社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

無作為判別構造解析を用いた日本語母音連結発声の自動認識

喬　　宇† 朝川　智† 峯松　信明†

† 東京大学大学院新領域創成科学研究科 〒 277–8561 千葉県柏市柏の葉 5–1–5
E-mail: †{qiao,asakawa,mine}@gavo.t.u-tokyo.ac.jp

あらまし　音声信号は様々な非言語的要因により変化し，音声認識システムはそれらに対処する必要がある．多くの非言語的変動は特徴量空間の変換により表現することができる．音声の構造的表象は特徴量空間の変換に対して不変であることが示されており，非言語的変動に対して頑健な音声認識が可能となる．しかし，構造的表象はその高い次元性が問題となる．これにより計算量が増えるだけでなく次元の呪い（curse of dimensionality）の問題も生じる．本研究では，この問題を解決する手法として Random Discriminant Structure Analysis（RDSA）を提案する．本手法は特徴量選択と判別分析とを用いることにより，高い次元性を持つ構造的表象のパラメータから冗長性を削減し，より低い次元の識別的な特徴量を計算する．さらに識別的特徴量を用いて複数の識別器を学習し，それらを統合することによって最終的な識別結果を出力する．連続的に発声された日本語 5 母音系列をタスクとした認識実験において，8 名の話者により学習した提案手法は 98.3% の認識率を示し，4,130 名による不特定話者 HMM（97.4% ）を上回る性能を示すことを確認した．

キーワード　無作為判別構造解析, 音声の構造的表象，音声認識，日本語母音系列

Recognition of Connected Japanese Vowel Utterances Using Random Discriminant Structure Analysis

Yu QIAO†, Satoshi ASAKAWA†, and Nobuaki MINEMATSU†

† Grad. School of Frontier Sciences, Univ. of Tokyo, 5–1–5, Kashiwanoha Kashiwa, Chiba, 277–8561 Japan
E-mail: †{qiao,asakawa,mine}@gavo.t.u-tokyo.ac.jp

Abstract　Automatic speech recognition has to deal with the non-linguistic variations of speech signals. Many non-linguistic variations can be modeled as the transformations of features. The universal structure of speech [12], [13], proves to be invariant to the feature transformations, and thus provides a robust representation for speech recognition. One of the difficulties of using the structure representation is due to its high dimensionality. This not only increases computational cost but also easily suffers from the curse of dimensionality [3], [8]. In this paper, we introduce Random Discriminant Structure Analysis (RDSA) to deal with this problem. Based on the observation that structural features are highly correlated and include large redundancy, the RDSA combines random feature selection and discriminative analysis to calculate several low dimensional and discriminative representations from an input structure. Then an individual classifier is trained for each representation and the outputs from each classifier are integrated for the final classification decision. Experimental results on connected Japanese vowel utterances show that our approach achieves a recognition rate of 98.3% based on the training data of 8 speakers, which is higher than that (97.4%) of HMMs trained with the utterances of 4,130 speakers.

Key words　random discriminant structure, structural representation of speech, speech recognition, Japanese vowel sequences

1. Introduction

Speech signals inevitablely exhibit non-linguistic variations, such as speaker,s communication channels, microphones and so on. One of the fundamental problems in speech recognition is to deal with these non-linguistic variations. Modern speech recognition studies largely make use of the statistical methods, for example GMM and HMM,

to solve this problem, which try to model the distributions of speech signals [4]. Extensive studies have shown that the statistical methods can achieve comparably high recognition rates when using proper models and sufficient training data. However, one of the disadvantages of these methods is that a large number of training samples must be prepared to estimate reliable distributions. The successful commercial speech recognition systems always make use of millions of data from thousands of speakers for training [7]. Contrary to this is human perception of speech. Thinking, a child doesn't need to hear the voice of thousands of persons before he (or she) could understand speech. This fact largely indicates that there may exist a robust representation of speech which is nearly invariant to non-linguistic variations. It is by this robust representation, we consider that children can learn speech with very biased training data called "mothers and fathers". This fact is also partly supported by recent advance in the neuroscience, which shows that the linguistic aspect of speech and the non-linguistic aspect are processed separately by the auditory cortex [18].

