Structural Representation of the Non-native Pronunciations

Satoshi ASAKAWA†, Nobuaki MINEMATSU‡, Toshiko ISEI-JAAKKOLA†, and Keikichi HIROSE†

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

Abstract

Acoustic representation of speech provided by phonetics, spectrogram, is noisy representation in that it shows every acoustic aspect of speech. Age, gender, shape, microphone, room, line, etc. are completely irrelevant to the pronunciation assessment. However, the spectrogram is affected inevitably by these factors. Recently, a novel acoustic representation of speech was proposed, where dimensions of these non-linguistic factors can hardly be seen[1, 2]. Every acoustic substance of speech is discarded and only their interrelations are extracted to represent the pronunciation structurally. Using this method, individual learners were described as distorted phonemic structures[3] and automatic scoring of the pronunciation was investigated[3, 4]. This paper describes two new analyses using the proposed method. The first analysis is done to examine whether the method can trace the development of a student's pronunciation appropriately using only a limited amount of speech. The second one focuses on the prosodic aspect of the pronunciation, especially stressed and unstressed vowels. The former indicates that the proposed method can show history of the student's development adequately and the latter clarifies that size of the pronunciation structure is highly correlated with the pronunciation proficiency.

1. Introduction

Few language teachers have good knowledge of acoustics and physics. Although spectrogram is provided as speech representation by acoustic phonetics, it is noisy because it shows many irrelevant things. Acoustic phonetics is phonetic acoustics. Is there any good method to represent the pronunciations purely?

Apart from semantics, linguistics provides two definitions of the phonemes[5]. 1) a phoneme is a class of phonetically-similar sounds 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 obvious that the first definition brought about the so-called speaker-independent HMMs widely used in speech recognition. However, collection of millions of /a/ sounds defines only its averaged distribution and never deletes dimensions indicating the unwanted non-linguistic factors. Recently, a novel acoustic representation of speech was proposed only based on the second definition and it is called the acoustic universal structure[1]. Dimensions of the static and inevitable non-linguistic features can hardly be found in the structural representation.

The authors already applied this new representation to pronunciation training[3, 4] and showed that the representation enabled much more robust and stable assessment of the pronunciation proficiency. In the previous studies, however, about 60 sentence utterances per student were required to describe how he/she is in the pronunciation development. To realize interactive applications of CALL, it is necessary to extract the structure with a limited amount of speech; stable estimation of the structure from only a few utterances. This paper tries to solve this difficult problem. Further, this paper focuses on the prosodic aspect of the pronunciation, which was not discussed at all in [3, 4], based on the structural representation of speech.

2. The acoustic universal structure

2.1. Acoustic modeling of the non-linguistic features

To delete the inevitable non-linguistic features from speech, they are modeled firstly, and then an algorithm for their deletion should be devised. Acoustic distortions caused by the non-linguistic features are often classified into three kinds; additive, multiplicative, and linear transformational distortions. The additive distortion(noise) is ignored here because it is not inevitable. Speakers can turn off a TV set if they want to. The other two distortions are, however, inevitable and their deletion has to be done not by hand but by an algorithm.

Acoustic characteristics of microphones and rooms are typical examples of the multiplicative distortion. If a speech event is represented by cepstrum vector c, the multiplicative distortion is addition of b and the resulting cepstrum is shown as $c'=c+b$. 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, where formant shifts are well approximated. According to [6], any monotonous frequency warping of the log spectrum is converted into multiplication of matrix A in cepstrum domain; $c'=Ac$. Various distortion sources are found in the speech communication channel. The total distortion due to the *inevitable* sources, A_i and b_i, is eventually modeled as $c'=Ac+b$, known as affine transformation.

2.2. The acoustic universal structure in speech

Deletion of the non-linguistic dimensions can be done by structuralizing speech acoustics. An N-point structure is geometrically determined uniquely by fixing length of all the ${}_NC_2$ lines including the diagonal lines(distance matrix). Then, a necessary and sufficient condition for the desired representation is that distance between any two points cannot be changed by any of a single affine transformation. This condition looks impossible because affine transformation distorts a structure unless it is of a special form. The solution can be obtained simply by distorting space so that the structure can be invariant.

