $i$. Note that values $\alpha_1$ to $\alpha_6$ are the constants which are calculated by the experiment[1] described in the next section. Here, we

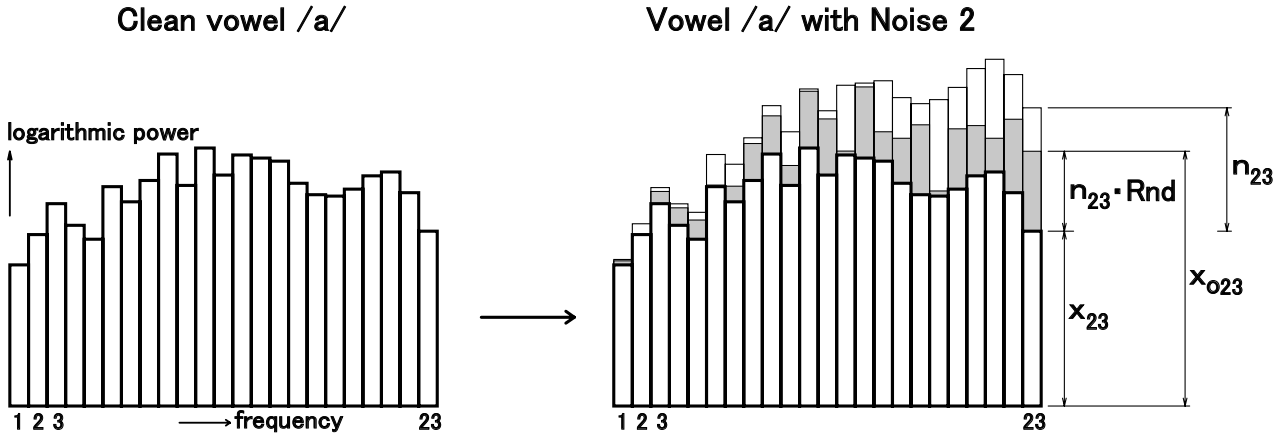


Figure 41. Addition of weighted random numbers to clean vowel.

create a noise pattern vector $\boldsymbol{n}$ having $n_i$ components, and represent it as follows.

$$\boldsymbol{n} = (n_1, n_2, \cdots, n_i, \cdots, n_m)^T \tag{46}$$

Next, if variable $Rnd$ is random numbers uniformly distributed within the range of 0.0 to 1.0, as shown in the following equations, we assign $s_{oi}$ to the component value $s_i$ of standard pattern vector, and assign $x_{oi}$ to the addition of the component value $x_i$ of input pattern vector and the weighted random numbers $n_i \cdot Rnd$.

$$\begin{aligned} s_{oi} &= s_i \\ x_{oi} &= x_i + n_i \cdot Rnd \qquad\qquad (i = 1, 2, 3, \cdots, m) \end{aligned} \tag{47}$$

Then, as shown in Eq. (17), we create an original standard pattern vector $\boldsymbol{s_o}$ having $s_{oi}$ components, and an original input pattern vector $\boldsymbol{x_o}$ having $x_{oi}$ components, and represent them as follows.

$$\begin{aligned} \boldsymbol{s_o} &= (\, s_{o1},\, s_{o2}, \cdots, s_{oi}, \cdots, s_{om}\,)^T \\ \boldsymbol{x_o} &= (x_{o1}, x_{o2}, \cdots, x_{oi}, \cdots, x_{om})^T \end{aligned} \tag{48}$$

$\boldsymbol{s_o}$ is the original standard pattern vector which has been created from the logarithmic power spectrum of clean standard vowel, and $\boldsymbol{x_o}$ is the original input pattern vector which has been created from the logarithmic power spectrum of clean input vowel, added by the weighted random numbers generated by the computer. Figure 41 shows the shape of the second formula of Eq. (47) using the noise pattern of Noise 2. The bar graph at the left of Figure 41 shows the shape of input pattern vector $\boldsymbol{x}$ given by Eq. (45), and the bar graph at the right of Figure 41 shows the shape of original input pattern vector $\boldsymbol{x_o}$ given by Eq. (48).

## 9.3 Calculation of component value $n_i$ of noise pattern vector

In Section 4.1, the microphone output signals of Babble, Car, Exhibition and Subway noise were added to those of the clean vowel with the 20 dB, 10 dB and 5 dB SNR, and the voice data was created. From these voice data, the logarithmic power spectrum was calculated, and the input pattern was created. Then, the shapes were compared between the standard and input patterns. On the other hand, in this section, as shown in Eq. (47), the input pattern is created by the direct addition of the weighted random numbers $n_i \cdot Rnd$ to the logarithmic power spectrum value $x_i$ of the clean vowel, and their shapes are compared. Therefore, we need to calculate each component value $n_i$ of the noise pattern vector that is equivalent to each SNR used in Section 4.1. In other words, in Figure 40 and on Table 8, we need to calculate values $\alpha_1$ to $\alpha_6$ that are equivalent to the above SNR. The following explains their calculation.

When the microphone output signal of the clean vowel is passed through the Mel filter bank with the $m$ frequency bands, we assume that the power spectrum array $X_i$ $(i = 1, 2, \cdots, m)$ is obtained. If the reference value of power spectrum is $X_0$, the logarithmic power spectrum array $x_i$ $(i = 1, 2, \cdots, m)$ that corresponds to $X_i$ can be calculated from the first formula of the following equation. Also, if the component value $n_i$ $(i = 1, 2, \cdots, m)$ of noise pattern vector is added to this logarithmic power spectrum array $x_i$ $(i = 1, 2, \cdots, m)$, value $x_i + n_i$ $(i = 1, 2, \cdots, m)$ is obtained. The relationship between the value $x_i + n_i$ and its corresponding power spectrum array $X_i + N_i$ $(i = 1, 2, \cdots, m)$ can be represented as the second formula of the following equation.

$$\begin{aligned} x_i &= 10 \log_{10} \frac{X_i}{X_0} && (n_i > 0) \\ x_i + n_i &= 10 \log_{10} \frac{X_i + N_i}{X_0} && (i = 1, 2, 3, \cdots, m) \end{aligned} \tag{49}$$

