# Pronunciation Portfolio
## — How were, are, and will be you? —

**N. Minematsu, S. Asakawa, K. Okabe, and K. Hirose**
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
{mine, asakawa, okabe, hirose}@gavo.t.u-tokyo.ac.jp

## Abstract

*No two students are the same. There are about 2 billion students of English on this planet and each student is always evolving through training. This means that there are about 2 billion types of English pronunciation. Despite the tremendous number of pronunciations, there has been no good method so far to represent each pronunciation individually. This study introduces a very novel method to represent the individual pronunciations. The method is based on physical implementation of structural phonology and the implementation can be regarded as a mathematical interpretation of Saussure's claim that language is a system of conceptual differences and phonic differences. Each student's pronunciation is acoustically and entirely represented as phonological structure with no dimensions to indicate non-linguistic features like age, gender, speaker, microphone, room, line, etc. This paper examines whether the structural representation can provide a good tool for pronunciation assessment. Results of experiments with good and intentionally-bad pronunciations of a single speaker showed that all the students used in the experiment are acoustically located between the two pronunciations, indicating that the students are judged to be acoustically closer to the speaker than the speaker himself is. This result shows that the proposed method can delete the irrelevant factors effectively and is extremely reliable in CALL.*

**Keywords**: *Student representation, structural phonology, pronunciation assessment, phonological distortion, non-linguistic features*

## 1 Introduction

Pronunciation training should be based upon articulatory phonetics because a speech sound is produced adequately only by the correct articulation. However, it is very difficult and expensive to measure movements of the articulators of students because the measurement requires specialized medical equipments. It is possible for good phoneticians to guess what is happening in students' mouthes only by listening to their speech not by measuring their articulators. Unfortunately, not all the language teachers have good knowledge of articulatory phonetics. As an alternative to articulatory phonetics, spectrograms, speech representation of acoustic phonetics, have been investigated to see whether they can be a good tool for pronunciation assessment. In one sense, spectrograms can show clearly the quality of the pronunciation because teachers can judge the quality of the segmental aspect of the pronunciation by *listening to* the spectrograms. In another, however, the spectrograms cannot clearly show the quality of the pronunciation because teachers cannot judge it by *looking at* the spectrograms. Teachers have expected speech engineering, i.e., computers, to judge the quality by looking at the spectrograms, which led to the development of CALL systems with speech recognition technologies. The question is whether computers can make reliable and pedagogically-sound enough judgment with the spectrograms. In the beginning, CALL systems were accepted by all the teachers and students because students were allowed to have virtual teachers anytime and anywhere with multimedia attractions(Hiller, 1993; Bernstein, 1990; Cucciarini, 1998). But recently, some papers have reported on the unreliability and instability of these systems(Ambra, 2003). Native speakers are sometimes judged to be worse than students, for example. The problem lies in that the spectrogram is a noisy representation of speech as it shows every acoustic aspect of speech, such as age, gender, size, speaker individuality, microphone difference, line difference, and so on. These factors are completely irrelevant to pronunciation assessment but are inevitably involved in a speech production and transmission process. Acoustic phonetics may be phonetic acoustics. It is the case with speech recognition, whose task is extracting lexical identity from speech. But the spectrogram can show things completely irrelevant to the task. In the past, speaker-independent and environment-independent acoustic models have been built by collecting a large amount of data, but they often require speaker and environment adaptation techniques. This fact implies that the models are not really speaker- or environment-independent. Collection of more data, i.e., a quantitative and naive solution, seems not to work pedagogically-sound enough.

A novel and qualitative solution was proposed by the first author. Deletion of the non-linguistic features was done not by collecting data but by deleting all the dimensions of the non-linguistic features from speech acoustics mathematically(Minematsu, 2004a; Minematsu, 2004b; Minematsu, 2005). The obtained acoustic representation of speech is regarded as physically-implemented structural phonology because only the interrelations of speech events are focused. The following section briefly introduces how to implement structural phonology on physics, where structuralization of speech events is carried out based upon information theory. After that, it is investigated whether the new representation is reliable enough for pronunciation assessment.

