interspeech 2008 incorporating SST 08

f-Divergence is a Generalized Invariant Measure Between Distributions

Yu Qiao and Nobuaki Minematsu

Graduate School of Engineering, The University of Tokyo, Tokyo, Japan
{qiao, mine}@gavo.t.u-tokyo.ac.jp

Abstract

Finding measures (or features) invariant to inevitable variations caused by non-linguistical factors (transformations) is a fundamental yet important problem in speech recognition. Recently, Minematsu [1, 2] proved that Bhattacharyya distance (BD) between two distributions is invariant to invertible transforms on feature space, and develop 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 forms of f-divergence. Many famous distances and divergences in information and statistics, such as Bhattacharyya distance (BD), KL-divergence, Hellinger distance, can be written into forms of f-divergence. As an application, we carried out experiments on recognizing the utterances of connected Japanese vowels. The experimental results show that BD and KL have the best performance among the measures compared.

Index Terms: f-divergence, invariant measure, invertible transformation, speech recognition

1. Introduction

Speech signals inevitably exhibit variations caused by non-linguistic factors, such as, gender, age, noise etc. The same text can be converted to different acoustic observations due to the differences of speaker and environments. Modern speech recognition methods deal with these variations largely by using the statistical methods (such as GMM, HMM) to model the distributions of the 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 [3]. 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 [4].

Recently, Minematsu found that Bhattacharyya distance (BD) is invariant to transformations (linear or nonlinear) on feature space [1, 2], 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 [5, 6, 7] and computer aided language learning (CALL) systems [8, 9].

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 [10, 11] 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.

Table 1: Examples of f-divergence

distance or divergence	corresponding $g(t)$ ($t = \frac{p_i(x)}{p_j(x)}$)
Bhattacharyya distance [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	$\frac{1}{2}(t\log\frac{2t}{t+1} + \log\frac{2}{t+1})$

2. Invariance of f-divergence

In probability theory, Csiszár f-divergence [10] (also known as Ali-Silvey distance [11]) 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. $g(t)$ is a convex function defined for $t > 0$, and $g(1) = 0$. X can be a n-dimensional space with coordinates $(x_1, x_2, ..., x_n)$. In this way, Eq. 1 is a multidimensional integration and $dx = dx_1 dx_2 ... dx_n$. Generally, it is required that $f_{div}(p_i(x), p_j(x)) \geq 0$ for any two distributions $p_i(x), p_j(x)$. It can be proved that $f_{div}(p_i(x), p_j(x)) = 0$, if and only if $p_i(x) = p_j(x)$ [12]. Many well known distances and divergences in statistics and information theory can be seen as special examples of f-divergence. Table 1 lists some examples.

Consider feature space X and two distributions $p_i(x)$ and $p_j(x)$ in X ($x \in X$). Let $h : X \rightarrow Y$ (linear or nonlinear) denote an invertible mapping (transformation) function, which

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

Accepted after peer review of full paper

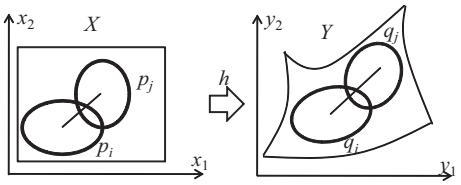

Figure 1: Invariance of f-divergence.

convert x into new feature y. In this way, distributions $p_i(x)$ and $p_j(x)$ is transformed to $q_i(y)$ and $q_j(y)$, 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,*

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

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

$$q_i(y) = p_i(h^{-1}(y))J(y), \quad (3)$$

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,

$$\begin{aligned} & f_{div}(p_i, p_j) \\ &= \int p_j(x) g(\frac{p_i(x)}{p_j(x)}) dx \\ &= \int p_j(h^{-1}(y)) g(\frac{p_i(h^{-1}(y))J(y)}{p_j(h^{-1}(y))J(y)}) J(y) dy \\ &= \int q_j(y) g(\frac{q_i(y)}{q_j(y)}) dy \\ &= f_{div}(q_i, q_j). \blacksquare \end{aligned} \quad (4)$$

