JAPANESE VOWEL RECOGNITION USING EXTERNAL STRUCTURE 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 is inevitably distorted by non-linguistic features such as vocal tract length, gender, age, microphone, room, line, hearing characteristics, and so on. Recently, a novel acoustic representation of speech was proposed, called the acoustic universal structure[1, 2]. It discards all the absolute properties of speech events and captures only their interrelations or contrasts to represent external structure of speech. Based on a mathematical model of the non-linguistic distortions, the new representation can remove the inevitable distortions as cepstrum smoothing can remove pitch information from speech. Recognition experiments using the external structure were conducted and it was shown that the proposed structure models trained from a *single* speaker with no normalization can outperform the conventional speaker-independent HMMs with CMN and SS. It was also found that lowpass filtering and white noise addition as preprocessing improved the performance because they can suppress the distortions rather well.

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 from speaker to speaker, microphone to microphone, room to room, line to line, and so on. Thus, the so-called speaker-independent HMMs can even have outlier speakers easily. Switching our attension to linguistics, it offers two definitions of the phonemes[3]. 1) a phoneme is a class of sounds that are phonetically similar and show certain characteristic patterns of distribution in the language or dialect under consideration 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 clear that the speaker-independent HMMs are based on the first definition. As far as the authors know, speech recognizers have never been built only based on the second definition.

Recently, a novel acoustic representation of speech was proposed based only on the second definition, called the acoustic universal structure[1]. It discards all the absolute properties of speech events and captures only their phonic differences or contrasts to extract a geometrical structure composed of the events. Based on distinctive features, some studies focused on *internal* structure of speech[4, 5, 6]. On the other hand, the acoustic universal structure focuses on *external* structure of speech, which can be interpreted linguistically as physical implementation of structural phonology and psychologically as speech Gestalt[2].

This study applies the external structure to speech recognition[7]. In order to discuss its fundamental characteristics, a very simple task was adopted; recognition of isolated vowel sequences in clean, noisy, and limited band environments.

2. THE ACOUSTIC UNIVERSAL STRUCTURE

2.1. Inevitable acoustic distortions in speech

There are three types of noises or distortions; additive, multiplicative and linear transformational. Background noise and music are typical examples of additive noise, which is not inevitable because a speaker can move to a quiet room. Then, the additive noise is not considered in this section.

Acoustic distortions caused by microphones, rooms and lines are examples of multiplicative distortion. A part of speaker individuality is also regarded as multiplicative distortion. This is because GMM-based speaker modeling represents speaker individuality by the average pattern of log spectrum. This distortion is inevitable because speech has to be produced by a certain human, transmitted by a certain media, 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; $c'=c+b$.

Vocal tract length difference and hearing characteristics difference are examples of linear transformational distortion and naturally inevitable. Both of them are often modeled as frequency warping of the log spectrum. Any monotonous frequency warping can be converted into multiplication of matrix A in cepstrum domain[8]; $c'=Ac$.

Various distortion sources are found in speech communication. The total distortion due to the *inevitable* sources is simply modeled as $c'=Ac+b$, i.e., affine transformation. The authors consider that this is the simplest mathematical model and its effectiveness may be limited. Some preprocessing will be examined in recognition experiments.

2.2. The acoustic universal structure in speech

Figure 1 shows Jakobson's phonological structure of French vowels and semi-vowels, proposed in structural phonology. He claimed that the phonological structure is invariant with speakers. The acoustic universal structure corresponds to physical and geometrical implementation of this structure in an acoustic space. Geometrically speaking, a 3-point structure, triangle, is determined fully by fixing length of the three lines. 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

