MULTI-STREAM PARAMETERIZATION FOR STRUCTURAL SPEECH RECOGNITION

Satoshi Asakawa[1], *Nobuaki Minematsu*[1] *and Keikichi Hirose*[2]

[1]Graduate School of Frontier Sciences, The University of Tokyo
[2]Graduate School of Information Science and Technology, The University of Tokyo
{asakawa,mine,hirose}@gavo.t.u-tokyo.ac.jp

ABSTRACT

Recently, a novel and structural representation of speech was proposed [1, 2], where the inevitable acoustic variations caused by non-linguistic factors are effectively removed from speech. This structural representation captures only microphone- and speaker-invariant speech contrasts or dynamics and uses no absolute or static acoustic properties directly such as spectrums. In our previous study, the new representation was applied to recognizing a sequence of isolated vowels [3]. The structural models trained with a single speaker outperformed the conventional HMMs trained with more than four thousand speakers even in the case of noisy speech. We also applied the new models to recognizing utterances of connected vowels [4]. In the current paper, a multiple stream structuralization method is proposed to improve the performance of the structural recognition framework. The proposed method only with 8 training speakers shows the very comparable performance to that of the conventional 4,130-speaker triphone-based HMMs.

Index Terms— speech recognition, robust invariance, the structural representation, multiple stream structuralization

1. INTRODUCTION

Every speech recognition system uses acoustic models based on phonetics, which observes acoustic events of speech directly and absolutely. The observations, however, inevitably vary according to the non-linguistic factors, such as age, gender, microphone, line, and so on. These non-linguistic variations are noises for extracting linguistic information from speech and, thus, the speaker-independent HMMs can still have outlier speakers easily even though they are trained with thousands of speakers. In contrast, humans, even children, can communicate orally without experiences of hearing the voices of thousands of speakers. We consider that this fact implies that there exists a robust representation of speech which is nearly invariant to the non-linguistic variations.

Apart from semantics, linguistics provides two definitions of the phoneme [6]. 1) A phoneme is a class of sounds that are phonetically similar. 2) A phoneme is one element in the sound system of a language having a characteristic set of interrelations with each of the other elements in that system. These two definitions correspond well to phonetics and phonology. The former discusses the absolute values of linguistic sounds and the latter does their relational values. It is clear that the HMMs are based on the first phonetic definition. As far as the authors know, however, speech recognizers have never been built only based on the second purely-phonological definition except for our previous studies.

Recently, a novel acoustic representation of speech was proposed, which is called the structural representation of speech [1]. It discards all the absolute properties of speech events because they inevitably transmit the non-linguistic information. The new representation captures only speech contrasts or dynamics to form an external structure composed of the acoustic events. Here, the speech contrasts are modeled in a distorted non-Euclidean space so that they are invariant with the non-linguistic variations. The authors already applied this new representation to speech recognition [3]. In order to discuss its fundamental characteristics when dealing with speech samples with high speaker dependency, a very simple task, recognition of isolated *vowel* sequences, was adopted. The proposed method trained only with a single speaker outperformed the conventional HMMs trained using more than four thousand speakers, even in the case of noisy speech. We also applied it to recognizing continuous speech, that is connected *vowels* [4]. It was shown that the contrast-based invariant representation could remove the speaker difference effectively on one hand, but different words were sometimes regarded as identical on the other hand. We called this undesirable by-effect as *problem of too strong invariance* and it is very critical.

In order to solve this problem, this paper introduces a multiple stream structuralization strategy as some constrains on the structural acoustic matching framework. Experiments of recognizing connected vowel utterances show that the proposed method only with 8 training speakers outperforms by far 260-speaker HMMs with CMN and provides the very comparable performance to 4,130-speaker HMMs with CMN.

2. STRUCTURAL REPRESENTATION OF SPEECH

2.1. Mathematical modeling of the non-linguistic variation

In speech recognition, three types of distortions or noises, additive, multiplicative and linear transformational, are often discussed. Background noise is a typical example of additive noise, but this is not inevitable because a speaker can move to a quiet room if needed. Speech recognition in noisy environments is surely an important issue, but, as we want to focus only on the *inevitable* distortions, additive noise is not considered here. For example, CALL systems need speaker-robust techniques more than noise-robust ones because a self-training system can be used in a student's private room [5].

The distortions caused by microphones and lines are examples of multiplicative distortion. GMM-based speaker modeling assumes that a part of speaker individuality is regarded as this type. These distortions are inevitable because speech has to be produced by a certain human and recorded by a certain device. If a speech event is represented by cepstrum vector $\boldsymbol{c}$, this type of distortion is addition of vector $\boldsymbol{b}$; $\boldsymbol{c'} = \boldsymbol{c} + \boldsymbol{b}$.