## 2 Physical implementation of structural phonology

### 2.1 Acoustic modeling of the non-linguistic features

In order to delete the non-linguistic features from speech, it is modeled firstly, and then an algorithm for its deletion is implemented. In speech recognition, distortions caused by the non-linguistic events are often classified into three kinds; additive, multiplicative, and linear transformational distortions. Out of the three, the additive distortion (noise) is ignored in this paper because it is not inevitable. Students can turn off a TV set before doing pronunciation practices. If they cannot for some reasons, they can move to another room to obtain clean environment. 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. Gaussian Mixture Modeling (GMM) of speakers indicates that a part of speaker individuality is also regarded as the multiplicative distortion. If a speech event is represented by cepstrum vector $c$, the multiplicative distortion is an addition of $b$ and the resulting cepstrum is shown as $c' = c + b$.

Vocal tract length difference is a typical example of the linear transformational distortion. The difference is often modeled as frequency warping of the log spectrogram, where formant shifts are well approximated. Strictly speaking, two listeners have two different hearing characteristics. Mel or Bark scaling is considered just the average pattern of the characteristics, which can be modeled as another frequency warping of the log spectrum. According to (Pitz, 2003), any monotonously continuous frequency warping of the log spectrum is converted into multiplication of matrix $A$ in cepstrum domain. The resulting cepstrum is shown as $c' = Ac$.

Various distortion sources are found in every step of speech communication. But the total distortion of speech caused by the inevitable sources, $A_i$ and $b_i$, is eventually modeled as $c' = Ac + b$, known as affine transformation.

### 2.2 From acoustic phonetics to structural phonology

In phonology, the non-linguistic features are ignored in researchers' brain and speech sounds are represented as abstract entities named phonemes. Phonology is a speech science to clarify a phonological sys-

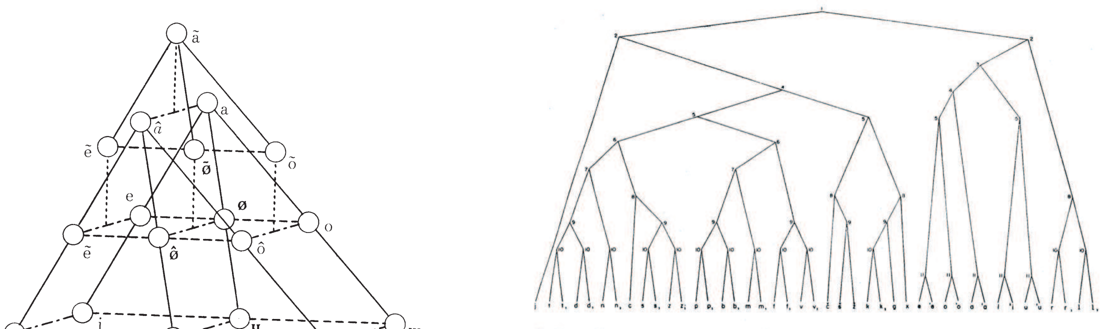

ã
a
ẽ
ɑ̃
œ̃
õ
e
ø
o
ɛ̃
ə
ɔ̃
j
ɥ
w
i
y
u
j t t, d d, n n, c s s, z z; p p, b b, m m, f f, v v, č š ž k k, g x e 'e o 'o a 'a i 'i u 'u r r, l l,

Fig. I-1. Branching diagram representing the morphonemes of Russian. The numbers with which each node is labelled refer to the different features, as follows: 1. vocalic vs. nonvocalic; 2. consonantal vs. nonconsonantal; 3. diffuse vs. nondiffuse; 4. compact vs. noncompact; 5. low tonality vs. high tonality; 6. strident vs. mellow; 7. nasal vs. nonnasal; 8. continuant vs. interrupted; 9. voiced vs. voiceless; 10. sharped vs. plain; 11. accented vs. unaccented. Left branches represent minus values, and right branches, plus values for the particular feature.