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 form $\int p_j(x) g(\frac{p_i(x)}{p_j(x)}) dx$.*

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,

$$\begin{aligned} & M(p_i, p_j) \\ &= \int p_j(x) G(p_j(x), p_i(x)) dx \\ &= \int p_j(h^{-1}(y)) G(p_i(h^{-1}(y)), p_j(h^{-1}(y))) J(y) dy \\ &= \int q_j(y) G(q_i(y) J(y)^{-1}, q_j(y) J(y)^{-1}) dy \\ &\equiv M(q_i(y), q_j(y) = \int q_j(y) G(q_i(y), q_j(y)) dy. \end{aligned} \quad (5)$$

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

$$G(q_i(y)J(y)^{-1}, q_j(y)J(y)^{-1}) \equiv G(q_i(y), q_j(y)). \quad (6)$$

Otherwise, we can find $q_j(y)$ that breaks Eq. 5.

Introduce functions $t(y) = q_i(y)/q_j(y)$ and $G'(t, q_j) = G(q_i, q_j)$. Thus Eq. 6 becomes:

$$G'(t(y), q_j(y)J(y)^{-1}) \equiv G'(t(y), q_j(y)). \quad (7)$$

Remind that we don't have any limitations on transformation h. Thus it is possible to set that $q_j(y) = J(y)$. Then, we have,

$$G'(t(y), q_j(y)) \equiv G'(t(y), 1). \quad (8)$$

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 $\int p_j(x) g(\frac{p_i(x)}{p_j(x)}) dx$. ■

Theorem 1 and Theorem 2 together show the sufficiency and necessary 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, $HD(p_i, p_j) = \int (\sqrt{p_i(x)} - \sqrt{p_j(x)})^2 dx$.

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

$$p_i(x) = \exp(\alpha_i + \eta_i^T x - \frac{1}{2} x^T \Lambda_i x), \quad (9)$$

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

$$p_j(x) = \exp(\alpha_j + \eta_j^T x - \frac{1}{2} x^T \Lambda_j x). \quad (10)$$

Then, Eq. 1 can be written into,

$$\begin{aligned} & f_{div}(p_i(x), p_j(x)) = \int \exp(\alpha_j + \eta_j^T x - \frac{1}{2} x^T \Lambda_j x) \\ & g(\exp(\alpha_i - \alpha_j + (\eta_i - \eta_j)^T x - \frac{1}{2} x^T (\Lambda_i - \Lambda_j) x)) dx. \end{aligned} \quad (11)$$

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:

$$\begin{aligned} & BD(p_i(x), p_j(x)) = \\ & \frac{1}{8}(\mu_i - \mu_j)^T (\frac{\Sigma_i + \Sigma_j}{2})^{-1} (\mu_i - \mu_j) + \frac{1}{2} \log \frac{|(\Sigma_i + \Sigma_j)/2|}{|\Sigma_i|^{1/2} |\Sigma_j|^{1/2}}. \end{aligned} \quad (12)$$

2) KL divergence:

$$KL(p_i(x), p_j(x)) = \frac{1}{2}(\log \frac{|\Sigma_j|}{|\Sigma_i|} + \mathrm{tr}(\Sigma_j^{-1}\Sigma_i) + (\mu_j - \mu_i)^T \Sigma_j^{-1}(\mu_j - \mu_i)). \quad (13)$$

3) Hellinger distance:

$$HD(p_i(x), p_j(x)) = 1 - \exp(-BD(p_i(x), p_j(x))). \quad (14)$$

In general case, we can use Monte-Carlo sampling to calculate f-divergence. But this is always computationally expensive, especially when x has a high dimension. When $p_i(x)$ and $p_j(x)$ are Gaussian mixtures, one may consider approximated techniques, such as, unscented transform [13] and variational approximation for fast calculation [14].

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 $f : X \rightarrow Y$ (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 [1, 5]. As shown in Fig. 2, 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 [1, 2], such as the length of vocal tract [15], can be modeled as the 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. 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 [16], 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.