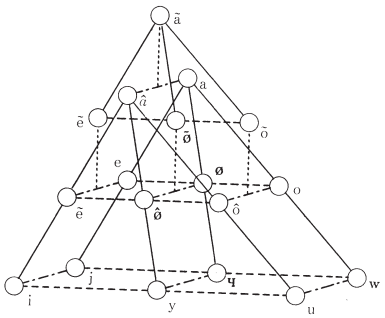

Fig. 1. Jakobson's phonological structure

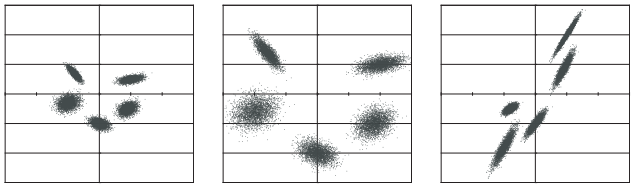

Fig. 2. Completely the same structure is found in the three.

because affine transformation always distorts a structure unless it is of a special form. How to make impossible possible? It can be done 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.

Let us explain with Bhattacharyya distance(BD). If the two distributions follow Gaussian, BD is formulated as follows.

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

μ_{12} is $\mu_1 - \mu_2$. If the common affine transformation is applied to the two distributions, BD cannot be changed. Figure 2 shows three structures of five distributions. Any two of the three structures can be converted to one another by multiplying A. Thus the three structures(matrices) are evaluated as completely the same. The structural invariance was realized because BD calculation distorts the space where the distributions are observed. Geometrical analysis of this distorted space is done in [2] based on differential geometry. In MLLR[9] and SAT[10], 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 acoustic universal structure can be applied to a *single* utterance. Figure 3 shows the structuralization process. After the utterance is converted into a sequence of distributions, the interrelations(distances) of any two of all the temporally-distant distributions are calculated to form a structure. In the distorted space, since A cannot change the matrix, any A is interpreted as rotation of the structure. Any b is interpreted as shift of the structure.

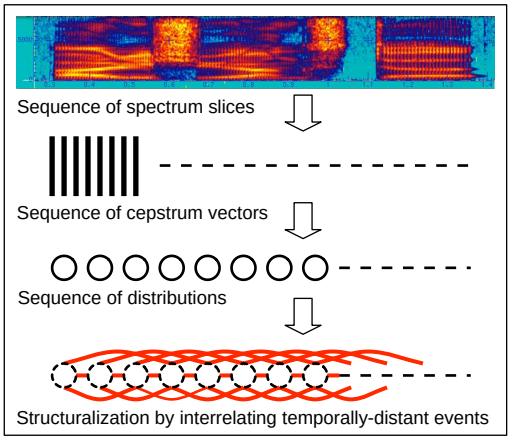

Fig. 3. Structuralization of a single utterance

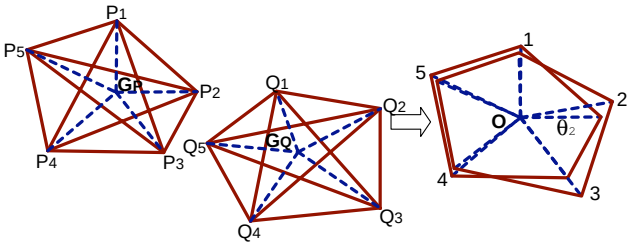

Fig. 4. Acoustic matching after shift(b) and rotation(A)

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 (see Figure 4). Suppose that there are two M-point structures in an N-dimensional *euclidean* space, where A allowing only rotation is an orthogonal matrix. The minimum of the total distance of the corresponding two events after the adaptation of b and A is formulated as follows.

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

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 $S^t T T^t S$. S and T are $(\vec{OP}_1, ..., \vec{OP}_M)$ and $(\vec{OQ}_1, ..., \vec{OQ}_M)$, respectively. It should be noted that acoustic matching score after the adaptation can be calculated without any absolute properties of the events and that without calculation of b and A. The score can be calculated only with the two distance matrices representing the two sets of speech events structurally. However, the above quantity cannot be used directly because triangular inequality is not always satisfied in the distorted space. In this study, approximate solution only with the two matrices is used and the solution will be explained below.

