STRUCTURAL REPRESENTATION OF THE PRONUNCIATION AND ITS USE FOR CALL

Nobuaki MINEMATSU†*, Satoshi ASAKAWA*†*, Keikichi HIROSE*‡

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

ABSTRACT

This paper applies the structural representation of the pronunciation for Computer-Aided Language Learning (CALL). This representation was proposed to remove non-linguistic features such as age, gender, speaker, etc from speech acoustics [1]. The removal was performed by extracting only the interrelations of speech events and discarding their absolute properties such as formants and spectrum envelopes. All the extracted interrelations mathematically form the external phonological structure of the events. Using this representation, in [2], the vowel structure of a language learner was extracted and it was shown that the structural development via training can be traced and visualized adequately. This structural visualization can be regarded as pronunciation portfolio [3]. This paper shows that the new representation can classify the language learners adequately and indicate which vowels should be corrected by priority.

Index Terms— AUS, CALL, portfolio, classification of learners

1. INTRODUCTION

Using the advanced speech technologies, many CALL systems have been developed and used in actual classrooms [4]. However, it is true that the technologies did not solve one of the most fundamental problems yet; the so-called "mismatch problem". If voice quality of a learner is different from that of the training speakers used to develop the system, the learner will inevitably receive a bad score from the system because his/her voice is an outlier to the system. This situation can happen rather easily and some papers indicate that CALL systems are not pedagogically sound enough [5]. Speaker adaptation techniques may avoid this situation but excessive adaptation of the acoustic models to the learner inevitably misjudge his/her pronunciation. This is because the adapted models tend to give a better score to bad pronunciation as they are adapted to the learner.

The most fundamental source of the mismatch problem lies in the speech representation used in every speech application, i.e., spectrogram. From it, we can guess not only phones but also age, gender, speaker, microphone, etc. Strictly speaking, the spectrogram is a very noisy representation. The current speech technologies have tried to extract the linguistic information from the spectrogram by a naive method, i.e., collecting data from thousands of speakers;

$$g(\text{linguistic}) = \sum_{\text{non-linguistic}} f(\text{linguistic}, \text{non-linguistic}). \quad (1)$$

In our previous study [1], using a mathematical model of acoustic variations caused by the non-linguistic factors, acoustically invariant properties were found and used to propose a new speech representation, called the acoustic universal structure (AUS). This representation is based on some findings of neurosciences claiming that the linguistic and non-linguistic features of speech can be separated in brains and that the former should be modeled as speech motions [6].

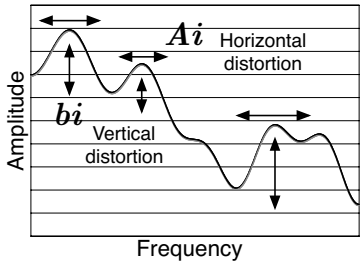

Fig. 1. Spectral distortions caused by A_i and b_i

2. THE ACOUSTIC UNIVERSAL STRUCTURE

2.1. Acoustic modeling of the non-linguistic speech variations

The non-linguistic speech variations are classified into three kinds; additive, convolutional, and linear transformational distortions. The additive distortion is ignored here because it is not inevitable.

Acoustic variations caused by microphones and rooms are typical examples of the convolutional distortion. If a speech event is represented by cepstrum vector c, this distortion is represented as addition of another vector b; c'=c+b. Vocal tract length difference and hearing characteristics difference are typical examples of the linear transformational distortion. They are often modeled as frequency warping of the log spectrum. Any monotonous frequency warping of the log spectrum is approximated well as multiplication of matrix A; c'=Ac [7]. Various distortion sources are found in speech communication but the total distortion due to the *inevitable* sources, A_i and b_i, is eventually modeled as c'=Ac+b, known as affine transformation. Figure 1 schematizes the spectral distortions due to A_i and b_i, which correspond to horizontal and vertical ones, respectively.

