Structural Assessment of Language Learners' Pronunciation

N. Minematsu†, K. Kamata†, S. Asakawa†, T. Makino‡, T. Nishimura†, 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, nt-tazuko@ams.odn.ne.jp

Abstract

Speaker-invariant structural representation of speech was proposed [1], where only the phonic contrasts between speech sounds were extracted to form their external structure. The acoustic substances were completely discarded. Considering a mapping function between speaker A's acoustic space and B's space, the speech dynamics was mathematically proven to be invariant between the two irrespective of the form of the function [2]. This structural and dynamic representation was applied to describe the pronunciation of learners [3]. Since the non-linguistic factors were removed effectively, the representation could highlighted the non-nativeness in the individual pronunciations. For vowel learning, it was automatically estimated for each of the learners which vowels to correct by priority [4]. Unlike the conventional approach, the estimation was done without the direct use of sound substances such as spectrums. In this paper, using the vowel charts of the learners plotted by an expert phonetician, the validity of this contrastive or relative approach is examined by comparing it with the conventional absolute approach. Results show the high validity of the proposed method.

Index Terms: phonic contrasts, pronunciation structure, CALL

1. Background and objective

One of the most fundamental and unsolved problems in speech technology is the mismatch problem. Speech systems trained by a specific group of speakers do not work well with other ones. This is because the widely-used speech representation, the spectrogram, carries not only linguistic information but also non-linguistic information. The spectrogram shows everything and it is very noisy. In most of the efforts made by speech engineering, however, the mismatch problem has been resolved through collecting speech data from thousands of speakers. But speaker adaptation or normalization techniques are still required.

In developmental psychology, infants are said to acquire language through imitating their parents' speech. But no children try to produce their parents' voices. They are not imitating the substances of the input speech. As their phonemic awareness is very immature, they cannot decode an input speech into a sequence of phonemes, and therefore they cannot speak by converting the individual phonemes into sounds. In this situation, what is imitated acoustically? Developmental psychology claims that they firstly acquire the holistic sound pattern of a word, i.e., word Gestalt. After that, they learn the segmental sound categories, i.e., phonemes [5]. The word Gestalt is considered as the skeleton of a spoken word and it has to be speaker-invariant because children don't control their speaker identity in their voices, whoever talks to them.

It is interesting that myna birds imitate the sounds themselves. Hearing a very good myna bird say something, one can guess its keeper easily. Hearing a very good child say something, however, it is impossible to guess its keeper. No human acquire language through imitating the substances of input sounds. However, most of the CALL systems directly compare an input utterance with the acoustic models trained with many native speakers. This fact simply claims that the systems assume that a learner is a myna bird to the averaged distributions over the native speakers. Is this assumption correct?

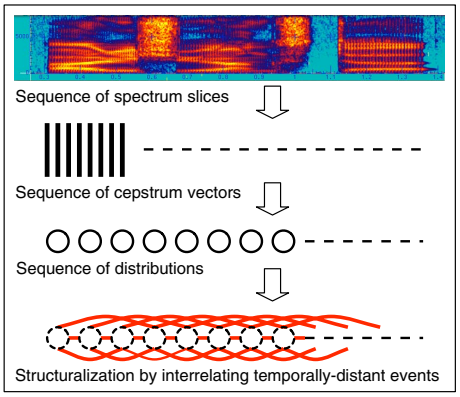

Figure 1: BD-based speaker-invariant structure of speech

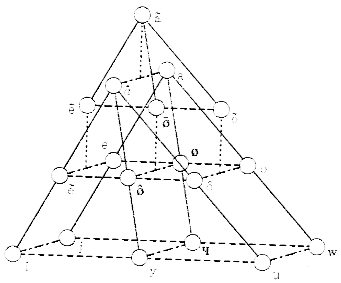

Figure 2: Jakobson's geometrical structure [6]

