Structural Representation of Pronunciation and its Application for Classifying Japanese Learners of English

N. Minematsu†*, K. Kamata*†*, S. Asakawa*†*, T. Makino*‡*, and K. Hirose*†

†The University of Tokyo, ‡Chuo University

{mine,k-kamata,asakawa,hirose}@gavo.t.u-tokyo.ac.jp, mackinaw@tamacc.chuo-u.ac.jp

Abstract

One of the most fundamental and unsolved problems in speech recognition is the mismatch problem. Speech systems trained by a specific group of speakers, e.g. adults, do not work well with another group, e.g. children. In the case of CALL, when a student receives a bad score from a system, it may be just because he is an outlier to the system. The problem is that he cannot know whether he is an outlier or not. Recently, a speaker-invariant structural and holistic representation of speech was proposed [1], where only the interrelations among speech sounds were extracted to form their external structure. Speech variation caused by speaker individuality was modeled mathematically and, based on the model, the speaker-invariance was guaranteed. This structural representation was already applied to describe the pronunciations of language learners [2]. Since the non-linguistic factors were well removed, the representation purely showed non-nativeness in the individual pronunciations. In this paper, using the new representation, language learners are automatically classified irrespective of speaker individuality. The classification is also done by an expert phonetician. High correlation is found between the two classifications.

1. Introduction

In most of all the speech systems, the spectrogram is used to represent acoustic features of speech. Since it contains not only linguistic information but also non-linguistic information, the mismatch problem more or less inevitably happens. To solve this problem, speech data have been collected from thousands of speakers to build the so-called speaker-independent models. But speaker adaptation or normalization techniques are often required. The speaker-independent models are not really speaker-independent.

A novel alternative was proposed [1]. The speech variation caused by the non-linguistic factors were modeled mathematically and, based on this model, the speech components corresponding to these factors were completely removed. Pitch information (harmonic structure) is deleted by smoothing a given spectrum slice. The proposed method removed the non-linguistic factors based on the model. No explicit adaptation or normalization was needed.

The new representation was already applied to speech recognition and in CALL. Since the speaker information can be removed from speech, speaker-independent speech recognition was implemented using only a single training speaker. The performance and the robustness of the proposed method was higher than the conventional speech recognizer trained by 4,130 speakers although the task was recognizing vowel sequences only [3]. In the CALL research, with the proposed method, non-nativeness was highlighted because speaker individuality was effectively suppressed [2]. In this paper, learner classification is examined, where the classification is expected to be done irrespective of gender and age.

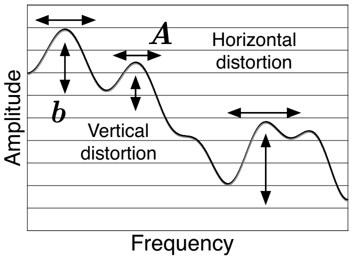

Figure 1: Spectral distortions caused by matrix A and vector b

2. Structural representation of speech

2.1. Modeling the non-linguistic speech variation

In speech recognition studies, the non-linguistic speech variation is classified into three kinds; additive, convolutional, and linear transformational distortions. Here, the first kind is ignored because it is not inevitable. Microphones and rooms are typical reasons of convolutional distortion. If a speech event is represented by cepstrum vector c, this distortion changes c into $c'=c+b$. A part of speaker individuality is also of this type. Vocal tract length difference is a typical factor to induce linear transformational distortion. This is often modeled as frequency warping of the spectrum and it changes c into $c'=Ac$ [4]. Figure 1 schematizes the spectral distortions due to matrix A and vector b, corresponding to horizontal and vertical ones, respectively. Although this model is very simple, it can change the speaker individuality of a speech sample easily.

2.2. Speaker-invariant structure embedded in speech

$c'=Ac+b$ is called affine transformation. If some acoustic features are found to be affine-invariant, they are speaker-invariant. Using these features, a speaker-invariant representation is possible. Every speech event is captured not as point but as distribution. All the event-to-event distances are calculated as Bhattacharyya distance (BD) to form a distance matrix among the events.

$$BD(p_1(c), p_2(c)) = -\ln \oint \sqrt{p_1(c)p_2(c)}dc. \qquad (1)$$

Since BD is affine-invariant, the distance matrix becomes speaker-invariant. Since a distance matrix determines a geometrical structure uniquely, the BD-based matrix defines its speaker-invariant structure. Figure 2 shows Jakobson's geometrical structure of the French vowels. He claimed that this structure should be observed irrespective of speakers. We consider that the proposed method implements structural phonology physically. The invariance indicates geometrically that multiplication of matrix A and addition of vector b mean rotation and shift of a structure, respectively. It should be noted that we showed recently that the transformation invariance of BD is satisfied also in the case of non-linear transformations [5]. Then, we call it robust and structural invariance.

