A Study on Invariance of $f$-divergence and Its Application to Speech Recognition

Yu Qiao, Member, IEEE, and Nobuaki Minematsu, Member, IEEE,

Abstract—Identifying features invariant to certain transformations is a fundamental problem in the fields of signal processing and pattern recognition. This paper explores a family of measures called $f$-divergences that are invariant to invertible transformations, and studies their application to speech recognition. We provide novel proofs for the sufficiency and necessity of the invariance of $f$-divergence. Several techniques to calculate or approximate $f$-divergences in general cases and for special distributions such as Gaussian and Gaussian mixture are reviewed. We show how to construct an invariant structural representation from sequence data through maximum likelihood decomposition, and prove the invariance of this decomposition. We demonstrate an application of this invariant representation to recognizing connected Japanese vowel utterances. In addition, we propose several techniques to improve the recognition performance. The experimental results show that the invariant structure achieves better performance than Hidden Markov Models, a widely used technique for acoustic modeling of speech sounds.

Index Terms—$f$-divergence, invariance to transformation, structural representation, speech recognition.

I. INTRODUCTION

Natural signals usually include variations irrelevant to the objective of analysis. In computer vision, the same object can result in different images due to viewpoint and illumination changes. In speech recognition, the same text is acoustically realized in different ways by different speakers, and thus leads to different acoustic signals [1]. Many of these variations can be modeled by certain transformations of input signals. For images, the change in viewpoint can be modeled as projective transformation. For speech, on the other hand, the change of vocal tract length corresponds approximately to linear transformations of cepstrum features [2]. In many signal processing tasks, one may want to reduce the effect of these unfavorable variations. Generally speaking, there are two approaches to this goal. One is to model the distribution of the variations. But estimating reliable distributions may require a large amount of data, especially for high-dimensional data. The other approach is to use features invariant to transformations.

This paper studies invariant measures between distributions. In previous works, the second author noted that the Bhattacharyya distance (BD) is invariant to transformations on feature space [3], and proposed an invariant structural representation of speech signal. The effectiveness of this structural representation has been demonstrated in automatic speech recognition (ASR) tasks [4], [5], computer-aided pronunciation training (CAPT) systems [6] and speech synthesis [7]. BD is an example of a family of invariant measures called $f$-divergence [8], [9]. Liese and Vajda [10] showed that $f$-divergence is invariant to sufficient statistics (invertible transformations are special examples). More generally, according to information processing theorem [8], [10], $f$-divergence never increases under stochastic kernel. Pardo and Vajda [11] showed that, for discrete distributions, the distances satisfying the information processing theorem must be written in the form of $f$-divergence. Although the invariance of $f$-divergence has been proven in the information and statistics community, to the best of our knowledge, this invariant property has been largely ignored by the pattern recognition community. The main objective of this paper is to explore the application of $f$-divergence invariance to speech recognition. In theory, we provide novel and direct proofs for the invariance of $f$-divergence. On practical side, this paper discusses how to calculate the $f$-divergence in general cases and for special distributions such as Gaussian and Gaussian mixture. To demonstrate the application of these ideas, we show how to use maximum likelihood decomposition to construct a structural representation from a sequence data and prove the invariance of the optimal decomposition. We apply this structural representation to the recognition of utterances of connected Japanese vowels. In addition, we introduce several techniques to improve recognition performance, namely, multi-stream division, dimension reduction and discriminant analysis. We carry out experiments on real and artificially transformed data. The experimental results indicate that the proposed method achieves better recognition rates than Hidden Markov Models, especially in mismatched conditions. A small portion of this work appeared in conference papers [12], [13]; more details are available in a technical report [14].

II. $f$-DIVERGENCE

In statistics and information theory, Csiszár $f$-divergence [8] (also known as Ali-Silvey distance [9]) measures the difference (dissimilarity) between two distributions. Formally,

$$D_f(p_i, p_j) = \int_X p_j(x) f\left( \frac{p_i(x)}{p_j(x)} \right) dx,$$

where $p_i(x)$ and $p_j(x)$ are density functions of two distributions on measurable space $X$. $f : (0, \infty) \to R$ is a real convex function and $f(1) = 0$. To deal with $p_j(x) = 0$, we set $0f(0) = 0$, $f(0) = \lim_{t \to 0} f(t)$ and $0f(0) = \lim_{t \to 0} tf'(t) = \lim_{u \to \infty} \frac{f(u)}{u}$ as those in [8]. Constraint $f(1) = 0$ is adopted to ensure that the $f$-divergence between two identical distributions is zero, $D_f(p, p) = 0$. But this constraint is not necessary in the following invariant analysis of $f$-divergence. For simplicity, in this paper, we limit our discussions to continuous distributions and assume that $x$ is a $d$-dimensional vector and $X$ is a manifold embedded in $\mathbb{R}^d$. However, many analysis and results can be generalized to discrete distributions.