THEOREM OF THE INVARIANT STRUCTURE
N 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.

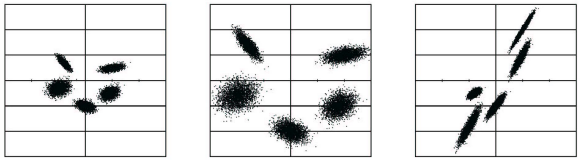

Figure 1: *Completely the same structure is found in the three.*

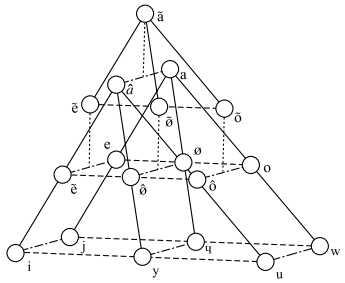

Figure 2: *Jakobson's geometrical structure*

Distribution means a Gaussian mixture. Bhattacharyya distance was adopted because it can be interpreted as normalized cross correlation between two PDFs $p_1(x)$ and $p_2(x)$.

$$BD(p_1(x), p_2(x)) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}dx, \quad (1)$$

where $0.0 \leq \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}dx \leq 1.0$ and name of unit of BD is bit because BD can be regarded as self-information. If the two distributions follow Gaussian, BD is formulated as follows.

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

μ_{12} is $\mu_1 - \mu_2$. Figure 1 shows three structures of five distributions. Any two of the three structures can be converted to one another by multiplying A. This fact means that the three structures(matrices) are evaluated as completely the same. Why this happens? Because BD calculation distorts the space where the distributions are observed. This distorted space is mathematically analyzed based on differential geometry[2] and this structure is named as the acoustic universal structure[1]. The structural representation can be interpreted as physical implementation of structural phonology. Figure 2 shows Jakobson's geometrical structure proposed for French vowels and he claimed that this structure is invariant with respect to speakers.

3. Structuralization with several utterances

3.1. Stable distribution estimation with a single instance

The vowel structure was focused on and a method of extracting the structure with several utterances, an instance per vowel, was devised. Since the acoustic universal structure is based on representing every speech event as distribution, a vowel has to be characterized by a distribution only with its single instance. Although the previous studies used ML(Maximum Likelihood) criterion to estimate the distribution, it works poorly when only a limited amount of data is available. Reference [7] solved this problem by introducing MAP(Maximum A Posteriori) criterion. The present study also used the MAP-based estimation.

In phonetics, the vowel structure is often shown as F_1-F_2 chart. However, it is known that shape of the chart depends on gender and age. The proposed method can remove the dependency mathematically. Then, the structure has to exist in an N-dimensional space and may cause difficulty of its visualization.

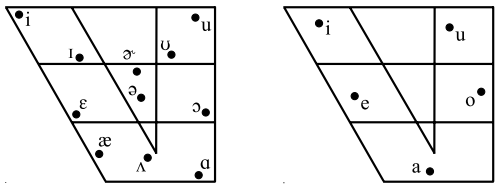

Figure 3: *Vowel charts of American English and Japanese*[9]

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: *Acoustic conditions of the analysis (Section 3)*

sampling	16bit / 16kHz
window	25 ms length and 4 ms shift
parameters	FFT cepstrum (1∼10)
HMMs	1-mixture monophones with diagonal matrices
topology	3 states and 1 distribution per HMM (GM)

3.2. Speech material used in the analysis

It would be ideal to take plenty of time to trace the pronunciation development of a student. Since this work aims at solving a technical problem, instead of tracing a student, speech material was obtained from a good speaker of American English and Japanese English. By mixing both Englishes, a variety of Englishes were simulated. An adult Japanese speaker, who had been an amateur actor on English stages, joined the recording. He can speak very good American English while he can speak English intentionally with a strong Japanese accent. A single utterance of /bVt/ was recorded for each vowel, where V was a monophthong of American English (/i, ɪ, ɛ, æ, ʌ, ɑ, ɔ, ʊ, u, ɚ, ə/). Five utterances of Japanese /bVt/ were recorded for each vowel, where V was a Japanese vowel (/a, i, u, e, o/).

