Dimension Reduction and Discriminant Analysis for Japanese Connected Vowel Recognition *

◎ Yu Qiao, Satoshi Asakawa, Nobuaki Minematsu and Keikichi Hirose (The Univ. of Tokyo)

1 Introduction

The aim of speech recognition is to extract only the linguistic information from speech signals. The acoustic variations caused by non-linguistic factors, such as speaker, communication channel and noise, pose a challenging problem for speech recognition. The same text can lead to different acoustic observations due to different speakers and different environments. 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 can achieve relatively high recognition rates when properly trained. However, they always require a large number of high quality data for training. This is very different from children spoken language acquisition, where the children mainly use very biased training data from mothers and fathers. This fact largely indicates that there may exist robust representations of speech which are nearly invariant to non-linguistic variations.

Along this line, the third author of this paper proposed an invariant structural representation of speech signals, which tries to remove the non-linguistic factors in speech signals [1]. Different from classical speech models, this structural representations focus on the dynamic motions in speech and discard the static features. Mathematically, the structural representations are made up of Bhattacharyya distances (BD), which are invariant to invertible transformations on feature space [2]. Our previous works have demonstrated the effectiveness and efficiency of this novel representation in both speech recognition tasks [3, 4] and computer aided language learning (CALL) systems [5].

However, there are two limitations for direct use of structural representation for speech recognition. 1) Its dimension is high, which not only increases the computational cost but also makes it easily suffer from the curse of dimensionality [3]. 2) The invariance can be too strong, such that two linguistically different speech signals may have similar structural representations [4]. In this paper, we introduce the techiniques of dimension reduction and discriminant analysis to address these two problems. As first, we build a structure for each sub-stream of the cepstrum features to overcome the too strong invariance. Then we calculate a reduced structure vector for each sub-stream and apply linear discriminant analysis for final classification. The new representation not only has a lower dimension but also is more discriminative. We carried out experiments on recognizing connected Japanese vowels. The experimental results show that the proposed method not only achieves higher recognition rate but also largely reduces the computational time of classification than the previous structure-based speech recognition methods [3, 4].

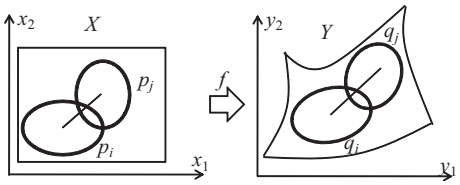

Fig. 1 Invariance of Bhattacharyya distance.

2 Invariant structure for speech representation

In this section, we will give a brief overview on invariant structure theory and on how to calculate structural representations from utterances [1, 3, 4].

2.1 Theory of invariant structure

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

Under invertible transformation f, it is not difficult to calculate that distribution $q_i(y)$ can be expressed by,

$$q_i(y) = p_i(f^{-1}(y))|J(y)|, \tag{1}$$

where f^{-1} denotes the inverse function of f, and $J(y)$ is the Jacobian matrix of function f^{-1}. Consider the Bhattacharyya distance (BD) defined by,

$$BD(p_i, p_j) = -\ln \int (p_i(x)p_j(x))^{1/2} dx. \tag{2}$$

The invariant structure theory [1] proves that BD keeps invariant under transformation f, that is, $BD(p_i, p_j) = BD(q_i, q_j)$ (Fig.1). If $p_i(x)$

*次元削減と判別分析を用いた日本語母音系列連続発声の認識. 喬宇, 朝川智, 峯松信明, 広瀬啓吉

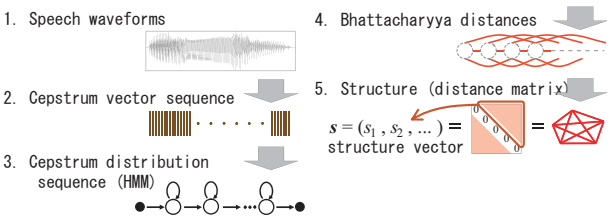

Fig. 2 Framework of structure construction.

and $p_j(x)$ are Gaussian with mean μ_i, μ_j and covariance Σ_i, Σ_j, we have $BD(p_i, p_j) = \frac{1}{8}(\mu_i - \mu_j)^T(\frac{\Sigma_i+\Sigma_j}{2})^{-1}(\mu_i - \mu_j) + \frac{1}{2}\ln\frac{|(\Sigma_i+\Sigma_j)/2|}{|\Sigma_i|^{1/2}|\Sigma_j|^{1/2}}$.