Figure 42 shows the relationship between $X_i$ and $x_i$ between $X_i + N_i$ and $x_i + n_i$ given by Eq. (49) for the $i$-th frequency band of the filter bank. This section aims to calculate the value $n_i$ that is equivalent to the SNR of 5 dB. The following equation can be obtained as an inverse function of Eq. (49).

$$\begin{aligned} X_i &= X_0 \cdot 10^{x_i/10} \\ X_i + N_i &= X_0 \cdot 10^{(x_i + n_i)/10} && (i = 1, 2, 3, \cdots, m) \end{aligned} \tag{50}$$

In Eq. (50), we can obtain the following equation by substituting the first formula into the

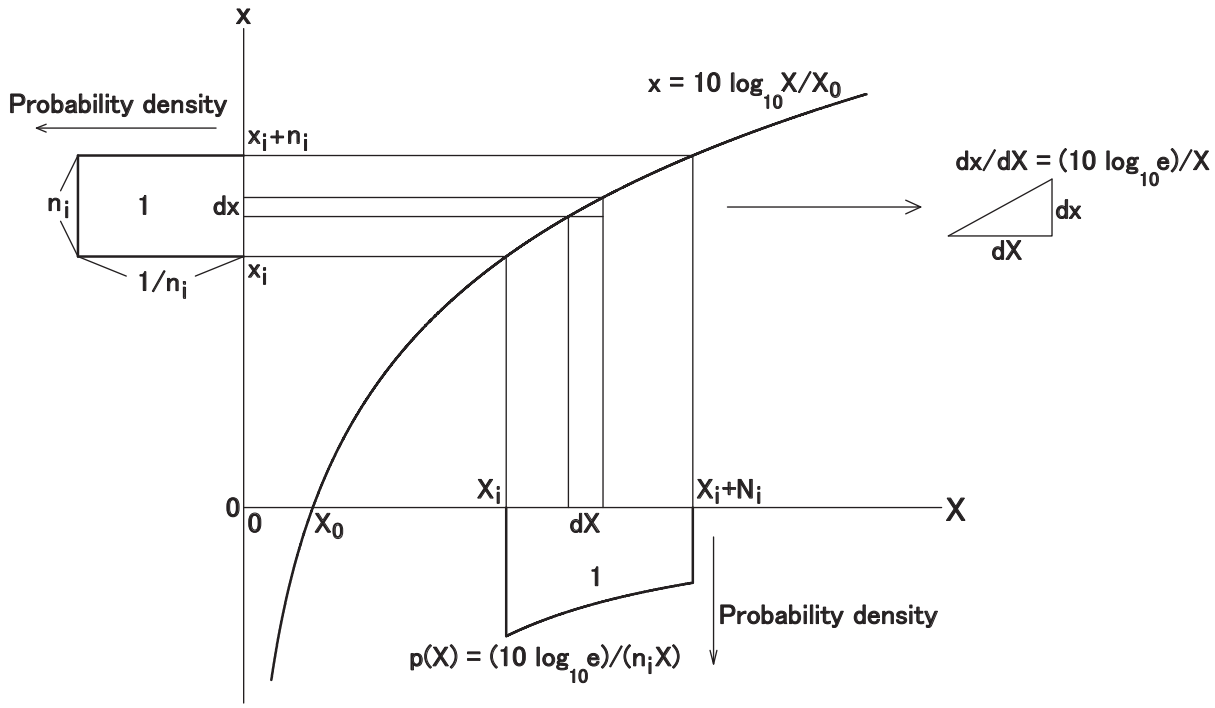


Figure 42. Relationship between power spectrum and logarithmic power spectrum.

second formula.

$$N_i = X_0 \cdot 10^{x_i/10}\,(10^{n_i/10} - 1) \qquad (i = 1, 2, 3, \cdots, m) \tag{51}$$

In Eq. (47), if the variable $Rnd$ is random numbers uniformly distributed within the range of 0.0 to 1.0, $x_{oi} = x_i + n_i \cdot Rnd$ and, therefore, $x_{oi}$ uniformly distributes within the range of $x_i$ to $x_i + n_i$. Figure 42 shows the probability density function of the flat shape which has function value $1/n_i$ in range $[x_i, x_i + n_i]$ on axis $x$. As shown in Figure 42, if we only focus on the $i$-th frequency band of the filter bank, it is appropriate to express the weighted random numbers $n_i \cdot Rnd$ as the uniformly distributed random numbers $n_i \cdot Rnd$. The weighted random numbers $n_i \cdot Rnd$ means the multiplication of different weight $n_i$ to each of the $i$-th frequency band. In this section, we use them in differently ways as necessary. Because the gradient of logarithmic curve $x = 10 \log_{10} X/X_0$ is $dx/dX = (10 \log_{10} e)/X$, the probability density function $p(X)$ on axis $X$, which corresponds to the probability density function $1/n_i$ on axis $x$, is described by the following equation.

$$p(X) = \frac{10 \log_{10} e}{n_i X} \qquad (i = 1, 2, 3, \cdots, m) \tag{52}$$

Thus, Figure 42 shows the probability density function which has function value $p(X) = (10 \log_{10} e)/(n_i X)$ in range $[X_i, X_i + N_i]$ on axis $X$. From the following equation, we can confirm that the total area of probability density function $p(X)$ is equal to 1. Here, we can

obtain the fifth formula of Eq. (53) by substituting Eq. (49) into the fourth formula of Eq. (53).

$$\begin{aligned}
\int_{X_i}^{X_i+N_i} p(X)\,dX &= \int_{X_i}^{X_i+N_i} \frac{10\log_{10} e}{n_i X}\,dX \\
&= \frac{10\log_{10} e}{n_i}\int_{X_i}^{X_i+N_i} \frac{1}{X}\,dX \\
&= \frac{10\log_{10} e}{n_i}\Big\{ \log_e(X_i+N_i) - \log_e X_i \Big\} \\
&= \frac{1}{n_i}\Big\{ 10\ \log_{10}\frac{X_i+N_i}{X_0} - 10\ \log_{10}\frac{X_i}{X_0} \Big\} \\
&= \frac{1}{n_i}\Big\{ (x_i+n_i) - x_i \Big\} \\
&= 1 \qquad\qquad (i=1,2,3,\cdots,m)
\end{aligned} \tag{53}$$