3. TOWARD STRUCTURE-BASED RECOGNITION

3.1. Acoustic matching between two distance matrices

In the structure-based speech recognition, acoustic matching between two structures is performed. Since the acoustic universal structure exists in a noneuclidean space, Equation (2) cannot be used directly. In [11], the following approximate equality was experimentally derived.

$$\frac{1}{M^2}\sum_{i<j}(\overline{P_iP_j} - \overline{Q_iQ_j})^2 \approx \frac{1}{M}\sum_i(\overline{P_iG_P} - \overline{Q_iG_Q})^2 \quad (3)$$

The right term approximates the average of distances between two corresponding phonemes of the two structures, 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 4) 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, a statistical model of a word can be built by using multiple utterances(structural entities) of the word.

3.2. For stabler estimation of the structures

The structure-based speech recognition presents a problem with respect to stability of estimating distributions of speech events. Parameters of a distribution have to be estimated from a single utterance, a very small number of frames. Therefore, ML(Maximum Likelihood) criterion is expected to work rather poorly. Then, 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. For prior knowledge, isolated vowel utterances were collected from a *single* speaker. As will be explained later, the proposed structure models were trained only with the single speaker. Each vowel utterance was represented as a diagonal Gaussian distribution of cepstrum parameters. The average vector and the diagonal covariance matrix for a new input utterance were estimated by referring to [12]. The following shows the procedure of the estimation, schematized in Figure 5. It should be noted that a single pair of μ_0 and Σ_0 was used commonly as prior knowledge for all the kinds of vowels.

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 $\{\mu_i\}$ $(= \frac{1}{m}\sum_{i=1}^{m}(\mathrm{DIAG}(\mu_i-\mu_0))^2)$
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, the average vector and the diagonal covariance matrix of a new input utterance can be derived as follows.

$$\mu_{MAP} = \hat{\mu}_0 \quad (4)$$
$$\Sigma_{MAP} = \hat{B}\hat{A}^{-1}, \quad (5)$$

where

$$\hat{\mu}_0 = \Omega(\Omega+nE)^{-1}\mu_0 + n(\Omega+nE)^{-1}\mu_{ML} \quad (6)$$
$$\hat{B} = B + \frac{n}{2}\Sigma_{ML} + \frac{n}{2}\Omega(\mathrm{DIAG}(\mu_{ML}-\mu_0))^2(\Omega+nE)^{-1} \quad (7)$$
$$B = E \quad (8)$$
$$\hat{A} = A + \frac{n}{2}E \quad (9)$$
$$A = \Sigma_0^{-1}. \quad (10)$$

μ_{MAP} takes a value between μ_0 and μ_{ML} and approaches μ_{ML} as n increases. Although n is originally the number of frames of an input vowel utterance, the influence of ML on MAP can be controlled by changing the value of n.

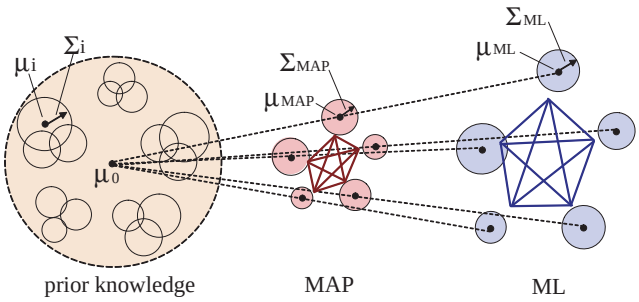

Fig. 5. MAP-based estimation of the parameters

Since a global affine transformation was adopted to characterize the inevitable non-linguistic features, its effectiveness may be limited. A previous work experimentally showed that the upper bands (above 2.2kHz) of spectrum of vowels carry a large portion of speaker identity[13]. Following this finding, two methods were examined as preprocessing; low-pass filtering and white noise addition. Although both the methods had been expected to reduce the performance, the experiments showed that they were effective.

