Japanese Vowel Recognition Based on Structural Representation of Speech

Takao MURAKAMI[†], *Kazutaka MARUYAMA*[†], *Nobuaki MINEMATSU*[‡] *and Keikichi HIROSE*[†]

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

Abstract

Speech acoustics varies from speaker to speaker, microphone to microphone, room to room, line to line, etc. Physically speaking, every speech sample is distorted. Socially speaking, however, speech is the easiest communication media for humans. In order to cope with the inevitable distortions, speech engineers have built HMMs with speech data of hundreds or thousands of speakers and the models are called speaker-independent models. But they often need to be adapted to the input speaker or environment and this fact claims that the speaker-independent models are not really speaker-independent. Recently, a novel acoustic representation of speech was proposed, where dimensions of the above distortions can hardly be seen. It discards every acoustic substance of speech and captures only their interrelations to represent speech acoustics structurally. The new representation can be interpreted linguistically as physical implementation of structural phonology and also psychologically as speech Gestalt. In this paper, the first recognition experiment was carried out to investigate the performance of the new representation. The results showed that the new models trained from a *single* speaker with no normalization can outperform the conventional models trained from 4,130 speakers with CMN.

1. Introduction

In every speech application, engineers have built acoustic models based on phonetics, which captures acoustic substances of the phonemes. The substances, however, inevitably vary and the well-known "sheep and goat" problem sometimes happens. Is there any other way to characterize the phonemes acoustically? Apart from semantics, linguistics provides two definitions of the phonemes[1]. 1) a phoneme is a class of phonetically-similar sounds and 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. It is obvious that the first definition brought about the so-called speaker-independent HMMs. As far as the authors know, no trials were made to implement speech recognition only based on the second definition.

Recently, a novel acoustic representation of speech, which is called the acoustic universal structure, was proposed[2]. This structural representation discards every acoustic substance of speech and captures only their interrelations. It can be interpreted linguistically as physical implementation of structural phonology and psychologically as speech Gestalt[3]. Collection of millions of /a/ sounds defines only the averaged distribution of /a/ and never deletes dimensions indicating speaker information. On the other hand, the structural representation can hardly have dimensions for speakers, microphones, rooms, lines, etc.

This study is the first trial of applying the new representation to speech recognition. In order to discuss its fundamental characteristics, simple vowel recognition was adopted as task.

2. The acoustic universal structure

2.1. Inevitable acoustic distortions in speech

Three types of noises or distortions are frequently discussed in speech recognition; additive, multiplicative, and linear transformational. Background noise is a typical example of the additive noise. But this is not inevitable because a speaker can move to a quiet room. Then, additive noise is ignored in this study.

Acoustic distortions caused by microphones, rooms, and lines are examples of the multiplicative distortion. GMM-based speaker modeling assumes that speaker individuality is represented rather well by the average pattern of log spectrum of the individual. This indicates that a part of speaker individuality is also regarded as the multiplicative distortion. This distortion is inevitable because speech has to be produced by a certain human and recorded by a certain acoustic device. If a speech event is represented by cepstrum vector c, the multiplicative distortion is addition of vector b and the resulting cepstrum is $c'=c+b$.

Two speakers have different vocal tract shapes and two listeners have different hearing characteristics. These are examples of the linear transformational distortion and naturally inevitable. Vocal tract length difference is often modeled as frequency warping of the log spectrum, where formant shifts are approximated. Hearing characteristics difference is also another frequency warping of the log spectrum. Any monotonous frequency warping can be converted into multiplication of matrix A in cepstrum domain[4]. The resulting cepstrum is $c'=Ac$.

Various distortion sources are found in speech communication. The total distortion due to the *inevitable* sources, A_i and b_i, is simply modeled as $c'=Ac+b$, i.e., affine transformation.