$f$-divergence has many appealing properties. Csiszár [8], [15] proved the reflexivity of $f$-divergence, if $f(t)$ is convex and is strictly convex at $t = 1$, $D_f(p_i, p_j) = 0$ only when $p_i = p_j$. Vajda [16] and Liese [17] showed the bounds on $f$-divergence, i.e., $0 \leq D_f(p_i, p_j) \leq \lim_{t \to 0} \{f(t) + tf'(\frac{1}{2})\}$. More properties of $f$-divergence can be found in [15], [17]. 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 or functions of the $f$-divergence measure.

Consider two distributions $p_i(x)$ and $p_j(x)$ in space $X$ ($x \in X$). Let $h : X \to \mathcal{Y}$ (linear or nonlinear) denote a differentiable and
invertible mapping function (diffeomorphism) that converts $x$ into $y$. Under transformation $h$, $p(x)dx = q(y)dy$ and $dy = |J(x)|dx$, where $J(x)$ denotes the determinant of the Jacobian matrix of $h$. Thus we have $q(y) = q(h(x)) = p(x)|J(x)|^{-1}$. In this manner, distributions $p_i(x)$ and $p_j(x)$ are transformed to $q_i(y)$ and $q_j(y)$, respectively. We have the following theorem.

**Theorem 1:** The $f$-divergence between two distributions is invariant under differentiable and invertible transformation $h$,

$$D_f(p_i, p_j) = D_f(q_i, q_j).$$

(2)

In [10], [17], Liese and Vajda stated that “$f$-divergences are invariants of sufficiency of stochastic kernels.” Transformations are special deterministic cases of general stochastic kernels and invertible transformations are examples of sufficient stochastic kernels. For them, we give a simple and direct proof.

**Proof:**

$$D_f(q_i, q_j) = \int_y q_j(y) f\left(\frac{q_i(y)}{q_j(y)}\right) dy$$

$$= \int_X p_j(x) |J(x)|^{-1} f\left(\frac{p_i(x)}{p_j(x)}|J(x)|^{-1}\right) |J(x)| dx$$

$$= \int_X p_j(x) f\left(\frac{p_i(x)}{p_j(x)}\right) dx = D_f(p_i, p_j).$$

(3)

Since $f$-divergences are invariant to invertible transformations, they can be used as robust features for signal analysis and classification. In the following, we consider a more general form of Eq. (1), $M(p_i, p_j) = \int_X G(p_i(x), p_j(x))dx$ where function $G : R^n \times R^n \rightarrow R$. We call $M(p_i, p_j)$ an integration measure. 1. Pardo and Vajda [11] proved that the distance satisfies the data processing conditions (more general than invertible condition) if and only if it is $f$-divergence for discrete distributions. Here we study the necessary conditions of invariance for continuous distributions. To begin with, we have

**Lemma 1:** Let $G(t)$ denote a real function $G : [0, +\infty) \rightarrow R$. If $\int_X p(x)G(p(x))dx = 0$ for all distributions $p(x)$, we have $G(t) = 0$ for all $t > 0$.

**Proof:**

Let $V = \int x dx$ denote the volume of $X$. We first prove $G(t) = 0$ for all $t \geq 1/V$. We can find a part of $X$, denoted by $X'$, whose volume is $1/t$. Introduce distribution $p'(x)$, where $p'(x) = t$ for $x \in X'$, otherwise $p'(x) = 0$. Then,

$$\int_X p'(x)G(p'(x))dx = \int_{X'} tG(t)dx = G(t) = 0.$$  (4)

If $V = \infty$, we have proved Lemma 1. Otherwise for $t < 1/V$, we divide $X$ into two parts, $X_1$ and $X_2$, whose volumes are $V_1$ and $V_2 (V_1 > 0, V_2 > 0$, and $V_1 + V_2 = V)$, respectively. Introduce distribution $p''(x)$, where $p''(x) = t$ for $x \in X_1$, and $p''(x) = t_2$ for $x \in X_2$. It is easy to discern that $t_2 = (1 - IV_1)/V_2 > 1/V$ and $G(t_2) = 0$. Therefore, we have

$$\int_X p''(x)G(p''(x))dx = \int_{X_1} tG(t_1)dx + \int_{X_2} t_2G(t_2)dx$$

$$= G(t)tV_1 = 0.$$  (5)

From Eqs. 4 and 5, we have $G(t) = 0$ for all $t > 0$.

**Theorem 2:** If an integration measure is invariant to all invertible and differential transformations, it must be written in the form of $\int_X p_j(x) f(\frac{p_i(x)}{p_j(x)}) dx$.

