# Distance Measure between Gaussian Distributions for Discriminating Speaking Styles

*Goshu Nagino and Makoto Shozakai*

Information Technology Laboratory, Asahi Kasei Corporation
Atsugi AXT Main Tower 22F, Okada 3050, Atsugi, Kanagawa, 243-0021, Japan
{g-nagino, makoto}@ljk.ag.asahi-kasei.co.jp

## Abstract

Discriminating speaking styles is an important issue in speech recognition, speaker recognition and speaker segmentation. This paper compares distance measures between Gaussian distributions for discriminating speaking styles. The Mahalanobis distance, the Bhattacharyya distance and the Kullback-Leibler divergence, which are in common use for a definition as a distance measure between Gaussian distributions, are evaluated in terms of an accuracy to discriminate speaking styles. In this paper, the accuracy is judged on a visualized map, where speaking style speech corpora are mapped onto two-dimensional space by utilizing a multidimensional scaling method. It is shown that speaking style clusters appear clearly grouped on the visualized map obtained by the Bhattacharyya distance and the Kullback-Leibler divergence. In addition, the visualized map corresponds to speech recognition performance, and the Kullback-Leibler shows higher sensitivity to recognition performance.

**Index Terms**: statistical MDS, gaussian distribution, distance measure, speaking style

## 1. Introduction

Fluctuation of speaking style has a strong impact on performance of speech recognition and speaker recognition [1][2] because of the change of Gaussian distribution shape in statistical models such as HMM and GMM. The Mahalanobis distance, the Bhattacharyya distance [3] and the Kullback-Leibler divergence [4] are known distance measures between Gaussian distributions. It is expected that these distance measures can represent the change of Gaussian distribution shape properly. This paper discusses their sensitivity to multiple speaking styles.

In the next section the three distance measures are introduced. In Section 3, this paper describes the overview of speaking style speech corpus. In Section 4, the procedure of evaluating each distance measure is described. In Section 5, a comparison of the distance measures is described. Finally, the summary of this paper is described in Section 6.

## 2. Distance Measure between Gaussian Distributions

In this section, three distance measures are introduced: the Mahalanobis distance, the Bhattacharyya distance and the Kullback-Leibler divergence. Each distance measure is in common use as a definition of distance measure between Gaussian distributions. The Gassuian distribution in one dimension is defined as follows:

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

Here, $\mu$ and $\sigma$ denotes model parameters, mean and variance value respectively. In a multidimensional case, covariance matrix is defined as diagonal in this paper. Each distance measure $D(p,q)$ between Gaussian distribution $p(x)$ and $q(x)$ is defined according to model parameters in formula (1).

### 2.1. Mahalanobis Distance

The Mahalanobis distance is similar to the Euclidean distance. It differs from the Euclidean distance in taking into account the correlations of the data set. The Mahalanobis distance is defined as follows:

$$D_{\_ma}(p,q) = \sqrt{\frac{\{\mu_p - \mu_q\}^2}{\sigma_p{}^2 + \sigma_q{}^2}} \tag{2}$$

### 2.2. Bhattacharyya Distance

The Bhatacharyya distance is a theoretical distance measure between two Gaussian distributions that is equivalent to an upper bound on the optimal Bayesian classification error probability. The Bhattacharyya distance is defined as follows:

$$D_{\_bh}(p,q) = \frac{1}{8}(\mu_p - \mu_q)^2 \left[\frac{\sigma_p{}^2 + \sigma_q{}^2}{2}\right]^{-1} + \frac{1}{2}\ln\frac{\left|(\sigma_p{}^2 + \sigma_q{}^2)/2\right|}{\sigma_p \sigma_q} \tag{3}$$

### 2.3. Kullback-Leibler Divergence

The Kullback-Leiber (KL) Divergence is the symmetric divergence between two classes. The KL Divergence represents a measure of degree of difficulty of discriminating between classes (the larger the divergence, the greater the separability between the classes). The KL Divergence is defined as follows:

$$D_{\_kl}(p,q) = \int p(x)\log\frac{p(x)}{q(x)} + q(x)\log\frac{q(x)}{p(x)}\,dx$$

$$\approx \frac{1}{n}\sum_{t=1}^{n} p(x)\log\frac{p(x_t)}{q(x_t)} + q(x_t)\log\frac{q(x\ )}{p(x_t)} \qquad (4)$$

Here, $x_1,\dots,x_n$ are synthetic samples from $p(x)$ to estimate the parameter of $p(x)$.