Along this line, the third author of this paper proposed a universal structure theory [12], [13] for speech, and proved that the structural representation is invariant to transformations (linear or nonlinear) in feature space [14]. To obtain a structural representation, an utterance is converted to a sequence of distributions (called events); then the structural representation is calculated as the Bhattacharyya distance matrix of these events. Our previous works [1], [15] have preliminarily exhibited the effectiveness of the structural representation in speech recognition. However, there is a difficulty for using the structural representation in recognition tasks: the dimensionality of structural representation is usually high. Let m denote the number of events in a structure. The dimensionality of the structural representation will be $O(m^2)$. It is well-known that the high dimensionality of input feature not only increases the computational time, but also makes it difficult to train robust classifiers (known as the curse of dimensionality [3], [8]). Moreover, we find that the structural features are highly related to each other and there exists large redundancy among them. Therefore, it is necessary to reduce the dimensionality for obtaining a more compact and discriminative representation.

This paper proposes Random Discriminant Structure Analysis (RDSA) for the structure-based speech recognition. Our approach makes use of random feature selection to preliminarily reduce the dimensionality of structures. The discriminative features are found as those with the largest ratios of between-class variance to within-class variance through Fisher Discriminant Analysis (FDA). The random feature selection can help to circumvent the overfitting and singularity problems of FDA. The classifier ensemble can reduce the variance and bias of single FDA classifier, thus to improve the recognition performance. Experimental results on connected Japanese vowel utterances show that our approach can achieve a recognition rate of 98.3% based on the training data of 8 speakers. This is higher than the recognition rates of all compared methods and than that (97.4%) of HMMs trained with 4,130 speakers. More details will appear in a coming conference paper [17].

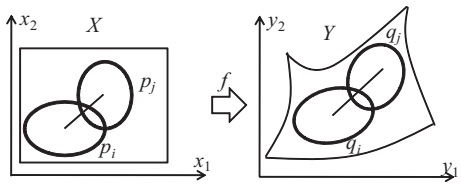

図 1 Invariance of Bhattacharyya distance.

2. Invariant Structure for Speech Representation

In this section, we will give a brief overview on invariant structure theory and on how to calculate structure representations from utterances [12], [13].

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 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. Thus if we can find invariant metrics in both space X and space Y, these metrics can serve as robust features for classification.

The universal structure theory shows Bhattacharyya distance (BD) between two distributions is an invariant metric Fig.1. BD is defined as,

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

It is not difficult to calculate that under transformation f, distribution $q_i(y)$ can be expressed by,

$$q_i(y) = p_i(f^{-1}(y))|J(y)|, \qquad (2)$$

where f^{-1} denotes the inverse function of f, and J is the Jacobian matrix of function f^{-1}. Then it can be proven that [14],

$$BD(p_i, p_j) = BD(q_i, q_j). \qquad (3)$$

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,

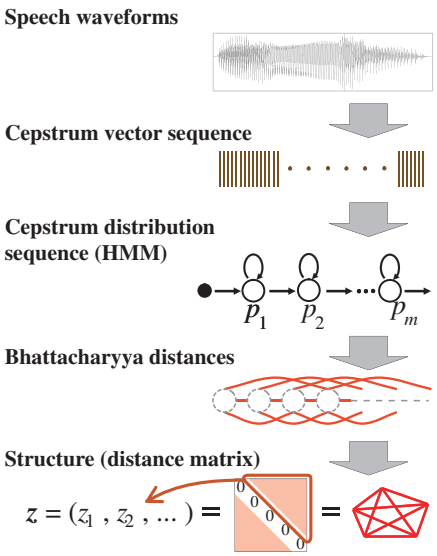

図 2　Framework of structure construction.