2.2. The acoustic universal structure in speech

What is desired is speech representation which is invariant to the inevitable acoustic distortions. This desire can be fulfilled by structuralizing speech acoustics. Geometrically speaking, an M-point structure is determined uniquely by fixing length of all the ${}_MC_2$ lines including the diagonal ones(distance matrix). Then, a necessary and sufficient condition for the invariant structure is that distance between any two points should not be changed by any of a single affine transformation. This condition is mathematically impossible to satisfy because affine transformation is transformation which distorts a structure. How to make impossible possible? The solution can be obtained simply by distorting space so that the structure can be invariant.

THEOREM OF THE INVARIANT STRUCTURE
M events are observed and every one is described not as point but as distribution. Distance between any two events is calculated as Bhattacharyya or Kullback-Leibler distance, which are based on information theory. A single affine transformation cannot change the distance matrix, i.e., the structure.

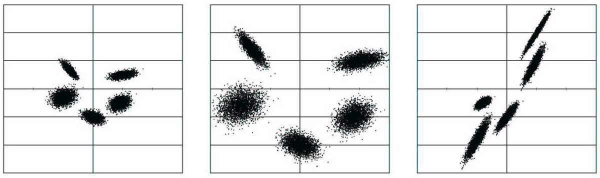
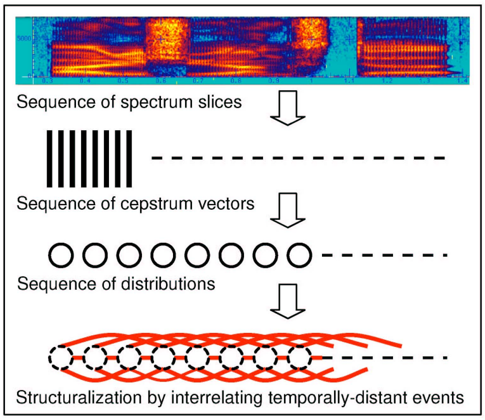


Figure 1: *Completely the same structure is found in the three.*



Sequence of spectrum slices

Sequence of cepstrum vectors

Sequence of distributions

Structuralization by interrelating temporally-distant events

Figure 2: *Structuralization of a single utterance*

Distribution means a Gaussian mixture. Bhattacharyya distance was adopted because it can be interpreted as normalized cross correlation between two PDFs $p_1(x)$ and $p_2(x)$.

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

where $0.0 \leq \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}dx \leq 1.0$ and name of unit of BD is bit because BD can be regarded as self-information. If the two distributions follow Gaussian, BD is formulated as follows.

$$\begin{aligned} &BD(p_1(x), p_2(x)) \\ &= \frac{1}{8}\mu_{12}^T \left(\frac{\sum_1 + \sum_2}{2}\right)^{-1} \mu_{12} + \frac{1}{2} \ln \frac{|(\sum_1 + \sum_2)/2|}{|\sum_1|^{\frac{1}{2}} |\sum_2|^{\frac{1}{2}}} \end{aligned} \quad (2)$$

μ_{12} is $\mu_1 - \mu_2$. Figure 1 shows three structures of five distributions. Any two of the three structures can be converted to one another by multiplying A. This fact indicates that the three structures(matrices) are evaluated as completely the same. Why this happens? Because BD calculation distorts the space where the distributions are observed. Geometrical analysis of this distorted space is done in [3] based on differential geometry. In MLLR[5] and SAT[6], speaker differences are characterized by affine transformations. The acoustic universal structure cannot be changed by a global affine transformation.

2.3. Structuralization of a single utterance

The structural representation can even be applied to a *single* utterance. Figure 2 shows the structuralization process. After the utterance is converted into a sequence of distributions, only the interrelations(distances) of any two of all the temporally-distant distributions are calculated to form the structure. Acoustic substances of every event are completely discarded. In the distorted space, since A cannot change the matrix, any A is interpreted as rotation. Then, acoustic matching between two M-point structures can be done by shifting(b) and rotating(A) a structure so that the two can be overlapped the best, shown in Figure 3. Suppose that there are two M-point structures in an N-dimensional *euclidean* space, where A allowing only rotation is an orthogonal matrix. In this case, the minimum of the total distance of the corresponding two events after the adaptation of b and A is formulated as