Figure 1: Jakobson's geometrical structure of some French phonemes (lefthand) and Halle's tree diagram of the Russian phonemes (righthand)

tem hidden in a language. Inspired by Saussure's claim(Saussure, 1916), Jakobson, Halle, and others discussed a system of the phonemes embedded in a language by using distinctive features(Jakobson, 1952), which were originally proposed by Jakobson. Figure 1 shows Jakobson's geometrical structure proposed for some French phonemes(Jakobson, 1975) and Halle's tree diagram of the Russian phonemes(Halle, 1959). They claimed that the structure is invariant and independent of speakers. Their structuralization of the phonemes is based on distinctive features of the phonemes and, for example, differences in the shape of line segments between two phonemes in Jakobson's structure represent differences of distinctive features between the corresponding two phonemes. In this paper, however, the distinctive features are not used because different linguists claim different sets of the features. Here, the linguists' consciousness of existence of the phonological structure is focused on and the consciousness was raised by a single claim of Saussure on language; "*Language is a system of conceptual differences and phonic differences.*"

The authors are interested in the acoustic aspect of language and only the phonic differences are considered here. Geometrically speaking, Saussure's claim that language is a system of phonic differences can be interpreted as a very simple definition of a structure. In an Euclidean space, an $n$-point structure is uniquely determined by fixing lengths of its ${}_nC_2$ diagonal lines, i.e., all the possible differences among the $n$ points. The differences are formulated by a distance matrix of the $n$ points and, with a bottom-up clustering algorithm, the matrix can produce a tree diagram of the structure. These considerations lead to the following. The distance matrix among the $n$ phonemes in an acoustic space can be regarded as mathematical and physical interpretation of the Saussure's claim and the matrix is geometrically equivalent to the structure itself and can produce the tree diagram shown in Figure 1. Viewing the $n$ elements as a structure indicates that the elements are observed only relatively. The structure extraction can be regarded as a process of ignoring some information in the elements. If it is possible to embed all the sources of the inevitable non-linguistic distortions in the ignored information, the resulting structure is expected to be the acoustic representation which the authors pursue.

### 2.3 Implementation of structural phonology on physics

Phonology claims that the structure is invariant with regard to all the kinds of non-linguistic features, which is mathematically translated that an $n$-point structure (distance matrix) is invariant with any affine transformation. This looks impossible because affine transformation is a transformation which distorts a structure. However, the above claim can be satisfied by the following procedure.

Let phoneme $x$ be represented as distribution $d_x(c)$ in a cepstrum space and distance between two elements (distributions) is calculated by Bhattacharyya distance (BD) measure(Bhattacharyya, 1943; Kailath, 1967).

$$BD(d_x(c), d_y(c)) = -\ln \int_{-\infty}^{\infty} \sqrt{d_x(c)d_y(c)}dc \tag{1}$$

This measure is derived based on information theory and can be interpreted as the amount of self-information of joint probability of the two independent distributions $d_x(c)$ and $d_y(c)$. If the two distributions follow Gaussians, the following is obtained.

$$BD(d_x(c), d_y(c)) = \frac{1}{8}\mu_{xy}^T \left(\frac{\sum_x + \sum_y}{2}\right)^{-1} \mu_{xy} + \frac{1}{2}\ln \frac{|(\sum_u + \sum_v)/2|}{|\sum_u|^{\frac{1}{2}}|\sum_v|^{\frac{1}{2}}} \tag{2}$$

$\mu_x$ and $\Sigma_x$ are the average vector and the variance-covariance matrix of $d_x(c)$, respectively. $\mu_{xy}$ is $\mu_x{-}\mu_y$. Although affine transformation of $c'{=}Ac{+}b$ modifies $\mathcal{N}(\mu, \Sigma)$ into $\mathcal{N}(A\mu + b, A\Sigma A^T)$, BD between $d_x(c)$ and $d_y(c)$ is not changed.

$$BD(A\mu_x + b, A\Sigma_x A^T, A\mu_y + b, A\Sigma_y A^T) = BD(\mu_x, \Sigma_x, \mu_y, \Sigma_y) \tag{3}$$

These facts mean that BD between any two distributions is not changed by any of an affine transformation and that the structure composed of the $n$ phonemes is not changed. Multiplication of $A$ and addition of $b$ are geometrically interpreted as rotation and shift of the structure, respectively. For example, acoustic changes of speech caused by increase of vocal tract length, i.e., human growth, is mathematically regarded as very slow rotation of the structure which takes about 15 years. When $d_x(c)$ and $d_y(c)$ are modeled as Gaussian mixtures, the invariance is still valid because the structure of all the component Gaussians is not changed at all. Now, the desired acoustic representation is gracefully derived.