Let $H : R \rightarrow R$ be a real function. Then, $H(D_f(p_i(x), p_j(x)))$ is also invariant. In this paper, we ignore the discussion on function $H$.

**Proof:** Assume $M(p_i, p_j) = \int_X G(p_i(x), p_j(x))dx$ to be an invariant integration measure. Thus, for any two distributions $p_i(x), p_j(x)$ and any invertible transformation $y = h(x)$ which converts $p_i(x), p_j(x)$ into $q_i(y), q_j(y)$, we have $M(p_i, p_j) \equiv M(q_i, q_j)$. Define $g(p_i(x), p_j(x)) = G(p_i(x), p_j(x))/p_j(x)$. Recall that $dy = |J(x)|dx$ and $p(x)|J(x)|^{-1} = q(h(x)) = q(y)$, we have

$$M(q_i, q_j) = \int_y q_j(y)g(q_i(y), q_j(y))dy$$

$$= \int_X p_j(x)g(p_i(x), |J(x)|^{-1}, p_j(x)|J(x)|^{-1})dx$$

$$= M(p_i, p_j) = \int_X p_j(x)g(p_i(x), p_j(x))dx.$$  (6)

Introduce $G'(p_i(x), p_j(x), |J(x)|^{-1}) = g(p_i(x), |J(x)|^{-1}, p_j(x)|J(x)|^{-1})$. Then,

$$\int_X p_j(x)G'(p_i(x), p_j(x), |J(x)|^{-1})dx \equiv 0.$$  (7)

Note that $p_i(x)$ and $|J(x)|$ are independent of $p_j(x)$. For any given $p_i(x)$ and $|J(x)|$, $G'$ can be seen as a function of only $p_j(x)$. According to Lemma 1, $G'(p_i(x), p_j(x), |J(x)|^{-1}) = 0$ must hold.

Introduce $t(x) = \frac{p_i(x)}{p_j(x)}$ and $g'(t(x), p_j(x)) = g(p_i(x), p_j(x))$. Then,

$$g'(t(x), p_j(x)) = g'(t(x), p_i(x)|J(x)|^{-1}) = g'(t(x), 1).$$  (8)

Therefore $g'(t(x), p_j(x))$ must be written in the form of $f(p_i(x)/p_j(x))$, and $M(p_i, p_j)$ has to be written in the form of $\int_X p_j(x) f(\frac{p_i(x)}{p_j(x)}) dx$.

We can generalize the invariant measure from two distributions to $n$ distributions. Using a similar analysis to the proofs for Theorems 1 and 2, we have

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

$$D_F(p_1, p_2, ..., p_n) = \int_{\Omega} p_n(x) F\left(\frac{p_1(x)}{p_2(x)}, ..., \frac{p_{n-1}(x)}{p_n(x)}\right) dx,$$  (10)

where function $F : R^n \rightarrow R$.

Theorems 1 and 2 together show the sufficiency and necessity of the invariance of $f$-divergence to a deterministic invertible transformation. In some problems, the transformation can be a probabilistic one due to noise or other uncertain factors. In this way, given input vector $x$ in space $\mathcal{X}$, the probabilistic transformation is described by a conditional distribution $p(y|x)$. Under $p(y|x)$, transformed distribution $q_i(y)$ can be calculated by $q_i(y) = \int_X p(y|x)p_j(x)dx$.

It can be shown that $f$-divergence never increases after probabilistic transformation: $D_f(p_i(x), p_j(x)) \geq D_f(q_i(y), q_j(y))$, which is also known as the information processing theorem [17], [11].

2We cannot find $h^*$ for all distribution $p_j(x)$. But our objective is to prove the necessity. The invariant form of $M$ must hold for any distribution and any transformation. Special cases will be enough here.
We need to calculate f-divergence in practice. Unfortunately, there exists no closed-form solution to calculate f-divergence in general. We noticed that, in spite of extensive efforts in estimating the integral functionals of a density with form \( \int_X p(x)f(p(x))dx \), there are only a few publications on how to calculate f-divergence. Furthermore, most of these works focus on nonparametric estimation of f-divergence. Here we are interested in a parametric calculation of f-divergence between Gaussian distributions or Gaussian mixtures, which will be applied in the next section.

Perhaps the most direct method would be to use Monte-Carlo sampling for approximation; however, this is always computationally expensive. Particularly when the dimension of \( x \) is high, it may be difficult to generate a sufficient number of i.i.d. samples. For Gaussian distributions, there can exist closed-form solutions for calculating f-divergence, such as BD, KL divergence, and Hellinger distance. In the reminder of this section, we will describe several techniques for parametric calculation or approximation of f-divergence between mixtures.