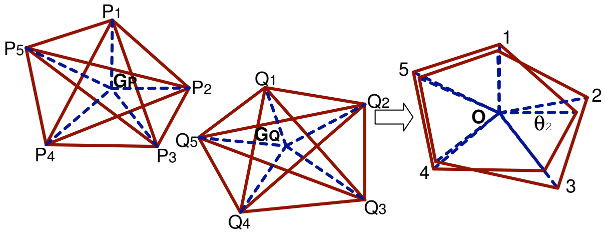


Figure 3: *Acoustic matching after shift(b) and rotation(A)*

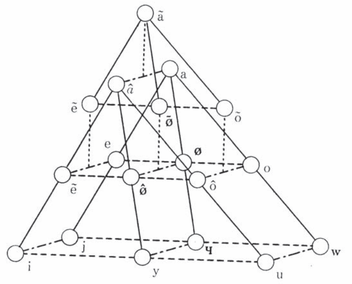


Figure 4: *Jakobson's geometrical structure*

$$\textstyle\sum_{i=1}^{M} \overline{OP_i}^2 + \overline{OQ_i}^2 - 2\sum_{i=1}^{N} \sqrt{\alpha_i}, \quad (3)$$

where O is the common gravity center of the two structures P and Q. α_i is the i-th eigen value of $N \times N$ matrix S^tTT^tS. S and T are $(\vec{OP}_1, ..., \vec{OP}_M)$ and $(\vec{OQ}_1, ..., \vec{OQ}_M)$ respectively. Acoustic matching score after adaptation can be calculated with no information of acoustic substances of the events and that without calculation of b and A. But the above quantity cannot be used directly because triangular inequality is not always satisfied in the distorted space. Then, approximate solution only with the two distance matrices should be used. The discussion implies that speech recognition may be possible with no direct use of acoustic substances of speech events.

2.4. Two kinds of interpretation of the structure[3]

Inspired by Saussure's claim on language, "*Language is a system of conceptual differences and phonic differences,*" Jackobson, Halle, and others developed structural phonology which discussed difference between two phonemes and structure composed of all the phonemic differences by using distinctive features. Figure 4 shows Jakobson's geometrical structure proposed for French vowels. The acoustic universal structure is also composed of differences(distances) between phonemes and invariant to affine transformation modeling the inevitable non-linguistic features. Thus the acoustic universal structure can be interpreted as physical implementation of structural phonology.

Perception of not sensory elements but the global quality produced through their interrelations is called Gestalt perception. Music perception and visual illusion are considered as its examples. When people hear music, anybody can identify name of the music without identifying its individual notes. A structurally-represented word contains not the information of its constituent phonemes but only their global quality. Once the word becomes known by the structural matching, however, all the phonemes in the word become known by referring to the mental lexicon. Some of the visual illusion, Titchener circles for example, can be mathematically interpreted as such that humans distort the space in their brains where the objects are found. The acoustic universal structure is composed of the interrelations of speech events and it is the global quality defined in a distorted space. This fact led us to define it as speech Gestalt.

3. Toward structure-based recognition

3.1. Acoustic matching between two distance matrices

In the structure-based speech recognition, every word is represented by its speech Gestalt, which is a distance matrix mathematically. If the Gestalt exists in an euclidean space, matching score can be calculated by the formula derived in Section 2.3. Since the Gestalt has to be in an noneuclidean space, however, some approximate solution is required. In [7], the following approximate equality was experimentally shown.

$$\frac{1}{M^2}\sum_{i<j}(\overline{P_iP_j}-\overline{Q_iQ_j})^2 \approx \frac{1}{M}\sum_{i}(\overline{P_iG_P}-\overline{Q_iG_Q})^2 \quad (4)$$

The right term approximates average of the total distance between two corresponding phonemes of the two structures after shift and rotation, where G_P and G_Q are put at a position(O) and structure Q is rotated so that the $\sum|\theta_i|$ (see Figure 3) should be minimized. The left term is euclidean distance between two matrices by viewing them as vectors. Since the above equality shows that euclidean distance between two matrices is physically definite and clear, a statistical model of a word can be built by using multiple utterances(structures) of the word.