## 3. Speaking Style Speech Corpus

In this section, a speaking style speech corpus used in evaluation is described. 126 Japanese females uttered lists of 175 words taken from 5240 isolated words set (called ATR5240 in Japan) in three or four speaking styles indicated in Table 1. Totally, 457 speaker dependent and speaking style dependent speech corpora are recorded. The speech data is overlaid with background noise recorded at an exhibition hall at a Signal-to-Noise ratio of 20 dB. Sampling frequency is 11.025kHz. In the following, the acoustic parameters consist of 10 MFCCs, 10 delta-MFCCs, and 1 delta-log power. Both noise cancellation by spectral subtraction and equalization by cepstrum mean normalization are applied.

Table 1 Speaking Styles

| Speaking style | Instructions provided for recording | Symbol |
|---|---|---|
| Normal | Read utterance list at normal speed of conversation. | □ |
| Fast | Read utterance list at faster than normal speed of speech. | × |
| High | Read utterance list at higher than normal tone of speech. | ○ |
| Whisper | Read utterance list at a level not to be overheard by near-by persons. | ● |
| Loud | Read utterance list at a level to be heard by persons at some distance. | Δ |
| Lombard | Read utterance list among an ambient car noise. | ∇ |
| Syllable enhanced | Read utterance list by enhancing the Japanese syllables. | ■ |

## 4. Procedure of Evaluating Distance Measure

In this section, the procedure of distance measures in Section 2 is described.

### 4.1. Create Distance Matrix

First, speaker and speaking style dependent acoustic models (SSD-models) are trained with the speech corpus described in Section 3. The acoustic model structure is a mono-phonemic HMM expressed by a single Gaussian distribution. Next, the distance between two SSD-models is computed. The distance $D_{ssd}(i,j)$ between SSD-model $i$ and $j$ is defined as follows:

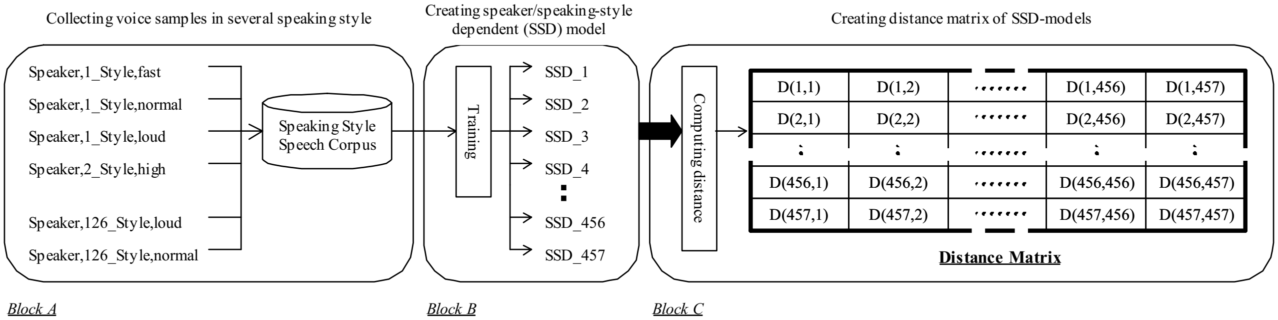
$$D_{ssd}(i,j) \equiv \sum_{k=1}^{K} d(i,j,k) * w(k) \Big/ \sum_{k=1}^{K} w(k) \qquad (5)$$

Here, $d(i,j,k)$ denotes the mutual distance between phoneme $k$ within SSD-model $i$ and phoneme $k$ within SSD-model $j$. $w(k)$ represents weight value such as an occurrence frequency rate of phoneme $k$. $K$ indicates total number of phonemes within SSD-model. Assuming all SSD-models ($i = 1,\dots,N$) share a common topology with one-on-one state alignment between respective SSD-models, $d(i,j,k)$ can be defined using formula (6).