A novel and speaker-invariant representation of speech was proposed [1], where only the phonic contrasts were extracted including the contrasts between long-distant sounds. As shown in Figure 1, an input speech stream was converted into a sequence of distributions and all the distances between any two of them were extracted as Bhattacharyya distance (BD). Since BD is invariant with any form of one-to-one mapping function (transformation-invariant) [2], a full set of BDs are invariant between two speakers if they utter linguistically the same content. As the shape of a triangle is uniquely determined by the length of the three segments, an $N \times N$ distance matrix determines an N point structure uniquely. Then, a full set of BDs are speaker-invariant and can represent a speech structure. This is the physical implementation of structural phonology and Figure 2 shows Jakobson's geometrical structure of the French vowels, i.e., his skeleton of the vowels [6]. Since the structure contains only the phonic differences in an utterance, it is interpreted to characterize the speech dynamics and to discard every speech substance. Therefore, only with this structural and dynamic representation,

it is impossible to identify separate sounds as phonemes. On the contrary, once a continuous utterance is given, the correct word recognition is possible enough by machines only with this structural and holistic representation. In [7], the vowel-based recognition performance was 98.6%. This result means that almost all the vowels in continuous utterances were correctly recognized without the direct use of their sound substances. Further, the speaker-independent speech recognition was possible only with a single training speaker even in a noisy environment [8].

Every baby acquires language mainly hearing a remarkably biased speech corpus, called mother and father. Half amount of the speech one hears in his whole life is his own speech. No human can experience a speaker-balanced corpus. Young children and many dyslexics cannot identify isolated sounds as sound symbols (phonemes) but they can enjoy oral communication using word Gestalt [9]. In spite of these undeniable facts, as far as we know, the current speech recognition technologies are based on the collectionism and the separate sound symbolization paradigm. We have to wonder whether these frameworks are so sound that they can be used securely for education.

"Or are they weird?"

To answer this critical question, through discussions with many speech and non-speech researchers, the holistic, structural, dynamic, and speaker-invariant representation of speech was proposed by the first author [1, 2]. For example, the recent progresses of brain sciences claim that linguistic information and non-linguistic information in speech are separated on the auditory cortex [10]. Some researchers consider that, on the cortex, a verbal message in speech should be encoded as motions in speech [11]. If the two kinds of information can be acoustically separated, a speaker-invariant representation has to exist. If the speech motions carry the linguistic information, the ability of identifying an isolated sound as phoneme is not needed for language competence. Speech communication without the ability of separate sound identification was experimentally verified [12, 13]. Technically speaking, speech samples of large people like giants and small people like fairies can be easily generated. It is interesting that their isolated vowels produced by machines could not be correctly identified by human listeners. With 65 [cm] people, the identification rate was chance level because their range of F_1 and F_2 was by far out of the range of real people. Once they uttered even a *meaningless* sequence of sounds continuously, however, the identification rate drastically improved. This is considered to be because the utterance has acoustic motions and the motions are speaker-invariant. Human speech *stream* perception may not be a process of sequential and separate identification of the incoming sounds. Young children firstly acquire the holistic sound pattern of a word [5, 9].

A series of experimental discussions have been done for applying this new framework to CALL [3, 4]. The pronunciation of a learner was represented as its structure as in Figure 2. For example, the vowel structure of a Japanese learner of American English (AE) was constructed by recording 11 monosyllabic word samples, each including a different vowel out of the entire 11 AE monophthongs. After estimating the 11 distributions from the 11 vowel segments, a vowel structure was calculated as BD-based distance matrix with speaker identity cancelled.

The pronunciation development was described as the structural change through training. It was automatically estimated for individual learners which vowels to correct by priority. Further, the learner classification was tentatively investigated. The learners were successfully classified purely based on their structural distortions, not based on their gender or age. These results

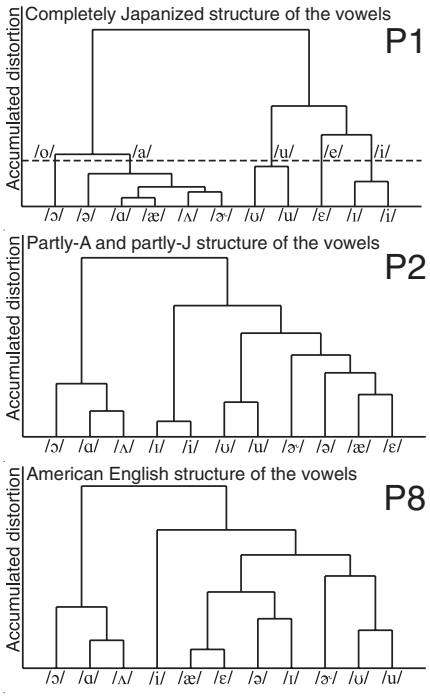

Figure 3: Japanized structure to American structure

were obtained not using sound substances but using sound contrasts only. In these works, however, the technical aspect of the proposed method was mainly examined. In this paper, a pedagogical investigation of the proposed framework is done. An expert phonetician is asked to plot a large number of vowel samples on the vowel chart. Using the results, two methods of assessing a learner are compared. One is the absolute comparison between a learner's chart and a teacher's one. The other is the relative or contrastive comparison only using sound contrasts.

