社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

変換不変性を有するダイバージェンスとその一般形

喬　　宇†　　峯松　信明†

† 東京大学大学院工学系研究科 〒 113–0033 東京都文京区本郷 7–3–1
E-mail: †{qiao,mine}@gavo.t.u-tokyo.ac.jp

あらまし　音声には性別や年齢、ノイズなどの非言語的情報が不可避的に含まれている。音声認識において、これらの非言語的要因に不変な特徴を抽出することは基本的かつ重要な問題である。近年峯松は [16], [17]、バタチャリヤ距離 (Bhattacharyya distance) が可逆的な変換に対する不変性を有することを証明し、音声の構造的不変表象を提案した。ここで、どのような特徴量が変換不変性を有するかが問題となる。本論文では、f-ダイバージェンスが変換に対して不変であること、および変換不変性を有する特徴は f-ダイバージェンスでの形で書ける事を示す。情報理論、統計理論においては、バタチャリヤ距離、KL-ダイバージェンス、Hellinger 距離、Pearson ダイバージェンスなどが知られているが、これらは全て f-ダイバージェンスに属している。本論文では、連続的に発声された日本語 5 母音系列を対象とした認識実験において、バタチャリヤ距離と KL-ダイバージェンスを用いた場合の認識率が最も高いことを示した。
キーワード　ダイバージェンス, 変換不変性, 構造表象, 音声認識

The general form of divergence invariant to transformations

Yu QIAO† and Nobuaki MINEMATSU†

† Grad. School of Engineering, Univ. of Tokyo 7–3–1, Hongo, Bunkyo-ku, Tokyo, 113–0033 Japan
E-mail: †{qiao,mine}@gavo.t.u-tokyo.ac.jp

Abstract　Speech signals inevitably include non-linguistic information, such as, gender, age, noise etc. Finding measures (or features) invariant to the inevitable variations caused by the non-linguistical factors (transformations) is a fundamental yet important problem in speech recognition. Recently, Minematsu [16], [17] proved that Bhattacharyya distance (BD) between two distributions is invariant to invertible transforms on feature space, and developed an invariant structural representation of speech based on it. There is a question: which kind of measures can be invariant? In this paper, we prove that f-divergence yields a generalized family of invariant measures, and show that all the invariant measures have to be written in the form of f-divergence. Many famous measures and divergences in information and statics theory, such as Bhattacharyya distance, KL-divergence, Hellinger distance, belong to f-divergence. We carried out experiments on recognizing utterances of connected Japanese vowels. The experimental results indicate that BD and KL have the best performance.
Key words　divergence, invariance, structural representation, speech recognition

1. Introduction

Speech signals carry information from multiple sources, which inevitably include variations caused by non-linguistic factors, such as, gender, age, noise etc. The same text can be converted to different acoustic observations by different speakers and by the same speaker but different time. Modern speech recognition methods deal with these variations largely by using statistical methods (such as GMM, HMM) to model the distributions of data. These methods can achieve relatively high recognition rates when using proper models and sufficient training data. However, to estimate reliable distributions, these methods always require a large number of samples for training. The successful commercial speech recognition systems always make use of millions of data from thousands of speakers for training [12]. However, it is very different from children's spoken language acquisition. A child does not need to hear the voices of thousands of people before he (or she) can understand speech. This fact largely indicates that there may exist robust measures of speech

which are nearly invariant to non-linguistic variations. It is by these robust measures, we consider that young children can learn speech by hearing very *biased* training data called "mother and father". This fact is also partly supported by recent advances in the neuroscience, which shows that the linguistic aspect of speech and the non-linguistic aspect are processed separately in the auditory cortex [25].

Recently, Minematsu found that Bhattacharyya distance (BD) is invariant to transformations (linear or nonlinear) on feature space [16], [17], and proposed an invariant structural representation of speech signal. Our previous works have demonstrated the effectiveness of invariant structural representation in both speech recognition task [2], [3], [23] and computer aided language learning (CALL) systems [18], [19].