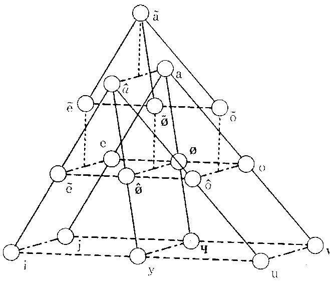

Figure 2: Jakobson's geometrical structure [6]

3. Development of the vowel structure

The new representation was applied to trace the development in vowel learning [2]. Various non-native pronunciations of the vowels were simulated by a Japanese speaker who can speak American English (AE) well. Each of the 11 AE vowels was recorded once as /bVt/ and each of the 5 Japanese vowels five times as /bVto/. Using the vowel segments, various vowel structures were constructed as matrices. For example, the completely Japanized English structure could be obtained by substituting Japanese /a/ sounds for /ʌ,æ,ɑ,ə,ɚ/ and the other Japanese vowels for the other corresponding AE vowels. Partly-American and partly-Japanese structures could be constructed by changing the substitution pattern. Figure 3 shows the completely Japanized structure, a partly-American and partly-Japanese one, and the AE one. Ward's clustering method was used here to convert a matrix into a tree diagram. The second tree diagram was obtained from the first one by correcting /ʌ,æ,ɑ,ə,ɚ/.

4. Classification of the learners

A learner, represented as a full set of vowel distances, was visualized as a tree diagram. If distance measure between two vowel matrices, i.e. two learners, is adequately derived, then, we can obtain a full set of learner-to-learner distances. This means that the learners can be classified purely based on their vowel structures, without any respect to age, gender, speaker, microphone, etc.

4.1. Speech material used in the experiment

Six male and six female high school or university students who were returnees from US joined the recording. The 11 AE vowels and the 5 Japanese vowels were recorded once as /bVt/ and five times as /bVto/, respectively. This was because five different American vowels, at most, were replaced by a Japanese vowel.

Considering the well-known Japanese habits of producing AE vowels, the substitution table was prepared, shown as Table 1. Using this table, 8 patterns of the vowel substitution were obtained, which is listed as Table 2. P1 and P8 correspond to the completely Japanized English and the good American English pronunciations, respectively. P2 to P7 are half-Japanese and half-American pronunciations. Now we have 8 different vowel structures per speaker and 96 vowel structures altogether. The aim of the experiment is to examine whether the 96 structures can be classified purely based on the vowel structures, not based on gender, age, or speaker.

4.2. Matrix-to-matrix distance measure

Suppose that two geometrical vowel structures, S and T, are given as two distance matrices. Then, structure-to-structure distance is obtained after shifting ($+b$) and rotating ($\times A$) one of the structures so that the two can be overlapped the best, shown in Figure 4. This operation naturally means implicit speaker adaptation. The distance is calculated as the minimum of the total distance between the corresponding two points after shift and rotation. In [1], it was experimentally shown that the minimum distance, D_1, could be approximately calculated as euclidean distance between the two matrices, where the upper-triangle elements form a vector;

$$D_1(S,T) = \sqrt{\frac{1}{M}\sum_{i<j}(S_{ij}-T_{ij})^2}, \qquad (2)$$

where S_{ij} is (i,j) element of matrix S and M is the number of the vowels. D_1 can be regarded as summation of differences of vowel contrasts between the two. For example, distance between /ʌ/ and