2. Development of the vowel structure

The vowel structure development was traced [3]. Various nonnative pronunciations of the vowels were simulated by an adult male speaker who can speak Japanese and 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 only, various vowel structures were estimated. For example, the totally Japanized English structure was obtained by substituting five Japanese /a/ sounds for /ʌ, æ, ɑ, ə, ɚ/ and the other Japanese vowels for the corresponding AE vowels. Partly-American and partly-Japanese vowel structures were constructed by changing the substitution pattern. Figure 3 shows the totally Japanized structure, a half-American and half-Japanese vowel structure, and the AE structure. Ward's hierarchical clustering method was adopted to visualize the structure. The second tree diagram was obtained from the first one by correcting /ʌ, æ, ɑ, ə, ɚ/.

3. Classification of the learners

A learner was visualized as tree diagram, which was generated as a full set of distances between any two vowels. If distance measure between two vowel matrices, i.e., two learners, is adequately derived, we can calculate a distance matrix of all the learners, which gives a classification tree of the learners.

6 male and 6 female university students who are returnees from US joined the recording. The same recording was done as

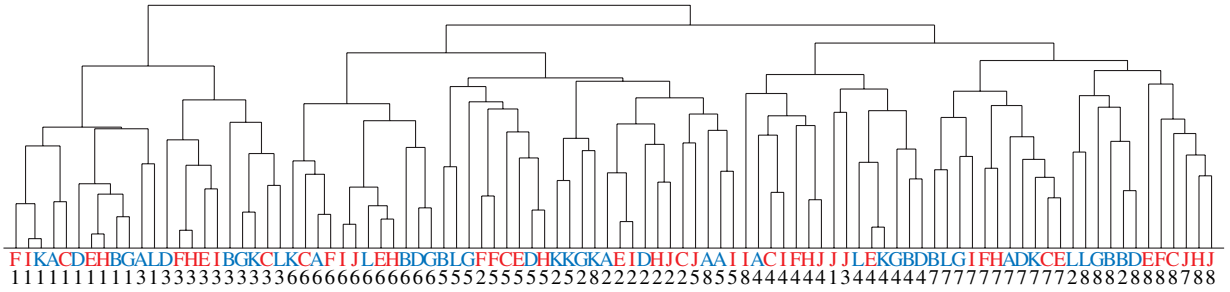

Figure 4: Classification of the 96 vowel structures

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.

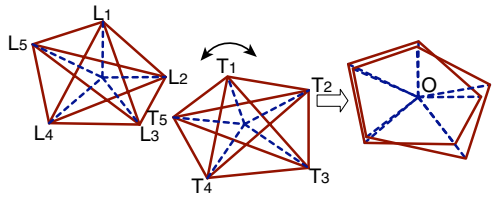

Figure 5: Distance calculation after shift and rotation

in the previous section. Considering the well-known Japanese habits, the substitution table was prepared, shown as Table 1. Using this table, 8 patterns of the vowel substitution were obtained, listed in Table 2. Then, we had 8 different vowel structures per speaker and 96 vowel structures altogether.