2.2. The acoustic universal structure in speech

If some acoustic properties are invariant with any kind of affine transformation, they will provide a speaker-invariant representation of speech. Every speech event is captured not as point but as distribution such as Gaussian mixture. Every event-to-event distance is calculated as Bhattacharyya distance, which is regarded as logarithm of normalized cross correlation between two PDFs;

$$BD(p_1(\boldsymbol{x}), p_2(\boldsymbol{x})) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(\boldsymbol{x}) p_2(\boldsymbol{x})}\, d\boldsymbol{x}. \quad (2)$$

It is known that BD is affine-invariant and this directly means that a full set of distances, which are calculated from all the distributions, is invariant with any kind of a single affine transformation. If we have N events, then, the distances are compactly represented as $N \times N$ distance matrix and the matrix is invariant with the non-linguistic factors. The distance matrix is mathematically corresponds to a geometrical structure as the shape of a triangle is determined uniquely

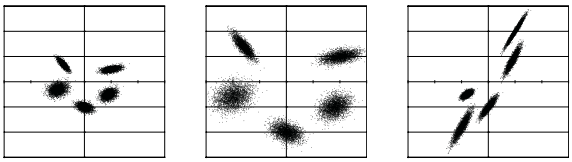

Fig. 2. The invariant underlying structure of a data set

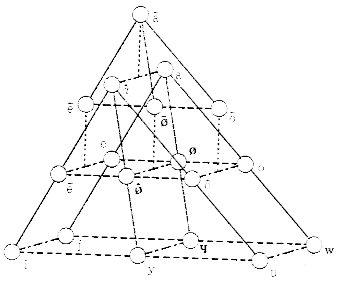

Fig. 3. Jakobson's geometrical structure of the French vowels

by length of all the lines. This invariance indicates that multiplication of matrix A and addition of vector b mean rotation and shift of a geometrical structure, respectively. Figure 2 shows three structures of five distributions. Any two of the three can be converted to one another by multiplying matrix A, meaning that the distance matrices are identical in the three. Some readers wonder whether the three structures are different. The differences can be seen when the structures are observed in an euclidean space. In other words, BD calculation warps the space so that the structures can be invariant. This invariant structure can be interpreted as physical realization of Jakobson's geometrical structure of phonemes (see Figure 3) [8] and it requires a noneuclidean geometry for its realization.

3. DEVELOPMENT OF THE VOWEL STRUCTURE

The structural representation was applied to trace the vowel learning [2]. In the experiment, various pronunciations of the vowels were simulated using a Japanese speaker who can speak American English (AE) very well. Each of the 11 AE vowels was recorded only once as /b V t/ and each of the 5 Japanese ones was done five times as /b V t o/. Using the vowel segments of these data, various vowel structures were generated. For example, the completely Japanized English structure can be obtained by substituting Japanese /a/ sounds for /ʌ, æ, ɑ, ə, ɚ/ and the other Japanese vowels adequately for the other AE ones. Partly-American and partly-Japanese English vowel structures can be generated by changing the substitution pattern. Figure 4 shows the completely Japanized structure, a partly-American and partly-Japanese structure, and the AE structure. Here, Ward's method was used for hierarchical clustering. The second tree diagram was obtained from the first one by correcting /ʌ, æ, ɑ, ə, ɚ/.

4. CLASSIFICATION OF LANGUAGE LEARNERS

A learner was visualized as tree diagram, which is generated by a full set of distances between any two of the vowels. If distance measure between two vowel matrices, i.e, two learners, is adequately derived, then, we can calculate a full set of distances between any two of the learners. This means that the learners can be classified purely based on their vowel structures, without any respect to age, gender, speaker, microphone, etc. This section investigates whether the learner classification will work well using various vowel structures simulated by using twelve Japanese returnees from US.