**Theorem 4:** Consider two mixtures \( p_i = \sum_{m=1}^{M} w_{i,m}p_i^m \) and \( p_j = \sum_{m=1}^{M} w_{j,m}p_j^m \). When function \( f \) is convex,

\[
D_f(p_i, p_j) \leq \sum_{m=1}^{M} D_f(w_{i,m}p_i^m, w_{j,m}p_j^m). \tag{11}
\]

We prove the above theorem with Jensen’s inequality (details can be found in [14]). Similarly, if \( f \) is concave, we have \( D_f(p_i, p_j) \geq \sum_{m=1}^{M} D_f(w_{i,m}p_i^m, w_{j,m}p_j^m) \). Do [18] obtained a similar result to Eq. 11 for KL divergence with the log-sum inequality. In practical calculations, we can find mapping relations \( n = T(m) \) between \( \{p_i^m(x)\}^M_m \) and \( \{p_j^m(x)\}^M_m \), which minimizes the following cost function,

\[
\min \sum_{m=1}^{M} D_f(w_{i,m}p_i^m, w_{j,T(m)}p_j^{T(m)}). \tag{12}
\]

This can be solved by the Hungarian algorithm for bipartite matching. A limitation of using Eq. 12 is that the two mixture models must have the same mixture numbers. We solve this limitation by dividing the mixtures. Consider \( p_i(x) = \sum_{m=1}^{N} w_{i,m}p_i^m(x) \) and \( p_j(x) = \sum_{n=1}^{N} w_{j,n}p_j^n(x) \) with different mixture numbers \( (M \neq N) \). We can divide \( m \)-th component \( p_i^m(x) \) into \( N \) components, \( w_{i,m}p_i^m(x) = \sum_{k=1}^{N} \alpha_{m,k}p_i^k(x) \). Similarly, for \( p_j^n(x) \), we have \( w_{j,n}p_j^n(x) = \sum_{l=1}^{N} \beta_{n,l}p_j^l(x) \). In this way, both the densities are divided into \( MN \) components, i.e., \( p_i(x) = \sum_{m=1}^{M} \sum_{n=1}^{N} \alpha_{m,n}p_i^n(x) \) and \( p_j(x) = \sum_{n=1}^{N} \sum_{m=1}^{M} \beta_{n,m}p_j^m(x) \). According to Theorem 4, we have \( D_f(p_i, p_j) \leq \sum_{m=1}^{M} \sum_{n=1}^{N} D_f(\alpha_{m,n}p_i^n, \beta_{n,m}p_j^m) \). Therefore, we can approximate the f-divergence by the following upper bound,

\[
\min_{\{\alpha_{m,n}, \beta_{n,m}\}} \sum_{m=1}^{M} \sum_{n=1}^{N} D_f(\alpha_{m,n}p_i^n, \beta_{n,m}p_j^m), \tag{13}
\]

subject to \( \alpha_{m,n} \geq 0, \beta_{n,m} \geq 0, \sum_{n=1}^{N} \alpha_{m,n} = w_{i,m}, \text{ and } \sum_{m=1}^{M} \beta_{n,m} = w_{j,n} \).

It is easy to see that the above equation provides a better approximation than Eq. 12. And we noticed similar idea has been introduced to estimate the KL divergence between GMMs [19]. The optimization of Eq. 13 can be difficult since \( D_f \) has an integration form. Fortunately, for many famous f-divergences, such as KL divergence, BD, and alpha-divergence, we can rewrite \( D_f(\alpha_{m,k}p_i^m, \beta_{n,l}p_j^n) = \beta_{n,l}f_{a}(\frac{\beta_{n,l}}{\alpha_{m,k}})D_f(p_i^m, p_j^n) + f_{b}(\frac{\alpha_{m,k}}{\beta_{n,l}}) \), where \( f_a, f_b \) are real functions independent of \( p_i^m(x), p_j^n(x) \). With this, Eq. 13 can be reduced to a function of \( \{\alpha_{m,n}, \beta_{n,m}\} \) without integration. This is actually an optimal problem with linear constraints, and there exist extensive studies and a large number of tools to address this problem.

