On invariant structural representation for speech recognition: theoretical validation and experimental improvement

Yu Qiao, Nobuaki Minematsu, and Keikichi Hirose

The University of Tokyo, Hongo, Bunkyo-ku, Tokyo, 113–0033, Japan

{qiao, mine, hirose}@gavo.t.u-tokyo.ac.jp

Abstract

One of the most challenging problems in speech recognition is to deal with inevitable acoustic variations caused by non-linguistic factors. Recently, an invariant structural representation of speech was proposed [1], where the non-linguistic variations are effectively removed though modeling the dynamic and contrastive aspects of speech signals. This paper describes our recent progresses on this problem. Theoretically, we prove that the maximum likelihood based decomposition can lead to the same structural representations for a sequence and its transformed version. Practically, we introduce a method of discriminant analysis of eigen-structure to deal with two limitations of structural representations, namely, high dimensionality and too strong invariance. In the 1st experiment, we evaluate the proposed method through recognizing connected Japanese vowels. The proposed method achieves a recognition rate 99.0%, which is higher than those of the previous structure based recognition methods [2, 3, 4] and word HMM. In the 2nd experiment, we examine the recognition performance of structural representations to vocal tract length (VTL) differences. The experimental results indicate that structural representations have much more robustness to VTL changes than HMM. w Moreover, the proposed method is about 60 times faster than the previous ones.

Index Terms: Speech recognition, invariant structure, PCA, discriminative analysis

1. Introduction

Speech recognition is a task to extract only the linguistic/text information from speech signals. However, speech signals inevitably include the acoustic variations caused by non-linguistic factors, such as speaker, communication channel and noise. The same text can lead to different acoustic observations due to different speakers and different enviroments. This poses a challenging problem for speech recognition. To deal with these variations, modern speech recognition approaches mainly make use of the statistical methods (such as GMM, HMM) to model the distributions of the acoustic features. These methods always require a large amount of data for training and can achieve relatively high recognition rates when there is a good match between training and testing data. But their performances always decrease significantly when mismatched. Contrary to this is children's spoken language acquisition. A child does not need to hear the voices of thousands of persons 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. We consider it is by these robust measures that children can learn speech with very biased training data from mothers and fathers. This 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 [5].

Inspired by these facts, the third author of this paper proposed an invariant structural representation of speech signals which aims at removing the non-linguistic factors in speech signals [1]. Different from classical speech models, the structural representations make use of invariant Bhattacharyya distances (or f-divergence in general [9]) to model the contrastive and dynamic aspects of speech and discard the static features. We have demonstrated the effectiveness of this novel representation in automatic speech recognition [2, 3], speech synthesis[6], and computer aided language learning (CALL) systems [7].

This paper describes our recent progresses on invariant structure for speech representation. To construct two structural presentations from a sequence and its transformed version, we also need to decompose them in an invariant way. In this paper, we prove that maximum likelihood estimation provides such an invariant decomposition. In addition, we introduce a method called discriminant analysis of eigen-structure to improve the performance of structure based speech recognition. We carried out experiments to examine the proposed methods on recognizing connected Japanese vowels. The results show that the proposed method not only achieves higher recognition rates but also largely reduces the computational time of classification than the previous structure based speech recognition methods [2, 3]. We also examined the performance of structural representations and HMM to the change of vocal tract length by using artificially warped data. The structure demonstrates much more robustness than HMM.

2. Invariant structure for speech

In this section, we firstly give a brief overview on structural representation for speech, and then prove how maximum likelihood decomposition can lead to invariant structures.

2.1. Theory of invariant structure

Consider feature space X and pattern P in X. Suppose P is composed of a sequence of K events $\{p_i\}_{i=1}^{K}$. Each event is described as a distribution $p_i(x)$ in the feature space. Note x can have multiple dimensions. Assume there is an invertible transformation $h : X \rightarrow Y$ (linear or nonlinear) which converts x into y. In this way, pattern P in X is mapped to pattern Q in Y, and event $p_i(x)$ is transformed to event $q_i(y)$ (Fig. 1). Thus if we can find invariant metrics in both space X and space Y, these metrics can yield robust features for classification.

Under transformation h, $p(x)dx = q(y)dy$ and $dy = |\Phi(x)|dx$, where $\Phi(x)$ denotes the determinant of the Jacobian matrix of h. Thus we have $q(y) = q(h(x)) = p(x)|\Phi(x)|^{-1}$.