4. RECOGNITION OF 5-VOWEL SEQUENCES

4.1. Experimental set-up

The task is recognizing Japanese vowel sequences and its length is 5; V_1-V_2-V_3-V_4-V_5. Each vowel is isolatedly produced and occurs once in the utterance ($V_i \neq V_j$). Since Japanese has five vowels (/a/, /i/, /u/, /e/ and /o/), 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 them spoke the 5 vowels 5 times. LPF was done with cut-off frequencies of 2kHz, 4kHz or 8kHz(full band). Cepstrum sequences were calculated from the LPF speech waveforms. Then, a cepstrum distribution was estimated from the central portion (140msec) 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($=8\times3{,}125$) structures in total. Reference [14] experimentally showed that size of the structure can be regarded as articulatory effort. Therefore, all of the 25,000 structures were normalized to have the same size. From each of the distance matrices, only the upper triangular elements were extracted to define a "structure vector." Structure vectors of other vowel sequences, e.g. /i/-/a/-/u/-/e/-/o/, was obtained by internally exchanging elements of structure vectors of /a/-/i/-/u/-/e/-/o/.

If the utterance-level structuralization can delete dimensions of the non-linguistic features, a *single* speaker is enough to train the proposed structure models. Thus, for training, a male speaker, different from the 8 test speakers, 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 them had 5 sets of the 5 vowels. Prior knowledge for MAP estimation was 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 (m=6×5×5). For testing, to estimate a distribution of any input, all the 7 groups of the 5 vowels were used (m=7×5×5). Each group can produce 3,125($=5^5$) structure vectors of /a/-/i/-/u/-/e/-

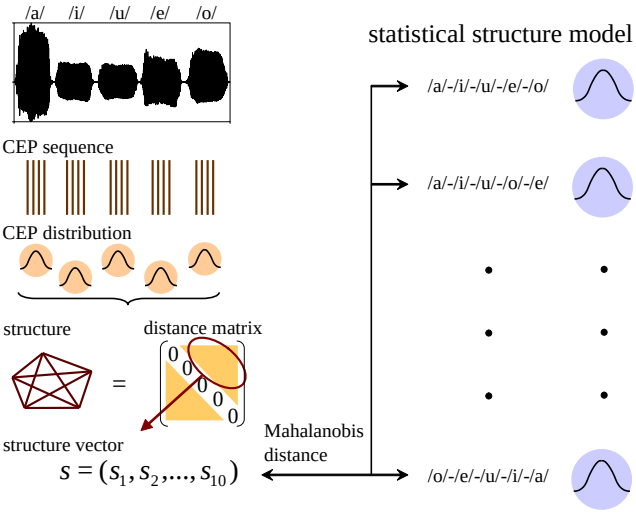

Fig. 6. 5-vowel recognition using the structure

/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($=7\times5^5$) structure vectors. The other 119 models were obtained by exchanging elements of the /a/-/i/-/u/-/e/-/o/ model. Distance between an input structure and a structure model was calculated as Mahalanobis distance. The overall procedure of the structural recognition is shown in Figure 6.

The recognition experiment with the conventional acoustic models was also done. Two kinds of speaker-independent HMMs were prepared; 4,130-speaker tied-mixture HMMs and 260-speaker tied-state HMMs. Besides, another HMM set, whose training data were the same as those of the structure models with 2kHz cut-off LPF, were prepared (the reason of preparing the LPF HMMs will be explained later). Parameters to build HMMs were MFCCs and their derivatives with CMN. The network grammar allowing only the 120 words was used as language model. Table 1 shows the acoustic conditions.