Any form of mapping function cannot change the matrix. This easily means that any mapping works geometrically as either of the two operations, rotation and shift. Suppose that a learner and a teacher are given as two distance matrices, L and T. Then, structure-to-structure distance is obtained after shifting and rotating one of the structures so that the two can be overlapped the best, shown in Figure 5. The distance is calculated as the minimum of the total distance between the corresponding two points after the two operations. The minimum distance D is approximately calculated as euclidean distance between the two matrices, where the upper-triangle elements form a vector;

$$D(L,T) = \sqrt{\sum_{i<j}(L_{ij} - T_{ij})^2}. \quad (1)$$

Figure 4 shows the result of classifying the 96 vowel structures. Numbers and alphabets represent the vowel patterns (1 to 8) and the speakers (A to L). We can say that reasonably good classification is done. Some different vowel patterns are found to

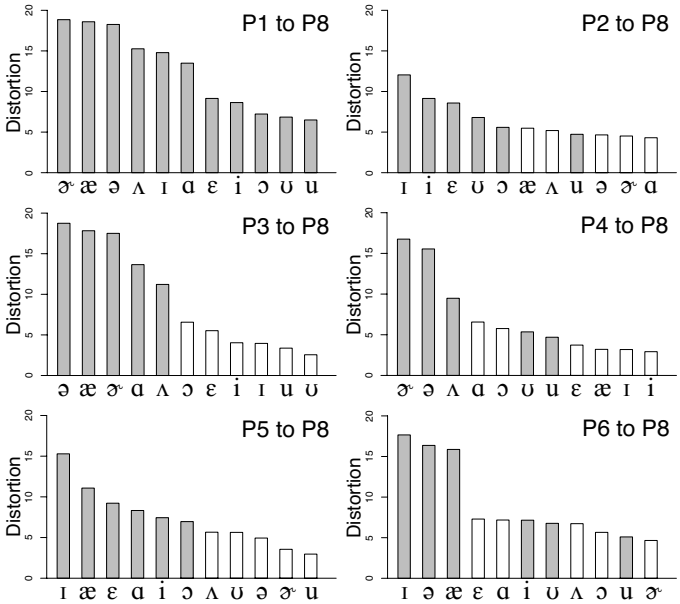

Figure 6: The vowel correction order estimated for P1 to P6

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, the length and the place of their stay in US are different from each other.

4. Estimation of the vowel correction order

Equation 1 shows the total distortion between L and T and it can be decomposed into components of the individual vowels;

$$d(L,T,v) = \sum_i |L_{vi} - T_{vi}|. \quad (2)$$

The vowel of the largest d should be corrected at first. 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 structure for each pattern. Using P8 as teacher, the vowel correction order was estimated for P1 to P6, shown in Figure 6. In the figure, bars represent d and gray bars mean that of the replaced vowels. The order for P1 is that for learners with the completely Japanized pronunciation. /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. It is often said in phonetics that /ɛ, i, ɔ, ʊ, u/ can be replaced with Japanese /e, i:, o, u, u:/. The result for P1 are very accordant with what phonetics tells. For P1 to P7, it is 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:/.

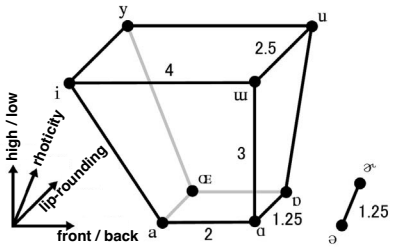

Figure 7: The four-dimensional vowel chart

5. Evaluation of the proposed framework

To assess a learner, the proposed framework ignores the sound substances and focuses on the sound contrasts only. The validity of this framework is evaluated by using the 96 vowel charts plotted by an expert phonetician, the fourth author.

If a teacher's chart and a learner's one are given, by overlapping the two trapezoids directly, we can estimate which vowel to correct at first, called direct assessment hereafter. In the previous sections, what we extracted were vowel structures without the trapezoidal axes. This is why we could not overlap the two structures directly and we estimated a good overlap by rotation and shift, called indirect assessment. With the 96 manually plotted charts, we can do both assessments although only the indirect one was possible with the structural speech analysis. In this section, the two methods are examined exactly on the same data and the correlation between the two is investigated.

5.1. Drawing vowel charts through listening

The 96 sets of the 11 vowels were presented to the phonetician through headphones. He was asked to draw the 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, two-dimensional trapezoidal vowel charts are usually used to show the structural relations among the vowels, where only the tongue position is focused on. In this work, 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 was for rhoticity of /ɚ/. In Figure 7, the four-dimensional framework adopted in the drawing software is shown, where the last dimension is separately added to the other three dimensions. Numbers on the segments indicate relative distance between two nodes. In the experiments, a two-dimensional trapezoidal framework was presented on a PC to specify the tongue position and values of the other two dimensions were separately asked to be entered.