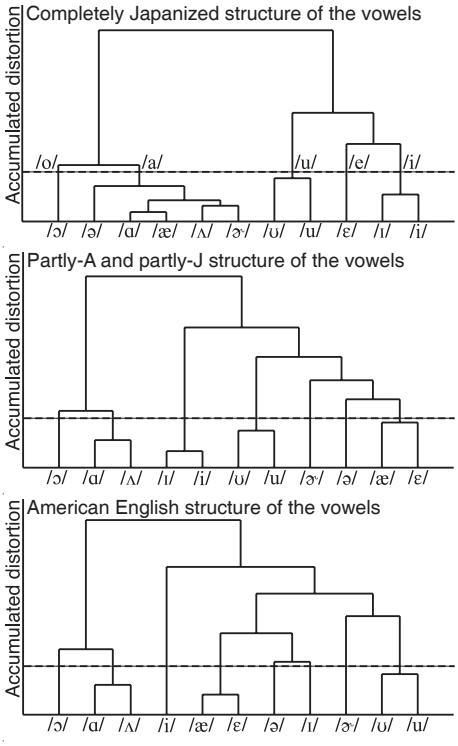

Fig. 4. From the Japanized structure to the American structure

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.

4.1. Speech material used in the experiment

Six male and six female high school or university students, who are returnees from US, joined the recording. The 11 AE vowels and the 5 Japanese vowels were recorded once as /b V t/ and five times as /b V t o/, respectively. This is because five different American vowels, at most, will be replaced by a Japanese vowel.

Considering well-konwn Japanese habits of producing AE vowels, the substitution table is prepared, shown as Table 1. Using this table, 8 patterns of the vowel substitution were obtained and listed in 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 struc-

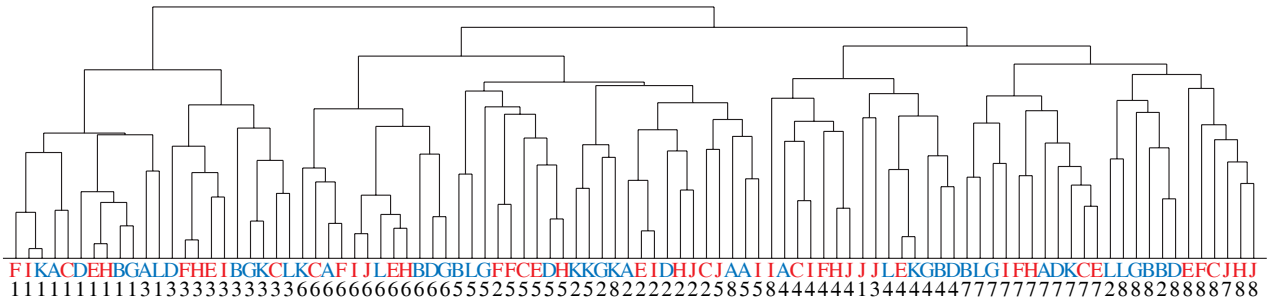

Fig. 5. Classification of the 96 vowel structures based on the *contrast-based* comparison (D_1)

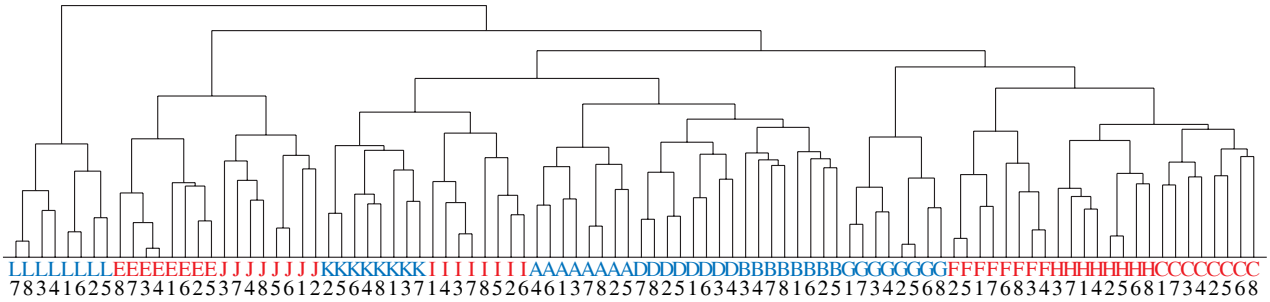

Fig. 6. Classification of the 96 vowel structures based on the *substance-based* comparison (D_2)