3.3. Tracing the simulated pronunciation development

Figure 3 shows vowel charts of American English and Japanese. Japanese learners often substitute Japanese vowels when they speak English. Table 1 shows typical examples of the vowel substitution. Japanese /a/ has very strong power and is substituted for five vowels of American English. In this analysis, pronunciation states were defined as the pronunciations with some vowel substitutions and the following states were considered.

S1: All the AE vowels are replaced with Japanese vowels.
S2: /ɑ, ʌ, æ, ɚ, ə/ are corrected.
S3: /i, ɪ/ are additionally corrected.
S4: /u, ʊ/ are additionally corrected.
S5: /ɛ/ are additionally corrected.
S6: /ɔ/ are additionally corrected.

Development of the pronunciation was simulated as transition from **S1** to **S6**. When multiple English vowels, e.g. /bʌt/ and /bæt/, were replaced with a Japanese vowel, different utterances of the vowel, two utterances of Japanese /bat/ in this case, were used. Acoustic conditions of the analysis is shown in Table 2.

Visualization of the structure, distance matrix mathematically, was done by drawing its tree diagram. Ward's method was adopted as bottom-up clustering. Figure 4 shows six tree diagrams and each tree was drawn by using 11 /bVt/ utterances. **S1** tree, structuralized from the intentionally Japanized pronun-