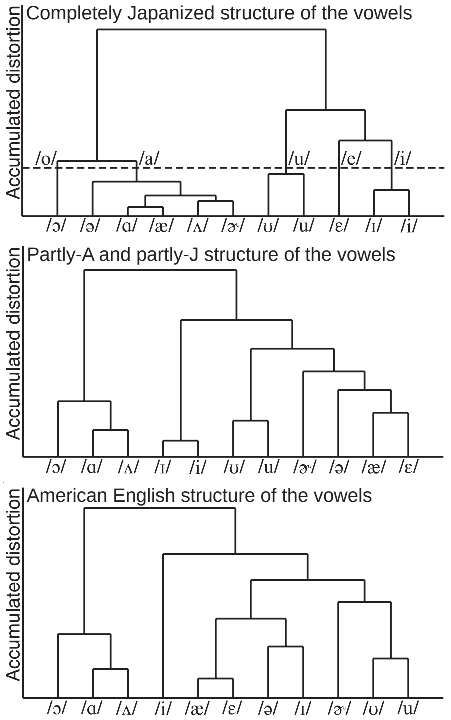

Figure 3: Japanized structure to American structure

Table 1: Vowel substitution table

Japanese vowels	↔	English vowels
a		ɑ, ʌ, æ, ɚ, ə
i		i, ɪ
u		u, ʊ
e		ɛ
o		ɔ

Table 2: 8 patterns of the vowel substitution

	ɑ	æ	ʌ	ə	ɚ	ɪ	i	ʊ	u	ɛ	ɔ
P1	J	J	J	J	J	J	J	J	J	J	J
P2	A	A	A	A	A	J	J	J	J	J	J
P3	J	J	J	J	J	A	A	A	A	A	A
P4	A	A	J	J	J	A	A	J	J	A	A
P5	J	J	A	A	A	J	J	A	A	J	J
P6	A	J	A	J	A	J	J	J	J	A	A
P7	J	A	J	A	J	A	A	A	A	J	J
P8	A	A	A	A	A	A	A	A	A	A	A

A : American English pronunciations are used.

J : Japanese vowels are substituted.

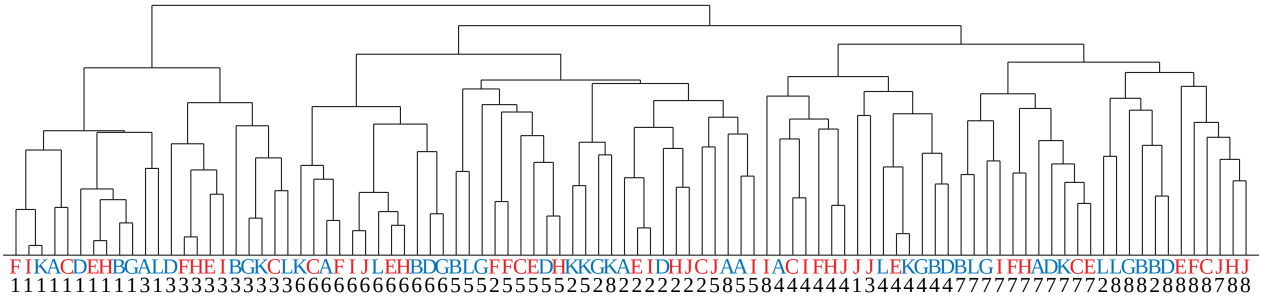

Figure 5: Classification of the 96 vowel structures based on the *contrast-based* comparison (D_1)

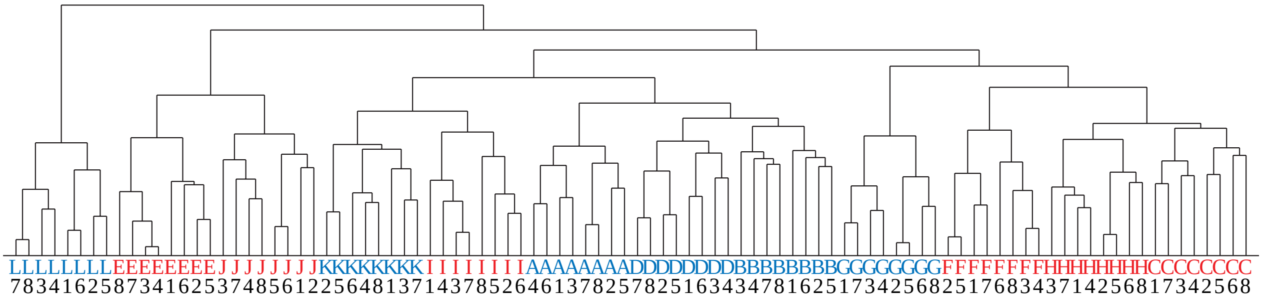

Figure 6: Classification of the 96 vowel structures based on the *substance-based* comparison (D_2)

