# Structural representation with a general form of invariant divergence *

◎ Yu Qiao, and Nobuaki Minematsu (The Univ. of Tokyo)

## 1 Introduction

There always exist variations caused by non-linguistic factors, such as, gender, age, noise etc in speech signals. The same phoneme sequences can be converted to various acoustic observations by different speakers and by the same speaker but at different times. Modern speech recognition methods deal with these variations mainly by taking advantage of 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 [1]. 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".

Recently, Minematsu found that Bhattacharyya distance (BD) is invariant to transformations (linear or nonlinear) on feature space [2, 3], 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 [4, 5, 6] and computer aided language learning (CALL) systems [7, 8].

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 [9, 10] provides a family of invariant measures and prove all invariant measures of integration type must be written in the form 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 [11].

## 2 Invariance of $f$-divergence

In this section, we give a brief introduction on $f$-divergence at first, and then discuss the invariant property of $f$-divergence. In probability theory, Csiszár $f$-divergence [9] (also known as Ali-Silvey distance [10]) measures the difference of two distributions. Formally,

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

where $p_i(x)$ and $p_j(x)$ are two distributions on feature space $X$. $g : [0, \infty) \rightarrow R$ 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 $dx = dx_1 dx_2 ... dx_n$. 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.

Consider two distributions $p_i(x)$ and $p_j(x)$ in feature space $X$ ($x \in X$). Let $h : X \rightarrow Y$ (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 $f$ invariant 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 $X$,*

$$f_{div}(p_i(x), p_j(x)) = f_{div}(q_i(y), q_j(y)). \quad (2)$$

*一般的な形式の不変ダイバージェンを用いたの構造表象. 喬宇, 峯松信明

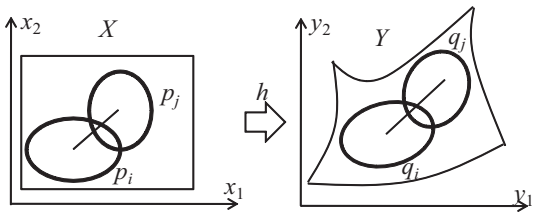


Fig. 1 Invariance of $f$-divergence.

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, $M(p_i(x), p_j(x)) = \int p_j(x) G(p_i(x), p_j(x))dx$, 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* $\int p_j(x) g(\frac{p_i(x)}{p_j(x)})dx$.

Theorem 1 and Theorem 2 together show the sufficiency and necessity of the invariance of $f$-divergence. We can generalize the invariant measure from two distributions to $n$ distributions. Using the similar analysis of the proofs for Theorem 1 and Theorem 2, we have,

**Theorem 3** *The invariant measure for n distributions* $p_1(x), p_2(x), ..., p_n(x)$ *have and must have the following* $F$*-measure form*

$$D_F(p_1(x), p_2(x), ..., p_n(x)) = \tag{3}$$

$$\int p_n(x) F(\frac{p_1(x)}{p_2(x)}, \frac{p_2(x)}{p_3(x)}, ..., \frac{p_{n-1}(x)}{p_n(x)})dx, \tag{4}$$

where $F : R^{n-1} \to R$. 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, $HD(p_i, p_j) = \int (\sqrt{p_i(x)} - \sqrt{p_j(x)})^2 dx$. More generally, it was shown that a subclass of $f$-divergence, named $f_\beta$-divergence, also satisfies the constraints of metric [12].

## 3 Calculation of $f$-divergence

There is a problem of how to calculate $f$-divergence. Unfortunately, in general cases, 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 general cases and for a few 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 [13]. 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

$$f_\alpha(p_i(x), p_j(x)) \approx \frac{1}{n} \sum_{k=1}^{K} g(\frac{p_i(x^k)}{p_j(x^k)}). \tag{5}$$

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 $\mu_i$ and $\mu_j$ and covariance matrix $\Sigma_i$ and $\Sigma_j$, respectively. Some examples are given as follows,

1) Bhattacharyya distance:

$$BD(p_i(x), p_j(x)) = \frac{1}{8}(\mu_i - \mu_j)^T (\frac{\Sigma_i + \Sigma_j}{2})^{-1} (\mu_i - \mu_j) + \frac{1}{2} \log \frac{|(\Sigma_i + \Sigma_j)/2|}{|\Sigma_i|^{1/2} |\Sigma_j|^{1/2}}. \tag{6}$$