4.2. Results of the experiments

Table 2 shows the recognition performance of the proposed method. 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. The recognition performance of MAP is better with the higher value of n in the lowest cut-off frequency condition. This is considered to be because the accuracy of ML is improved in that condition. It should be noted that MAP and 2kHz cut-off LPF gave the 100% performance. This fact indicates that

- Speech recognition without any direct use of absolute acoustic properties of the individual phonemes
- Speech recognition only with acoustic models (structure models) of a *single* speaker
- Speech recognition without any explicit use of normalization or adaptation techniques

are completely possible in this specific task. In most of the cases, input speech of different acoustic conditions can be converted to LPF speech with 2kHz cut-off. Thus, the proposed method was expected to show higher robustness than the conventional HMMs,

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(n=10)	42.9%	62.7%	100.0%
MAP(n=1)	42.6%	62.1%	100.0%
MAP(n=0.1)	45.7%	60.8%	99.9%
MAP(n=0.01)	70.3%	65.4%	96.7%

Table 3. Recognition performance of the four methods

methods	full band	4kHz	2kHz
full band HMM(260)	100.0%	93.8%	72.3%
full band HMM(4,130)	100.0%	95.2%	87.5%
limited band HMM(1)	88.8%	88.8%	88.8%
Proposed(1)	100.0%	100.0%	100.0%

whose performance often degrades due to the mismatch problem. For fair comparison, another set of HMMs trained with the 2kHz LPF speech samples used for training the structural models were examined. Table 3 shows the performance of the conventional HMMs and the proposed method for different input speech conditions(full band and limited band up to 4kHz and 2kHz). The parenthesized numbers are those of training speakers. The full band HMMs and the limited band HMMs were trained with full band speech and limited band (up to 2kHz) speech with CMN, respectively. With the limited band HMMs and the proposed method, 2kHz LPF was always done as preprocessing. With the full band HMMs, CMN was always done for acoustic mismatch cancellation. Although the proposed method showed 100% performance for every condition, the performance of the full band HMMs were degraded clearly even with CMN because the training and testing conditions were mismatched. The performance of the limited band HMMs was shown to be inferior to that of the proposed method. All the recognition errors with the limited band HMMs were caused by two specific female speakers. This indicates that LPF can remove speaker individuality rather well but it is not perfect. The remaining speaker information can be removed by structuralizing speech events.

Although the adopted task was very primitive and some problems about continuous speech including consonant sounds remain to be solved, the authors consider that the experimental results show the very high potential of the proposed method.

5. RECOGNITION IN NOISE

5.1. Structural distortion caused by additive noise

The proposed method was devised focusing on only multiplicative and linear transformational distortions. What about additive noise? Suppose that $|Y(f)|^2$, power spectrum of noisy speech, is

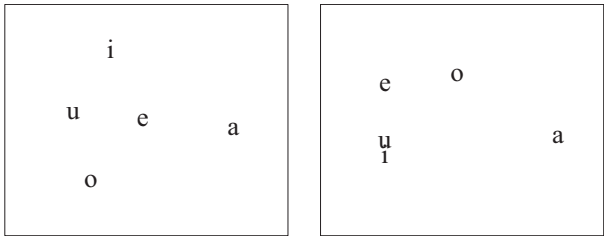

Fig. 7. Two 5-vowel structures of a male speaker (lefthand: clean speech, righthand: noisy speech)

approximated as

$$|Y(f)|^2 \approx |X(f)|^2 + |N(f)|^2, \tag{11}$$

where $|X(f)|^2$ and $|N(f)|^2$ are power spectrum of clean speech and noise, respectively. Then, log power spectrum of noisy speech ($y(f) = \log(|Y(f)|^2)$) is derived as

$$y(f) \approx \log(\ \exp(x(f)) + \exp(n(f))\ ). \tag{12}$$

It is clear that additive noise has non-linear effects on cepstrum and inevitably distorts a structure. Figure 7 is an example of such distortion, where two structures of Japanese 5 vowels are visualized with multidimensional scaling. The lefthand structure is from a 5-vowel utterance in clean environment. The righthand one is from the same utterance in white noise (SNR = 10 [dB]). It is clearly found that additive noise causes structural distortion. Especially, the distance between /i/ and /u/ becomes shorter. This is considered to be because the 1st formants of /i/ and /u/ are closer to each other and the other formants were covered by noise.