5.2. Correlation between the two assessments

In each vowel chart, each of the 11 vowels had coordinate values in the four dimensional space. A teacher's chart and a learner's chart can be directly compared by using these values. Then, a priority score of correcting vowel v was defined as euclidian distance between the teacher's v and the learner's v. With the coordinate values of the 11 vowels, their distance matrix was also obtained. And using the teacher's matrix and the learner's one, another priority score for correcting v was calculated by using Equation 2. For each vowel chart, two kinds of priority scores, direct and indirect, were assigned to the individual vowels. The correlation between the two scores is in question.

Table 3 shows the results. For each speaker, P8 was treated as teacher. The two priority scores were assigned to every vowel in P2 to P7 and the correlation was calculated separately for each P_i. The overall correlation was 0.78. The priority scores

Table 3: Correlations between the two assessment methods

P1	P2	P3	P4	P5	P6	P7
0.68	0.82	0.82	0.89	0.80	0.75	0.85

can also be defined by using P_i ($i \neq 8$) as teacher. Using every P_i as teacher and the others, P_j ($j \neq i$), as learners, the overall correlation was calculated and it was 0.78. We consider that very high correlation is found between the two assessment methods, whatever target pronunciation the learners are aiming at.

This finding claims that, only with the phonic contrasts, learners can be assessed adequately and some good instructions on which vowels should be corrected can be given to them. However, it is found that the correlation of P1 is relatively low and we have a clear reason for that. The indirect priority score is based on Equation 2 and it is calculated by accumulating v's difference to each of the other sounds. Even if v is pronounced as correct, it will be judged as incorrect if all the other sounds are pronounced as wrong. In P1, as every vowel was replaced by Japanese vowels, this adverse effect is considered to be able to happen. In the structural analysis, this effect is basically unavoidable because we don't have any absolute anchoring point in the structure. Putting it another way, we can say that the degree of freedom in the overlapping operations of rotation and shift is too high. Without any constraints, unrealistic operations can be performed for a given structure. In [7], this problem was solved by deriving some geometrical constraints and we're planning to use them for the pronunciation assessment.

6. Conclusions

In this paper, after pointing out some critical problems in the current framework of speech recognition, a novel framework was proposed where the structural and speaker-invariant representation of speech was introduced. Although this framework uses only the phonic contrasts and the data was generated by simulation in the experiments, the comparison between the direct and the indirect assessments of language learners' pronunciation showed the high correlation between the two.

7. References

[1] N. Minematsu, *et al.*, "Theorem of the invariant structure and its derivation of speech Gestalt," Proc. SRIV, 47–52, 2006.

[2] N. Minematsu *et al.*, "Linear and non-linear transformation invariant representation of information and its use for acoustic modeling of speech," Proc. Spring Meeting ASJ, 147–148, 2007.

[3] S. Asakawa, *et al.*, "Structural representation of the non-native pronunciations," Proc. InterSpeech, 165–168, 2005.

[4] N. Minematsu *et al.*, "Structural representation of the pronunciation and its use for CALL," Proc. SLT, 126–129, 2006.

[5] P. W. Jusczyk, The discovery of spoken language, Bradford Books, 1997.

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

[7] S. Asakawa, *et al.*, "Automatic recognition of connected vowels only using speaker-invariant representation of speech dynamics," Proc. InterSpeech, 2007 (accepted).

[8] T. Murakami, *et al.*, "Japanese vowel recognition using external structure of speech," Proc. ASRU, 203–208, 2005.

[9] S. E. Shaywitz, Overcoming dyslexia, Random House Inc. 2005.

[10] S. K. Scott, *et al.*, "The neuroanatomical and functional organization of speech perception," Trends in Neurosciences, 26, 2, 100–107, 2003.

[11] P. Belin, *et al.*, "'What', 'where' and 'how' in auditory cortex," Nature neuroscience, 3, 10, 965–966, 2000.

[12] D. Smith *et al.*, "The processing and perception of size information in speech," J. Acoust. Soc. Am., 117(1), 305–318, 2005.

[13] Y. Hayashi *et al.*, "Comparison of perceptual characteristics of scaled vowels and words," Proc. Spring Meeting ASJ, 473–474, 2007.