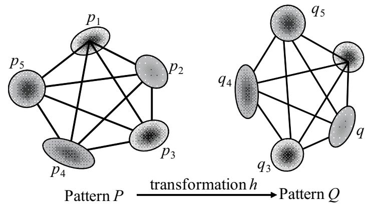

Figure 1: Invariant structures $\mathcal{D}^Q = \mathcal{D}^P$.

Consider f-divergence [8] defined as,

$$D_f(p_i, p_j) = \oint p_j(x) f\left(\frac{p_i(x)}{p_j(x)}\right) dx, \quad (1)$$

where $f : (0, \infty) \to R$ is a real convex function and $f(1) = 0$. 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 the functions of f-divergence measure. We can examine the invariance of f-divergence [9] as

$$\begin{aligned} D_f(q_i, q_j) &= \oint q_j(y) f\left(\frac{q_i(y)}{q_j(y)}\right) dy \\ &= \oint p_j(x)|\Phi(x)|^{-1} f\left(\frac{p_i(x)|\Phi(x)|^{-1}}{p_j(x)|\Phi(x)|^{-1}}\right) |\Phi(x)| dx \\ &= \oint p_j(x) f\left(\frac{p_i(x)}{p_j(x)}\right) dx = D_f(p_i, p_j). \quad (2) \end{aligned}$$

We also proved that all invariant integration measures $\oint M(p_i, p_j)dx$ must be written into the form of f-divergence [9]. We can obtain an $K \times K$ divergence matrix $\mathcal{D}^P$ with $\mathcal{D}^P(i, j) = D_f(p_i, p_j)$ and $\mathcal{D}^P(i, i) = 0$. Then $\mathcal{D}^P$ provides a structural representation of pattern P. Similarly, we can obtain structure representation $\mathcal{D}^Q$ for pattern Q. Then we have that $\mathcal{D}^Q = \mathcal{D}^P$, which indicates that the structural representation based on f-divergence is invariant to transformations.

2.2. Construction of structural representation

In order to calculate a structural representation from a sequence, we need to decompose it into a set of distributions at first. For continuous speech signals, there don't exist explicit marks for sequence segmentation (decomposition). In this paper, we make use of maximum likelihood (ML) estimation of HMM to decompose a sequence into a set of distributions. Let $X = [x^1, x^2, ..., x^T]$ denote a sequence of speech signals, 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 a set of the initial probabilities of k-th state, $A = \{a_{ij}\}$ represents the transition probability from i-th state to j-th state, $B = \{b_i\}$ represents the parameters of the output distribution $p(x|b_i)$ for i-th state. We calculate f-divergences between every two distributions in set $\{p(x|b_i)\}$ for constructing an invariant structure. The objective of ML estimation is to determine the parameters which maximize likelihood $\hat{\Lambda} = \arg\max_\Lambda L(X, \Lambda)$. The likelihood function is given by,

$$\begin{aligned} L(X, \Lambda) =& p(X|\Lambda) = \sum_{S \in \mathbb{S}} p(X, S|\Lambda) \\ =& \sum_{S \in \mathbb{S}} \pi_{s_1} \prod_{t=1}^{T-1} a_{s_t, s_{t+1}} \prod_{t=1}^{T} p(x^t|b_{s_t}), \quad (3) \end{aligned}$$

where $S = [s_1, s_2, ..., s_T]$ denotes a sequence of states, and $\mathbb{S}$ denotes a set of possible state sequences. Let $Y = [y^1, y^2, ..., y^T]$ 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 1 *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. 3) 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,*

$$\hat{\pi}^X = \hat{\pi}^Y, \ \hat{A}^X = \hat{A}^Y, \ \text{and}\ p(x|\hat{b}_i^X)|\Phi(x)|^{-1} = p(y|\hat{b}_i^Y). \quad (4)$$