Where, if the uniformly distributed random numbers $n_i \cdot Rnd$ are added to the logarithmic power spectrum $x_i$ of the clean vowel on axis $x$, we assume that the power spectrum on axis $X$, which corresponds to $x_i+n_i \cdot Rnd$, is $X$. Now, expected value $E_i[X]$ of the power spectrum $X$ can be calculated by the following equation.

$$\begin{aligned}
E_i[X] &= \int_{X_i}^{X_i+N_i} X \cdot p(X)\,dX \\
&= \int_{X_i}^{X_i+N_i} X \cdot \frac{10\log_{10} e}{n_i X} dX \\
&= (10\log_{10} e) \cdot \frac{1}{n_i} \cdot N_i \qquad\qquad (i=1,2,3,\cdots,m)
\end{aligned} \tag{54}$$

We can obtain the following equation by substituting Eq. (51) into Eq. (54).

$$E_i[X] = (10\log_{10} e) \cdot X_0 \cdot 10^{x_i/10} \cdot \frac{10^{n_i/10}-1}{n_i} \qquad (i=1,2,3,\cdots,m) \tag{55}$$

On axis $X$ of Figure 42, the average energy of power spectrum of the clean vowel is $X_i$, and the average energy of power spectrum, which corresponds to the uniformly distributed random numbers $n_i \cdot Rnd$, is $E_i[X]-X_i$. Therefore, the signal-to-noise ratio (SNR) of the entire frequency band can be calculated by the following equation.

$$\begin{aligned}
SNR &= 10\ \log_{10}\frac{\sum_{i=1}^{m} X_i}{\sum_{i=1}^{m}(E_i[X]-X_i)} \\
&= 10\ \log_{10}\frac{\sum_{i=1}^{m} X_i}{\sum_{i=1}^{m} E_i[X] - \sum_{i=1}^{m} X_i}
\end{aligned} \tag{56}$$

We can obtain the following equation by substituting Eqs. (50) and (55) into Eq. (56).

$$
\begin{aligned}
SNR &= 10\ \log_{10} \frac{X_0 \sum_{i=1}^{m} 10^{x_i/10}}{(10\ \log_{10} e) \cdot X_0 \sum_{i=1}^{m} 10^{x_i/10} \cdot \dfrac{10^{n_i/10}-1}{n_i} - X_0 \sum_{i=1}^{m} 10^{x_i/10}} \\
&= 10\ \log_{10} \sum_{i=1}^{m} 10^{x_i/10} \\
&\quad -10\ \log_{10} \left\{ (10\ \log_{10} e) \sum_{i=1}^{m} 10^{x_i/10} \cdot \frac{10^{n_i/10}-1}{n_i} - \sum_{i=1}^{m} 10^{x_i/10} \right\}
\end{aligned}
\tag{57}
$$

Furthermore, we assign $\psi(n_1, n_2, \cdots, n_m)$ to the right side of Eq. (57) that is subtracted by the left side, and represent it as follows.

$$
\begin{aligned}
\psi(n_1, n_2, \cdots, n_m) &= 10\ \log_{10} \sum_{i=1}^{m} 10^{x_i/10} \\
&\quad -10\ \log_{10} \left\{ (10\ \log_{10} e) \sum_{i=1}^{m} 10^{x_i/10} \cdot \frac{10^{n_i/10}-1}{n_i} - \sum_{i=1}^{m} 10^{x_i/10} \right\} - SNR
\end{aligned}
\tag{58}
$$

In Eq. (58), $x_i$ is the logarithmic power spectrum value of the clean vowel, and we can set its value using the voice data. Therefore, Eq. (58) is the function of $n_i$ ($i=1, 2, \cdots, m$).

Next, we show that $\psi(n_1, n_2, \cdots, n_m)$ decreases monotonically when each $n_i$ ($i=1, 2, \cdots, m$) increases. For that purpose, we assign $\phi_1(n_i)$ to term $(10^{n_i/10}-1)/n_i$ of Eq. (58) as follows, and we check its increase or decrease.

$$
\phi_1(n_i) = \frac{10^{n_i/10}-1}{n_i} \qquad (i = 1, 2, 3, \cdots, m) \tag{59}
$$

Here, we can obtain the following equation by differentiating Eq. (59) by $n_i$.

$$
\begin{aligned}
\phi_1'(n_i) &= \left( \frac{10^{n_i/10}-1}{n_i} \right)' \\
&= \frac{(\log_e 10^{1/10})\ n_i\, 10^{n_i/10} - 10^{n_i/10} + 1}{n_i^2} \\
&= \frac{\phi_2(n_i)}{n_i^2} \qquad (i = 1, 2, 3, \cdots, m)
\end{aligned}
\tag{60}
$$

Furthermore, we assign $\phi_2(n_i)$ to the numerator of Eq. (60) as follows, and we check its positive or negative.

$$
\phi_2(n_i) = (\log_e 10^{1/10})\ n_i\, 10^{n_i/10} - 10^{n_i/10} + 1 \qquad (i = 1, 2, 3, \cdots, m) \tag{61}
$$

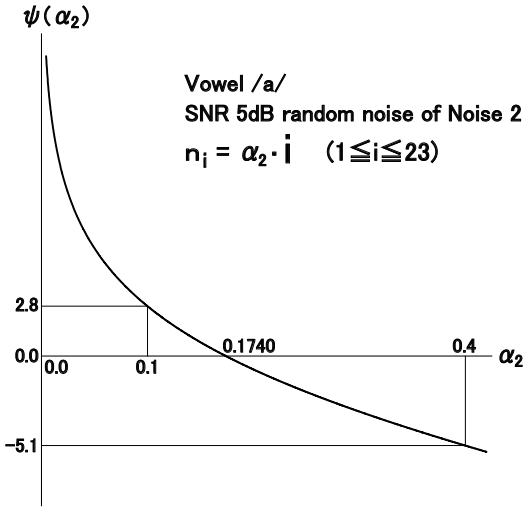


Figure 43. Graph of function $\psi(\alpha_2)$.