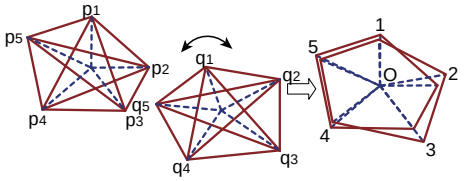

図 3　Utterance matching by shift and rotation.

we calculate a sequence of cepstrum from input speech waveforms. Then an HMM is trained based on a single cepstrum sequence and each state of HMM is regarded as an event p_i. Thirdly we calculate the Bhattacharyya distances between each pair of p_i and p_j. These distances will form a $m \times m$ 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 a 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.

It can be shown that many non-linguistic variances [12], [13], such as the length of vocal tract [16], 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 f 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 adaption research of speech recognition and the speaker conversion research of speech synthesis. In Maximum Likelihood Linear Regression (MLLR) based speaker adaption [10], a linear transformation: $y = Hx + d$ 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. H and d), which lead to the minimum difference (Fig.3). This minimum difference serves as a matching score of utterances. It has been shown that, using structural representation, we can approximate the minimum difference without explicitly estimating transformation parameters [13].

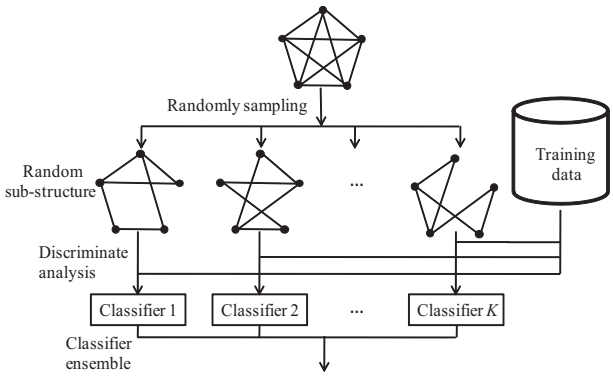

図 4　Random discriminant structure analysis.

3. Random Discriminant Structure Analysis

One of the difficulties of using BDs for classification is its high dimensionality. Let m denote the number of events. Then, the dimensionality of structural representation will be $m(m-1)/2$. The high dimensionality will increase the computational cost and make it difficult to train robust classifiers (known the Curse of Dimensionality [8]). Moreover, the BDs 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}). If we consider the space of BD distances, only a small part (a low dimensional subspace) of this high dimensional space should contain discriminative information. Based on these observations, we think it is essential to reduce the input structure into a compact (low dimension) yet discriminative representation for obtaining a better recognition rate.

In this paper, we will develop a method called Random Discriminant Structure Analysis, which combines feature selection and feature transformation for estimating a low-dimensional discriminative representation of structures. This method includes three steps. Firstly, we randomly sample the edges from an input structure to obtain several random sub-structures. Then discriminative analysis is applied on each random sub-structure to train a classifier for that structure. Finally, the outputs of each classifier are combined to reach the final decision. The flow chart of RDSA is shown in Fig. 4. And the details will be explained as in 3. 1.

3. 1 Construction of Random Structure

In the first step, we construct K random sub-structures $\{E_k\}_{k=1}^{K}$, each E_k is obtained by randomly sampling S edges $\{e_i^k\}_{i=1}^{S}$ from E. This can also be seen as randomly selecting a small number of dimensions from vector z. In the next,

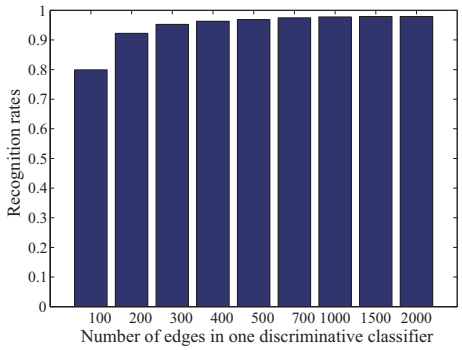

図 5 Recognition rates vs numbers of edges.

we will apply discriminant analysis on each sub-structure E_k independently. The random sub-structure construction can reduce the dimensionality of original structures while the number of training data remains the same.