Proof Let $\Lambda^Y = \{\pi^Y, A^Y, B^Y\}$ denote a set of parameters of an HMM for Y. For any $p(y|b_i^Y)$, there is a corresponding $p(x|\tilde{b}_i^X)$ such that $p(y|b_i^Y) = p(x|\tilde{b}_i^X)|\Phi(x)|^{-1}$. We can calculate the likelihood of Λ^Y as follows,

$$\begin{aligned} L(Y, \Lambda^Y) &= \sum_{S \in \mathbb{S}} \pi_{s_1}^Y \prod_{t=1}^{T-1} a_{s_t, s_{t+1}}^Y \prod_{t=1}^{T} p(y^t|b_{s_t}^Y) \\ &= \sum_{S \in \mathbb{S}} \pi_{s_1}^Y \prod_{t=1}^{T-1} a_{s_t, s_{t+1}}^Y \prod_{t=1}^{T} p(x^t|\tilde{b}_{s_t}^X)|\Phi(x^t)|^{-1} \\ &= \prod_{t=1}^{T} |\Phi(x^t)|^{-1} L(X, \tilde{\Lambda}^X), \quad (5) \end{aligned}$$

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

$$\max_{\Lambda^Y} L(Y, \Lambda^Y) = \max_{\tilde{\Lambda}^X} \prod_{t=1}^{T} |\Phi(x_t)|^{-1} L(X, \tilde{\Lambda}^X). \quad (6)$$

We can examine that the parameter set $\hat{\Lambda}^Y$ given by Eq. 4 maximizes the likelihood, $L(Y, \hat{\Lambda}^Y) = \prod_{t=1}^{T} |\Phi(x_t)|^{-1} L(X, \hat{\Lambda}^X) = \max_{\Lambda^Y} L(Y, \Lambda^Y)$. Thus $\hat{\Lambda}^Y$ is a set of optimal parameters.

A quick inference of Theorem 1 is that for two sequences under a transformation, their structural representation constructed by ML decomposition must be the same, too. It is noted that to find global ML estimation can be difficult, since the famous EM training of HMM (Baum-Welch algorithm) can lead to local optimization. In our experiments, the structures obtained by local optimization proved to be sufficiently effective.

In practice, we construct a structural representation from an utterance by the following procedure (Fig. 2). At first, we calculate a sequence of cepstrum from input speech waveforms. Then an HMM is trained from a single cepstrum sequence and each state of HMM is regarded as an event p_i. Thirdly we calculate the f-divergences between any two events. These divergences will form a $K \times K$ symmetric distance matrix $\mathcal{D}$ with zero diagonal. We can expand the non-diagonal (nonzero) parts of $\mathcal{D}$ into to a structural vector z, which yields the structural representation.

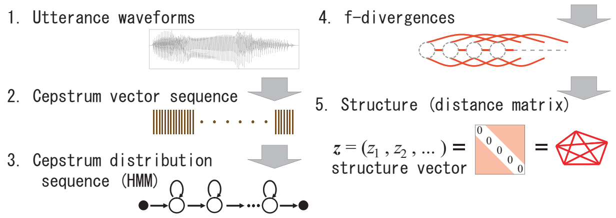
1. Utterance waveforms
2. Cepstrum vector sequence
3. Cepstrum distribution sequence (HMM)
4. f-divergences
5. Structure (distance matrix)
$z = (z_1, z_2, \ldots) =$
structure vector
0
0
0
0
=

Figure 2: Framework of structure construction.

3. Discriminant analysis of eigen-structure

The most attractive property of structural representation is its invariance to transformation on feature space, which allows us to remove the non-linguistic factors in speech recognition. However, there are two limitations for directly using structural representations for speech recognition: too strong invariance and too high dimension. In the next, we propose a method of discriminant analysis of eigen-structure method to deal with both the limitations. The diagram of this method is shown in Fig. 3.

The invariant structures discard the non-linguistic information in speech signals. 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 is called “too strong invariance problem”, which decreases the recognition performance of structural representation [3]. To overcome the first limitation, we need to reduce the too strong invariance and to find a rich representation which provides sufficiently discriminative information for classification. Our previous work [3] 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. 3. Geometrically speaking, this equals to decomposing the feature space into several sub-spaces and construct a structural representation in each subs-space [3].

The structure has a high dimension. Let K denote the number of distributions and n is the number of streams. Then, the dimensionality of its structural representation is $O(nK^2)$. The high dimensionality not only increases the computational cost and but also makes it difficult to train robust classifiers (known as the curse of dimensionality problem [10]). On the other hand, the f-divergences are highly correlated features (thinking d_{p_i,p_j} can be largely effected by d_{p_i,p_k} and d_{p_k,p_j}). This fact makes dimension reduction possible. We applied Principal Componenet Analysis (PCA) to the structure vector of each stream to obtain a low dimensional ($\frac{1}{10}$ of the original dimension in our experiments) vector for it. Then we joint these low dimensional vectors of all the streams to a eigen-representation.