There is a question: are there invariant measures other than BD, or, more generally, which kind of measures can be invariant? In this paper, we show that f-divergence [1], [6] provides a family of invariant measures and prove all invariant measures of integration type must be written as the forms of f-divergence. f-divergence family includes many famous distances and divergences in information and statistics, such as, Bhattacharyya distance, KL-divergence, Hellinger distance, Pearson divergence, and so on. We also carried out experiments to compare several well-known forms of f-divergence through a task of recognizing connected Japanese vowel utterances. The experimental results show that BD and KL have the best performance among the measures compared. A portion of this work will appear in [24].

2. Invariance of f-divergence

In this Section, we gives a brief introduction on f-divergence at first, and then discuss the invariant property of f-divergence. In probability theory, Csiszár f-divergence [6] (also known as Ali-Silvey distance [1]) measures the difference of two distributions. Formally,

$$f_{div}(p_i(x), p_j(x)) = \int p_j(x) g(\frac{p_i(x)}{p_j(x)}) dx, \qquad (1)$$

where $p_i(x)$ and $p_j(x)$ are two distributions on feature space X. $g : (0, \infty) \to R$ is a convex function and $g(1) = 0$. X can be an n-dimensional space with coordinates $(x_1, x_2, ..., x_n)$. In this way, Eq. 1 is a multidimensional integration and $dx = dx_1 dx_2 ... dx_n$.

f-divergence has found applications in decision theory [20], and channel and source coding [5], [26], and pattern recognition [4]. f-divergence has many beautiful properties. Csiszár [6], [7] proved the reflexivity of f-divergence,

[Lemma 1] $f_{div}(p_i(x), p_j(x)) = 0$, if and only if $p_i(x) = p_j(x)$.

Vajda [27] and Liese [15] showed the boundaries of f-divergence as

表 1 Examples of f-divergence

distance or divergence	corresponding $g(t)$ $(t = \frac{p_i(x)}{p_j(x)})$
Bhattacharyya distance (注1)	$\sqrt{t}$
KL-divergence	$t \log(t)$
Symmetric KL-divergence	$t \log(t) - \log(t)$
Hellinger distance	$(\sqrt{t} - 1)^2$
Total variation	$\lvert t - 1 \rvert$
Pearson divergence	$(t - 1)^2$
Jensen-Shannon divergence	$\frac{1}{2}(t \log \frac{2t}{t+1} + \log \frac{2}{t+1})$

[Lemma 2]

$$0 \leqq f_{div}(p_i(x), p_j(x)) \leqq \lim_{t \to 0} \{g(t) + t g(\frac{1}{t})\}. \qquad (2)$$

More properties of f-divergence can be found in [7], [15]. Many well known distances and divergences in statistics and information theory such as KL-divergence, Bhattacharyya distance, Hellinger distance etc., can be seen as special cases of f-divergence. Table 1 lists some examples.

Consider two distributions $p_i(x)$ and $p_j(x)$ in feature space X ($x \in X$). Let $h : X \to Y$ (linear or nonlinear) denote an invertible mapping (transformation) function, which convert x into new feature y. In this way, distributions $p_i(x)$ and $p_j(x)$ are transformed to $q_i(y)$ and $q_j(y)$ (Fig. 1), respectively. We wish to find measures invariant f to transformation h, $f(p_i, p_j) = f(q_i, q_j)$. The invariant measures can serve as robust features for speech analysis and classification. We have the following theorem as shown in Fig. 1.

[Theorem 1] The f-divergence between two distributions is invariant under invertible transformation h on feature space X,

$$f_{div}(p_i(x), p_j(x)) = f_{div}(q_i(y), q_j(y)). \qquad (3)$$

Proof Under transformation $y = h(x)$, distribution $q_i(y)$ is calculated by,

$$q_i(y) = p_i(h^{-1}(y)) J(y), \qquad (4)$$

where h^{-1} denotes the inverse function of h, and $J(y)$ is the absolute value of the determinant of the Jacobian matrix of function $h^{-1}(y)$.