## 3 Description of the individual students

### 3.1 Speech database used in the analysis

ERJ (English Read by Japanese) database(Minematsu, 2004c) was used, which contains English sentences read by 202 Japanese students, Japanese English (JE), and 20 native speakers of General American (GA). Table 1 shows conditions for the acoustic analysis. Mathematically speaking, the variance-covariance matrix of an HMM should be a full matrix to allow rotation of the structure. This condition might cause some distortions in results of the experiments. Phoneme-to-phoneme distance is defined as average distance over the three state-to-state BDs between two phonemes.

### 3.2 Structural representation of the individual students

Figure 2 shows two examples of structural description of an American teacher of English and a poor student. The Japanese tree clearly shows the well-known Japanese habits of English pronunciation. Confusions of /r/&/l/, /s/&/th/, /z/&/dh/, /f/&/h/, /iy/&/ih/, /v/&/b/, and so on are found. Mid and low vowels are closely located to each other because there is only one mid and low vowel in Japanese. Schwa is also found close to them. Technically speaking, it is possible enough to describe about 2 billion English pronunciations individually based upon their phonological structures. The proposed method can

Table 1: *Conditions for the acoustic analysis*

| | |
|---|---|
| sampling | 16bit / 16kHz |
| window | 25 ms length and 10 ms shift |
| parameters | FFT-based cepstrums and their derivatives |
| speakers | 202 Japanese and 20 Americans |
| training data | 60 sentences per speaker |
| HMMs | speaker-dependent, context-independent, and 1-mixture monophones with diagonal matrices |
| topology | 5 states and 3 distributions per HMM |
| monophones | b,d,g,p,t,k,jh,ch,s,sh,z,zh,f,th,v,dh,m,n,ng,l,r, w,y,h,iy,ih,eh,ae,aa,ah,ao,uh,uw,er,ax |

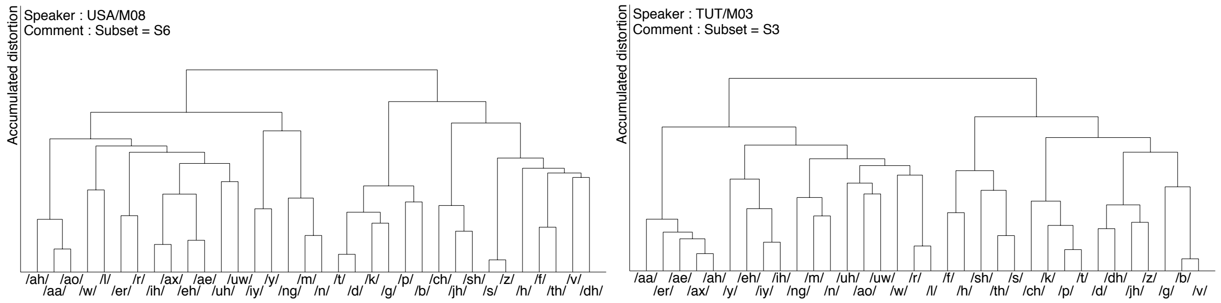

Speaker : USA/M08
Comment : Subset = S6
Accumulated distortion
/ah/ /ao/ /l/ /r/ /ax/ /ae/ /uw/ /y/ /m/ /t/ /k/ /p/ /ch/ /sh/ /z/ /f/ /v/
/aa/ /w/ /er/ /ih/ /eh/ /uh/ /iy/ /ng/ /n/ /d/ /g/ /b/ /jh/ /s/ /h/ /th/ /dh/
Speaker : TUT/M03
Comment : Subset = S3
Accumulated distortion
/aa/ /ae/ /ah/ /eh/ /ih/ /m/ /uh/ /uw/ /r/ /f/ /sh/ /s/ /k/ /t/ /dh/ /z/ /b/
/er/ /ax/ /y/ /iy/ /ng/ /n/ /ao/ /w/ /l/ /h/ /th/ /ch/ /p/ /d/ /jh/ /g/ /v/