Table 9. Solution $\alpha_2$ of $\psi(\alpha_2)=0$ ( Noise 2 : $n_i = \alpha_2 i$ ).

| $\alpha_2$ | /a/ | /i/ | /u/ | /e/ | /o/ |
|---|---|---|---|---|---|
| SNR 5 dB | 0.1740 | 0.1642 | 0.1769 | 0.1701 | 0.1843 |
| SNR 3 dB | 0.2484 | 0.2340 | 0.2519 | 0.2426 | 0.2623 |
| SNR 1 dB | 0.3421 | 0.3216 | 0.3457 | 0.3337 | 0.3595 |

For that purpose, we calculate Eq. (61) if $n_i=0$ and its derived function as follows.

$$\phi_2(\,0\,) = 0 \tag{62}$$

$$\phi_2'(n_i) = (\log_e 10^{1/10})^2\; n_i\, 10^{n_i/10} > 0 \qquad (\,n_i > 0\,) \qquad (i = 1, 2, 3, \cdots, m) \tag{63}$$

From Eqs. (62) and (63), it is clear that $\phi_2(n_i)>0$. Then, from Eq. (60), it is clear that $\phi_1'(n_i)>0$ and, therefore, Eq. (59) is a monotonically increasing function. From the above, it is clear that the value of Eq. (58) decreases monotonically when each $n_i$ $(i=1,2,\cdots,m)$ increases.

In this paper, each $n_i$ $(i = 1, 2, \cdots, m)$ is related to each other by the parameter $\alpha_k$ $(k=1,2,\cdots,6)$ as shown on Table 8. In the case of Noise 1 to Noise 6 shown on Table 8, each $n_i$ increases monotonically when each $\alpha_k$ increases and, therefore, the value of Eq. (58) decreases monotonically. In particular, $n_i{=}\alpha_2 i$ $(i=1,2,\cdots,m)$ for Noise 2, and Eq. (58) can be rewritten as follows.

$$\psi(\alpha_2) = 10\,\log_{10}\sum_{i=1}^{m} 10^{x_i/10} - 10\,\log_{10}\left\{(10\,\log_{10} e)\sum_{i=1}^{m} 10^{x_i/10}\;\cdot\;\frac{10^{\alpha_2 i/10}-1}{\alpha_2 i} - \sum_{i=1}^{m} 10^{x_i/10}\right\} - SNR \tag{64}$$

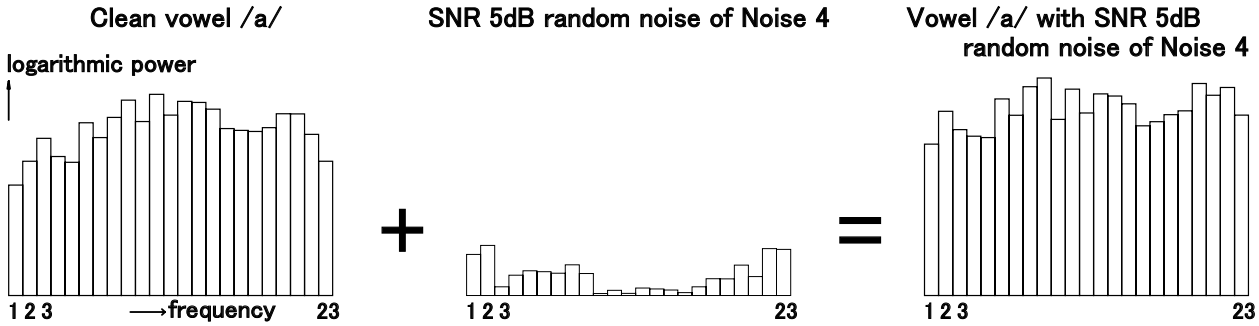


Figure 44. Addition of random noise to clean vowel.

Figure 43 shows a relational graph between $\alpha_2$ and $\psi(\alpha_2)$ obtained through numerical analysis of Eq. (64). Note that we assumed that SNR=5 in Eq. (64). Also, we have substituted the mean value of each logarithmic power spectrum, calculated from the central 100 frames of the clean vowel /a/, into $x_i$ $(i = 1, 2, \cdots, m)$. As shown in Figure 43, Eq. (64) is a monotonically decreasing function, and it is clear that we can uniquely determine a solution $\alpha_2$ of $\psi(\alpha_2)$=0 through numerical analysis. As described above, we could obtain solution $\alpha_2$=0.1740 of $\psi(\alpha_2)$=0 from Figure 43. Table 9 shows the values of $\alpha_2$ which are obtained for each vowel and for each SNR when SNR=5, SNR=3 and SNR=1 and if the noise pattern of Noise 2 and "01Clean" (shown in Section 4.1) of each vowel are used. "01Clean" is the first "clean vowel" that was produced among 72 sounds in 12 weeks.

The above calculation procedure is summarized below. First, in Eq. (50), power spectra $X_i$ and $X_i+N_i$ on axis $X$ shown in Figure 42 are expressed by logarithmic power spectra $x_i$ and $x_i+n_i$ on axis $x$. Also, in Eq. (55), expected value $E_i[X]$ of power spectrum $X$ on axis $X$, which corresponds to $x_i+n_i \cdot Rnd$ on axis $x$, is expressed by $x_i$ and $n_i$. Then, we calculate the SNR on axis $X$ using Eq. (56), substitute Eqs. (50) and (55) into Eq. (56). Therefore, the SNR is expressed by $x_i$ and $n_i$ in Eq. (57). We substitute the mean value of the logarithmic power spectra of the clean vowel into $x_i$. Now, Eq. (57) is an equation of $m$ variables with unknowns $n_i$ $(i=1, 2, \cdots, m)$. In this paper, each $n_i$ $(i=1, 2, \cdots, m)$ is related by the parameter $\alpha_k$ $(k=1, 2, \cdots, 6)$ as shown on Table 8. Therefore, Eq. (57) is rewritten by Eq. (64). Eq. (64) is an equation of single variable with unknown $\alpha_2$. And we calculate solution $\alpha_2$ and obtain value $n_i$ that is equivalent to the SNR of 5 dB.