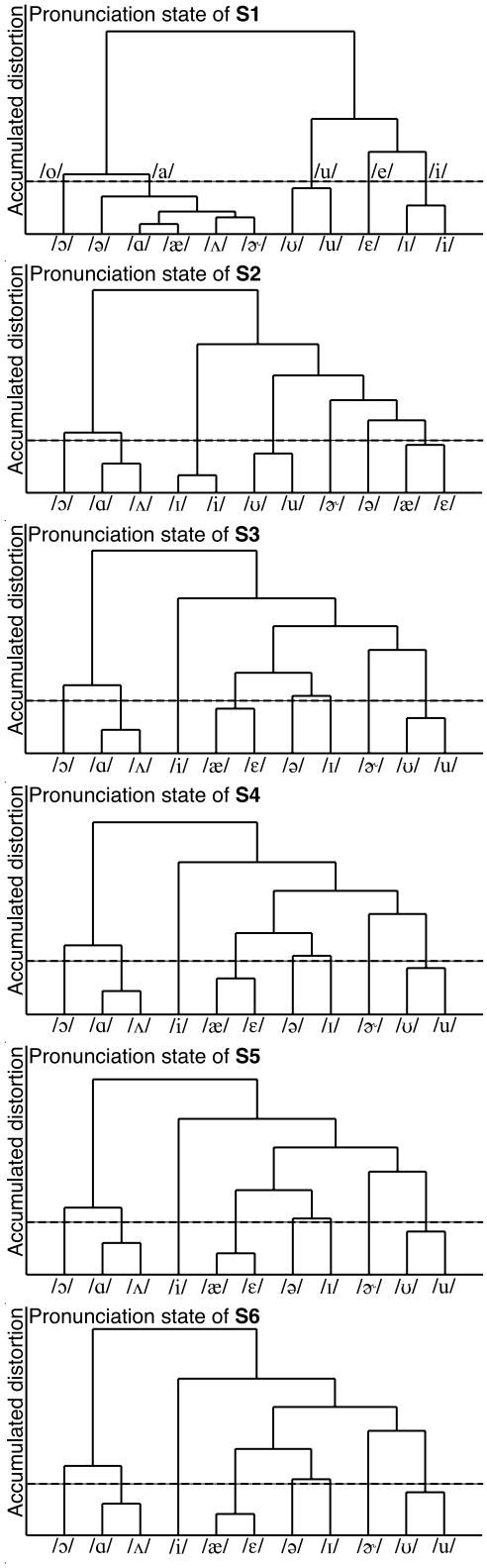

Figure 4: *Six tree diagrams of the pronunciation*

ciation, shows the well-known Japanese habits and clear separation of the vowels into 5 Japanese vowels. **S6** tree, from the good pronunciation, is very accordant to phonetics of American English. These two trees have good correspondence to the two vowel charts(Figure 3). Gradual changes are found from **S1** to **S6**. For example, correction of /ɑ, ʌ, æ, ɚ, ə/, i.e. **S1** to **S2**, 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**, /æ/ and /ɛ/ get closer. **S5** and **S6**, however, show almost no difference. These results show that the structuralization with a single utterance per vowel can describe the pronunciation adequately and that it is possible enough to record history of a student's development with several utterances. Although the analysis was done only with a single speaker, since reference [7] showed that the utterance-level structuralization can delete dimensions of speaker differences effectively, this method can be applied to other speakers as it is. It is very interesting that the structural acoustic models can recognize speech with no direct use of acoustic substances of the phonemes[7]. Further, it was found surprisingly that the structural models trained by a single speaker outperformed the conventional HMMs trained by 4,130 speakers in a specific task[7].

Table 3: *Acoustic conditions of the analysis (Section 4)*

sampling	16bit / 16kHz
window	25 ms length and 10 ms shift
parameters	Mel cepstrum (1∼12)
HMMs	1-mixture monophones with full matrices
topology	3 states and 1 distribution per HMM (GM)

4. Stressed vowels and unstressed vowels

4.1. Some correlates of the structure with prosody

In our previous studies[3, 4], the structural representation of the pronunciation was examined whether it can assess the segmental aspect. The previous section used the representation to measure goodness of the vowel production. In this section, however, some correlates of the structural representation with prosody are discussed. As is known in phonetics, schwa is the most fundamental vowel in that it is located at the center of the vowel chart (see Figure 3). It is also known that schwa is generated with a sound tube of a *fixed* cross-sectional area, indicating that schwa is produced with the least articulatory effort. It is often said that, if vowels are reduced, they get closer to schwa[10]. These considerations 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 an American female speaker, the above prediction was experimentally investigated. Table 3 shows acoustic conditions of the analysis. To train the acoustic models, phonemic and stress labeling was required and this was done by a semi-automatic method. PRONLEX dictionary was referred to for determining the initial labels and they were modified with the speaker-dependent models trained so far. Namely, the acoustic models and the phonemic and stress labeling were simultaneously trained and adjusted. In the rest of this paper, 'æ1' and 'æ0' mean stressed and unstressed 'æ's, respectively.

4.2. Size of the vowel structure

In Ward's method, two elements are merged into one sequentially so that the accumulated distortion should be minimized. The accumulated distortion is represented by height of the tree grown so far. Finally, all the elements are integrated into a single element(centroid) and height of the final tree is mathematically equal to VQ(Vector Quantization) distortion when all the data is represented by the single centroid. This quantity can be interpreted as radius or size of the structure.

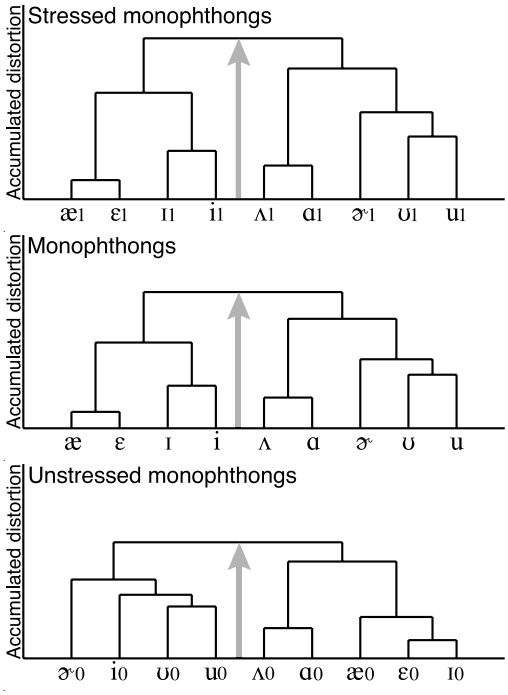

Figure 5: *Tree diagrams of American English vowels*

Figure 5 shows three tree diagrams; stressed vowels, vowels, and unstressed vowels. In the vowel tree, approximately 60% of the vowels are stressed. For a few vowels, a very strong bias between occurrences as stressed and those as unstressed was found and these vowels were not used. /i, ɪ, u, ʊ, ɛ, æ, ʌ, ɑ, ɚ/ were used. Clearly shown in the figure, the vowel tree is lower than the stressed vowel tree and higher than the unstressed vowel tree. The stressed tree is 1.4 times higher than the unstressed one. Although shape of the tree is similar between the vowel tree and the stressed one, some differences are found between the unstressed tree and the other two ones. Two reasons are possible. One is differences of phonemic environments between the stressed vowels and the unstressed ones. It is known that acoustic properties of phonemes depend on their phonemic environments. It is also likely that some unstressed vowels were acoustically realized as schwa sounds completely. These results indicate validity of our prediction that size of the structure corresponds to magnitude of the articulatory effort.