Figure 2: *Structural representation of an American teacher of English(lefthand) and a poor student(righthand)*

also represent a change found in a student's structure before and after a pronunciation lesson given to the student. What kind of pronunciation changes into what kind of pronunciation with what kind of pronunciation lesson? The proposed method enables recording history of changes found in a student's pronunciation. If history of the changes of many students is once stored, a new student may get better instructions by referring to the history of the old students.

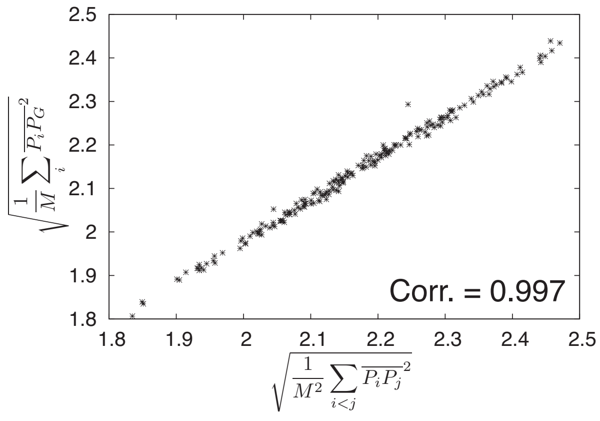
It should be noted that the representation contains only the acoustic interrelations of speech events with no absolute acoustic properties of the individual events. Even if all the phones are modeled with this method, the entire model cannot recognize a single phone input because it does not have any absolute information on the individual phones. For the same reason, the entire model cannot synthesize any phones. What's possible, then? In the following discussions, it is shown that the interrelational model of all the speech events can do a very good job.

## 4 Distance measure between two structures

In this section, distance measure between two structures, namely, two speakers, is investigated. If an $M$-point structure, $P$, exists in an Euclidean space, the following equation is true, where $P_G$ is a gravity center of $\{P_i\}$.

$$\sqrt{\frac{1}{M^2}\sum_{i<j}\overline{P_iP_j}^2} = \sqrt{\frac{1}{M}\sum_{i}\overline{P_iP_G}^2} \qquad (4)$$

If BD is used for Euclid distance, the above equation is not satisfied. But $\sqrt{\mathrm{BD}}$ satisfies the equation approximately. Figure 3 shows values of the left and the right terms of the above equation calculated