2) KL divergence:

$$KL(p_i(x), p_j(x)) = \frac{1}{2}(\log \frac{|\Sigma_j|}{|\Sigma_i|} + \mathrm{tr}(\Sigma_j^{-1}\Sigma_i) + (\mu_j - \mu_i)^T \Sigma_j^{-1} (\mu_j - \mu_i)). \tag{7}$$

3) Hellinger distance:

$$HD(p_i(x), p_j(x)) = 1 - \exp(-BD(p_i(x), p_j(x))). \tag{8}$$

### 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 [14, 15] to calculate the $f$-divergence. 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

$$x_m^k = \mu_m + \sqrt{\lambda_m^k} U_m^k, \tag{9}$$

$$x_m^{k+n} = \mu_m - \sqrt{\lambda_m^k} U_m^k, \tag{10}$$

where $(k = 1, 2, ..., n)$, $\lambda_m^k$ and $U_m^k$ are the $k$-th eigenvalue and eigen vector of $\Sigma_m$, respectively. It is

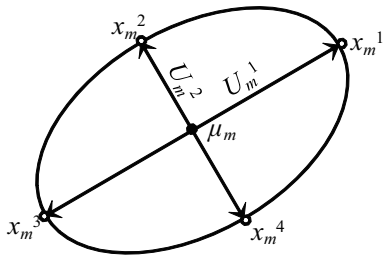


Fig. 2 An examples of a set of sigma points.

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,

$$f_{div}(p_i(x), p_j(x)) \approx \frac{1}{2n} \sum_{m=1}^{M} w_m \sum_{k=1}^{2n} g(\frac{p_i(x_m^k)}{p_j(x_m^k)}). \tag{11}$$

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 [16].

## 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, 4]. As shown in Fig. 3, at first, we

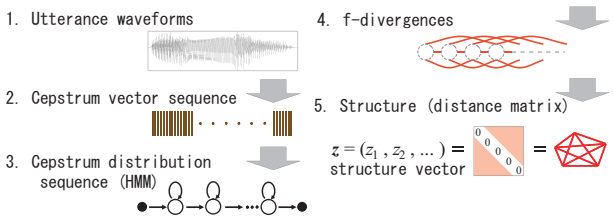


Fig. 3 Framework of structure construction.

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 $m(m-1)/2$. The Euclidean distance between two structural representations serves as a matching score of utterances. It is shown that, using structural representation, we can approximate the difference without explicitly estimating transformation parameters [17].

## 5 Experiments

To compare the performance of various forms of $f$-divergence on speech recognition, we used the connected Japanese vowel utterances [4] in experiments. 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 provided 5 utterances for each word. So the total number of utterances is 16×120×5=9,600. Among them, we used 4,800 utterances from 4 male and 4 female speakers for training and the other 4,800 utterances for testing. 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 [18]. Each state (event) of an HMM is described by a 13-dimension Gaussian distribution with a diagonal covariance matrix.

Table 1 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%** |

Following [4], 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 [4, 5].

We calculate the structural representations by using Bhattacharyya distance (BD), Hellinger distance (HD) and symmetric KL-divergence (SKL), 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) [5]. For NN and NM, Euclidean distance is used. For GM, we used diagonal covariance matrices. For RDSA [5], we used 20 randomly selected substructures with each including 700 edges. The results are summarized in Table 1. We can find that the performances of symmetric KL-divergence and Bhattacharyya distance are similar. And Hellinger distance has the lowest recognition rates.

## 6 Conclusions

One of the basic problems in speech recognition is to deal with the non-linguistic 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 yields a family of measures 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 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 experiments, 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.

## 参考文献


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

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

[3] N. Minematsu, S. Asakawa, and K. Hirose, "Linear and non-linear transformation invariant representation of information and its use for acoustic modeling of speech," *Proc. Spring Meeting Acoust. Soc. Jpn.*, pp. 147–148, 2007.

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

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

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

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

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

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

[10] 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)*, vol. 28, no. 1, pp. 131–142, 1966.

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

[12] 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*, vol. 55, no. 3, pp. 639–653, 2003.

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

[14] 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*, vol. 3, 1997.

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

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

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

[18] J. L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate GM observations of Markov chains," *IEEE Trans. SAP*, vol. 2, no. 2, pp. 291–298, 1994.