3.2. Some improvements for stabler structure estimation

To implement the structure-based speech recognition, a problem has to be solved with respect to stability of estimating distributions. A word has to be represented as structure and the structure has to be composed by distributions. This means that parameters of a distribution have to be estimated from a very small number of frames. Although ML(Maximum Likelihood) criterion is the simplest to estimate the parameters, for the above reason, ML was expected to estimate the parameters rather poorly. Therefore, besides ML criterion, MAP(Maximum a Posteriori) criterion was investigated. As described in Section 1, the task adopted in this study was recognition of sequences of isolated vowels. Then, for prior knowledge, isolated vowel utterances were collected from a *single* speaker. This is because, as will be explained, the proposed acoustic models were trained only with the speaker. Each vowel utterance was represented as a diagonal Gaussian distribution of cepstrum parameters. An average vector and a diagonal covariance matrix for a new input utterance was estimated by referring to [8]. The following shows the procedure of the estimation, which is schematized in Figure 5.

μ_i : average vector of the i-th utterance

Σ_i : diagonal covariance matrix of the i-th utterance

μ_0 : average of $\{\mu_i\}$ $(=\frac{1}{m}\sum_{i=1}^{m}\mu_i)$

Σ_0 : average of $\{\Sigma_i\}$ $(=\frac{1}{m}\sum_{i=1}^{m}\Sigma_i)$

S_μ : diagonal covariance matrix of $\{\mu_i\}$ $(=\frac{1}{m}\sum_{i=1}^{m}(\text{DIAG}(\mu_i-\mu_0))^2)$

Ω : $=\Sigma_0 S_\mu^{-1}$

μ_{ML} : average vector of an input utterance

Σ_{ML} : diagonal covariance matrix of the input utterance,

where m is the total number of vowel utterances obtained for prior knowledge and DIAG(α) is a diagonal matrix whose diagonal elements are those of vector α. By using the above quantities, an average vector and a diagonal covariance matrix of a new input utterance can be derived as follows:

$$\mu_{MAP} = \hat{\mu}_0 \quad (5)$$

$$\Sigma_{MAP} = \hat{B}\hat{A}^{-1} \quad (6)$$

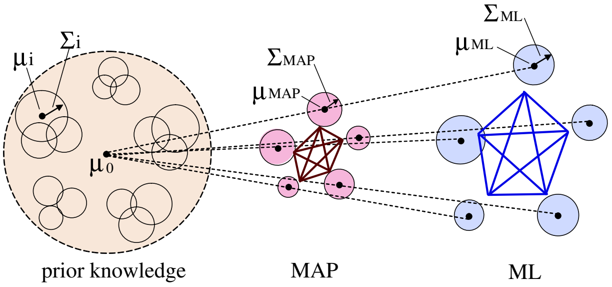


Figure 5: *MAP-based estimation of the parameters*

$$\hat{\mu}_0 = \Omega(\Omega+nE)^{-1}\mu_0 + n(\Omega+nE)^{-1}\mu_{ML} \quad (7)$$

$$\hat{B} = B + \frac{n}{2}\Sigma_{ML} + \frac{n}{2}\Omega(\text{DIAG}(\mu_{ML}-\mu_0))^2(\Omega+nE)^{-1} \quad (8)$$

$$B = E \quad (9)$$

$$\hat{A} = A + \frac{n}{2}E \quad (10)$$

$$A = \Sigma_0^{-1} \quad (11)$$

μ_{MAP} takes a value between μ_0 and μ_{ML} and approaches μ_{ML} as n increases. n is the number of frames of a new input vowel utterance and it was fixed to 14 because only the central portion (14 frames) was used from each vowel in this study.

In this study, a global affine transformation was adopted to characterize the non-linguistic features. Since this is the simplest model, its effectiveness may be limited for speech recognition. A previous work experimentally showed that upper bands (above 2.2kHz) of spectral envelopes of vowels carry speaker information[9]. Following this finding, LP filtering was introduced as preprocessing to remove the individuality effectively.