Recall $dx = J(y)dy$, we have,

$$\begin{aligned} & f_{div}(p_i, p_j) \\ &= \int p_j(x) g(\frac{p_i(x)}{p_j(x)}) dx \\ &= \int p_j(h^{-1}(y)) g(\frac{p_i(h^{-1}(y)) J(y)}{p_j(h^{-1}(y)) J(y)}) J(y) dy \\ &= \int q_j(y) g(\frac{q_i(y)}{q_j(y)}) dy \\ &= f_{div}(q_i, q_j). \blacksquare \end{aligned} \qquad (5)$$

(注1) : Bhattacharyya distance is a function of a f-divergence: $BD(p_i, p_j) = -\log \int (p_i(x) p_j(x))^{1/2} dx = -\log f_{div}(p_i, p_j)$.

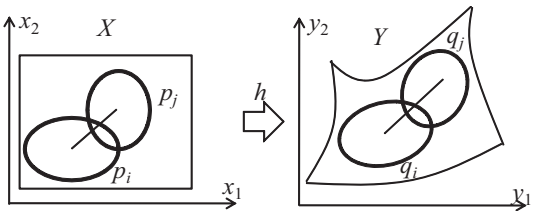

図 1 Invariance of f-divergence.

Let $F : R \to R$ denote any real value function. It is easy to see that $F(f_{div}(p_i(x), p_j(x)))$ is also invariant to transformation. In the next, we consider a more general form of Eq. 1, $M(p_i(x), p_j(x)) = \int G(p_i(x), p_j(x))p_j(x)dx$, which we call *integration measure*. There is a question, whether or not there exist invariant integration measures other than f-divergence? The answer is NO.

[Theorem 2] All the invariant integration measures have to be written in the form of $\int p_j(x)g(\frac{p_i(x)}{p_j(x)})dx$.

Proof Assume $M(p_i, p_j) = \int p_j(x)G(p_i(x), p_j(x))dx$ be an invariant integration measure, $M(p_i(x), p_j(x)) = M(q_i(y), q_j(y))$. We have,

$$\begin{aligned} & M(p_i, p_j) \\ &= \int p_j(x)G(p_j(x), p_i(x))dx \\ &= \int p_j(h^{-1}(y))G(p_i(h^{-1}(y)), p_j(h^{-1}(y)))J(y)dy \\ &= \int q_j(y)G(q_i(y)J(y)^{-1}, q_j(y)J(y)^{-1})dy \\ &\equiv M(q_i(y), q_j(y) = \int q_j(y)G(q_i(y), q_j(y))dy. \end{aligned} \tag{6}$$

Remind that $q_j(y)$ can be any distribution function. Thus the following equation must always holds,

$$G(q_i(y)J(y)^{-1}, q_j(y)J(y)^{-1}) \equiv G(q_i(y), q_j(y)). \tag{7}$$

Otherwise, we can find $q_j(y)$ that breaks Eq. 6.

Introduce functions $t(y) = q_i(y)/q_j(y)$ and $G'(t, q_j) = G(q_i, q_j)$. Thus Eq. 7 becomes:

$$G'(t(y), q_j(y)J(y)^{-1}) \equiv G'(t(y), q_j(y)). \tag{8}$$

Remind that we have no limitation on transformation h other than it is invertible. Thus it is possible to set that $q_j(y) = J(y)$. Then, we have,

$$G'(t(y), q_j(y)) \equiv G'(t(y), 1). \tag{9}$$

Therefore $G'(t(y), q_j(y))$ can be written into the form of $G'(t(y)) = g(q_i(y)/q_j(y))$. In this way, we prove that $M(p_i(x), p_j(x))$ has to be written in the form of $\int p_j(x)g(\frac{p_i(x)}{p_j(x)})dx$. ■