Two speakers have different vocal tract shapes and two listeners have different hearing characteristics. Mel scaling is just the average pattern of the hearing characteristics. These are typical examples of linear transformational distortion, which is naturally inevitable.

Vocal tract length difference as well as hearing characteristics difference is often modeled as frequency warping of the spectrum. Any monotonic frequency warping in the spectral domain can be converted into multiplication of matrix $\boldsymbol{A}$ in the cepstral domain [7]; $\boldsymbol{c}' = \boldsymbol{A}\boldsymbol{c}$.

Although various distortion sources can be found in speech communication, the total distortion due to the *inevitable* sources is simply modeled as $\boldsymbol{c}' = \boldsymbol{A}\boldsymbol{c} + \boldsymbol{b}$, i.e., affine transformation.

2.2. The structural representation of speech

In order to obtain a speaker invariant representation, we focused on a *structure* composed of acoustic events. An n-point structure is determined uniquely by fixing the length of all the ${}_nC_2$ lines including the diagonal ones. In other words, a geometrical structure is completely represented as its *distance matrix*. Then, a necessary and sufficient condition for the invariant structure is that distance between any two points is invariant with any of a single affine transformation. This condition seems to be mathematically impossible to satisfy because affine transformation always distorts a structure unless it is of a special form. However, it can be satisfied by distorting the space so that the distance can be invariant.

Let us consider Bhattacharyya distance (BD), one of the distance measures between two distributions $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$,

$$BD(p_1(\boldsymbol{x}), p_2(\boldsymbol{x})) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(\boldsymbol{x})p_2(\boldsymbol{x})}d\boldsymbol{x}. \quad (1)$$

It was mathematically proven that, when a linear or non-linear one-to-one transformation is done on two distributions commonly, BD is not changed before and after the transformation [8]. Considering that the speaker conversion technique in speech synthesis applies an adequate mapping function on source speech samples and that BD cannot be changed by any of a linear or non-linear transformation, we can regard this transform invariance as *robust* invariance. When the two distributions are Gaussian, BD is formulated as follows,

$$\begin{aligned} &BD(p_1(\boldsymbol{x}), p_2(\boldsymbol{x})) \\ &= \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \left(\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2}\ln \frac{|(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)/2|}{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}|\boldsymbol{\Sigma}_2|^{\frac{1}{2}}} \end{aligned} \quad (2)$$

and in this case, BD is invariant to any common affine transformation. When acoustic events are described as cepstral distributions, an affine invariant structure can be obtained by calculating a BD-based distance matrix from cepstral distributions, and this structural representation will be invariant to the inevitable non-linguistic variations. Since any $\boldsymbol{A}$ and any $\boldsymbol{b}$ cannot change a structure, $\boldsymbol{A}$ is interpreted as rotation of the structure, and $\boldsymbol{b}$ is interpreted as its shift. As told above, the structural invariance is realized because BD calculation distorts the space where the distributions are observed. Analysis of this distorted space is done in [2] based on differential geometry.

2.3. Acoustic matching between two structures

Acoustic matching between two n-point structures is done by shifting ($\boldsymbol{b}$) and rotating ($\boldsymbol{A}$) a structure so that the two can be overlapped the best (see Figure 1). It was experimentally shown that the minimum of the total distance of the corresponding two events after the adaptation with $\boldsymbol{b}$ and $\boldsymbol{A}$ can be approximately calculated as Euclidean distance between the two distance matrices, where the upper-triangle elements form a vector [9],

$$D = \sqrt{\frac{1}{n}\sum_{i<j}(p_{ij} - q_{ij})^2}, \quad (3)$$

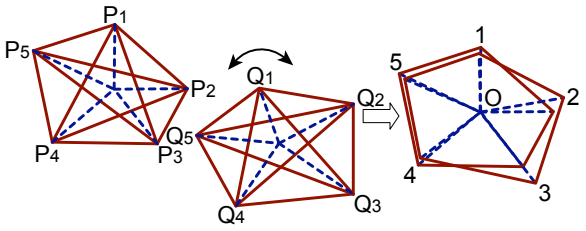

Fig. 1. Acoustic matching after shift($\boldsymbol{b}$) and rotation($\boldsymbol{A}$)