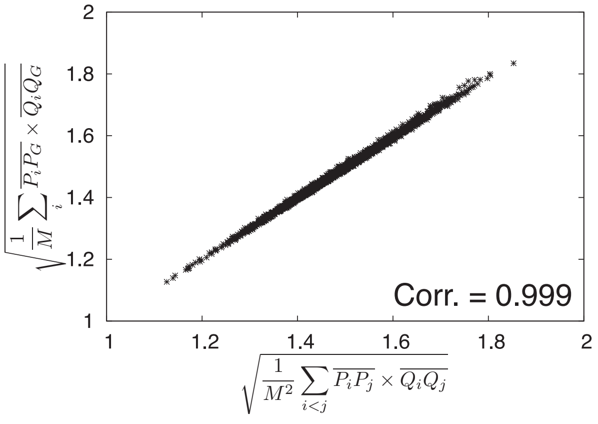


Figure 3: $\sqrt{BD}$ *approximately satisfies Eqn. (4).*

Figure 4: $\sqrt{BD}$ *approximately satisfies Eqn. (5).*

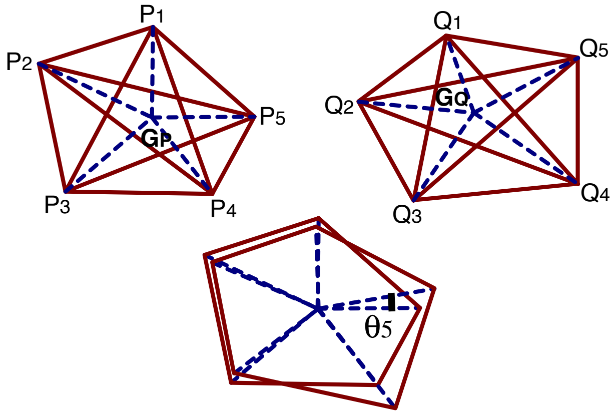


Figure 5: *Two structures and their shift & rotation for fitting*

from all the individual students with their English vowel models. The same tendency was found with the consonant HMMs. In the following discussions, BD denotes $\sqrt{\text{BD}}$. Now, let us consider two structures, $P$ and $Q$. If $M$ points are phones in a cepstral space with their distributions, then the following equation is approximately true for JE phones.

$$\sqrt{\frac{1}{M^2}\sum_{i<j}\overline{P_iP_j}\times\overline{Q_iQ_j}} \approx \sqrt{\frac{1}{M}\sum_{i}\overline{P_iP_G}\times\overline{Q_iQ_G}} \qquad (5)$$

Figure 4 shows the both terms calculated from any two of the students with their vowel models. It is the case with the consonant models. The above two equations lead to the following.

$$\sqrt{\frac{1}{M^2}\sum_{i<j}(\overline{P_iP_j}-\overline{Q_iQ_j})^2} \approx \sqrt{\frac{1}{M}\sum_{i}(\overline{P_iP_G}-\overline{Q_iQ_G})^2} \qquad (6)$$

The right term is approximation of averaged cepstrum distance over all the corresponding phone pairs between the two structures *after shift and rotation*, where the two gravity centers are put at a position and one of the two structures is rotated so that the $\sum|\theta_i|$ (see in Figure 5) should be minimized. The left term is Euclid distance between two distance matrices by viewing a matrix as a vector. In brief, Euclid distance between two matrices, structural distortion, approximates cepstrum distance averaged over all the corresponding phone pairs of the two structures after full adaptation with regard to $A$ and $b$.

## 5 Automatic scoring of the proficiency

### 5.1 Preliminary discussions of the structural comparison

Figures 6 and 7 show the structural distortion, which is defined in the previous section, and the positional distortion, which is defined as the averaged cepstrum distance with no shift or rotation, for two cases. One is the distortion between two GA speakers (GA-GA) and the other is that between a GA and a Japanese speakers (GA-JE). In Figure 6, only the vowels are used and, in Figure 7, all the phones but diphthongs are used. In the former, while GA-JE and GA-GA distributions are overlapped in the positional distortion, they are clearly separated in the structural distortion. This was much to be expected because the two distortions differ in whether adaptation is done or not. In the latter, however, the structural distortion shows less clear separation. The authors consider two reasons. One is that phoneme-to-phoneme distance is simply defined as average of the three state-to-state distances although a dominant state is expected to exist among the three states. The other is in the form of the variance-covariance matrix, which should have been a full matrix to allow rotation of the structure. The better conditions will be examined in future works and in this paper, for automatic assessment, an adequate selection of the phone pairs is done.