Theorem 1 and Theorem 2 together show the sufficiency and necessity of the invariance of f-divergence. Generally, f-divergence may not be a metric, since it may not satisfy symmetry rule ($f_{div}(p_i(x), p_j(x)) \neq f_{div}(p_j(x), p_i(x))$) and subadditivity triangle inequality ($f_{div}(p_i(x), p_j(x)) + f_{div}(p_j(x), p_k(x)) < f_{div}(p_i(x), p_k(x))$). But there exist special forms of f-divergence, which is also a metric. Hellinger distance is such an example,

$$\begin{aligned} HD(p_i, p_j) &= \int p_i(x)(\sqrt{\frac{p_j(x)}{p_i(x)}} - 1)^2 dx \\ &= \int (\sqrt{p_i(x)} - \sqrt{p_j(x)})^2 dx. \end{aligned} \tag{10}$$

More generally, it was shown that a subclass of f-divergence, named f_β-divergence, also satisfies the constraints of metric [21].

3. Calculation of f-divergence

There is a problem of how to calculate f-divergence. Unfortunately, in general case, there exists no closed-form solution for f-divergence of Eq. 1. In the next, we will discuss several techniques to calculate f-divergence for the general case and for the special types of distributions.

3.1 Calculation of f-divergence using Monte-Carlo sampling

Since the direct calculation of f-divergence is intractable, we can consider approximate methods based on Monte-Carlo sampling [8]. This method draws a set of independent samples $\{x^k\}_{k=1}^K$ from the distribution $p_j(x)$ at first. Assume K is large enough. Then, f-divergence can be approximated by

$$f_\alpha(p_i(x), p_j(x)) \approx \frac{1}{n}\sum_{k=1}^{K} g(\frac{p_i(x^k)}{p_j(x^k)}). \tag{11}$$

But this can be always computationally expensive. Especially when x has a high dimension, we need a huge number of random vectors for approximating f-divergence.

3.2 f-divergence of Gaussian distributions

When the distributions are Gaussian, there may exist closed-form solutions. Assume $p_i(x)$ and $p_j(x)$ are two Gaussian distributions with mean μ_i and μ_j and covariance matrix Σ_i and Σ_j, respectively. The canonical parametrization of $p_i(x)$ is,

$$p_i(x) = \exp(\alpha_i + \eta_i^T x - \frac{1}{2}x^T \Lambda_i x), \tag{12}$$

where $\Lambda_i = \Sigma_i^{-1}$, $\eta_i = \Sigma_i^{-1}\mu_i$ and $\alpha_i = -0.5(d\log 2\pi - \log|\Lambda_i| + \eta_i^t \Lambda_i \eta_i$. Similarly, we have

$$p_j(x) = \exp(\alpha_j + \eta_j^T x - \frac{1}{2}x^T \Lambda_j x). \tag{13}$$

where $\Lambda_j = \Sigma_j^{-1}$, $\eta_j = \Sigma_j^{-1}\mu_j$ and $\alpha_j = -0.5(d\log 2\pi - \log|\Lambda_j| + \eta_j^t \Lambda_j \eta_j)$. Then, Eq. 1 can be written into,

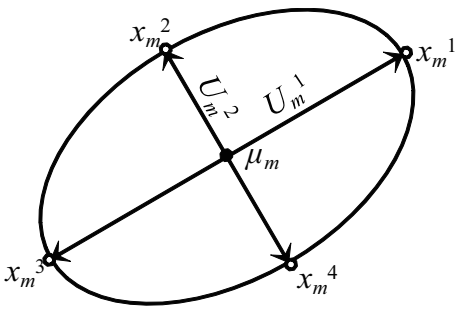

図 2 Example of sigma points.

$$f_{div}(p_i(x), p_j(x)) = \int \exp\left(\alpha_j + \eta_j^T x - \frac{1}{2}x^T \Lambda_j x\right)$$
$$g(\exp(\alpha_i - \alpha_j + (\eta_i - \eta_j)^T x - \frac{1}{2}x^T(\Lambda_i - \Lambda_j)x))dx. \quad (14)$$