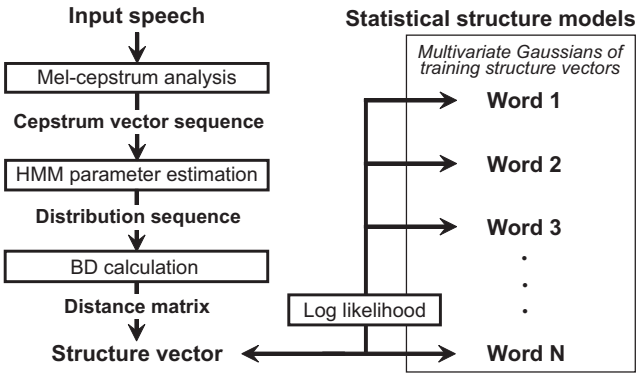

Fig. 2. Framework of the structural recognition

where p_{ij} is (i, j) element of distance matrix P. It should be noted that the acoustic matching score after the adaptation can be calculated without any absolute properties of the events and that without formulating $\boldsymbol{b}$ and $\boldsymbol{A}$ explicitly. In other words, a structural acoustic matching process implicitly includes an adaptation process but, for the adaptation, no additional operations are required. Further, the score can be calculated only with the two distance matrices representing the two sets of speech events structurally.

3. STRUCTURE-BASED SPEECH RECOGNITION

3.1. Framework of the structure-based recognition

The overall framework is shown in Figure 2. The left side of the figure shows the procedure to extract a structure from an input utterance. First, a cepstrum vector sequence was obtained from an input utterance by acoustic analysis. Then, to convert the vector sequence to a distribution sequence, an HMM was trained with the single vector sequence. Here, its transition probabilities were discarded. Since all the distributions had to be estimated from a single utterance, the MAP-based estimation was adopted. After that, a distance matrix was obtained by calculating BDs between any two of the distributions. Finally, the upper-triangle elements of the distance matrix were used as feature vector, called *structure vector*.

The right side of the figure is a reference template database. Here, training samples of each word were converted into structures and using them, a statistical structure model was trained for each word. We adopted multivariate Gaussian distribution for the modeling. Acoustic similarity between an input structure vector and a statistical structure model was calculated as log likelihood. The template showing the maximum log likelihood is the result of recognition.

3.2. A problem of too strong invariance

With any transformation, linear or non-linear, the structural invariance is satisfied. This robust invariance is very effective to remove the non-linguistic variations from speech acoustics. However, it will be so strong that it should cause a critical problem, where a word and another linguistically different word are treated as identical. This too strong invariance should decrease the performance easily. Some constraints have to be introduced to restrict allowable transformations.

We focused on $\boldsymbol{A}$, the rotation, and any $\boldsymbol{A}$ cannot change the structure. What kind of $\boldsymbol{A}$ is required to be considered if we want to model only the speech variations caused by the vocal tract length difference. In [10], to model the effect of the vocal tract length, $\boldsymbol{A}$ is formulated as follows.

$$\boldsymbol{A} = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \cdots \\ 0 & 1-\alpha^2 & 2\alpha-2\alpha^3 & \cdots & \cdots \\ 0 & -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}, \quad (4)$$

where α is a warping parameter and $|\alpha| < 1$. If α is sufficiently small, α^n with high order n can be ignored, and matrix $\boldsymbol{A}$ has non-zero elements only in and near the diagonal. If we adopt the structural recognition framework of Figure 2 as it is, an utterance and its transformed version with a completely different matrix from Equation (4) are judged as identical. This is the critical problem and we have to solve it.

If a structure in a space is projected into one of its sub-spaces, the projected structure will naturally change through transformation. By hypothesizing that the structural invariance is still satisfied in the sub-space, geometrically speaking, the allowable transformations are restricted. This hypothesis is easily introduced into the structural matching procedure by separating a cepstrum stream into multiple independent sub-streams. Then, a structure is constructed for each sub-stream, called multiple stream structuralization.

Why do we consider multiple stream structuralization? We have a very good and strong reason. $(\boldsymbol{c}^T, \boldsymbol{\Delta c}^T)^T$ is a feature vector here, where $\boldsymbol{c} = (c_1\ ,\ c_2\ ,\ \cdots\ ,\ c_M)^T$ is a cepstrum vector and $\boldsymbol{\Delta c} = (\Delta c_1\ ,\ \Delta c_2\ ,\ \cdots\ ,\ \Delta c_M)^T$ is its derivative. BD is invariant to any common affine transformation;

$$\begin{pmatrix} \boldsymbol{c'} \\ \boldsymbol{\Delta c'} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{\Delta c} \end{pmatrix} + \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} \quad (5)$$

where any of $\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}$ satisfies the invariance.