$$d(i,j,k) = \frac{1}{S(k)}\sum_{s=0}^{S(k)-1}\frac{1}{L}\sum_{l=0}^{L-1} dd(i,j,k,s,l) \qquad (6)$$

$S(k)$ represents the number of states in phoneme $k$. $L$ stands for the number of dimension of the acoustic feature parameters. $dd(i,j,k,s,l)$ is the distance between the $l$-dimension Gaussian distribution of the $s$-th state in phoneme $k$, SSD-model $i$, and the equivalent Gaussian of the SSD-model $j$. In this paper, each distance measure described in formula (2), (3) and (4) are applied to $dd(i,j,k,s,l)$. The distance is computed to all combination of SSD-models.

The block diagram of creating distance matrix is shown in Figure 1. Total number of a combination of distance between two SSD-models is $457 \times 457 = 208849$. Three distance matrices are created in each distance measure.



| | | | | |
|---|---|---|---|---|
| D(1,1) | D(1,2) | ....... | D(1,456) | D(1,457) |
| D(2,1) | D(2,2) | ....... | D(2,456) | D(2,457) |
| : | : | ....... | : | : |
| D(456,1) | D(456,2) | ....... | D(456,456) | D(456,457) |
| D(457,1) | D(457,2) | ....... | D(457,456) | D(457,457) |

Figure 1 *Block Diagram of Creating Distance Matrix*

### 4.2. Visualization by utilizing MDS

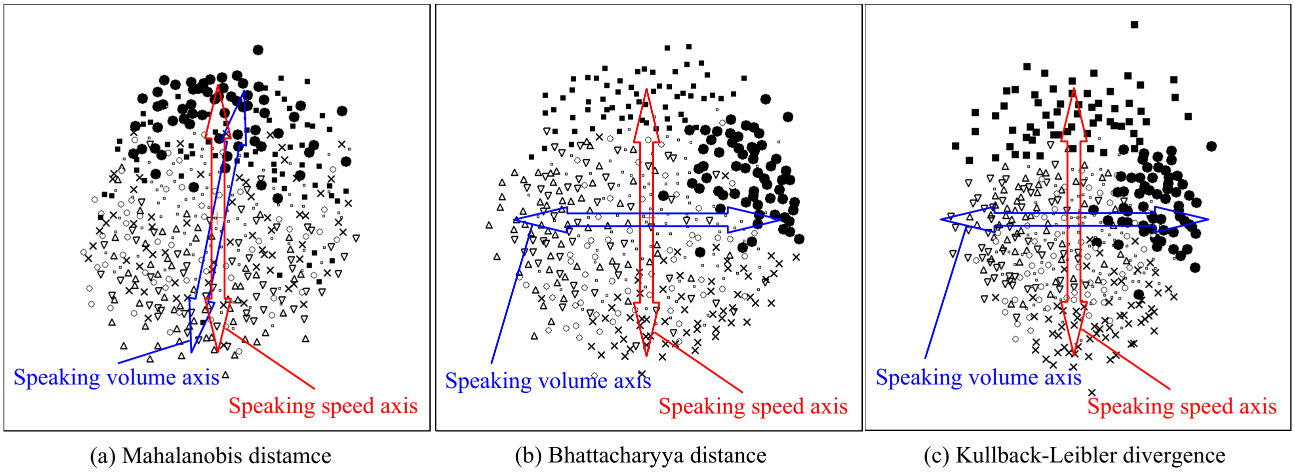
The multidimensional scaling (MDS) method [5] featuring a visual mapping of multidimensional information onto visible space of low order (one to three) dimension is extremely effective in enhancing perceptibility of multidimensional data space such as acoustic space expressed by speech data. In this paper, speaking style speech corpus is visualized by utilizing Sammon mapping [6], a conventional MDS method, with the distance matrix created in section 4.2. If a distance measure can detect the change of Gaussian distribution shapes depending on speaking style fluctuations, it would be expected of the SSD-models belonging to the same speaking style to appear clustered together on the visualized acoustic space.