The above form is near to Fourier transform or bilateral Laplace transform which has been widely studied. Many forms of g can lead to closed form solutions of the integrations of f-divergence. Some examples are given as follows,

1) Bhattacharyya distance:

$$BD(p_i(x), p_j(x)) =$$
$$\frac{1}{8}(\mu_i - \mu_j)^T(\frac{\Sigma_i + \Sigma_j}{2})^{-1}(\mu_i - \mu_j) + \frac{1}{2}\log\frac{|(\Sigma_i + \Sigma_j)/2|}{|\Sigma_i|^{1/2}|\Sigma_j|^{1/2}}. \quad (15)$$

2) KL divergence:

$$KL(p_i(x), p_j(x)) =$$
$$\frac{1}{2}(\log\frac{|\Sigma_j|}{|\Sigma_i|} + \mathrm{tr}(\Sigma_j^{-1}\Sigma_i) + (\mu_j - \mu_i)^T\Sigma_j^{-1}(\mu_j - \mu_i)). \quad (16)$$

3) Hellinger distance:

$$HD(p_i(x), p_j(x)) = 1 - \exp(-BD(p_i(x), p_j(x))). \quad (17)$$

3.3 f-divergence of Gaussian Mixtures

When $p_i(x)$ and $p_j(x)$ are Gaussian mixtures, there exist fast approximation techniques other than Monte Carlo sampling. For example, one can use unscented transform [10], [13] to calculate the f-divergence. The procedure is described as follows,

Let Gaussian mixture $p_j(x) = \sum_{m=1}^{M} w_m N(x|\mu_m, \Sigma_m)$. For each Gaussian distribution $N(x|\mu_m, \Sigma_m)$, we can calculate a set of $2n$ "sigma" points as

$$x_m^k = \mu_m + \sqrt{\lambda_m^k}U_m^k, \quad (18)$$
$$x_m^{k+n} = \mu_m - \sqrt{\lambda_m^k}U_m^k, \quad (19)$$

where $(k = 1, 2, ..., n)$, λ_m^k and U_m^k are the k-th eigenvalue and eigen vector of Σ_m, respectively. It is not hard to see that these points could capture the mean and covariance information of $N(x|\mu_m, \Sigma_m)$. Examples of sigma points are depicted in Fig. 2.

Using unscented transform, f-divergence can be approximated by the following formula,

$$f_{div}(p_i(x), p_j(x)) \approx \frac{1}{2n}\sum_{m=1}^{M} w_m \sum_{k=1}^{2n} g(\frac{p_i(x_m^k)}{p_j(x_m^k)}). \quad (20)$$

Although the above calculation resembles the Monte-Carlo sampling, it doesn't require random sampling, and it only needs a small number of points. Therefore, it is much faster than the Monte-Carlo sampling. One may also consider the variational approximation techniques to calculate the f-divergence between two Gaussian mixtures [11].

4. Invariant structural representation using f-divergence

f-divergence can be used to construct the invariant structural representation of a pattern. Consider pattern P in feature space X. Suppose P can be decomposed into a sequence of m events $\{p_i\}_{i=1}^m$. Each event is described as a distribution $p_i(x)$. We calculate the f-divergence d_{ij}^P between two distributions $p_i(x), p_j(x)$, and construct an $m \times m$ divergence matrix D^P with $D^P(i,j) = d_{ij}^P$ and $D^P(i,i) = 0$. Then D^P provides a structural representation of pattern P. Assume there is a map $f : X \to Y$ (linear or nonlinear) which transforms X into a new feature space Y. In this way, pattern P in X is mapped to pattern Q in Y, and event p_i is transformed to event q_i. Similarly, we can calculate structure representation D^Q for pattern Q. From Theorem 1, we have that $D^Q = D^P$, which indicates that the structural representation based on f-divergence is invariant to transformations on feature space.