We carried out experiments to examine the performance of Matching Upper Bound (MUB) given by Eq. 12, Decomposed Upper Bound (DUB) given by Eq. 13, and Unscented Transform (UT) [20] on calculating BD and KL divergence between Gaussian mixtures. In the first experiment, we changed the dimension of \( x \) from 6, 9, ..., to 18 and fixed the mixture number as 10. In the second experiment, we fixed the dimension as 5 and changed the mixture number from 4, 8, ..., to 20. For each dimension and mixture number, we generated 30 pairs of GMMs with random means and covariance matrices. The ground truth f-divergence \( D_f \) was calculated by Monte-Carlo sampling with enough samples. Let \( D_f^m \) denote a value of f-divergence estimated by an approximate method. We calculated ratios \( \frac{D_f^m}{D_f^m} \) as evaluation criterion. The results are shown in Fig. 1, where the solid lines represent the average of ratios and the upper/lower error bars correspond to the mean differences between the average and the estimated ratios. One can find that MUP is not a good approximation of KL divergence, since it can lead to values a few times larger than the ground truth of f-divergence. The performance of MUP on BD is even worse, so we had to omit its curves in Fig. 1. DUB and UT showed similar performances in calculating KL divergence. DUB calculation always gives larger values than ground truths, while UT usually leads to smaller values. But UT showed poor performances in estimating BD. We also noticed that DUB was robust to the increase of dimensions. But the performance of all the methods dropped as the mixture number increased.

**IV. INVARIANT STRUCTURAL REPRESENTATION OF SPEECH**

This section shows how to construct an invariant structural representation of a pattern based on f-divergence and discusses its application to speech recognition. Consider pattern \( P \) in feature space \( X \), which can be decomposed into \( m \) distributions \( \{p_i(x)\}^{m}_{i=1} \). We calculate the f-divergence \( D_f^m \) between two distributions \( p_i(x) \) and \( p_j(x) \), and construct an \( m \times m \) divergence matrix \( D^P \) with
\[D^P(i, j) = D_f(p_i, p_j)\] and \[D^P(i, i) = 0.\] Then \(D^P\) provides a structural representation of pattern \(P\). Consider transformation \(h\), which converts \(p_i(x)\) to \(q_j(y)\), and pattern \(P\) to \(Q\) (Fig. 2). Similarly, we can calculate structure representation \(D^Q\) for pattern \(Q\). From Theorem 1, we have \(D^Q = D^P\), which indicates that the structural representation based on \(f\)-divergence is invariant to transformations.

A. Construction of structural representation

To construct a structural representation from a sequence, we first need to decompose the sequence into a set of distributions. For continuous speech signals, explicit marks for segmentation (decomposition) do not exist. In this paper, we make use of maximum likelihood (ML) estimation of HMM to decompose a sequence into several distributions. Let \(X = [x^1, x^2, \ldots, x^T]\) denote sequence data, where \(x^t\) represents the \(t\)-th frame vector, and \(T\) is the length of \(X\). Assume the HMM contains \(K\) states and its parameters are denoted by \(\Lambda = \{\pi, A, B\}\), where \(\pi = \{\pi_k\}\) denotes the initial probability of \(k\)-th state, \(A = \{a_{ij}\}\) where \(a_{ij}\) represents the transition probability from \(i\)-th state to \(j\)-th state, \(B = \{b_i\}\) where \(b_i\) represents the parameters of the output distribution \(p(x|b_i)\) for \(i\)-th state. We calculate \(f\)-divergence between each distribution pair in \(\{p(x|b_i)\}\) for constructing an invariant structure. The objective of ML estimation of HMM is to determine the parameters which maximize the following likelihood,

\[
\hat{\Lambda} = \arg \max_{\Lambda} L(X, \Lambda). \quad (14)
\]

The likelihood function \(L(X, \Lambda) = p(X|\Lambda)\) can be decomposed by,

\[
L(X, \Lambda) = \sum_{S \in S} p(X, S|\Lambda) = \sum_{S \in S} \prod_{t=1}^{T-1} a_{s_t, s_{t+1}} \prod_{t=1}^{T} p(x^t|b_{s_t}). \quad (15)
\]

\(S = [s_1, s_2, \ldots, s_T]\) denotes a sequence of states, and \(S\) denotes a set of possible state sequences. Let \(Y = [y^1, y^2, \ldots, y^T]\) denote the transformed sequence of \(X\), where \(y^t = h(x^t)\). We can also apply the HMM decomposition on \(Y\). There is a question whether the HMM decomposition of \(X\) and \(Y\) will lead to the same structure or not. Actually, this can be ensured by the following theorem.

**Theorem 5:** Consider two sequences \(X\) and \(Y\) with invertible transformation \(h\) between their frame vectors \(y^t = h(x^t)\). Let \(\hat{\Lambda}^X = \{\hat{\pi}^X, \hat{A}^X, \hat{B}^X\}\) denote a set of optimal parameters of ML estimation (Eq. 15) for sequence \(X\). Then, there must exist a set of optimal parameters \(\hat{\Lambda}^Y = \{\hat{\pi}^Y, \hat{A}^Y, \hat{B}^Y\}\) for sequence \(Y\), that satisfies the following equations,

\[
\hat{\pi}^X = \pi^Y, \quad \hat{A}^X = \hat{A}^Y, \quad \text{and} \quad p(x|b_i^X), |J(x)|^{-1} = p(y|b_i^Y), \quad (16)
\]

where \(J(x)\) is the determinant of the Jacobian matrix of transformation \(h\).