The framework where aggregate of statistical model as an approximation of speech corpus is projected onto visible space by applying distance measure between statistical models to the conventional MDS method is called the COSMOS (COmprehensive Space Map of Objective Signal) method [7][8]. A visualized map itself is called COSMOS map. Respective whole acoustic model projected onto the COSMOS map is called a STAR.

In this paper, three distance measures are compared by observing how speaking styles are gathered on this visualized map.

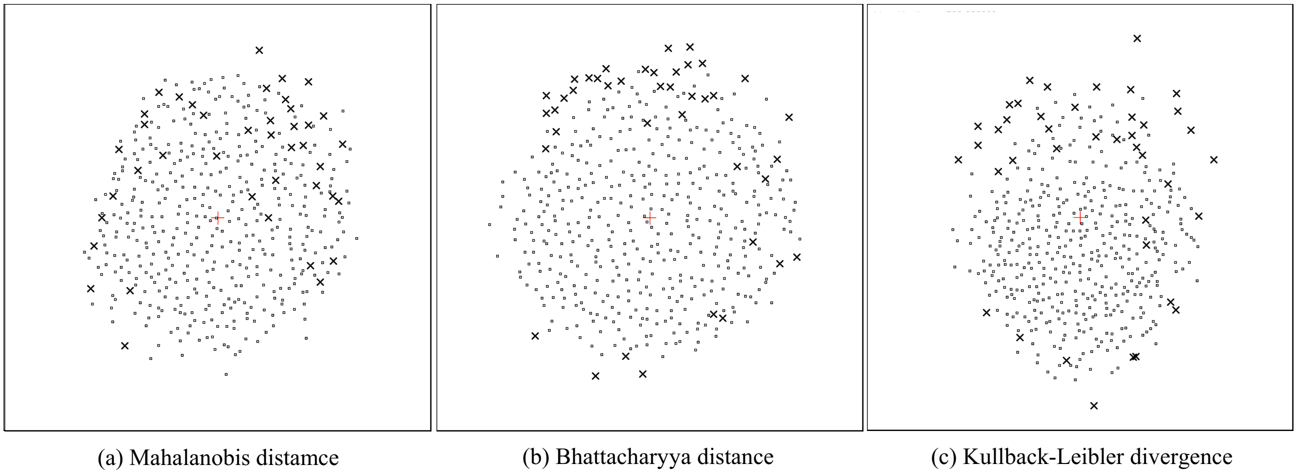
## 5. Experiment

Figure 2 shows the COSMOS map by the Mahalanobis distance, the Bhattacharyya distance and the KL divergence. Each STAR (point) corresponds to SSD-model and is symbolized according to Table 1. In Figure 2, "Normal" and "High" speaking style are not clearly gathered in all COSMOS maps. SSD-models in both speaking style are located in center on all COSMOS maps. Therefore, these two speaking styles do not have impact on the change of Gaussian distribution shape. And each distance measure between Gaussian distributions has low sensitivity for discriminating these two speaking style. "Syllable-enhanced", "Fast", "Loud", "Lombard" and "Whisper" speaking style are



(a) Mahalanobis distamce (b) Bhattacharyya distance (c) Kullback-Leibler divergence

Figure 2 *Sensitivity for Speaking Styles*



(a) Mahalanobis distamce (b) Bhattacharyya distance (c) Kullback-Leibler divergence