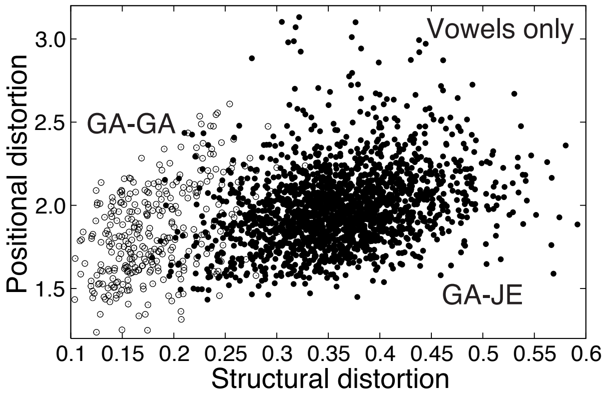


Figure 6: *Structural and positional distortions for vowels*

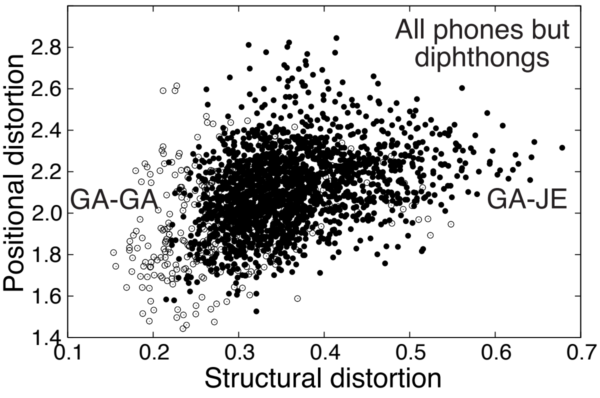


Figure 7: *Structural and positional distortions for all the phones but diphthongs*

### 5.2 Automatic scoring of the pronunciation proficiency

Student $i$ in ERJ has his/her pronunciation score $s_i$ rated by 5 American teachers ($1 \leq s_i \leq 5$). Then, the phone pair selection was done so that correlation between $5 - s_i$ and the structural distortion between student $i$'s structure and the teacher's one should be maximized. The number of the selected phone pairs is 52. Figure 8 shows results of automatic scoring of the pronunciation proficiency based upon the structural distortion. Good correlation is obtained between the two quantities.

## 6 Two different pronunciations of a single speaker

An interesting experiment was carried out with two different pronunciations of a single speaker. The first author is a Japanese and was an amateur actor of an English drama club. On the stage, he was requested to pretend to be an American and mastered how to control muscles around the mouth and the belly and how to control air flow from the lung. The first author considers that the English way of control of the air flow is rather different from the Japanese way of control. Four pronunciations were prepared, shown in Table 2. Two are from a male (M) and a female (F) Americans. The other two are the first author's normal pronunciation (A) and his intentionally Japanized pronunciation (B). Speaker-dependent HMMs were built from (F), (M), and (B). Acoustic similarity between samples of (A) and the individual models was calculated in the following three ways.

- With the normal likelihood score of $P(o|M)$.

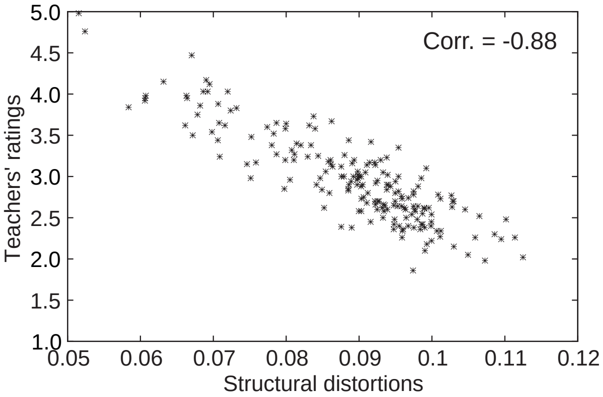


Figure 8: *Proficiency assessment with the structural distortion*

Table 2: *Four kinds of pronunciations used in the experiment*

| spk | USA(F) | USA(M) | author(A) | author(B) |
|---|---|---|---|---|
| gender | F | M | M | M |
| age | 50 | 46 | 36 | 36 |
| mic | SEN | SEN | cheap | cheap |
| room | SP | SP | living | living |
| AD | DAT | DAT | laptop | laptop |
| pron. | perfect | perfect | good | Japanized |