If the feature vector is divided into two streams, $\boldsymbol{c}$ and $\boldsymbol{\Delta c}$, BD is invariant in each sub-space.

$$\boldsymbol{c'} = \boldsymbol{A}_{11}\boldsymbol{c} + \boldsymbol{b}_1 \quad (6)$$

$$\boldsymbol{\Delta c'} = \boldsymbol{A}_{22}\boldsymbol{\Delta c} + \boldsymbol{b}_2 \quad (7)$$

In this case, $\boldsymbol{A}_{12}$ and $\boldsymbol{A}_{21}$ can be regarded as $\boldsymbol{0}$ in Equation (5), and the rotation of structure is considered separately in each sub-space. Dividing the feature vector into lower-dimensional sub-vectors assumes more of the upper and lower triangular elements of $\boldsymbol{A}$ to be zero. The multiple stream structuralization is regarded as good constraints on the allowable transformations of the structure.

If the proposed framework in Figure 2 is adopted as it is, it will cause a problem of too strong invariance. To solve this, by translating the algebraic constrains of Equation (4) into its corresponding geometrical constraints, they are introduced into the structural speech recognition. In this paper, only uniform division is tentatively examined. A feature vector is divided into a group of sub-vectors of the same number of dimensions. The total distance between two structures is calculated by accumulating structural sub-distances obtained in the individual sub-spaces.

Table 1. Acoustic conditions for the analysis

sampling	16bit / 16kHz
window	25 ms length and 4 ms shift
parameters	Mel cepstrum (1 to 12) + Δ (1 to 12)
distribution	1-mixture Gaussian with a diagonal matrix

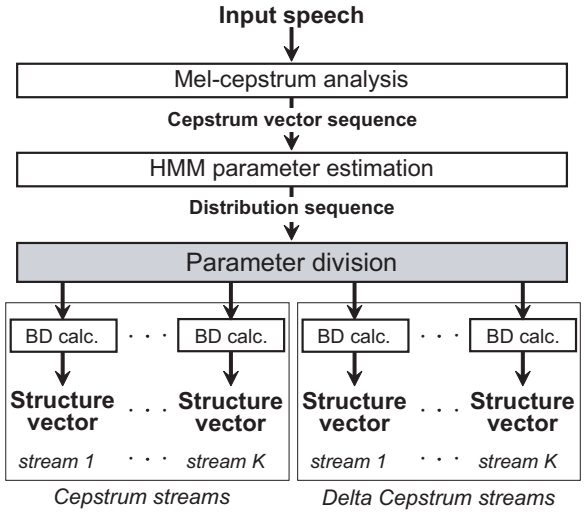

Fig. 3. Structuralization with parameter division

4. EXPERIMENTS

4.1. Experimental set-up

In order to investigate the fundamental characteristics of the proposed framework, utterances of connected vowels were adopted as recognition task. Vowel sounds are known to be much more dependent on speakers acoustically than consonant sounds. The number of vowels in an utterance was set to 5; V_1-V_2-V_3-V_4-V_5, where $V_i \neq V_j$. Since Japanese has 5 vowels, /aiueo/, Perplexity was ${}_5P_5$ (120). 8 male and 8 female adult speakers joined the recording and 5 utterances were recorded for each of the 120 words. The total number of utterances was 9,600. The samples from 4 males and 4 females were used for training and the others for testing. In our previous study [3], only a single speaker was used for training. In this work, however, as the required number of utterances was so large, multiple speakers were used for training. The conditions for acoustic analysis and HMM parameter estimation are shown in Table 1.

4.2. Parameter division

The parameter division discussed in the previous section was carried out after estimating the distribution sequence (Figure 3).

The left side of Figure 2 is replaced by Figure 3. If a speech stream was treated as two separate sub-streams of cepstrum and its Δ, two structures were always calculated. The parameter division was further carried out to introduce additional constraints, where the

