Structural representation of pronunciation and its use in pronunciation training

N. Minematsu*, S. Asakawa*, K. Hirose*, and T. Makino**
*The University of Tokyo, Japan
**Chuo University, Japan

1 Introduction No two students are the same. Pronunciation teaching should be started after teachers know exactly how individual students are in their development. In one-to-fifty situations such as classrooms, it is difficult for a teacher to know their pronunciations precisely. Computers have provided us with tools to visualize the pronunciations based on acoustics. In this case, however, good knowledge of acoustics is needed. Further, the acoustic representation inevitably shows many things irrelevant to the proficiency, such as speaker individuality, gender, age, microphone differences and so on. The first author already proposed a unique method of representing speech acoustics, where dimensions of the non-linguistic factors are well removed, Minematsu (2004a). This representation can be regarded as physical implementation of structural phonology, where only the interrelations among speech sounds are focused. In this paper, the non-native pronunciations are described based on the new method and the development of a student's pronunciation is traced. Further, automatic assessment of the pronunciation is investigated experimentally.

2 Structural representation of speech acoustics Spectrogram is noisy representation in that it inevitably shows things completely irrelevant to the pronunciation proficiency. The first author proposed a method to represent speech acoustics where the static non-linguistic factors can hardly be seen, Minematsu (2004a). Since explanation of this method requires good knowledge of mathematics, only its short introduction is done here. This new method was inspired by Jakobson's phonological structure in Figure 1, where French vowels and semi-vowels are structurally represented and it was claimed that this structure is invariant with speakers. In acoustic phonetics, the vowel structure is often represented as F1-F2 formant chart. It is known that this representation clearly shows gender and age difference. With the proposed method, these non-linguistic factors can be effectively suppressed. What is geometrical definition of a structure? A triangle is determined by fixing length of all the three segments. An n-point structure, in turn, is determined by fixing length of all the segments including its diagonal lines. This means that an n-point structure is fully represented by its distance matrix among the n points. If cepstrum parameters are used to represent envelopes of the spectrogram and if a speech sound is represented as point in a cepstrum space, an acoustic structure of n speech sounds is an n-point structure in the space. The acoustic structure can become invariant if the non-linguistic factors cannot change distance between any two points. Mathematically speaking, however, the structure has to be variant due to the non-linguistic factors. How to make inevitably variant structures invariant? The invariant structure can be obtained by applying theorem of the invariant structure, which was proposed by the first author Minematsu et al (2005). Here, every speech sound is represented as cepstrum distribution, not point. Distance between two distributions is calculated by distorting space so that the structure can be invariant. The space distortion can easily make variant structures invariant.

3. Structural representation of the non-native pronunciation Using the proposed method, individual students are described as distorted speech structures of the target language. Since the new method extracts an acoustic structure as distance matrix,

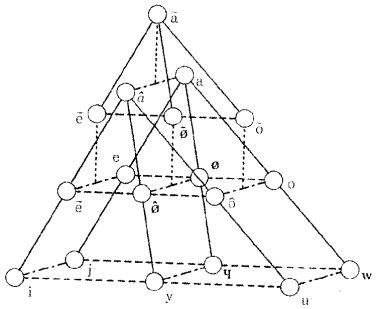

Figure 1. Jakobson's phonological structure of French vowels and semi-vowels

visualization of the structure using absolute properties, e.g. formants, is impossible. Here, tree diagram was adopted to visualize the matrix. Figure 2 shows two trees; one from utterances of an American speaker and the other from utterances of a Japanese student learning American English, both reading the same 60 sentences. The Japanese tree clearly shows the well-known Japanese habits of English pronunciation. Confusions of /r/ & /l/, /s/ & /θ/, /z/ & /ð/ /i/ & /ɪ/, /v/ & /b/, etc are observed. Mid and low vowels are located very close to each other. Schwa is found to be very close to the above vowels.

Since the proposed method only extracts distance matrix of the speech sounds, nothing is known about physical properties of the individual sounds such as formant frequencies. This strategy of the analysis is contradictory to that of acoustic phonetics, where physical properties of the individual speech sounds are intensively measured. The authors consider that the conventional strategy only gives the noisy representation of speech, spectrogram, and that an alternative method, which is stable and reliable, has to be devised especially for educational use. A candidate answer this paper shows was provided by implementing structural phonology physically. Comparison between the new method and the conventional one was done in Minematsu (2004b) with respect to reliability.