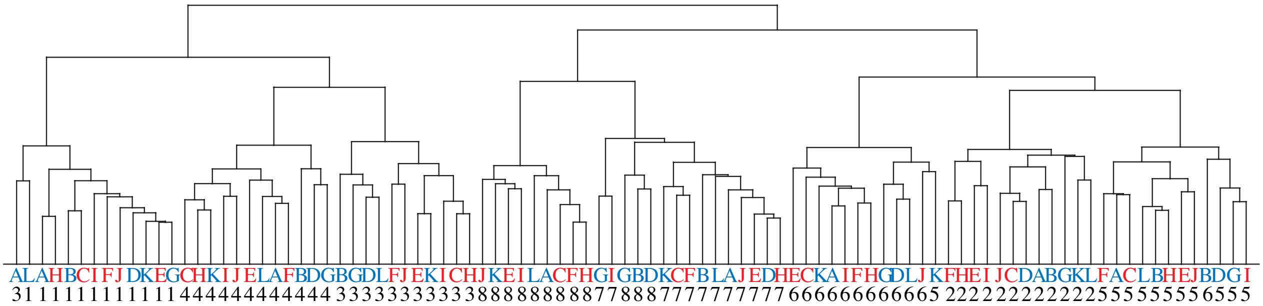

Figure 7: Classification of the 96 vowel charts plotted by an expert phonetician

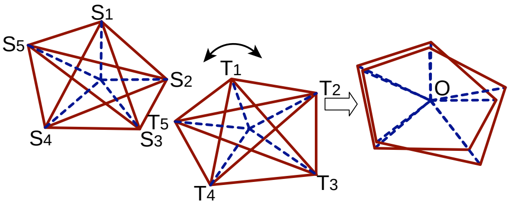

Figure 4: Distance calculation after shift and rotation

/ɛ/ is compared between the two structures and the difference of the two distances is summed. In the conventional framework such as DTW and HMM, vowel substance /ʌ/ of a structure and that of another was directly compared acoustically. In this framework, distance between two vowel structures, D_2, is formulated as

$$D_2(S,T) = \sqrt{\tfrac{1}{M}\sum_i BD(v_i^S, v_i^T)}. \qquad (3)$$

v_i^S is vowel i of S. Table 3 shows the acoustic conditions. Each vowel is modeled as diagonal Gaussian distribution and the parameters are estimated using MAP (Maximum A Posteriori) criterion.

4.3. Results and discussions

Figures 5 and 6 show the results of classifying the 96 vowel structures in two different ways. Numbers and alphabets at the leaf nodes represent the vowel patterns (1 to 8) and the speakers (A to L), respectively. The two colors indicate the two genders. If vowel contrasts are compared in Figure 5, good pronunciation classification is done. On the other hand, if vowel substances are compared directly in Figure 6, which is often done in DTW, it leads to the complete speaker classification. It should be noted that the two tree diagrams of Figures 5 and 6 were obtained from the same data set and that the structural difference between the two trees is solely caused by difference in defining the two distance measures D_1 and D_2. Most of the speech applications were built based on the substance-based comparison of sounds. We consider that this is one of the reasons why CALL is sometimes criticized not to be pedagogically-sound enough. These criticisms are reviewed in [7].

Table 3: Acoustic conditions of the analysis

sampling	16bit / 16kHz
window	25 ms length and 1 ms shift
parameters	FFT cepstrum (1∼10)
HMMs	1-mixture monophones with diagonal matrices
topology	3 states and 1 distribution per HMM (GM)

In Figure 5, some different vowel patterns are found to belong to a single subtree, e.g. P2, P5, and P8. This is considered due to differences of the language background among the 12 speakers. Although they are returnees from US, length and place of their stay in US are different from each other. The vowel structure strongly depends on the speaker's regional accent [8]. If returnees with the same language background both for the two languages can be used, a more coherent classification tree will be obtained.

5. Comparison with a phonetician's tree

Figures 5 and 6 clearly show that the proposed representation is remarkably valid to capture the non-nativeness in the individual pronunciations irrespective of the non-linguistic factors. However, it is still difficult to claim that every step of binary and gradual separation of the learners in Figure 5 is adequate enough. To investigate the adequacy of the classification, a similar tree diagram is generated from manual plotting by an expert phonetician. The manual tree is compared with the automatic tree.

5.1. Drawing vowel charts through listening

Figure 5 was generated with the 96 distance matrices and the matrices were automatically calculated by the structural speech analysis. This means that, if the distance matrices are obtained by a phonetician's listening to all the vowel samples, his own tree diagram can be generated through the same procedures used in Section 4.

The 96 sets of the 11 vowels were presented to the phonetician through headphones. He was asked to draw 96 vowel charts. To facilitate this task, a vowel chart drawing software was developed and, by clicking a mouse, the position of each vowel was specified. In phonetics, a two-dimensional trapezoidal vowel chart is usually used to show the structural relations among the vowels, where only the tongue position is focused on. In this work, however, a four-dimensional chart was adopted. The first two dimensions were used to specify the tongue position. The third one was for lip-rounding and the last one for rhoticity of /ɚ/. In Figure 8, the four-dimensional framework adopted in the software is shown, where the last dimension is separately added to the other three ones. Numbers on the segments are relative distances between two nodes. In the experiments, a two-dimensional trapezoidal framework was presented on a PC monitor to specify the tongue position and values of the other two dimensions were separately asked.