In the next, we describe a brief introduction on how to obtain a structural representation from an utterance [2], [16]. As shown in Fig. 3, at first, we calculate a sequence of cepstral features from input speech waveforms. Then an HMM is trained based on that cepstrum sequence and each state of HMM is regarded as event p_i. Thirdly we calculate the f-divergences between each pair of p_i and p_j. These distances will form an $m \times m$ distance matrix D with zero diagonal, which is the structural representation. For convenience, we can expand D into a vector z with dimension $m(m-1)$. If the f-divergence used satisfies the symmetry rule $f_{div}(p_i, p_j) = f_{div}(p_j, p_i)$ (for examples, Bhattacharyya distance, Hellinger distance, total variations), D is a symmetric matrix. In this case, we only need use the upper triangle of D and z has dimension $m(m-1)/2$.

It can be shown that many non-linguistic variations [16], [17], such as the length of vocal tract [22], can be modeled as transformation of feature space. Suppose that X and Y represent the acoustic spaces of two speakers A and B, and P and Q represent two utterances of A and B, respectively.

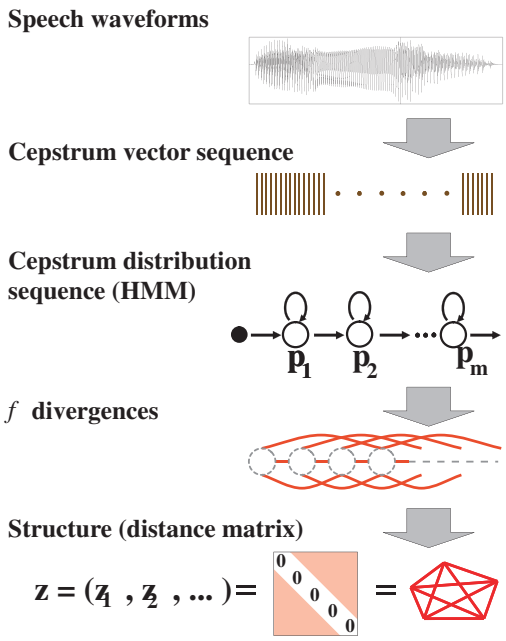

図 3　Framework of structure construction.

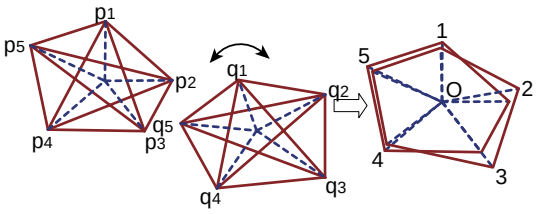

図 4　Utterance matching by shift and rotation.

Then h can be seen as a mapping function from A's utterance to B's. In fact, this problem has been widely addressed in the speaker adaptation of speech recognition research and the speaker conversion of speech synthesis research. In Maximum Likelihood Linear Regression (MLLR) based speaker adaption [14], a linear transformation: $y = h(x) = Hx + d$ is used, where H and d denote rotation and translation parameters respectively. For matching utterances P and Q, the speaker adaption methods need to explicitly estimate transformation parameters (i.e. H and d), which lead to the minimum difference. This minimum difference serves as a matching score of utterances. [17] showed that the acoustic matching score of two utterances after shift and rotation (Fig.4) can be approximated only with the difference of the two structures of the utterances without explicitly estimating transformation parameters.

5. Experiments

To compare the performance of various forms of f-divergence on speech recognition, we used the connected Japanese vowel utterances [2] in experiments. It is known that acoustic features of vowel sounds exhibit larger between-speaker variations than consonant sounds. Each word in the data set corresponds to a combination of the five Japanese vowels 'a','e','i','o' and 'u', such as 'aeiou','uoaie', ... So there are totally 120 words. The utterances of 16 speakers (8 males and 8 females) were recorded. Every speaker provides 5 utterances for each word. So the total number of utterances is 16×120×5=9,600. Among them, we use 4,800 utterances from 4 male and 4 female speakers for training and the other 4,800 utterances for testing.