Here we use random feature selection instead of greedy selection methods. This is because, our structural features (BDs) are highly correlated features. The greedy selection methods can only reach the local optimal combination of some of the features, which makes it unsuitable for our task. Moreover, our method includes a classifier ensemble strategy. This requires the independence among individual classifiers, which can be largely satisfied through random selection. The efficiency of random feature selection in recognition had been exhibited in [6]. It was shown in [19] that a random subspace method (similar to random feature selection) is useful for discriminant analysis . We found that only a small number of edges can include sufficient information for an individual discriminative analysis. This can be verified by our experimental results given in Fig. 5 that shows the average recognition rates for using different number of edges (features). The detailed setting of the experiments are described in Section 4. The original pattern includes 3,900 edges. It is easy to see that when the number of edges is larger than 400 (about 10% edges), the increase of edge numbers in an individual classifier cannot improve the recognition rates very much.

3. 2 Discriminant Analysis

We use Fisher Discriminant Analysis (FDA) for discriminant analysis due to its simplicity and effectiveness. FDA is a classical method to find the discriminant linear transformation W of features z [3]: $t = W^T z$, where t denotes the discriminant features and usually has lower dimension than z. 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 features. Assume we have M training samples $\{z_i\}_{i=1}^M$ belonging to N classes $\{C_j\}_{j=1}^N$. 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:

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

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

where m_j is the mean of class C_j and m is the mean of all the training samples. W can be computed as the eigenvectors of $S_w^{-1} S_b$. Once W is known, we can determine the discriminative features as $W^T z$ for sample z. For each random set E_k, we apply FDA on it to obtain W_k. Then the nearest mean classifier F_k can be built by using the discriminant features:

$$\arg\min_j |W_k^T z^k - W_k^T m_j^k|, \tag{7}$$

where z^k denotes the distance vector of edges in E_k and m_j^k denotes the mean distance vector of edges in E_k for j-th class.

FDA can be used to determine the discriminative structure. However, it is well-known that FDA may suffer from overfitting when the dimensionality of the features is high and the number of training samples is limited [3]. This fact can influence the performance of FDA. Another serious problem of FDA is that the within-class scatter matrix S_w can be singular and have no inverse. In our approach, these problems can be largely circumvented through the use of random edge selection which reduces the dimensionality of input features. The final performance is further improved through classifier combination.

3. 3 Classifier Ensemble

In the final phase, we integrate the outputs from each classifier to reach the final classification decision. It has been shown that classifier ensemble is an efficient method to reduce the variance and bias of an individual classifier [2]. There are two typical strategies for classifier ensemble: summation and voting. Assume the outputs of each individual classifier is a vector containing the confidence score for each category. For the summation method, the output vectors are added together and the final class is decided as the one with the highest summarized confidence. This can be expressed by

$$\arg\min_j \sum_k |W_k^T z^k - W_k^T m_j^k|. \tag{8}$$

In voting, the final decision is identified as the category supported by the largest number of individual classifiers. We experimentally compared the two ensemble methods and found that summation has better performance. In the experiments, we will use summation for classifier for ensemble without special notification.

4. Experiments

To examine the performance of random discriminant structure analysis, we use the connected vowel utterances [1] for experiments. It is known that acoustic features of vowel sounds exhibit larger between-speaker variations than consonant sounds. The data used includes all combinations of five Japanese vowels 'a','e','i','o' and 'u', such as 'aeiou','aeiuo', ... So there are totally 120 words. The samples of 16 speakers (8 males and 8 females) are recorded. Every speaker provides 5 utterances for each word. So the total number of utterances is 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.

We calculate twelve Mel-cepstrum features and one power feature for every frame of an utterance. HMM training is used to convert cepstrum vector sequence into events (distributions). Since we have only one training sample, an MAP-based learning algorithm [5] is adopted. The trained HMM includes 25 states, and each state has a 13-dimension Gaussian distribution with a diagonal covariance matrix. Following [1], we divide a cepstrum feature steam into 13 multiple sub-streams, and calculate the structures for each sub-stream. So an utterance is represented as a set of ${}_{25}C_2 \times 13 = 3,900$ edges. More details can be found in our works [1], [13]. We use a regularized version of Fisher discriminant analysis (RDA) [11] to train an individual classifier of each random structure. It can be shown that the regularization can reduce the unfavorable effects of noisy samples and overfitting problem.