5.2. Results and discussions

Figure 7 shows a tree diagram generated from the 96 vowel charts plotted by the phonetician. As in Figure 5, it is clearly shown that classification of the pronunciations, not the speakers, is done effectively. In comparison between the two trees, although the relative location of P4 is different, we can say that the two trees are very similar structurally. In Figure 5, the 8 vowel patterns are firstly divided into [1,3]+[6,5,2,4,7,8] and then into [1,3]+[6,5,2]+[4,7,8]. In Figure 7 they are clustered into [1,4,3]+[8,7,6,2,5] and then into [1,4,3]+[8,7]+[6,2,5]. Another analysis was done to investigate the structural similarity quantitatively, which was correlation analysis between the two sets of ninety-six 11×11 vowel matrices. Figure 9 shows the correlation between the two sets of 5,280 ($96 \times {}_{11}C_2$) vowel distances. High correlation of 0.72 is obtained and this result shows good validity of the proposed method. However, the above results were obtained from only a single phonetician and the comparison with others should be done in the near future.

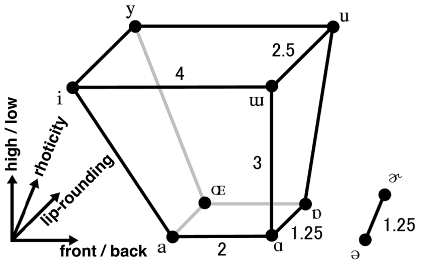

Figure 8: Four-dimensional vowel chart

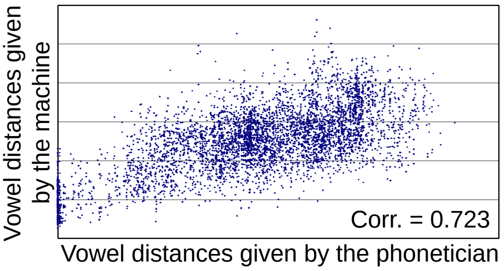

Figure 9: Correlation between the 2 sets of vowel distances

6. Conclusions

The speaker-invariant, structural and holistic representation of speech was effectively applied to classify language learners independently of the inevitable non-linguistic variations in speech. The representation only captured the phonic contrasts and discarded the speech substances to remove these variations. Results showed that the contrast-based comparison effectively classified the pronunciations not the speakers and it was very similar structurally to the manual classification done by an expert phonetician. On the other hand, the substance-based comparison showed the complete speaker classification. In the proposed method, absolute properties such as formant frequencies were completely discarded. Interestingly enough, it means that only the contrasts of speech sounds may be sufficient enough to be used for assessing learners' pronunciations. In [9], without absolute acoustic properties, it was correctly estimated for each learner which vowel should be corrected by priority. Interested readers should refer to that work.

7. References

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

[2] S. Asakawa *et al.*, "Structural representation of the non-native pronunciations," Proc. INTERSPEECH, pp.165–168 (2005)

[3] N. Minematsu *et al.*, "Theorem of the invariant structure and its derivation of speech Gestalt," Proc. Int. Workshop on Speech Recognition and Intrinsic Variations, pp.47–52 (2006)

[4] M. Pitz *et al.*, "Vocal tract normalization equals linear transformation in Cepstral space," IEEE Trans. Speech and Audio Processing, vol. 13, pp.930–944 (2005)

[5] N. Minematsu *et al.*, "Linear and non-linear transformation invariant representation of information and its use for acoustic modeling of speech," Proc. Spring Meeting Acoust. Soc. Jpn., pp.147–148 (2007)

[6] R. Jakobson *et al.*, Notes on the French phonemic pattern, Hunter, N.Y. (1949)

[7] A. Neri *et al.*, "Automatic speech recognition for second language learning: how and why it actually works", Proc. ICPhS, pp.1157–1160 (2003)

[8] M. Huckvale, "ACCDIST: a metric for comparing speakers' accents," Proc. ICSLP, pp.29–32 (2004)

[9] N. Minematsu *et al.*, "Structural representation of the pronunciation and its use for CALL," Proc. Int. Workshop on Spoken Language Technology, pp.126–129 (2006)