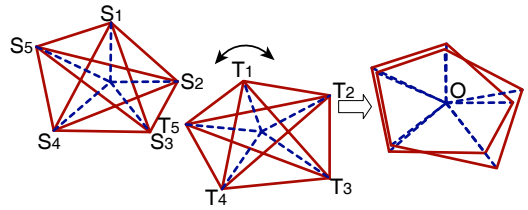

Fig. 7. Distance calculation after shift and rotation

tures all together. 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 nor 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 ($\times A$) a structure so that the two can be overlapped the best, shown in Figure 7. The distance is calculated as the minimum of the total distance between the corresponding two points after the shift and the rotation. In [1], it was experimentally shown that the minimum distance, D_1, can be approximately calculated as euclidean distance between the two distance matrices, where the upper-triangle elements form a vector;

$$D_1(S,T) = \sqrt{\frac{1}{M}\sum_{i<j}(S_{ij} - T_{ij})^2}, \quad (3)$$

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 /ɛ/ is compared between the two structures. In the conventional acoustic matching framework such as DTW and HMM, however, 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 follows.

$$D_2(S,T) = \sqrt{\frac{1}{M}\sum_{i} BD(v_i^S, v_i^T)}, \quad (4)$$

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)

where v_i^S is vowel i of structure S. Table 3 shows the acoustic conditions. Each vowel is modeled as diagonal Gaussian distribution and, since it has to be estimated from a single sample, the parameter estimation was done 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. If vowel contrasts are compared in Figure 5, rather good pronunciation classification is done. On the other hand, if vowel substances are compared directly, which is often done in DTW, Figure 6 shows that the comparison leads to complete speaker classification. It should be noted that the two tree diagrams were obtained from the same data set and that the structural difference between the two trees is purely caused by difference in distance measures D_1 and D_2. Many speech applications were built on substance-based comparison of sounds. We consider that this is why CALL softwares are sometimes criticized not to be pedagogically sound enough [5].

In Figure 5, under some subtrees, different vowel structure 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. It is true that the vowel structure strongly depends on the speaker's regional accent [9]. If returnees with the same language background both for Japanese and American English can be used, we consider that a more coherent classification tree can be obtained.

5. WHICH VOWELS TO CORRECT AT FIRST?

In Section 3, the structural representation was shown to effectively trace the structural development of the vowels. Looking at the vowel structure only, however, it may be difficult to determine which vowels to correct at first. In this section, an algorithm is devised to estimate the order of vowel correction only with the distance matrices of the learner and the teacher. It should be noted that the proposed algorithm does not refer to any absolute properties of the vowels directly such as formants or spectrums but use only the vowel contrasts.

5.1. The vowel generating the largest structural distortion

Matrix-to-matrix distance was derived as D_1 in Equation 3. D_1 indicates the total distortion between the two structures and it can be decomposed into components of the individual vowels. The *local* structural distortion caused by vowel v, $d(v)$, is defined simply as

$$d(v) = \sum_{i=1}^{M} |S_{vi} - T_{vi}|. \quad (5)$$

If S and T correspond to a learner matrix and a teacher's one, the vowel giving the largest $d(v)$ should be corrected at first.

5.2. Estimation of the order of vowel correction

The same speech samples as those of Section 4.1 were used. The 96 vowel structures were divided into 8 patterns (P1 to P8) and 12 structures (A to L) of each pattern were averaged to define the averaged vowel structure for each pattern. P8 is regarded as distance matrix of a teacher and we have 7 learners, one of which has the complete Japanese accent (P1) and the others have partly Japanese accented pronunciations. Using Equation 5, the order of vowel correction is estimated for each learner. It should be examined whether the replaced AE vowels (see Table 2) are ranked as higher.

5.3. Results and discussions

The estimated orders for P1 to P6 are shown in Figure 8. In the figure, bars represent $d(v)$ and gray bars mean that of the replaced vowels. The order for P1 is that for learners with the completely Japanized pronunciation. Considering $d(v)$ for the individual vowels, the 11 vowels can be classified into 3; high, middle, and low priority of correction. /ɚ, æ, ə/ should be corrected by the highest priority. /ʌ, ɪ, ɑ/ are in the second group and /ɛ, i, ɔ, ʊ, u/ are in the last group, showing the lowest priority. In some textbooks of American English pronunciation for Japanese beginners [10], it is often said that /ɛ, i, ɔ, ʊ, u/ can be replaced with Japanese vowels of /e, i:, o, u, u:/. This means that the result for P1 are very accordant with what Japanese phonetics and American English phonetics tell.