表 2　Comparisons of recognition rates

Method	NN	NM	GM	RDSA
Bhattacharyya dis.	93.0%	95.6%	96.4%	98.2%
Hellinger dis.	89.0%	95.1%	56.6%	96.0%
symmetric KL-div.	93.2%	95.6%	96.4%	**98.4%**

For each utterance, we calculate twelve Mel-cepstrum features and one power coefficient. Then HMM training is used to convert a cepstrum vector sequence into 25 events (distributions). Since we have only one training sample, we used an MAP-based learning algorithm [9]. Each state (event) of an HMM is described by a 13-dimension Gaussian distribution with a diagonal covariance matrix. Following [2], we divided the 13D cepstrum feature steam into 13 multiple sub-streams and calculated the structures for each sub-stream. So an utterance is represented as a set of $25 \times 24 \times 13 = 7,800$ edges. When using symmetric f-divergence, such as BD and HD, only half of the edges (3,900) are necessary. More details can be found in our previous works [2], [23].

We calculated the Bhattacharyya distance (BD), Hellinger distance (HD) and symmetric KL-divergence (SKL) for building structures, respectively. As for classification, we used the following classifiers: nearest neighbors (NN), nearest mean (NM), Gaussian distribution model (GM) and random discriminant structure analysis (RDSA) [23]. For NN and NM, Euclidean distance is used. For GM, we used diagonal covariance matrices. For RDSA [23], we used 20 randomly selected sub-structures with each including 700 edges. The results are summarized in Table 2. We can find that the performances of symmetric KL-divergence and Bhattacharyya distance are similar. And Hellinger distance has the lowest recognition rates.

We reduces the numbers of speakers in training data. We randomly selected k ($1 \leqq k \leqq 7$) speakers from the 8 training speakers and use their data for learning the classifiers. For each k, we repeat this procedure 8 times and calculate the average recognition performance. The RDSA classifier is used for classification due to its good performance. The results are given in Fig. 5.

6. Conclusions

Speech recognition faces the difficulty of non-linguistic

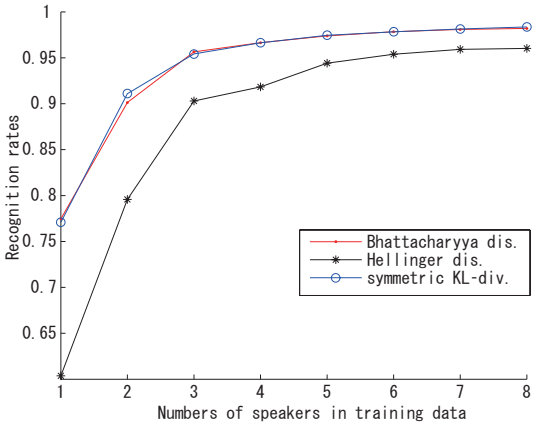

図 5 Comparison of the recognition rates of different distances and different numbers of speakers in training data.

variations exhibited by speech signals. Recently, an invariant representation for speech has been proposed for speech recognition, which is composed by Bhattacharyya distances invariant to transformation. So there is a question which kind of measure can be invariant. This paper proves that f-divergence between two distributions is invariant to invertible transformation (linear and nonlinear) on feature space, and shows all invariant integration measures have to be written in the form of f-divergence. We discuss the properties of f-divergence and study how to calculate f-divergence for the general case and for Gaussian and Gaussian mixture distributions. We described a short review on how to construct an invariant structural representation of an utterance by using f-divergences. In the experiment, we compare the performance of several well-known forms of f-divergences through recognizing utterances of Japanese vowels. The results show that Bhattacharyya distance and symmetric KL-divergence achieve the best performance among all the measures compared. It is noted that the invariance of f-divergence is very general, and doesn't limit to speech signal. The proposed theories may have applications in other signal analysis and pattern recognition tasks.