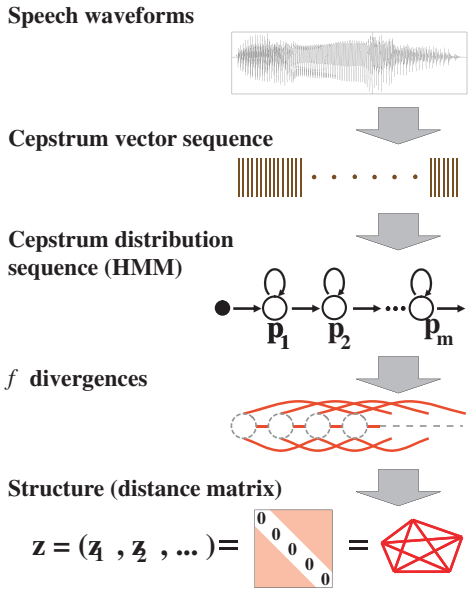

Figure 2: Framework of structure construction.

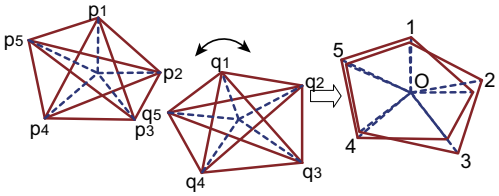

Figure 3: Utterance matching by shift and rotation.

H and d), which lead to the minimum difference. This minimum difference serves as a matching score of utterances. [2] showed that the acoustic matching score of two utterances after shift and rotation (Fig.3) 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 [5] 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.

For each utterance, we calculate the 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 [17]. Each state (event) of a HMM is described by a 13-dimension Gaussian distribution with a diagonal covariance matrix. Following [5], we divided

Table 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%**

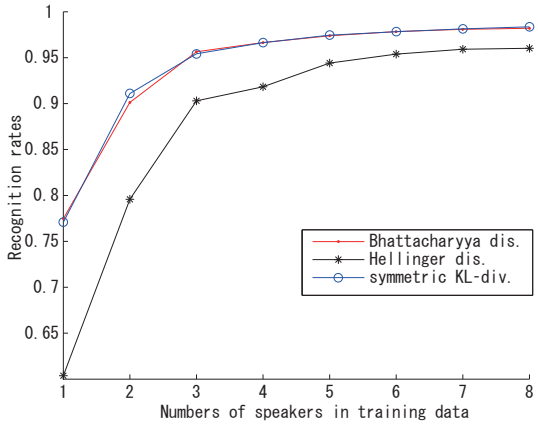

Figure 4: Comparison of the recognition rates of different distances and different numbers of speakers in training data.

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. More details can be found in our previous works [5, 6].

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) [6]. For NN and NM, Euclidean distance is used. For GM, we used diagonal covariance matrices. For RDSA [6], we used 20 randomly selected sub-structures with each structure 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 ($1 \leq k \leq 7$) 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. 4.

6. Conclusions

This paper proves that f-divergence between two distributions is invariant to invertible transformation (linear and nonlinear) on feature space, and show all invariant integration measures have to be written in the forms of f-divergence. We discuss how to construct an invariant structural representation of an utterance by using f-divergences. We compare the recognition performance of several well-known forms of f-divergences through speech recognition experiments. The results show that Bhattacharyya distances and symmetric KL-divergence achieve the best performance. 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 pattern analysis and recognition tasks.

7. References

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

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

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

[4] S. K. Scott and I. S. Johnsrude, "The neuroanatomical and functional organization of speech perception," *Trends in Neurosciences*, vol. 26, no. 2, pp. 100–107, 2003.

[5] 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.

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

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

[8] 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.

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

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

[11] SM Ali and SD 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.

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

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

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

[15] M. Pitz and H. Ney, "Vocal Tract Normalization Equals Linear Transformation in Cepstral Space," *IEEE Trans. SAP*, vol. 13, no. 5, pp. 930–944, 2005.

[16] CJ Leggetter and PC Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[17] 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.

[18] T. Kawahara and et. al., "Recent progress of open-source LVCSR engine Julius and Japanese model repository," *Proc. ICSLP*, pp. 3069–3072, 2004.