By using the above calculation procedure, the value of each $\alpha_k$ $(k=1, 2, \cdots, 6)$ is calculated for the noise patterns of Noise 1 to Noise 6, and Table 9 of each noise pattern is obtained. Then, the weighted random numbers $n_i \cdot Rnd$, which is equivalent to the SNR, is generated by the computer. Figure 44 shows the process where the weighted random numbers $n_i \cdot Rnd$

$(i=1,2,\cdots,m)$ equivalent to the SNR of 5 dB are added to the logarithmic power spectrum $x_i$ $(i=1,2,\cdots,m)$ of the clean vowel /a/, using Noise 4 and Eq. (47), and then the component value $x_{oi}$ $(i=1,2,\cdots,m)$ of the original input pattern vector is created. It is clear that the shape of the weighted random numbers, shown at the center of Figure 44, is similar to the difference pattern of the actual noise shown in Figure 39.

Finally in this section, we discuss the relationship between the area (or energy) of the weighted random numbers generated by the computer and that of the difference pattern of actual noise. After calculating the average area of the weighted random numbers of 5 dB SNR and that of the difference pattern of 5 dB SNR, using the central 100 frames of each vowel produced for a period of 2 seconds, we have found that the former value is 16.2% greater than the latter value. We suppose that there are two causes for that as follows. First, in the calculation of the weighted random numbers, we substituted the mean value of the logarithmic power spectra, calculated from the central 100 frames of each vowel produced for a period of 2 seconds, into Eq. (64), and obtained solution $\alpha_k$ $(k=1,2,\cdots,6)$. These frames are overlapped for the 25 msec frame width and 10 msec frame period. In the calculation of the difference pattern, we calculated the SNR using the microphone output signal of the entire interval of 2-second vowel. We suppose that those average areas are different because the calculation intervals of SNR differ between them. Second, we obtained the logarithmic power spectrum value $x_i$ $(i=1,2,\cdots,m)$ of the clean vowel using the Hamming window, and substituted this value into Eq. (64) in order to obtain solution $\alpha_k$. Therefore, we suppose that an effect of the Hamming window appears as described at the beginning of Chapter 9. In Section 10.2, based on our experiments, we will discuss the estimation error of optimum value caused by the above area difference.

## 9.4 Creation of original pattern vectors

Here, we use the $\alpha_k$ $(k=1,2,\cdots,6)$ values obtained in the previous section, and create the original standard pattern vector and original input pattern vector given by Eq. (48), by applying the $\alpha_k$ values to the same voice data as those used in Chapter 4. Note that the original standard pattern vector is abbreviated as "the standard pattern", and the original input pattern vector is abbreviated as "the input pattern" hereafter. Table 10 shows the type and the number of the 23rd dimensional logarithmic power spectrum that has been used for the standard and input patterns in the optimization experiment. The logarithmic power spectra, each consisting of 100 frames shown on the first row of Table 10, have been

Table 10. Logarithmic power spectra for optimizing normal distribution.

| | | /a/ | /i/ | /u/ | /e/ | /o/ |
|---|---|---|---|---|---|---|
| | 01 Clean | 100 | 100 | 100 | 100 | 100 |
| | Standard pattern | 1 | 1 | 1 | 1 | 1 |
| {1} | 01 Clean with SNR 5dB random noise of Noise 1 | | | | | |
| | Input pattern | 100×50 | 100×50 | 100×50 | 100×50 | 100×50 |
| {2} | 01 Clean with SNR 5dB random noise of Noise 2 | | | | | |
| | Input pattern | 100×50 | 100×50 | 100×50 | 100×50 | 100×50 |
| : | : | ... | ... | ... | ... | ... |
| {6} | 01 Clean with SNR 5dB random noise of Noise 6 | | | | | |
| | Input pattern | 100×50 | 100×50 | 100×50 | 100×50 | 100×50 |

extracted from "01Clean" of each vowel. Then, as described in Section 3.10, the median is determined from the above 100 frames and it is used as the standard pattern of each vowel. The logarithmic power spectra, each consisting of one frame shown on the second row of Table 10, are the standard patterns that have been determined for each vowel.

Also, the logarithmic power spectra, each consisting of 100×50 frames shown in {1} to {6} of Table 10, have been created by adding the weighted random numbers to the logarithmic power spectra, each consisting of 100 frames of the above "01Clean", using Eq. (47) and the noise patterns of Noise 1 to Noise 6 shown in Figure 40 when SNR=5. During this time, the uniformly distributed random numbers *Rnd* are generated repeatedly and the logarithmic power spectra, each consisting of 100×50 frames, are created. Then, the logarithmic power spectra of these 6×100×50×5 frames are used as the input patterns.

As described above, in the optimization experiment, we create the standard pattern and the input pattern by using the weighted random numbers generated by the computer and five patterns of "clean vowel 01Clean".