Although PCA can significantly reduce the dimensionality of structure vectors, it doesn't take account of the category information. Sometimes PCA actually smears the classes together so that eigen structure vectors are not linearly separable. For this reason, we apply Fisher discriminant analysis (FDA), also known as linear discriminant analysis (LDA) [11] to calculate a more discriminative representation. The final classification is made with this compact and discriminative representation.

4. Experiments

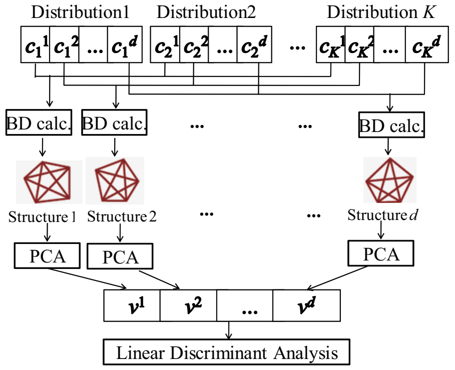

Figure 3: Multiple stream structuralization.

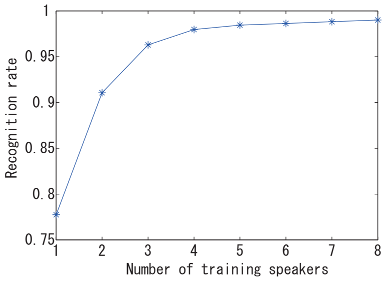

Figure 4: Comparison of the recognition rates of different numbers of speakers in training data.

We carried out experiments on the connected Japanese vowel utterances database [3] to evaluate the performances of the proposed Discriminant Analysis of Eigen-Structure (DAES) in Section 3. Each word in the database corresponds to a combination of the five Japanese vowels ‘a’,‘e’,‘i’,‘o’ and ‘u’, such as ‘aeiou’,‘uoaei’, ... So there are totally 120 words. It is noted that compared with consonant sounds, vowel sounds usually exhibit larger between-speaker acoustic variations. The utterances of 16 speakers (8 males and 8 females) were recorded. Every speaker provided 5 utterances for each word. Totally the number of utterances is 16×120×5=9,600. Each structure includes 25 distributions, and each distribution is described by a 13D Gaussian distribution with a diagonal covariance matrix. Following [3], we divide the 13D cepstrum+ 13D delta cepstrum feature vectors into 13 multiple sub-streams with block size 2. We calculate the structural vectors for each sub-stream with BD. (We also conducted experiments on KL-divergence. The results are very similar and thus omitted.) Each structural vector before PCA has a dimensionality of 25×24/2=300. We change the number of training speakers from 1 to 8, and the results are shown in Fig. 4. Although the recognition rate slightly drop as the number of speakers decreases, we obtain a recognition rate 98.0% with only four training speakers.

We compare the recognition rate of our method with those of the previous structure-based recognition methods, such as, multiple stream structuralization modeling (MSS) [3], two stage LDA (2-LDA)[4], random discriminant structure analysis (RDSA) [2], and word HMM. For each method, we use 4,800

Table 1: Comparisons of recognition rates

Method	DAES	2-LDA[4]	MSS[3]	RDSA[2]	HMM
Rate	**99.0%**	98.6%	95.3%	98.3%	98.3%

utterances from 4 male and 4 female speakers for training and the other 4,800 utterances for testing. Results are given in Table 1. The proposed method can achieve the highest recognition rates among them. Moreover, it is much faster than the previous structure-based recognition methods. The computational time for classification is only about 1/60 of MSS and 1/65 of RDSA.