4.1 Experiment 1

In the first experiment, we examine the performance of various numbers of edges used in sub-structures and various numbers of discriminative classifiers. We set the numbers of edges as 100, 200, 300, ..., and 2,000, and the numbers of discriminative classifiers from 1 to 30. For each combination of edge number and classifier number, we repeat the training procedure 20 times to get 20 sets of RDSA classifiers (注1) and calculate their average recognition rates.

The results are summarized in Fig. 6. It can be seen that when the number of discriminative classifiers is larger than 10 and the edge numbers is larger than 300, the increase of classifier number can only improve the recognition rates very little. Also when edge number is larger than 700 and classifier number is larger than 10, there is no improvement of the recognition rates observed if we increase the edge numbers

(注1)：20 is a small number if we consider there exists millions of possible combination of edges and classifiers. However, due to the time limitation, it is impossible for us to test the experiments on large numbers.

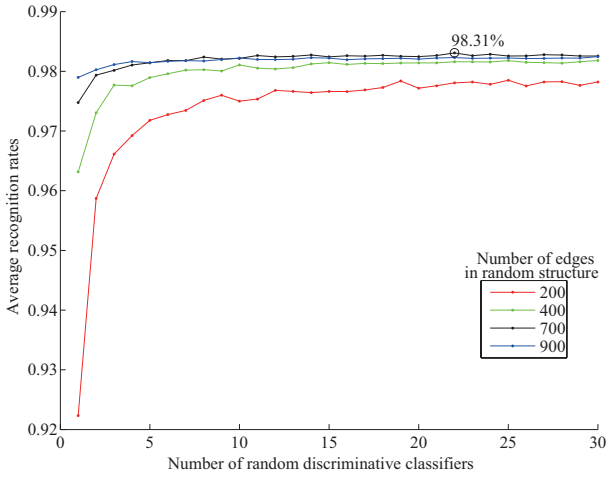

図 6 Recognition rate vs. number of edges.The black circle represent the highest recognition rate.

of individual classifier. In fact, as we can find in Fig. 6, the highest average recognition rate is achieved when edge number equals to 700 and classifier equals to 22. This is because: although, for an individual classifier, the addition of edges can increase recognition rates, this may reduce the independence among different classifiers and impair the performance for classifier ensemble. These results indicate that we need not to use a large edge number and a large classifier number for achieving a good recognition rate. This is important in practice, since for small edge number (1/4-th of all the edges) and classifier number (about 20), we don't have to do large computation in both training and testing procedures.

4.2 Experiment 2

We also make comparisons with other classical classification methods: nearest neighbors (NN), nearest mean (NM), Gaussian distribution model (GM) and Mahalanobis distances (MD). For nearest neighbors and nearest mean, Euclidean distance is used. For Gaussian distribution and Mahalanobis distances, we use diagonal covariance matrices. The results of using 8 speakers' data for training are summarized in table 1. We can see the proposed method achieves the highest recognition rate. We also examine the effect of using smaller numbers of speakers for training data. We randomly selected k ($1 \leqq k \leqq 7$) 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 their average performance. Note that testing data are the same and no testing data is used in training. For all the experiments, the proposed method always has the best performance and is less influenced by the reduction of training speakers. With the training utterances from only 5 speakers, the proposed RDSA can achieve a higher recognition rate than that of HMM (97.4%) trained by the utterances of 4,130 speakers [9]. (HMM trained by the 260-speakers has a

表 1 Comparisons of recognition rates. The 2nd row shows the numbers of training speakers. The first five methods use the structural representation as input.

method	NN	NM	MD	GM	RDSA	HMM
#speaker	8					4,130
rate	93.1%	95.2%	94.1%	96.2%	**98.3%**	97.4%