4. Recognition of 5-vowel sequences

If the utterance-level structuralization can delete dimensions of the non-linguistic features completely, the authors would like to address the following three *crazy* questions.

- Is speech recognition possible with no direct use of acoustic substances of the individual phonemes?
- Is speech recognition possible only with acoustic models (structure models) of a *single* speaker?
- Is speech recognition possible without any normalization or adaptation techniques?

An experiment was designed to answer these questions.

The recognition task is sequences of 5 isolated Japanese vowels (/a/, /i/, /u/, /e/, and /o/). Since each vowel occurs once in the utterance, a word is represented by V_1-V_2-V_3-V_4-V_5 ($V_i \neq V_j$). The vocabulary size, PP, is ${}_5P_5$=120. Isolated vowels were recorded for training and testing the proposed structure models. For testing, the recording was done with 4 male and 4 female adult speakers, each of whom spoke the 5 vowels 5 times. Three types of LPF were done with cut-off frequencies of 2kHz, 4kHz and 8kHz(full-band). Cepstrum sequences were calculated from the three types of speech waveforms. Then, a cepstrum distribution was estimated from the central portion (140msec, n=14) of each vowel based on ML or MAP. 3,125(=5^5) structures of /a/-/i/-/u/-/e/-/o/ were obtained from each speaker, 25,000(=8×3,125) structures in total. Reference [10] experimentally showed that size of the structure can be regarded as articulatory effort. Thus, all of the 25,000 structures were normalized to have the same size. From the resulting distance matrix, only the upper triangular elements were extracted

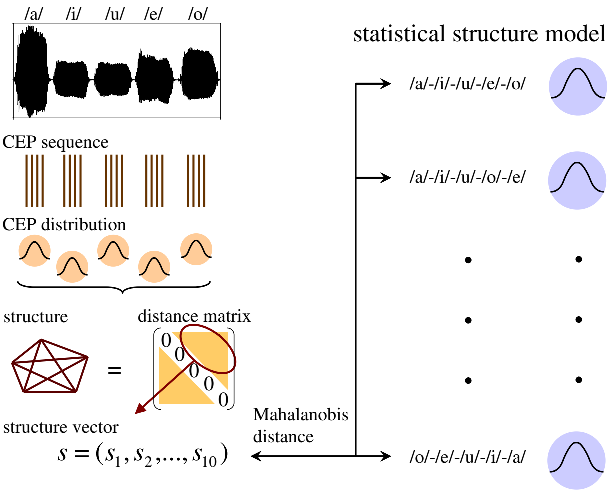


Figure 6: *5-vowel recognition using the structure.*

to define a "structure vector." Structure vectors of other vowel sequences, e.g. /i/-/a/-/u/-/e/-/o/, can be obtained by internally exchanging elements of structure vectors of /a/-/i/-/u/-/e/-/o/.

For training the proposed structure models, another *single* male speaker spoke the Japanese 5 vowels isolatedly and the recording was repeated 35 times. The same LPF was done. The 35 vowel sequences were divided into 7 groups, each of which had 5 sets of the 5 vowels. μ_0 and Σ_0, which are prior knowledge of MAP estimation, were obtained irrespective of the kind of vowel in the following manner. To estimate a distribution of an input vowel of a group, samples excluding those of the group were used and m=6×5×5. For testing, to estimate a distribution of any input, all the 7 groups of the 5 vowels were used and m=7×5×5. Each group can produce 3,125(=5^5) structure vectors of /a/-/i/-/u/-/e/-/o/ with the normalized size. Then, the statistical structure model of /a/-/i/-/u/-/e/-/o/ was calculated as a single multivariate Gaussian distribution using 21,875(=7×5^5) structure vectors. The other 119 models were obtained by exchanging elements of the /a/-/i/-/u/-/e/-/o/ model. Finally, the structure models were trained separately for three cases of LPF cut-off frequencies. Distance between an input structure and a structure model was calculated as Mahalanobis distance. The overall procedure of the experiment is shown in Figure 6.