5.2. Removal of speaker individuality by additive noise

In the previous section, as preprocessing, LPF was carried out because a large portion of speaker individuality was observed in the upper band of spectrum. In Figure 8, spectral envelopes of /a/ observed for 5 different speakers are shown. In clean speech, speaker individuality is easily found in the upper band. With LPF of 2kHz cut off, the upper band of the spectrum is modified to have a rather uniform shape with speaker individuality suppressed. Due to LPF, the amplitude of the upper band spectrum is very low but low amplitude is not necessary and the uniform shape is allowed to be realized with high amplitude. The uniform shape of the upper band spectrum with high amplitude among speakers can be realized by *adding noise*. In Figure 8, spectral envelopes of /a/ of 5 speakers are shown with white noise (SNR=10[dB]). It seems that, as LPF, additive noise can suppress speaker individuality rather well and additive noise is expected to raise the performance of the proposed method as LPF did in the previous section.

In the previous section, the proposed method showed 100% performance when LPF with 2kHz cut off was done commonly in training and testing. In the following section, as in LPF, the same level of white noise was added commonly in training and testing, namely no mismatch. In this case, the performance is expected to be improved. Considering that the proposed method needs speech samples of only a single speaker for reference, an interesting discussion is possible about the mismatch problem with respect to additive noise. If a system has an extremely high-quality text-to-speech synthesizer to build reference patterns on-line and the system can detect the level of environmental noise correctly, then the

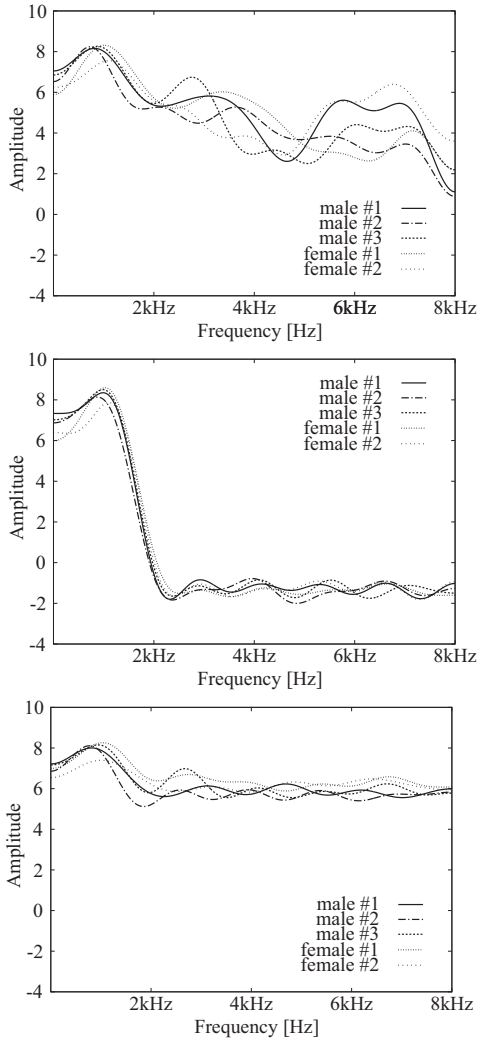

Fig. 8. Spectral envelopes of /a/ of 5 speakers (top: clean speech, middle: LPF speech, bottom: noisy speech)

system can generate the reference patterns on-line with the noise matched to the actual environment. In the case of HMMs trained with thousands of speakers, the complete re-training with speech samples with the matched noise takes a very long time. Parameter-level adaptation or modification of the HMMs is considered to work worse compared to the HMMs generated with the complete re-training. Since the proposed method uses only a single speaker to generate reference patterns, the complete re-training is possible enough if a perfect text-to-speech synthesizer exists. At least, a human listener has a perfect synthesizer if he is not handicapped, namely structure-based motor theory.