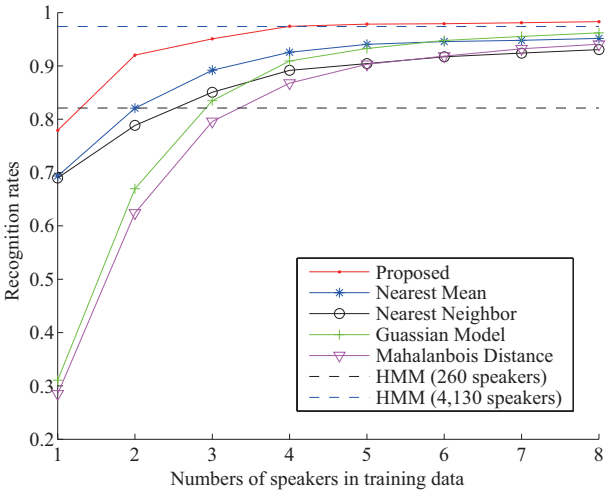

図 7 Comparison of the recognition rates among different methods and different numbers of speakers in training data

recognition rate of 82.1%.) Two facts should be noted here. 1) The structural representations are reliable features, since the simple classifiers such as nearest neighbors and nearest mean can achieve relatively high recognition rates with limited training data. 2) The reduction of training speakers can lead to significant decrease of recognition rates. This means that we still depend on sufficient training data (although it is much less than that of HMM in our experiments) for achieving a good performance.

5. Conclusions

This paper proposed a novel method, Random Discriminant Structure Analysis (RDSA) for universal structure based speech recognition. RDSA has the advantages of random structure construction, discriminant analysis and classifier ensemble. Compared with the original structural representation, the representation calculated by RDSA has lower dimensions and is more discriminative. It also preserves the desirable invariant property of input structure. In RDSA, the random structure construction can circumvent the overfitting and singularity problem of FDA. For classification, RDSA makes use of discriminant analysis and classifier ensemble to improve the recognition rates. In the experiments, the proposed method achieved a recognition rate of 98.3% on the connected vowel utterances based on the training speech of 8 speakers, which is higher than all compared methods, and the HMM trained by the utterances of 4,130 speakers. The proposed method is more robust to the reduction of the numbers of training speakers. For future work, we are considering to develop a mechanism which can integrate edge selection and classifier ensemble in a more effective way, and to evaluate the proposed methods on larger utterance databases that includes both vowels and consonants.

6. Acknowledgment

The first author would like to thank the Japan Society for the Promotion of Science (JSPS) for the financial support under contract P07078.

文　献

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

[2] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley-Interscience, 2000.

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

[4] S. Furui. *Digital Speech Processing, Synthesis and Recognition*. Marcel Dekker, 2001.

[5] J.L. Gauvain and C.H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixtureobservations of Markov chains. *IEEE Trans. SAP*, 2(2):291–298, 1994.

[6] T.K. Ho. The random subspace method for constructing decision forests. *IEEE Trans. PAMI*, 20(8):832–844, 1998.

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

[8] A.K. Jain. Statistical Pattern Recognition: A Review. *IEEE Trans. PAMI*, 22(1):4–37, 2000.

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

[10] CJ Leggetter and PC Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–185, 1995.

[11] D. Lin and X. Tang. Recognize High Resolution Faces: From Macrocosm to Microcosm. *Proc. CVPR*, pages 1355–1362, 2006.

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

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

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

[15] T. Murakami, K. Maruyama, N. Minematsu, and K. Hirose. Japanese Vowel Recognition using External Structure of Speech. *Proc. ASRU*, pages 203–208, 2005.

[16] M. Pitz and H. Ney. Vocal Tract Normalization Equals Linear Transformation in Cepstral Space. *IEEE Trans. SAP*, 13(5):930–944, 2005.

[17] Y. Qiao, S. Asakawa, and N. Minematsu. Random Discriminant Structure Analysis for Automatic Recognition of Connected Vowels. *Proc. ASRU*, 2007.

[18] S. K. Scott and I. S. Johnsrude. The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, 26(2):100–107, 2003.

[19] M. Skurichina and R.P.W. Duin. Bagging, Boosting and the Random Subspace Method for Linear Classifiers. *Pattern Analysis & Applications*, 5(2):121–135, 2002.