---

**Proof:** Let \(\Lambda^Y = \{\pi^Y, A^Y, B^Y\}\) denote a set of parameters of an HMM for \(Y\). For any \(p(y|b_i^Y)\), we introduce its corresponding distribution \(p(x|b_i^X)\) such that \(p(y|b_i^Y) = p(x|b_i^X), |J(x)|^{-1}\). We can calculate the likelihood of \(\Lambda^Y\) as follows,

\[
L(Y, \Lambda^Y) = \sum_{S \in S^Y} \prod_{t=1}^{T} a_{s_t, s_{t+1}} \prod_{t=1}^{T} p(y^t|b_{s_t})
\]

\[
= \sum_{S \in S^Y} \prod_{t=1}^{T} a_{s_t, s_{t+1}} \prod_{t=1}^{T} p(x^t|b_{s_t}), |J(x)|^{-1}
\]

\[
= \prod_{t=1}^{T} |J(x^t)|^{-1} L(X, \hat{\Lambda}^X), \quad (17)
\]

where \(\hat{\Lambda}^X = \{\pi^X, A^X, \hat{B}^X\}\) and \(\hat{B}^X = \{\hat{b}_i^X\}\). Note that the first term \(\prod_{t=1}^{T} |J(x^t)|^{-1}\) only depends on transformation \(h\) and is independent of parameters \(\hat{\Lambda}^X\). Since \(\hat{\Lambda}^X = \arg \max_{\Lambda^X} L(X, \hat{\Lambda}^X)\) is an optimal parameter set, we have

\[
\max_{\Lambda^Y} L(Y, \Lambda^Y) = \max_{\hat{\Lambda}^X} \prod_{t=1}^{T} |J(x^t)|^{-1} L(X, \hat{\Lambda}^X)
\]

\[
= \prod_{t=1}^{T} |J(x_t)|^{-1} L(X, \hat{\Lambda}^X). \quad (18)
\]

We can examine that the parameter set \(\hat{\Lambda}^Y\) given by Eq. 16 maximizes the likelihood,

\[
L(Y, \hat{\Lambda}^Y) = \prod_{t=1}^{T} |J(x_t)|^{-1} L(X, \hat{\Lambda}^X) = \max_{\Lambda^Y} L(Y, \Lambda^Y). \quad (19)
\]

Thus, \(\hat{\Lambda}^Y\) is a set of optimal parameters.

A quick inference of Theorem 5 is that for two sequences under certain transformation, their structural representation constructed by ML decomposition must be the same. In practice, we can obtain the invariant structure of an utterance with the procedure shown in Fig. 3. At first, we calculate a sequence of cepstral features from an utterance. Then, HMM training is used to convert a cepstrum vector sequence into a sequence of \(m\) distributions. Thirdly, we calculate the \(f\)-divergences between each pair of distributions to obtain the structural representation. For convenience, we can expand \(D\) into a structural vector \(z\) with dimension \(m(m - 1)/2\). If the \(f\)-divergence used satisfies the symmetry rule, we only need to use the upper triangle of \(D\) and \(z\) has a dimension of \(m(m - 1)/2\).

B. Invariant structures for speech recognition

In this section, we apply invariant structures to recognizing utterances of connected Japanese vowels. A fundamental problem in speech recognition is dealing with non-linguistic variations, such as speaker, communication channels, microphones, and so on. It can be shown that many of these speech signal variations [3], such as the difference of vocal tract length and channel distortion, can be
approximated as linear transformations of cepstrum feature space [2]. The invariant structural representation can discard these unfavorable variations and thus provide robust features to speaker differences. It can be shown that the similarity score between two utterances after a structure is shifted and rotated so that the two structures are overlapped best, can be approximated as a difference between the two structure vectors \( ||D_P - D_Q|| \) (Fig. 2) [21], and thus this difference provides a matching score for speech recognition.

However, there are two limitations to directly use structural representations for speech recognition: 1) the invariance can be too strong, such that two linguistically different speech signals may have similar structural representations [5]; 2) its dimension may be too high, which not only increases the computational cost but also increases the liability to “the curse of dimensionality” [4]. In the next section, we develop a method to circumvent these limitations, which includes three steps: multiple stream structuralization, dimension reduction and discriminant analysis.