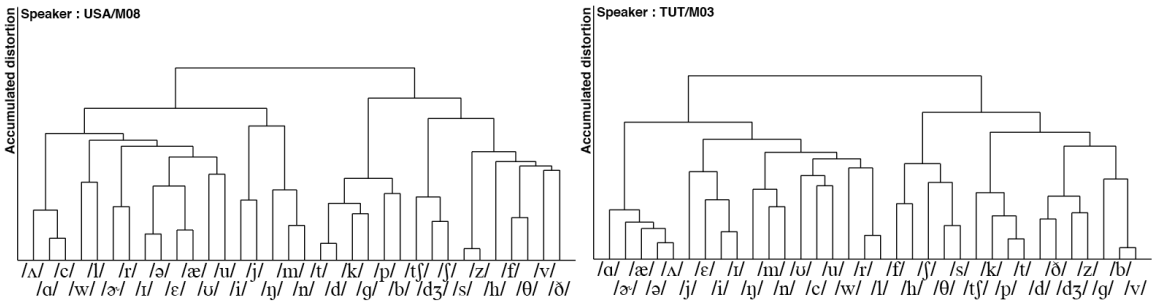

Figure 2. Tree diagrams of American English (left) and Japanese English (right)

3. Tracing the development of the pronunciation Different students show different distortions in their pronunciation structures. Within a student, the structure even changes easily through training. In this section, various Japanese English pronunciations are simulated and each of them is represented structurally. An adult male Japanese speaker, who had been an amateur actor on English stages, spoke /bVt/ words, where V is a

monophthong of American English (/ɪ i ɛ æ ʌ ɑ ʊ ɔ u ɚ ə/) or a Japanese vowel (/a ɪ u e o/). For American English, only one utterance was recorded per vowel and, for Japanese, five utterances were recorded per vowel. By mixing both the types of pronunciations, a variety of English pronunciations were simulated. Figure 3 shows vowel charts of American English and Japanese. Japanese learners often substitute Japanese vowels when they speak English and Table1 shows typical examples of the vowel substitution. Japanese /a/ has very strong power and is substituted for five vowels of American English. In the current analysis, pronunciation states were defined as the pronunciations with some vowel substitutions and the following states were considered.

S1: All the American English vowels are replaced with Japanese vowels.

S2: / æ ʌ ɑ ɚ ə/ are corrected.

S3: /i ɪ/ are additionally corrected.

S4: / ʊ u/ are additionally corrected.

S5: /ɛ / is additionally corrected.

S6: /ɔ/ is additionally corrected and all the vowels are pronounced correctly.

When multiple English vowels, e.g. /bʌt/ and /bæt/, were replaced with a Japanese vowel, different utterances of the vowel, two utterances of /bat/ in this case, were used.

S1 tree, the intentionally Japanized pronunciation, shows clear separation of the vowels into the 5 Japanese vowels. S6 tree, the good pronunciation used on English stages is accordant to American English phonetics. These two trees have very good correspondence to the two vowel charts. Gradual changes are found from S1 to S6. For example, correction of /æ ʌ ɑ ɚ ə / destroys the Japanese vowel system embedded in S1. Transition from S2 to S3 separates /i/ and /ɪ/. That from S3 to S4 enlarges the separation of /u/ and /ʊ/. In S5, /ɛ/ and /æ/ get closer. S5 and S6, however, show almost no difference. These results show that structuralization with a single example per vowel can describe the pronunciation effectively and efficiently and that it is possible to log a student's development with a small amount of utterances. Although this analysis was done using a single speaker, since Minematsu (2005) showed that the structuralization can delete dimensions of speaker differences effectively, this method can be used for other speakers as it is. It is interesting that structural acoustic models can recognize speech automatically with no direct use of acoustic substances of speech sounds, Murukami et al (2005).