## 9.5 Variance optimization of normal distribution

We determine the optimum value of the variance $\sigma^2$ of the normal distribution (the optimum value of $\omega$) using both the standard and input patterns created in the previous section and the algorithm of the geometric distance $d_A$ shown in Eq. (42). Similar to the vowel recognition experiments in Chapters 4 and 7, the value $\omega$ is incremented by 0.2 from 3.0 to 23.0, and

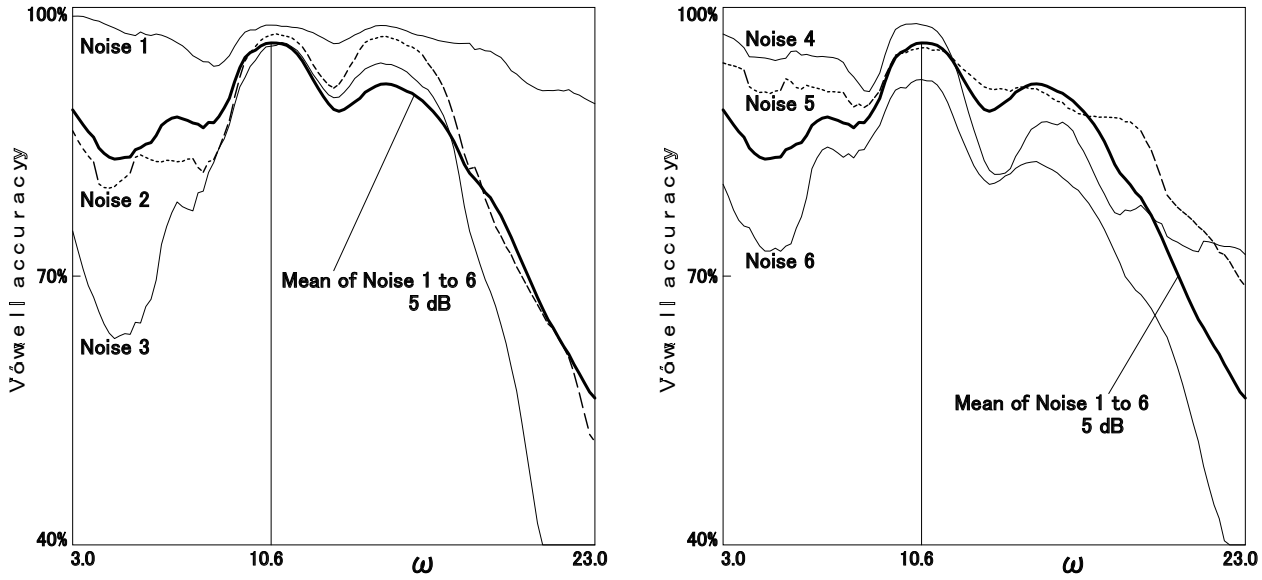


Figure 45. Vowel recognition accuracy and optimum value $\omega$.

the recognition accuracy of the input pattern is calculated by using 100×50×5-frame input patterns shown in {1} to {6} of Table 10. Figure 45 shows the calculated relationship between the value $\omega$ and the recognition accuracy by six thin lines, respectively. Also, these six curves are averaged and the average recognition accuracy is shown by thick lines in Figure 45. From Figure 45, it is discovered that the recognition accuracy curve of Noise 1 is higher than each curve of Noise 2 to Noise 6 in the all $\omega$ value range. We suppose the cause as follows. Within the geometric distance algorithm, the "wobble" caused by the random numbers is replaced by the shape change of the reference pattern having the initial shape of the normal distribution. During this time, the shape of the noise pattern of Noise 1 is flat (or uniform) as shown in Figure 40 and, therefore, we suppose that the "wobble" is absorbed effectively. Furthermore, from Figure 45, it is discovered that the peak of recognition accuracy is at the same location for each of the Noise 1 to Noise 6 curves. We can see that the average recognition accuracy of Noise 1 to Noise 6 becomes maximum if $\omega$=10.6. Thus, we determine $\omega$=10.6 as the optimum value and use it in the following evaluation experiments. When we have performed the optimization experiment using the input pattern, each consisting of 100×10 frames, instead of the input pattern, each consisting of 100×50 frames shown in {1} to {6} of Table 10, we could obtain almost the same curves as the recognition accuracy curves shown in Figure 45. The optimum value was $\omega$=10.6. This shows that we can reduce the processing overhead to obtain the optimum value.

# Chapter 10

# Evaluation Experiments of Vowel Recognition

To check the effectiveness of optimization method described in the previous section, we have performed the evaluation experiments for the "clean vowel" and the "vowel with actual noise" using the value $\omega$=10.6 determined in the previous section and the algorithm of the geometric distance $d_A$ shown in Eq. (42). The value $\omega$=11.0 is used in Section 7.1, but the value $\omega$=10.6 is used for the evaluation experiments in this section. Except for this value, we have performed the evaluation experiments of vowel recognition using the same voice data and the method as those used in Chapter 7.

## 10.1 Evaluation experiments and their results

In the optimization experiment of the previous section, we determined the optimum value (estimated value) of $\omega$=10.6 by using only the "clean vowel 01Clean" that was produced first among 72 sounds in 12 weeks as shown on Table 10. Similar to the vowel recognition experiments in Chapter 7, in the evaluation experiments of this section, the median was determined from 100 frames of the above "clean vowel 01Clean" and it was used as the standard pattern of each vowel. On the other hand, the "clean vowel 02Clean to 72Clean" produced in the 2nd to 72nd sounds were used as the input patterns. In addition, the actual Babble, Car, Exhibition and Subway noises were added to these "clean vowel 02Clean to 72Clean" with the 20 dB, 10 dB and 5 dB SNR, and the input patterns were created.

Table 11 shows the result of evaluation experiments. As shown on Table 11, the average recognition accuracy of the "vowel with actual noise of 5 dB SNR" is 80.28% in the evaluation experiment where the optimum value (estimated value) of $\omega$=10.6 is used.