1) Multiple stream structuralization: Invariant structures can cancel the non-linguistic variations of speech signals caused by transformations. On the other hand, since the structure is invariant to any invertible linear or nonlinear transformations, some linguistic information, which is useful for recognition, may also be discarded. This diminishes the recognition performance of structural representation. To overcome this limitation, we need to relax the overly strong invariance and to find a rich representation which provides discriminative information for classification. In other words, we wish to balance the invariant property and the discriminative ability of structural representation. Our previous work [5] introduced a multiple stream structuralization method to deal with this problem. We divide a speech stream into several sub streams according to the dimensionality of cepstrum features, and calculate Bhattacharyya distances for each sub stream, as shown in Fig. 4. Geometrically speaking, this is equivalent to decomposing the feature space into several subspaces of the whole cepstrum space, and construct a structural representation in each subspace.

2) Dimension reduction: The dimension of structure representation is usually high. Let \( m \) denote the number of distributions. Then, the dimensionality of its structural representation is \( O(m^2) \). When using multiple stream structuralization, the dimensionality rises to \( O(km^2) \), where \( k \) is the number of sub streams. The high dimensionality not only increases the computational cost but also makes it difficult to train robust classifiers (known as “the curse of dimensionality problem”). On the other hand, the BDs are highly correlated features (thinking \( D_f(p_i, p_j) \) can be largely related to \( D_f(p_i, p_k) \) and \( D_f(p_k, p_j) \)). This fact makes dimension reduction possible. Let \( z_j \) denote a structure vector of the \( j \)-th sub stream. We apply principal component analysis (PCA) to the structure vectors of each sub stream and find that the first \( 10\% \) of eigenvectors with the largest eigenvalues contain \( 80\%-90\% \) energy of the whole structure vectors. Let \( E_j = [e_1^j, e_2^j, ..., e_t^j] \) denote the first \( t (t < m(m-1)/2) \) eigen vectors with the largest eigen values of the covariance matrix of \( j \)-th sub stream. Since the eigen vectors with small eigen values usually correspond to unimportant and noisy directions, we can obtain reduced structure vector (RSV) \( v_j \) by projecting \( z_j \) into the subspace spanned by \( E_j \),

\[
v_j = E_j^T(z_j - \bar{z}_j),
\]

where \( \bar{z}_j \) is the mean structure vector of the \( j \)-th sub stream.

3) Discriminant analysis: After dimension reduction, we combine the reduced structure vectors from each sub stream into a single one and make use of LDA for classification. Let \( v = [v_1, v_2, ..., v_k] \) denote an augmented structure vector (ASV) with dimension \( tk \), where \( k \) is the number of distributions. For convenience, we use \( v^i \) to represent the ASV of \( i \)-th utterance. LDA aims at finding a discriminant linear transformation \( W \) to calculate the discriminative features \( W^T v \). Mathematically, this is achieved by maximizing the following ratio (generalized Rayleigh quotient),

\[
\hat{W} = \arg \max_W \frac{||W^T S_w W||}{||W^T S_b W||},
\]

where \( S_b \) is the between-class scatter matrix, and \( S_w \) is the within class scatter matrix of the ASVs. Assume we have \( M \) training samples \( \{v^i\}_{i=1}^M \) belonging to \( N \) categories \( \{C_j\}_{j=1}^N \). Let \( n_j \) denote the number of training samples in \( C_j \). Then \( S_b \) and \( S_w \) can be calculated by the following equations:

\[
S_w = \sum_{j=1}^N \sum_{v^i \in C_j} (v^i - \mu_j)(v^i - \mu_j)^T,
\]

\[
S_b = \sum_{j=1}^N n_j (\mu_j - \mu)(\mu_j - \mu)^T,
\]

where \( \mu_j \) is the mean of the ASVs of class \( C_j \) and \( \mu \) is the mean of all the training samples. Optimal \( \hat{W} \) can be computed as the eigenvectors of \( S_w^{-1} S_b \). For vector \( v \) with unknown category, we classify it by using the discriminative features:

\[
\arg \min_j |W^T v - \hat{W}^T \mu_j|.
\]

One may suggest applying LDA directly to structure vector \( z \) without applying PCA in step 2. However, \( z \) has a high dimensionality, which causes LDA to suffer easily from the singular covariance matrix problem and overfit the training data [22].