In the next, we examine the robustness of structural representations with respect to the change of vocal tract lengths (VTL). The difference of VTL is a major cause of the non-linguistic variations. And this difference can be modeled by warping the frequency axis of the power spectrum of the speech signals [12]. Let ω denote angular frequency of a base speaker and $\hat{\omega}$ angular frequency of another (warped) speaker ($0 \leq \omega, \hat{\omega} \leq \pi$). One popular warping function has the following form [12],

$$e^{j\hat{\omega}} = \frac{e^{j\omega} - \alpha}{1 - e^{j\omega}\alpha}, \tag{7}$$

where α represents a warping parameter ($-1 < \alpha < 1$). With negative/positive values of α, the VTL is lengthened/shortened. $\alpha = -0.4/+0.4$ approximately doubles/halves the VTL. As it is very difficult to gather speech corpus with large VTL variances in practice , we artificially generate utterances with various VTLs by applying the warping function Eq. 7 on each utterance in the above Japanese vowel word database. We set warping parameter α as -0.4, -0.35, ..., 0,..., 0.4. For each α, we conduct matched and mismatched experiments. In the matched experiment, both training and testing data are warped under the same α, while in the mismatched one, only testing data are warped. Since the warping function Eq. 7 are applied for FFT cepstrum features, we made use of 17 dimensional cepstrum vectors in this experiment. KL-divergence is used to calculate structural representations. (Experiments with BD show very similar results.) We compared the recognition performance between structural representations and HMM. The results are shown in Fig. 5. We carried out another experiment with speaker-independent triphone HMMs, which were trained with 4,130 speakers [13] and were tested against the utterances warped with $\alpha = 0.33$. The rate was 1.4%, by far lower than that of the structures (90.0%), trained only with 8 speakers. As one can see, structural representations obtain higher recognition rates in matched and mismatched experiments for every α. Especially, in the mismatched case, the recognition rates of HMM will drop significantly when $|\alpha|$ is large; on the other hand, structural representations show significant better rates when compared with HMM. This indicates that structural representations are much more robust with the change of VTLs.

5. Conclusions

We study the invariant structural representation for speech recognition. The contributions of this paper are two aspects. Theoretically, we prove that maximum likelihood decomposition will lead to the same structures for a sequence and its transformed version. This result yields an important basis for constructing invariant structures. Practically, this paper proposes the discriminant analysis of eigen-structure for speech recognition. In this method, too strong invariance of structural representation is relaxed adequately by constructing structures for each stream of speech signal. PCA and FDA are used to reduce the dimension and to obtain a discriminative representation. Experiments show that our method achieves a recognition rate (99.0%) on a connected Japanese vowel database, which is higher than the results of our previous structure based methods [2, 3, 4], and word HMMs trained with the same database. We also found that the proposed structural representation show much higher robustness to the change of vocal tract length when compared with HMM.

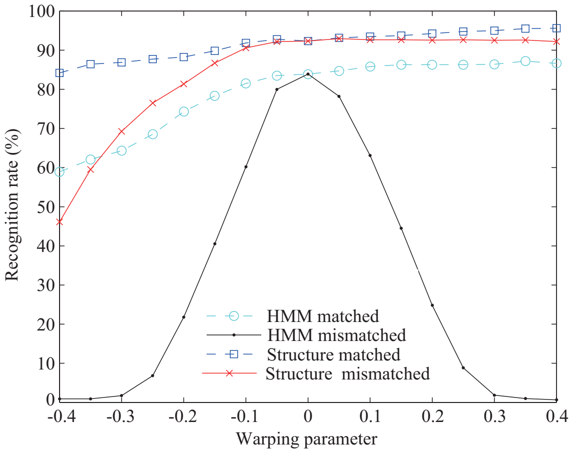

Figure 5: Word recognition rates for warped utterances.

6. References

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

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

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

[4] S. Asakawa, Y. Qiao, N. Minematsu, and K. Hirose, “Speech recognition with super robustness to speaker variability based on discriminant analysis and speech structures,” *Proc. Autumn Meeting of Acoust. Soc. Jpn.*, pp. 113–116, 2008.

[5] 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.

[6] S. Saito, D. Asakawa, N. Minematsu, and K Hirose, “Structure to speech – speech generation based on infant-like vocal imitation,” *Proc. INTERSPEECH*, pp. 1837–1840.

[7] N. Minematsu, S. Asakawa, and K. Hirose, “Structural representation of the pronunciation and its use for CALL,” *Proc. of IEEE Spoken Language Technology Workshop*, pp. 126–129, 2006.

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

[9] Y. Qiao and N. Minematsu, “f-divergence is a generalized invariant measure between distributions,” *Proc. INTERSPEECH*, 2008.

[10] A.K. Jain, “Statistical Pattern Recognition: A Review,” *IEEE Trans. PAMI*, vol. 22, no. 1, pp. 4–37, 2000.

[11] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

[12] 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.

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