5.3. The recognition experiment in noise

White noise was added on every vowel sample of the 8 testing speakers (SNR: 0[dB], 10[dB] or 20[dB]). Two types of LPF were examined (cut-off: 2kHz or 8kHz). Both of the distortions were conducted commonly in training and testing. Table 4 shows the results. In the case of MAP, the best performance is listed out of the results with n = 10, 1, 0.1, and 0.01. As expected in the

Table 4. Recognition performance of the proposed method in noise

	full band		2kHz	
SNR	ML	MAP	ML	MAP
∞	24.7%	70.3%	86.8%	100.0%
20[dB]	73.9%	92.9%	67.9%	99.8%
10[dB]	77.4%	99.1%	68.1%	86.7%
0[dB]	73.9%	87.0%	71.1%	85.1%

Table 5. Recognition performance of the three methods in noise

SNR	HMM(260)	HMM(4,130)	Proposed(1)
∞	100.0%	100.0%	100.0%
20[dB]	100.0%	98.8%	99.8%
10[dB]	94.3%	97.2%	99.1%
0[dB]	83.0%	86.8%	87.0%

previous section, the accuracy with full band was drastically improved by adding noise. This result clearly indicates that white noise has a similar effect to LPF; making the upper band spectrum envelopes uniform to cancel speaker individuality. However, the performance with 2kHz LPF got worse in noisy environment (SNR=10[dB]). That with full band also got worse in the lowest condition (SNR=0[dB]). The authors are interested in the optimal combination of LPF and additive noise and some other techniques for speaker individuality cancellation as preprocessing.

The noisy speech samples were recognized by the conventional methods of full band HMMs with spectral subtraction(SS). Estimation of the power spectrum in noisy segments was done averaging the spectrum of the beginning portion (300ms) of each utterance. Table 5 shows the performance of the conventional methods (with SS) and the proposed method. In the case of lower SNR, the complete and on-line re-training with a single speaker is superior to the conventional methods with SS, although the re-training is currently done with natural speech.

6. CONCLUSIONS

This paper applied the acoustic universal structure to speech recognition in clean, limited band, and noisy conditions. Although the task is very primitive and it is recognition of sequences of isolated Japanese vowels, the experiments showed the high potential of the proposed method. The proposed structure models trained only with a single speaker outperformed the conventional HMMs trained with thousands of speakers with CMN and SS. Since the mathematical model of the inevitable non-linguistic distortions is very simple, preprocessing was introduced to effectively suppress the distortions; LPF and additive noise. It is very interesting that the recognition performance was drastically improved with the preprocessing. It was also found that the only preprocessing could not remove the speaker differences completely and still had outlier speakers. Since only a single speaker is used to prepare reference patterns (acoustic models) in the proposed method, if an extremely high-quality text-to-speech synthesizer is provided, the on-line generation of the patterns is possible enough with the always-matched noise and distortion. For future work, the authors are implementing an algorithm to estimate a structure from continuous speech including consonant sounds. The integration of the conventional methods and the proposed method is also interesting to the authors.

7. REFERENCES

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

[2] N. Minematsu *et al.* "Theorem of the invariant structure and its derivation of speech Gestalt," Proc. ASRU, (2005, submitted)

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

[4] A. Gutkin *et al.* "Structural representation of speech for phonetic classification," Proc. ICPR, vol.3, pp.438-441 (2004)

[5] T. Fukuda *et al.* "Orthogonalized distinctive phonetic feature extraction for noise-robust automatic speech recognition," IEICE transactions, vol.E87-D, no.5, pp.1110-1118 (2004)

[6] L. Deng *et al.* "Production models as a structural basis for automatic speech recognition," Speech Communication, vol.33, no.2-3, pp.93-111 (1997)

[7] T. Murakami *et al.* "Japanese vowel recognition based on structural representation of speech," Proc. EUROSPEECH (2005, accepted)

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

[9] 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)

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

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

[12] 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)

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

[14] 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)