Figure 3 *Sensitivity for recognition performance*

clearly gathered in all COSMOS maps. “Loud” and “Lombard” are overlapped each other because they are similar speaking styles. In the COSMOS map with the Mahalanobis distance in Figure 2 (a), “Syllable-enhanced” and “Whisper” speaking styles are overlapped. It means that the Mahalanobis distance has low sensitivity to these two speaking styles. In the same way, the Mahalanobis distance has low sensitivity to “Loud”, “Lombard” and “Fast” speaking styles. On the other hand, the Bhattacharyya distance and the KL divergence have high sensitivity for discriminating “Syllable-enhanced”, “Fast”, “Loud”, “Lombard” and “Whisper” speaking style in Figure 2 (b) and (c). In addition, SSD-models of "Syllable-enhanced" speaking style are located in the opposite side of the space where SSD-models of "Fast" speaking style are located. The axis of speaking speed is clearly shown in the COSMOS maps with the Bhattacharyya distance and the KL divergence. Furthermore, SSD-models for “Loud” and “Lombard” styles are located opposite to “Whisper”. The speaking volume axis is clearly shown. These axes are crossed in the COSMOS map with the Bhattacharyya distance and the KL divergence. In the COSMOS map with the Mahalanobis distance, they are adjacent. Therefore, the Bhattacharyya distance and the KL divergence have higher sensitivity to speaking styles than the Mahalanobis distance. In order to discriminate speaking style, the Bhattacharyya distance and the KL divergence are better. In other words, regarding the sensitivity to speaking styles, the three distance measures between Gaussian distribution are ordered as follows:

$$D_{\_ma} << D_{\_bh} \cong D_{\_kl} \tag{7}$$

In Figure 3, closed evaluation based on the speaker and speaking style independent acoustic model (called as SSI-model) was conducted for all speakers of different speaking styles. The evaluation is executed by isolated word recognition utilizing a network consisting of the 175 words contained in the vocabulary of voice samples of each speaker. In Figure 3, speakers with speech recognition performance below 90% are symbolized as “×” and speakers with recognition performance above 90% are symbolized as “□”. According to Figure 3, speakers having a lower recognition performance tend to be distributed in the periphery of all COSMOS maps. However, both Bhattacharyya distance and KL divergence have higher sensitivity than the Mahalanobis distance. In this speech recognition experiment, speaking speed has more impact on the recognition performance than speaking volume. A speaking style of almost lower performance speakers seems “Fast” and “Syllable-enhanced”. It is suggested that the variation of speaking style along speaking speed axis is bigger than that of speaking style along speaking volume axis. In the COSMOS map with the KL divergence, the variance of speaking speed is bigger than that of speaking volume. In the COSMOS map with the Bhattacharyya distance, the variance of speaking speed and speaking volume is almost the same. Therefore, the KL divergence has higher sensitivity to recognition performance than the Bhattacharyya distance. Finally, with respect to the sensitivity to recognition performance, these three distance measures between Gaussian distribution are orderd as follows:

$$D_{\_ma} < D_{\_bh} < D_{\_kl} \tag{8}$$

## 6. Summary

This paper discussed distance measures between Gaussian distributions for discriminating speaking style. A representation process of different speaking styles onto a two dimension map is introduced, and the distance measures between Gaussian used for the multidimensional scaling method, regarding the accuracy of the speaking style discrimination, is also judged.. In the experiment, the Bhattacharyya distance and the KL divergence have higher sensitivity to speaking styles than the Mahalanobis distance. In addition, the KL divergence has higher sensitivity to recognition performance.

In the future work, the three distance measures will be evaluated in a set of speaker clustering technique to build more consistent and precise acoustic models. Furthermore, sensitivity to other issues, such as noise will be investigated.

## 7. References

[1] M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass, “Effect of speaking style on LVCSR performance,” ICSLLP-96, pp. S16-S19, 1996.

[2] Adami et al., “Modeling Prosodic Dynamics for Speaker Recognition,” ICASSP-2003, Hong Kong, 2003.

[3] K. Fukunaga, “Introduction to statistical pattern recognition (Second edition),” Academic Press, Inc., San Diego, 1990.

[4] S. Kullback, “Information theory and statistic,” Dover Publications, New York, 1968.

[5] A. K. Jain et al., “Statistical pattern recognition: a review,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp.4-37, 2000.

[6] J. W. Sammon, “A nonlinear mapping for data structure analysis,” IEEE Trans. Computers, vol.C-18, no.5, pp.401-409, May 1969.

[7] M. Shozakai et al., “Acoustic space analysis method utilizing statistical multidimensional scaling technique,” Proc. NSIP-05, Sapporo, Japan, May 2005.

[8] G. Nagino et al., “Building an effective corpus by using acoustic space visualization (COSMOS) method,” IEEE ICASSP, vol. I, pp.449-452, 2005.