Another 5-vowel recognition experiment was done with the conventional acoustic models. 2 kinds of speaker-independent HMMs with CMN were prepared, 4,130-speaker tied-mixture HMMs and 260-speaker tied-state HMMs. Parameters used for both HMM sets were fixed to full-band MFCCs and their derivatives. CFG allowing only the 120 words was used as language model. Table 1 shows the acoustic conditions.

Table 2 shows the recognition performance of the proposed method for three cases of training and testing conditions. Since PP is 120, the chance level is 0.8%. The recognition accuracy of ML and full-band is much better than the chance level but it is still very low. The accuracy was drastically improved by MAP and LPF. It should be noted that MAP and 2kHz cut-off LPF gave the 100% performance. The proposed method was expected to show higher robustness than the conventional HMMs. This is because, most of the cases, input speech of different acoustic conditions can be converted to LPF speech with 2kHz cut-off. Table 3 shows the comparison with the conventional HMMs. The parenthesized numbers are those of training speakers. In the proposed method, 2kHz LPF was always done as preprocessing. In the conventional methods, however, CMN was always done as environment normalization. The performance of the HMMs degraded clearly even with CMN when training and testing conditions were mismatched. On the other hand, the proposed method achieved 100% performance for every case. For more strict comparison, however, the HMMs trained with LPF speech with 2kHz cut-off should be prepared.

Table 1: *Acoustic conditions used in the experiments*

Sampling	16bit/16kHz
Window length / shift	25msec / 10msec
Parameters for structures	MCEP (α=0.55)
Estimation of distributions	ML or MAP
Parameters for HMMs	MFCC + ΔMFCC + ΔE (with CMN)
Cut off freq. of LPF	2kHz, 4kHz, or full-band

Table 2: *Recognition performance of the proposed method*

	full-band	4kHz	2kHz
ML	24.7%	47.9%	86.8%
MAP	43.0 %	62.8%	100.0%

Table 3: *Recognition performance of the three methods*

methods	full-band	4kHz	2kHz
HMM(260)	100.0%	93.8%	72.3%
HMM(4,130)	100.0%	95.2%	87.5%
Proposed(1)	100.0%	100.0%	100.0%

5. Conclusions

This paper firstly introduced a novel and structural representation of speech acoustics, and then applied the representation to speech recognition. Although the task was simple and it was only the recognition of 5-vowel sequences, the results clearly showed that the proposed structural models trained only with a single speaker outperformed the conventional HMMs trained with more than 4,000 speakers. The authors believe that the answer to the three *crazy* questions is "Definitely yes in this specific task!!" The proposed method is based on a completely different definition of the phoneme, and therefore, it can be integrated effectively with the conventional HMMs. For that, some major problems have to be solved concerning distribution estimation from continuous speech including consonant sounds.

6. References

[1] H. A. Gleason, "An introduction to descriptive linguistics," New York: Holt, Rinehart & Winston (1961)

[2] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," Proc. ICASSP, pp.889-892 (2005)

[3] N. Minematsu *et al.* "Theorem of the invariant structure and its derivation of speech Gestalt," Technical Report of IEICE, SP2005-12, pp.1-8 (2005, in Japanese)

[4] M. Pitz *et al.*, "Vocal tract normalization as linear transformation of MFCC," Proc. EUROSPEECH, pp.1445-1448 (2003)

[5] C. J. Leggetter *et al.*, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol.9, pp.171-185 (1995)

[6] T. Anastasakos *et al.*, "A compact model for speaker-adaptive training," Proc. ICSLP, vol.2, pp.1137-1140 (1996)

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

[8] C.H. Lee *et al.*, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," IEEE Trans. Signal Processing, vol.39, no.4, pp.806-814 (1991)

[9] T.Kitamura *et al.*, "Speaker individualities in speech spectral envelopes", JASJ(E), Vol.16, No.5 (1995)

[10] N. Minematsu *et al.*, "The acoustic universal structure in speech and its correlation to para-linguistic information in speech," Proc. IWMMS'2004, pp.69-79 (2004)