7. Acknowledgment

The first author would like to thank the Japan Society for the Promotion of Science (JSPS) for the financial support under contract P07078.

文　　献

[1] S. M. Ali and S. D. Silvey. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142, 1966.

[2] S. Asakawa, N. Minematsu, and K. Hirose. Automatic recognition of connected vowels only using speaker-invariant representation of speech dynamics. *Proc. INTERSPEECH*, pages 890–893, 2007.

[3] S. Asakawa, N. Minematsu, and K. Hirose. Multi-stream parameterization for structural speech recognition. *Proc. ICASSP*, pages 4097–4100, 2008.

[4] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349–369, 1989.

[5] T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley New York, 1991.

[6] I. Csiszar. Information-type measures of difference of probability distributions and indirect. *Stud. Sci. Math. Hung.*, 2:299–318, 1967.

[7] I. Csiszar and P.C. Shields. *Information Theory And Statistics: A Tutorial*. Now Publishers Inc, 2004.

[8] G.S. Fishman. *Monte Carlo: Concepts, Algorithms, and Applications*. Springer, 1996.

[9] J. L. Gauvain and C.H. Lee. Maximum a posteriori estimation for multivariate GM observations of Markov chains. *IEEE Trans. SAP*, 2(2):291–298, 1994.

[10] J. Goldberger and H. Aronowitz. A Distance Measure Between GMMs Based on the Unscented Transform and its Application to Speaker Recognition. *Proc. of Eurospeech*, pages 1985–1989, 2005.

[11] J. R. Hershey and P. A. Olsen. Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models. *Proc. ICASSP*, pages 317–320, 2007.

[12] http://tepia.or.jp/archive/12th/pdf/viavoice.pdf.

[13] S.J. Julier and J.K. Uhlmann. A new extension of the Kalman filter to nonlinear systems. *Int. Symp. Aerospace/Defense Sensing, Simul. and Controls*, 3, 1997.

[14] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–185, 1995.

[15] F. Liese and I. Vajda. On Divergences and Informations in Statistics and Information Theory. *Information Theory, IEEE Transactions on*, 52(10):4394–4412, 2006.

[16] N. Minematsu. Yet another acoustic representation of speech sounds. *Proc. ICASSP*, pages 585–588, 2004.

[17] N. Minematsu. Mathematical Evidence of the Acoustic Universal Structure in Speech. *Proc. ICASSP*, pages 889–892, 2005.

[18] N. Minematsu and et. al. Structural representation of the pronunciation and its use for CALL. *Proc. of IEEE Spoken Lan. Tech. Workshop*, pages 126–129, 2006.

[19] N. Minematsu and et. al. Structural assessment of language learners' pronunciation. *Proc. INTERSPEECH*, pages 210–213, 2007.

[20] F. Österreicher and I. Vajda. Statistical information and discrimination. *Information Theory, IEEE Transactions on*, 39(3):1036–1039, 1993.

[21] F. Österreicher and I. Vajda. A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3):639–653, 2003.

[22] M. Pitz and H. Ney. Vocal Tract Normalization Equals Linear Transformation in Cepstral Space. *IEEE Trans. SAP*, 13(5):930–944, 2005.

[23] Y. Qiao, S. Asakawa, and N. Minematsu. Random discriminant structure analysis for automatic recognition of connected vowels. *Proc. ASRU*, pages 576–581, 2007.

[24] Y. Qiao and N. Minematsu. f-divergence is a generalized invariant measure between distributions. *Proc. INTERSPEECH*, 2008 (accepted).

[25] S. K. Scott and I. S. Johnsrude. The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, 26(2):100–107, 2003.

[26] F. Topsøe. Information theoretical optimization techniques. *Kybernetika, Praha*, 15:8–27, 1979.

[27] I. Vajda. On the f -divergence and singularity of probability measures. *Periodica Math. Hungar*, 2:223–234, 1972.