2.2 Structuralization of an utterance

In the next, we show how to calculate a structural representation from an utterance. As shown in 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 event p_i. Thirdly we calculate the Bhattacharyya distances between any two events. These distances will form an $m \times m$ symmetric distance matrix M_{BD} with zero diagonal, which can be seen as the structural representation. For convenience, we can expand the upper triangle of M_{BD} into *structure vector* z of dimension $m(m-1)/2$. It is easy to see that this structural representation must be invariant to transformations in feature space. In speech engineering, many non-linguistic variations can be modeled as affine transformations on cepstrum-feature space [1]. And thus the structural vector should be approximately invariant to the non-linguistic variations. The Euclidean distance between the structural vectors of two utterances can be used a matching score for speech recognition [1]. More details of this procedure can be found in [1, 3, 4].

3 Dimension Reduction and Discriminant analysis

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: 1) the invariance can be too strong, such that two linguistically different speech signals may have similar structural representations [4]; 2) its dimension is high, which not only increases the computational cost but also makes it easily suffer from the curse of dimensionality [3]. In the next, we describe a method to deal with both the limitations, which includes three steps: multiple stream structuralization, dimension reduction and discriminant analysis.

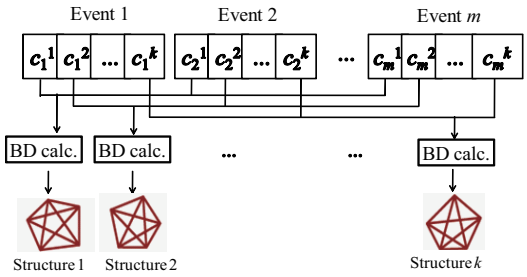

Fig. 3 Multiple stream structuralization.

3.1 Multiple stream structuralization

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 [4]. To overcome the first limitation, we need to release the too 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 discriminate ability of structural representation. Our previous work [4] 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 decompose the feature space into several sub-spaces and construct a structural representation in each subs-space. The multiple stream structuralization allows us to preserve more discriminative information of speech signals without too much effecting the invariance of structures [4].

3.2 Dimension reduction

The dimension of structure representation is usually high. Let m denote the number of events. Then, the dimensionality of its structural representation is $m(m-1)/2$. When using multiple stream structuralization, the dimensionality raises to $km(m-1)/2$, where k is the number of streams. 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 [6]). On the other hand, the BDs are highly correlated features (thinking d_{p_i,p_j} can be largely influenced by d_{p_i,p_k} and d_{p_k,p_j}). This fact makes dimension reduction possible. Let z_j denote a structure vector of the j-th stream. We apply principal component analysis (PCA) on the structure vectors of each stream and find that the first 10% eigen vectors with the largest eigen val-

ues 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 the j-th stream. Consider space S_j of the structure vectors of the j-th stream. S_j has a dimensionality of $m(m-1)/2$. $e_1^j, e_2^j, ...$ and, e_t^j span a subspace of S_j, which best accounts for the distribution of the j-th stream's structure vectors. The eigen vectors with small eigen values usually corresponds to unimportant and noisy directions in S_j. We can obtain *reduced structure vector* (RSV) v_j by projecting z_j into the subspace spanned by E_j,

$$v_j = E_j^T(z_j - \hat{z_j}), \tag{3}$$

where $\hat{z_j}$ is the mean structure vector of the j-th stream. v_j has a dimension of t, which is much less than $m(m-1)/2$.

Instead of PCA, the second author independently proposed a linear discriminant analysis (LDA) for dimension reduction [7]. LDA, also known as Fisher discriminant analysis (FDA) [8], aims at finding a linear transformation like Eq. 3 to reduce dimensionality. The difference comes from how to calculate the transformation matrix E_j. LDA calculates E_j in a supervised way by finding the maximally discriminant features.

3.3 Discriminant analysis

After dimension reduction, we combine the reduced structure vectors from each stream into a single one and make use of LDA for classification. Although LDA can be used for dimension reduction in step 2 [7], it is noted that LDA severs here as a classifier.

Let $v = [v_1, v_2, ..., v_k]$ denote an augmented structure vector (ASV), where k is the number of substreams. 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),

$$\hat{W} = \arg\max_W \frac{|W^T S_b W|}{|W^T S_w W|}, \tag{4}$$

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:

$$S_w = \sum_{j=1}^{N} \sum_{v^i \in C_j} (v^i - \mu^j)(v^i - \mu^j)^T, \tag{5}$$

$$S_b = \sum_{j=1}^{N} n_j(\mu^j - \mu)(\mu^j - \mu)^T, \tag{6}$$

where μ^j is the mean of the ASVs of class C_j and μ is the mean of all the training samples. $\hat{W}$ 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:

$$\arg\min_j |\hat{W}^T v - \hat{W}^T \mu^j|. \tag{7}$$

One may suggest to apply LDA directly on structure vector z without applying PCA in step 2. However, z has a high dimensionality, which makes LDA easily suffer from the singular problem of covariance matrix and overfit the training data [8]. Finally, it is noted that discriminant analysis of eigen structure resembles a very successful face recognition method, usually called Fisherface [9]. One of the big differences between this method and ours is that we apply PCA on each sub-stream not the whole feature vector. This is because that, in our problem, the correlations between different sub-stream (cepstrum features) are generally very small.