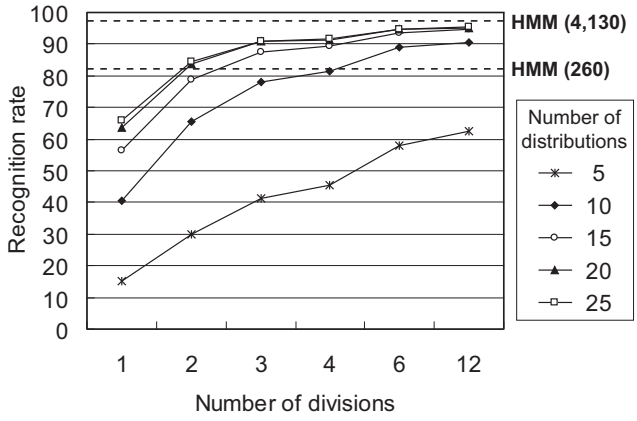

Fig. 4. Recognition performance with constraints

number of division K was 1 (no division), 2, 3, 4, 6, and 12 for each of the two streams. Finally, $2K$ distance matrices were obtained from one utterance. In this experiment, the number of the length of distribution sequences was set to 5, 10, 15, 20 and 25. For comparison, two sets of speaker-independent HMMs were tested for the same task; 260-speaker tied-state HMMs and 4,130-speaker tied-mixture HMMs, both of which were distributed by Continuous Speech Recognition Consortium in Japan [11] and were trained using MFCC with CMN applied. As language model, CFG allowing only the testing 120 words was used.

Results with the parameter division are shown in Figure 4 as function of the number of divisions. It can be seen that, with a larger number of divisions, the better performance was obtained, and when the number of distributions is larger than 20, the increase of distributions can hardly improve the recognition rates. Considering that an input utterance is connected five vowels, we can say that transient segments have to be modeled as separate distributions from those corresponding to stationary segments. The best performance was 95.3% in 12 divisions with 25 distributions. Figure 4 also shows the results of the conventional HMMs; 82.1% and 97.4% for 260- and 4,130-speaker HMMs, respectively. These results indicate that the proposed method only with 8 training speakers has much higher accuracy and robustness compared to the 260-speaker HMMs and provides the very comparable performance to the 4,130-speaker HMMs.

5. CONCLUSIONS

This paper showed the results of applying the speaker-invariant and structural representation of speech to recognizing utterances of connected Japanese vowels. By translating the algebraic constraints posed by transformation matrix $\boldsymbol{A}$ into its geometrical ones, a novel technique of multiple stream structuralization was proposed. With this technique and without any direct use of absolute speech features, the statistical structure models trained only with 8 speakers achieved 95.3% as recognition rate. This performance is by far better than that of 260-speaker HMMs and very comparable to that of 4,130-speaker HMMs. Since the structural representation of an utterance is obtained by extracting speech contrasts, the proposed method cannot identify an isolated sound. Considering the current speech recognition technology is based on absolutely identifying individual sounds, i.e. phonetics, the proposed framework, i.e. phonology, is completely opposite to the conventional framework. We consider that this strategic difference correspond to reductionism vs. holism [8] and the integration of both the strategies is very interesting to us.

Although the current paper did not describe this at all, the further performance improvement was already realized by another technique. One of the difficulties of the multiple stream structuralization is high dimensionality of feature parameters, which not only increases computational cost but also makes it difficult to train a classifier. Another new technique, called Random Discriminant Structure Analysis (RDSA) [12], which combines random feature selection, discriminative analysis, and classifier ensemble, successfully reduced the number of dimensions and, at the same time, improved the performance. The same recognition task was adopted and the statistical structure models proposed in this paper with RDSA achieved the recognition rate of 98.3%, which is higher than that of 4,130-speaker HMMs with CMN (97.4%).

For future work, the optimal parameter division should be obtained to realize more appropriate constraints on the allowable geometrical transformations for broader speaker variability. Further, we are now implementing an algorithm to estimate a structure from an utterance including consonant sounds to build a word recognizer for practical use.

6. REFERENCES

[1] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," *Proc. ICASSP*, pp.889–892, 2005.

[2] N. Minematsu, T. Nishimura, K. Nishinari, and K. Sakuraba, "Theorem of the invariant structure and its derivation of speech Gestalt," *Proc. SRIV*, pp.47–52, 2006.

[3] T. Murakami, K. Maruyama, N. Minematsu, and K. Hirose, "Japanese vowel recognition using external structure of speech," *Proc. ASRU*, pp.203–208, 2005.

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

[5] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," *Proc. Speech and Language Technology in Education*, 2007.

[6] H. A. Gleason, *An introduction to descriptive linguistics*, Holt, Rinehart & Winston, N.Y., 1961.

[7] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech and Audio Processing*, 13, 5, pp.930–944, 2005.

[8] N. Minematsu, S. Asakawa, and K. Hirose, "Linear and nonlinear transformation invariant representation of information and its use for acoustic modeling of speech," *Proc. Spring Meeting Acoust. Soc. Jpn.*, pp.147–148, 2007.

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

[10] T. Emori and K. Shinoda, "Rapid vocal tract length normalization using maximum likelihood estimation," *Proc. Interspeech*, pp. 1649–1652, 2001.

[11] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, "Recent progress of open-source LVCSR engine Julius and Japanese model repository," *Proc. ICSLP*, pp.3069–3072, 2004.

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