Table 11. Vowel recognition accuracy with geometric distance $d_A$. ($\omega = 10.6$)

| | Babble | Car | Exhibition | Subway | Mean |
|---|---|---|---|---|---|
| Clean | | | | | 99.98% |
| SNR 20 dB | 99.92% | 99.86% | 99.22% | 99.56% | 99.64% |
| SNR 10 dB | 98.52% | 97.94% | 88.16% | 93.97% | 94.65% |
| SNR 5 dB | 91.68% | 82.13% | 66.85% | 80.44% | 80.28% |

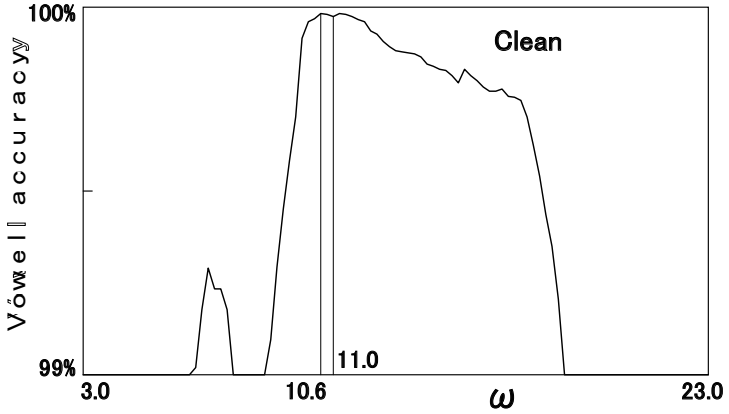


Figure 46. Recognition accuracy of clean vowel.

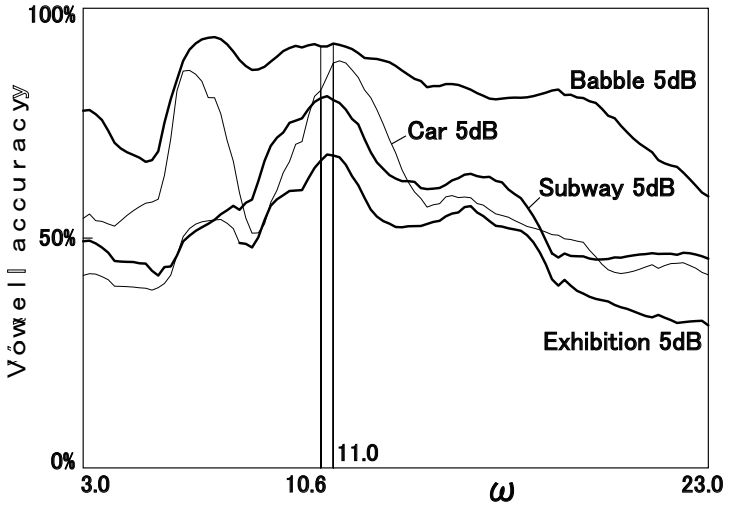


Figure 47. Recognition accuracy of vowel with actual noise.

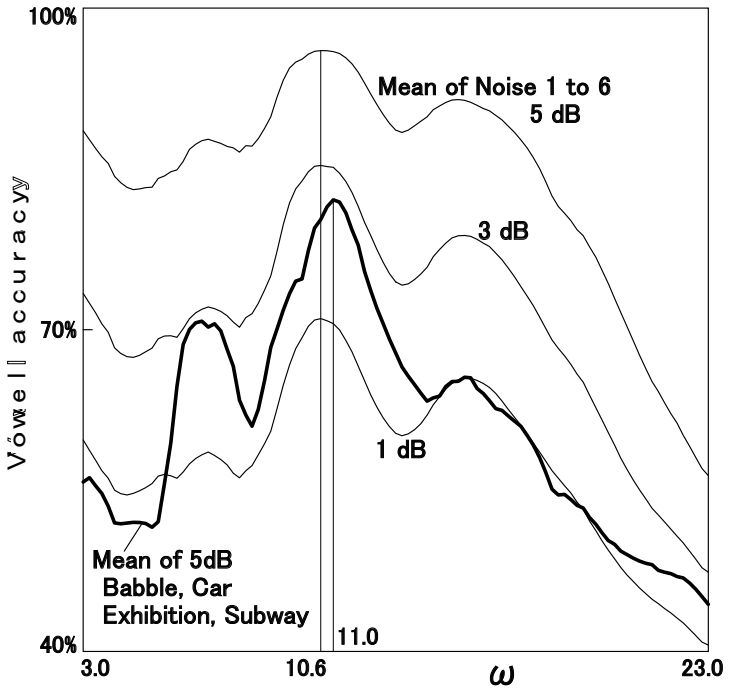


Figure 48. Vowel recognition accuracy and optimum value $\omega$.

## 10.2 Verification of optimum value

Table 11 shows the result of recognition accuracy using the optimum value (estimated value) of $\omega$=10.6 that we have determined from Figure 45. Here, in order to verify that the value $\omega$=10.6 is truly the optimum value, the value $\omega$ is incremented by 0.2 from 3.0 to 23.0 and the recognition accuracy of the "clean vowel" and the "vowel with actual noise of 5 dB SNR" is calculated. Figures 46 and 47 show the calculated relationship between the value $\omega$ and the recognition accuracy for the input patterns of the "clean vowel" and the "vowel with Babble 5dB, Car 5dB, Exhibition 5dB, and Subway 5dB", respectively. From Figures 46 and 47, we can find that the recognition accuracy is almost maximum in the value $\omega$=10.6.

Furthermore, the four curves of actual noise, shown in Figure 47, are averaged and this average recognition accuracy is shown by a thick line in Figure 48. In the calculation of the average recognition accuracy for Noise 1 to Noise 6 shown by thick lines in Figure 45, the values of SNR=5, SNR=3 and SNR=1 are used respectively, and their results are shown by three thin lines in Figure 48. Note that the average recognition accuracy curves of 5 dB SNR shown by the thick lines in Figure 45, are the same as that shown by the thin line in Figure 48. In Figure 48, the recognition accuracy curves of the optimization experiments using the "vowel with weighted random numbers" are shown by three thin lines, but the recognition accuracy curve of the evaluation experiment using the "vowel with actual noise" is shown by one thick line. From Figure 48, it is clear that the four curves of recognition accuracy have the same features and that the locations of the maximum recognition accuracy almost match each other. This means that we can estimate the optimum variance value of the normal distribution, using the "vowel with weighted random numbers" instead of the "vowel with actual noise". From Figure 48, it is also clear that the weighted random numbers of 3 dB SNR is equivalent to the actual noise of 5 dB SNR for the average recognition accuracy. We suppose the cause as follows. Within the geometric distance algorithm, the "wobble" of input pattern is replaced by the shape change of the reference pattern having the initial shape of the normal distribution. During this time, the "wobble" caused by the random numbers is more random than the actual noise and, therefore, we suppose that the "wobble" is absorbed effectively.