4.3. Proficiency estimation based on size of the structure

Unstressed vowels are produced by reducing their structure, i.e. by making less articulatory efforts. As shown in Figure 3, the Japanese vowel system does not have the central vowel. It is very difficult for Japanese learners to produce this central vowel adequately, and therefore to create English rhythm correctly. Goodness of producing stressed vowels and unstressed vowels is expected to be automatically measured by considering ratio of the stressed structure size to the unstressed structure size.

60 sentences were read by 19 Japanese students (10 males and 9 females), which were speakers of set-6 of ERJ(English Read by Japanese) database[8]. Procedures to calculate the two kinds of sizes of the vowel structure were the same as those in Section 4.2. Each student in ERJ has his/her pronunciation proficiency score rated by 5 American teachers of English.

Figure 6 shows the correlation between the ratios of the structure sizes and the proficiency scores rated by the five teachers. The vertical axis and the horizontal one indicate the ratios of the sizes and the proficiency scores, respectively. Very high correlation is clearly found and this result indicates very high validity of using ratios of the two vowel structure sizes for automatic assessment of the pronunciation. The averaged ratio of four American speakers who read the set-6 in ERJ was 1.17.

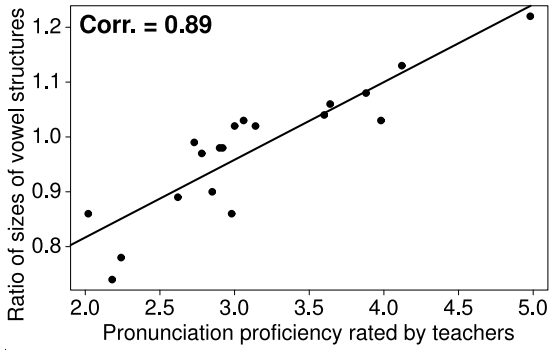

Figure 6: *Correlation between structure sizes and human scores*

5. Conclusions

This paper described two new analyses of the non-native pronunciation based on the acoustic universal structure. The first analysis showed that stable estimation of the vowel structure is possible only with a single instance per vowel and that history of a student's development of the pronunciation can be recorded appropriately. The second analysis focused on the prosodic aspect of the pronunciation; stressed vowels and unstressed ones. Here, size of the vowel structure was examined to know whether it can reflect goodness of producing stressed or unstressed vowels. The results showed that size of the acoustic universal structure can be regarded as magnitude of the articulatory efforts and that ratio of the stressed vowel structure size to the unstressed structure size can be a good indicator of the pronunciation proficiency. The authors are currently integrating the two analyses to realize automatic assessment of the proficiency based on the structure size estimated with a single utterance per vowel.

6. 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," *Technical Report of IEICE*, SP2005-12, pp.1–8 (2005, in Japanese)

[3] N. Minematsu, "Pronunciation assessment based upon the phonological distortions observed in language learners' utterances," *Proc. ICSLP*, pp.1669–1672 (2004)

[4] N. Minematsu, "Pronunciation assessment based upon the compatibility between a learner's pronunciation structure and the target language's lexical structure," *Proc. ICSLP*, pp.1317–1320 (2004)

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

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

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

[8] N. Minematsu *et al.*, "Development of English speech database read by Japanese to support CALL research," *Proc. ICA*, pp.557–560 (2004)

[9] International Phonetic Association, *Handbook of the International Phonetic Association*, Cambridge University Press (1999)

[10] J. Clark and C. Yallop, *An introduction of phonetics and phonology*, 2nd edition, Blackwell Publishers Inc. (1995)