4 Experiments

We carried out experiments on the connected Japanese vowel utterances database [4] to evaluate the performances of the proposed method. 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 acoustical variations. 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 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 events (distributions). Since we have only one training sample, we used an MAP-based learning algorithm [10]. The trained HMM includes 25 states, and each state is described by a 13-dimension Gaussian distribution with a diagonal covariance matrix. Following [4], we divide the 13D cepstrum+ 13D delta cepstrum feature vectors into 13 multiple substreams with block size 2. Each sub-streams contains two cepstrum and two delta cepstrum features. We calculate the structural vectors for each substream. Each structural vector has a dimensionality of 25×24/2=300. We make comparisons between PCA+LDA (P-LDA) and 2 phase LDA (2-LDA) [7] in the following experiments.

The dimension t of reduced structure vector v_i is an important parameter. We change t from 5 to

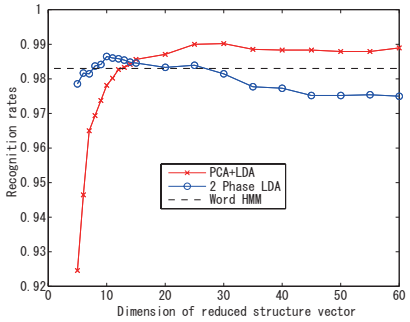

Fig. 4 Recognition rates vs. the dimension of reduced structure vector.

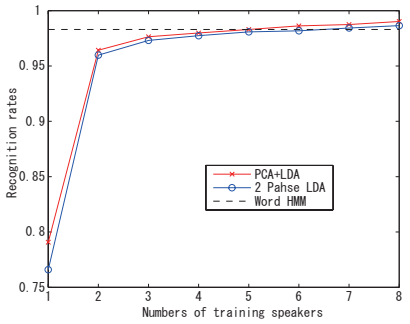

Fig. 5 Comparison of the recognition rates of different numbers of speakers in training data.

60 and see how it influences the recognition performance. The results are summarized in Fig. 4. The highest recognition rate 99.0% is achieved at $t = 30$ by P-LDA. One can also find that to increase t from 30 to 60 leads to a slight decrease of the recognition rates. This is because that the eigen vectors associated with small eigen vectors (large index) usually correspond to noisy directions and don't include much information for classification. The best rates are a bit higher than the recognition rate 98.3% of word HMMs trained with the same data.

We reduce the number of training speakers, and the results are shown in Fig. 5. Although the recognition rates slightly drop as the number of speakers decreases, we obtain a recognition rate 98.0% with only four training speakers. Generally, we can see that the performances of P-LDA and 2-LDA are very close. In another study [7], we have found that the structural representation achieved higher performance than HMMs on the artificially warped utterances, corresponding to those of speakers with different vocal tract length.

We compare the recognition rate of our method with those of the previous structure-based recognition methods, such as, multiple stream structuralization modeling (MSS) [4], and random discriminant structure analysis (RDSA) [3] . 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 MSP and 1/65 of RDSA.

Table 1 Comparisons of recognition rates

Method	P-LDA	2-LDA[7]	MSS[4]	RDSA[3]	HMM
Rate	**99.0%**	98.6%	95.3%	98.3%	98.3%

5 Conclusions

This paper proposes the method of dimension reduction and discriminant analysis for structure-based speech recognition. The proposed method deals with two limitations of invariant structural representation, too strong invariance and high dimensionality. The too strong invariance is released by constructing structures for each stream of speech signal. PCA and LDA 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 [3, 4], and word HMMs trained with the same data set. Moreover, the proposed method is about sixty times faster than the previous ones [3, 4] in classification. As future work, we are now investigating structure based methods for recognizing utterances including consonant sounds. We will study how to compare structures with different numbers of events.

参考文献

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

[2] 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 ASJ*, pp. 147–148, 2007.

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

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

[5] 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.

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

[7] S. Asakawa, *A study on word speech recognition based on structural representation of speech*, Ph.D. thesis, The Univ. of Tokyo, 2008, to appear.

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

[9] P.N. Belhumeur et al., "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Trans. PAMI*, vol. 19, no. 7, pp. 711–720, 1997.

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