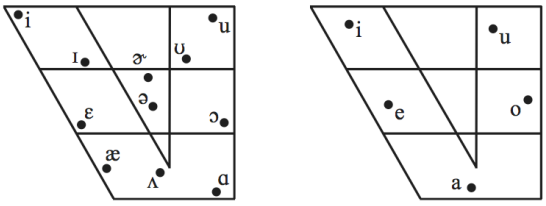

Figure 3. Vowel charts of American English and Japanese

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

Table 1. Vowel substitution table

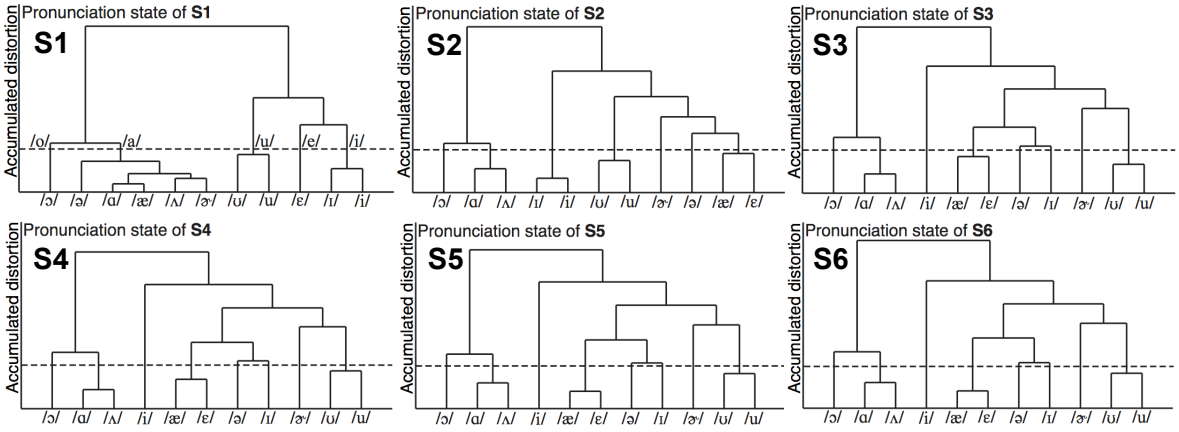

Figure 4. Visualization of various non-native pronunciations as tree diagrams

3. Assessment of the pronunciation based on size of the vowel structure The structural representation of the pronunciation can show not only its segmental aspect but also its prosodic aspect. Schwa is the most fundamental vowel in that it is located at the center of the vowel chart and that it is produced with the least articulatory effort. It is often said that unstressed vowels get closer to schwa sounds. These predict that the vowel structure gets larger if they are stressed and smaller if they are unstressed. Size of the structure may be interpreted as magnitude of the articulatory effort.

Using 709 sentences read by a female American, the prediction was examined experimentally. Here, only /ɪ i ɛ æ ʌ ɑ ʊ u ɚ/ were focused because the other two vowels had a strong bias between occurrences as stressed and those as unstressed. Figure 5 shows a tree diagram of the stressed vowels and that of the unstressed vowels, where /æ1/ and /æ0/ mean stressed æ and unstressed æ. Height of the tree corresponds to radius of the structure. The stressed tree is 1.4 times higher than the unstressed tree and the above prediction was verified experimentally.

Size of the vowel structure was used to assess the non-native pronunciations. 60 sentences were read by 19 Japanese students (10 males and 9 females) and two kinds of the vowel structures were extracted from each of them. The same sentences were read by 4 Americans and the vowel structures were extracted in the same way. Figure 6 shows the correlation between the ratios of the structure sizes (stressed to unstressed) and the proficiency scores rated by 5 American teachers. Very high correlation is found and this result shows high validity of using size of the vowel structure for automatic assessment of the pronunciation. The averaged ratio of the 4 Americans was 1.17.

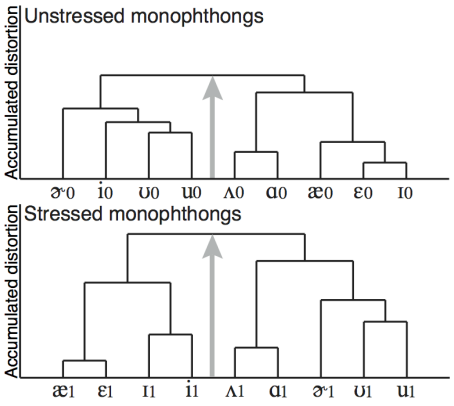

Figure 5. Unstressed and stressed tree diagrams

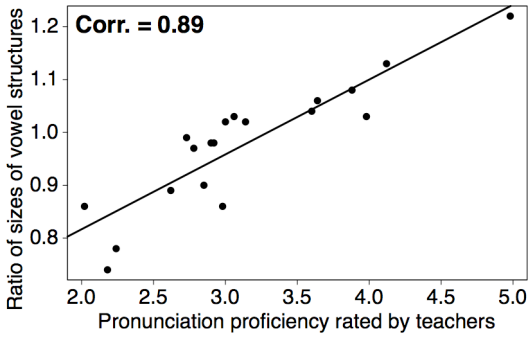

Figure 6. Correlation between ratios of structure sizes and human scores

4. Conclusions This paper effectively introduced the structural representation of speech sounds to the pronunciation training. Although some experimental results were shown, large portions of technical discussions were intentionally omitted. Interested readers in these issues should make contact to the first author (mine@gavo.t.u-tokyo.ac.jp).

5. References

N. Minematsu, (2004a) “Yet another acoustic representation of speech sounds,” Proc. ICASSP, pp.585-588

N. Minematsu, (2004b) “Pronunciation assessment based upon the phonological distortions observed in language learners’ utterances,” Proc. ICSLP, pp.1669-1672

N. Minematsu, (2005) “Mathematical evidence of the acoustic universal structure,” Proc. ICASSP, pp.889-892 (2005)

N. Minematsu, T. Nishimura, K. Nishinari, and K. Sakuraba, (2005) “Theorem of the invariant structure and its derivation of speech Gestalt,” English version is submitted to Interspeech’ 2005 and Japanese version is published as IEICE technical report, SP2005-12, pp.1-8

T. Murakami, K. Maruyama, N. Minematsu, and K. Hirose, (2005) “Japanese vowel recognition based on structural representation of speech,” submitted to Interspeech 2005