C. Experiments

We used a database of connected Japanese vowel utterances [5] to evaluate the proposed method. It is known that acoustic features of vowel sounds exhibit larger between-speaker variations than consonant sounds. Each word in the data set corresponded to a combination of the five Japanese vowels ‘a’, ‘e’, ‘i’, ‘o’ and ‘u’, such as ‘aieou’, ‘uoaei’, etc, so there were 120 words in total. The utterances of 16 speakers (8 males and 8 females) were recorded. Every speaker uttered each word 5 times. The total number of utterances was \( 16 \times 120 \times 5 = 9,600 \). Of these, 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 calculated 12 mel-cepstrum features and one power coefficient. Then, ML-based decomposition was used to convert a cepstrum vector sequence into 25 Gaussian distributions.
TABLE I
Comparisons of recognition rates

<table>
<thead>
<tr>
<th>Method</th>
<th>NN</th>
<th>NM</th>
<th>GM</th>
<th>EDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bhattacharyya dis.</td>
<td>93.0%</td>
<td>95.6%</td>
<td>96.4%</td>
<td>98.8%</td>
</tr>
<tr>
<td>Hellinger dis.</td>
<td>89.0%</td>
<td>95.1%</td>
<td>56.6%</td>
<td>97.0%</td>
</tr>
<tr>
<td>symmetric KL div.</td>
<td>93.2%</td>
<td>95.6%</td>
<td>96.4%</td>
<td>99.0%</td>
</tr>
</tbody>
</table>

Fig. 5. Word recognition rates for warped utterances.

We divided the 13D cepstrum feature stream into 13 multiple sub streams and calculated a structure for each sub stream.

We made use of the BD, HD and symmetric KL div. for building structures. As for classification, we used the following classifiers: nearest neighbors (NN), nearest mean (NM), Gaussian distribution model (GM) and eigen discriminant analysis (EDA) proposed in Section IV-B. For NN and NM, Euclidean distance was used. For GM, we used diagonal covariance matrices. The results are summarized in Table I. We found that the performances of symmetric KL div. and BD were similar. Hellinger distance had the lowest recognition rates. The EDA results were higher than the recognition rate (98.35%) of the word HMMs trained with the same utterances.

In the next, we examine the robustness of structural representations with respect to the change of vocal tract lengths (VTL). Difference in VTL is a major cause of non-linguistic variations. This difference can be modeled by warping the frequency axis of the power spectrum of speech signals [2]. Let \( \omega \) and \( \hat{\omega} \) denote angular frequencies of a base speaker and another (transformed) speaker \( (0 \leq \omega, \hat{\omega} \leq \pi) \). One popular warping function has the following form [2],

\[
e^{j\hat{\omega}} = \left( \frac{e^{j\omega} - \alpha}{1 - e^{j\omega \alpha}} \right), \tag{25}
\]

where \( \alpha \) represents a warping parameter \((-1 < \alpha < 1)\). With negative/positive values of \( \alpha \), the VTL is lengthened/shortened. \( \alpha = -0.4/0.4 \) approximately doubles/halves the VTL. In practice, it is very difficult to gather a speech corpus with large VTL variances. For this reason, we artificially generated utterances with various VTLs by applying the warping function Eq. 25 to each utterance in the above Japanese vowel word database. We set warping parameter \( \alpha \) as \(-0.4, -0.35, ..., 0.0, 0.4\). For each \( \alpha \), we conducted matched and mismatched experiments. In the matched experiment, both training and testing data were warped under the same \( \alpha \), while in the mismatched one, only testing data were warped. Since the warping function was only applied to FFT cepstrum features, we made use of 17 dimensional cepstrum vectors in this experiment. KL divergence was used to calculate structural representations and EDA was adopted for classification. We compared the recognition performance between structural representations and HMM. The results are shown in Fig. 5. As one can see, structural representations yield higher recognition rates in matched and mismatched experiments for every \( \alpha \). Especially in the mismatched case, the recognition rates of HMM drop significantly when \( |\alpha| \) is large; on the other hand, structural representations show much better rates when compared with HMM. This indicates that structural representations are much more robust to changes in VTLs than HMMs.

V. Conclusions

This paper examined \( f \)-divergence, a family of invariant measures between distributions, and its application to utterance recognition. On theoretical side, we provided new and direct proofs for the necessity and sufficiency of the invariance of \( f \)-divergence. We discussed the properties of \( f \)-divergence and studied how to calculate \( f \)-divergence in general cases and for special distributions, such as Gaussian and Gaussian mixture. We carried out experiments to compare the performances of approximate methods for calculating KL divergence and Bhattacharyya distance between GMNs. On practical side, we demonstrated how to construct an invariant structural representation for utterances using \( f \)-divergence and maximum likelihood based decomposition, and proved the invariance of this decomposition. We made use of this structural representation for utterance recognition, and developed several techniques to improve recognition performance. The experimental results exhibited that structural representation can achieve better recognition performance than HMM. Particularly in the mismatched experiments, the proposed method demonstrated significant improvements compared with HMM. These analyses indicate that the proposed invariant structures based on \( f \)-divergence provide robust representation of speech.

VI. Acknowledgement

The authors would like to thank the anonymous reviewers for their valuable comments on this manuscript.

REFERENCES