For P2 to P7, it is easily found that the replaced vowels tend to have higher priority for correction. Although some of the replaced vowels are ranked lower than some of the others in P2, P4, and P6, these vowels are /ʊ, u, i/, which are known to be especially closer to Japanese vowels of /u, u:, i:/. Considering these facts, the estimated vowel correction orders can be said to be remarkably valid. Although each pattern was obtained by averaging 12 structures, the algorithm can be used when a learner's structure is compared with a specific teacher's one. If speech samples can be obtained from a movie star of the target language, the system may show some hints to move the learner's mouth closer to that of the admired star. In Section 4.3, it was shown that the conventional substance-based comparison may tend to focus not on vowel identity but on speaker identity. In this sense, if the movie star is female and the learner is male, the substance-based CALL systems may claim "You have to mimic the teacher precisely. You have to produce female voice."

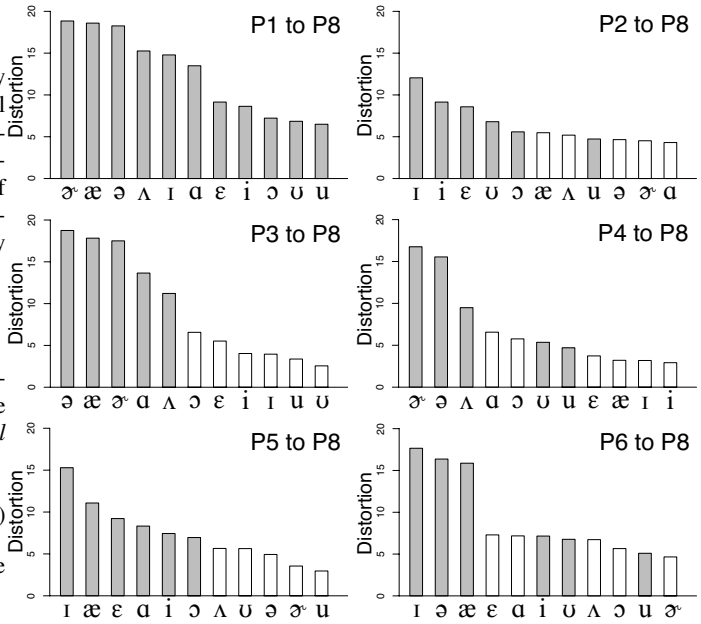

Fig. 8. The order of the vowel correction estimated for P2 to P6

6. CONCLUSIONS

This paper carried out two new experiments using the structural representation of speech, where absolute properties of speech are discarded and only the phonic contrasts are used directly. The first experiment classified language learners without any respect to non-linguistic features of speech and the second one estimated which vowels to correct by priority. The results showed the very high validity of the proposed representation for both the tasks. As the experiments were based on simulation, we're currently collecting speech samples from real learners and have completed the collection from more than two hundred learners so far. Some new results using the real learners' data will be presented in the near future.

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. EUROSPEECH, pp.165–168 (2005)

[3] N. Minematsu *et al.*, "Pronunciation portfolio; how were, are, and will be you?" Proc. Int. Workshop on Language e-Learning (IWLeL), pp.87–95 (2004)

[4] T. Kawahara *et al.*, "Practical use of English pronunciation system for Japanese students," Proc. ICSLP, pp.1689–1692 (2004)

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

[6] P. Belin *et al.*, "'What', 'where' and 'how' in auditory cortex," *Nature neuroscience*, vol.3, no.10, pp.965–966 (2000)

[7] 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)

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

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

[10] H. Saito *et al.*, "English pronunciation lessons for adults," NHK Pub., Tokyo (2003, in Japanese)