At the end of Section 9.3, we described the difference between the area (or energy) of the weighted random numbers of 5 dB SNR and that of the difference pattern of actual noise of 5 dB SNR. Next, we discuss this. In Figure 48, we can obtain the value $\omega$=10.6 even when

we use any of the recognition accuracy curves, shown by three thin lines, in the optimization experiment. Now, on Table 9, the value $\alpha_2$ of 1 dB SNR is almost 2 times that of 5 dB SNR. In other words, the area of noise pattern of 1 dB SNR is almost 2 times larger than that of 5 dB SNR case. This is similar to other $\alpha_k$ values. When compared with this change, the 16.2% difference shown in Section 9.3 is small. They show that the difference between their areas does not affect the estimation of optimum value.

From the average recognition accuracy curve of the "vowel with actual noise of 5 dB SNR" shown by thick line in Figure 48, it is discovered that the recognition accuracy becomes maximum if $\omega$=11.0. It is discovered that the recognition accuracy is 80.28% if $\omega$=10.6 and the recognition accuracy is 82.11% if $\omega$=11.0. The difference between them is 1.83% and it is small. From the recognition accuracy curve of the "clean vowel" shown in Figure 46, it is discovered that the recognition accuracy is 99.98% if $\omega$=10.6 and the recognition accuracy is 99.97% if $\omega$=11.0. The difference between them is small. This shows that we can determine the optimum value of $\omega$ using the "vowel with weighted random numbers".

In this paper, as shown in Figures 45 and 48, we have used the vowel recognition accuracy as the objective function in order to estimate the optimum variance value. Meanwhile, we used a statistic $T$ of "Welch's $T$-test" as the objective function and performed the optimization experiment for bird vocalisations.[25] If we compare the two results, we find that the former objective function curves and the latter objective function curve have the same features.

# Chapter 11

# 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.

Furthermore, we have used the weighting vector that consists of the rate of change of the moment ratio, and created two weighted pattern vectors by performing the product-sum operation using the weighting vector and the standard pattern vector and the product-sum operation using the weighting vector and the input pattern vector. Then, we have proposed a new algorithm that uses the angle between these weighted pattern vectors as the geometric distance. At this time, we have evaluated the processing overhead and the computational memory required for the new algorithm. Also, we have performed the vowel recognition experiments, and confirmed that the recognition performance becomes stable.

Moreover, we have proposed a new optimization method of the geometric distance to determine the optimum variance value of the normal distribution, using the weighted random numbers generated by the computer and five patterns of vowels. At this time, we have performed the vowel recognition experiments using the "vowel with weighted random numbers" and the "vowel with actual noise", respectively, and checked the relationship between the variance of the normal distribution and the vowel recognition accuracy. The results have shown that the curves of their vowel recognition accuracy have the same features and that the locations of the maximum recognition accuracy almost match each other. This means that we can estimate the optimum variance value of the normal distribution using the "vowel with weighted random numbers" instead of the "vowel with actual noise". Then, we have used the estimated value obtained from the "vowel with weighted random numbers" and performed the evaluation experiments for the "vowel with actual noise of 5 dB SNR", and verified the effectiveness of our proposal.

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. Also, this paper uses the geometric distance for one-dimensional patterns. We extend the concept of Geometric Distance so that it can be applied to two-dimensional patterns such as voice prints and images.[23]

This paper shows that we have obtained the estimated value of $\omega$=10.6 using each noise pattern of Noise 1 to Noise 6. On the other hand, we have found that the true optimum value is $\omega$=11.0 in the evaluation experiments where we used four types of actual noises of Babble, Car, Exhibition, and Subway. In order to reduce the difference between them, we will perform the optimization experiments using more types of noise patterns and will perform the evaluation experiments using more types of actual noises. We will compare the results of those experiments, find out the type of noise pattern to be required at minimal for optimization, and improve our optimization method so that we can determine a more accurate estimation value and reduce the processing overhead by using less types of noise patterns. We will apply the results of the algorithm proposed in this paper and the emotional expression analysis of text[26, 27] to our project named Recognizing Human Emotion and Creating Machine Emotion.[28, 29] Also, we will perform the optimization experiments using the normal random numbers, instead of the uniformly distributed random numbers, and will compare the results of these experiments.

# 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 scale 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/

[23] M. Jinnai, N. Boucher, J. Robertson and S. Kleindorfer. *Design Considerations in an Automatic Classification System for Bird Vocalisations using the Two-dimensional Geometric Distance and Cluster Analysis*, Proc. 20th Int. Congr. Acoustics, 130, 2010.

[24] M. Jinnai, Y. Akashi, S. Nakaya, F. Ren and M. Fukumi. *Recognition of Abnormal Vibrational Responses of Signposts using the Two-dimensional Geometric Distance and Wilcoxon Test*, Proc. 6th Int. Congr. IEEE NLP-KE, pp.166-173, 2010.

[25] M. Jinnai, N. Boucher, M. Fukumi and H. Taylor. *A New Optimization Method of the Geometric Distance in an Automatic Recognition System for Bird Vocalisations*, Proceedings of the Acoustics 2012 Nantes Conference, 105

[26] F. Ren. *From Cloud Computing to Language Engineering, Affective Computing and Advanced Intelligence*, IJAI, Volume 2, Number 1, pp.1-14, July, 2010.

[27] C. Quan and F. Ren. *Sentence Emotion Analysis and Recognition Based on Emotion Words Using Ren-CECps*, IJAI, Volume 2, Number 1, pp.105-117, July, 2010.

[28] F. Ren. *Invited paper, Robotics Cloud and Robotics School*, Proc. 7th Int. Congr. IEEE NLP-KE, pp.1-8, 2011.

[29] F. Ren. *Affective Information Processing and Recognizing Human Emotion*, Electronic Notes in Theoretical Computer Science, Vol.225, No.2009, pp.39-50, 2009.