SEN = Senheiser, SP = Sound-proof

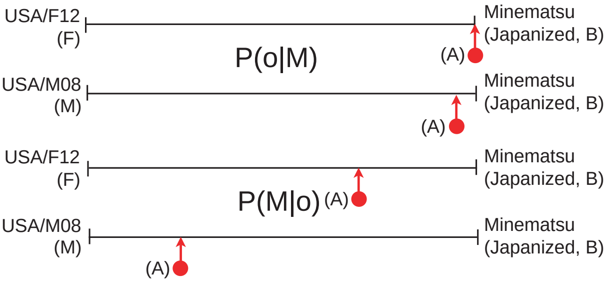


Figure 9: *Proficiency rating with $P(o|M)$ and $P(M|o)$*

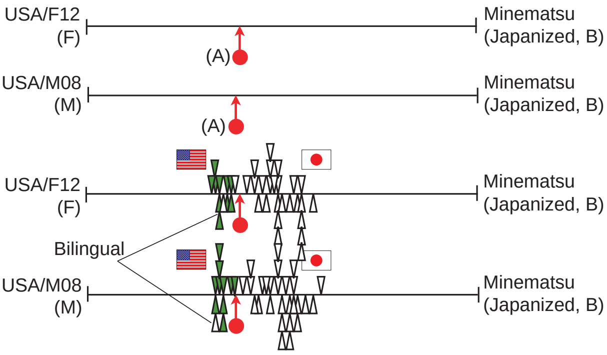


Figure 10: *Proficiency rating with the structural distortion*

- With the posteriori probability score of $P(M|o)$.
- With the proposed structural distortion score.

$P(o|M)$ is acoustic likelihood score between an input speech sample ($o$, observation) and an acoustic model ($M$, model). This score is usually used in speech recognition and mathematically interpreted such that observation $o$ is generated as output of model $M$ with a probability of $P(o|M)$. $P(M|o)$ is often used in CALL systems to normalize differences in compatibility between an input speaker and the acoustic models. If the first author's normal pronunciation (A) should be pedagogically judged to be closer to (F) than to (B), the authors can claim that Table 2 is the most difficult condition for speech technology. This is because, between (A) and (F), everything is mismatched except for the proficiency and because, between (A) and (B), everything is matched except for the proficiency.

Figure 9 shows results with $P(o|M)$ and $P(M|o)$, where (A) is placed between the two models proportionally to the similarity scores. With $P(o|M)$, (A) is almost completely the same as (B), which was much to be expected because (A) and (B) are from the same speaker. Although $P(M|o)$ is often used for compatibility normalization, Figure 9 shows that it does not always work. This sometimes happens in actual classrooms and this is why the conventional CALL systems are sometimes criticized. Figure 10 shows results with the structural distortion. The other Japanese and Americans (set 6 in ERJ) are also plotted. White and green(or gray) triangles represent Japanese and Americans, respectively. Above and below the line represent female and male speakers, respectively. It is surprising that all the Japanese students but the only bilingual speaker are judged *acoustically* closer to the author (B) than the author himself (A) is. This can never happens if direct spectrogram-to-spectrogram matching is done. This is because the spectrogram shows every acoustic aspect of an event and can be considered as rather noisy for pronunciation assessment. The authors believe that the education should be supported only by the reliable and stable technology.

## 7 Conclusions

This paper firstly introduces a novel representation method of speech acoustics, where static and inevitable non-linguistic features are well discarded. Speech events are modeled probabilistically as distributions, distance between any two of the events is calculated based upon information theory, and the events are relatively captured as a structure. The resulting structure is mathematically invariant with any static non-linguistic features. The proposed structural representation of speech is regarded as physical implementation of structural phonology and also as mathematical interpretation of Saussure's claim on language. This paper shows that the proposed method can describe the individual pronunciations, possibly about 2 billion pronunciations, and also examines whether the method can realize good and reliable assessment of the pronunciation. Experiments showed that the method can assess the pronunciation proficiency accurately and its reliability and stability is remarkably high. The authors are further trying to increase the reliability and examining the method with children's voices.

## References

S. Hiller, E. Rooney, J. Laver, and M. Jack, "SPELL: an automated system for computer-aided pronunciation teaching," Speech Communication, vol.13, pp.463–473 (1993)

J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, "Automatic evaluation and training in English pronunciation," Proc. Int. Conf. Spoken Language Processing(ICSLP), pp.1185–1188 (1990)

C. Cucchiarini, F. de Wet, H. Strik, and L. Boves, "Assessment of Dutch pronunciation by means of automatic speech recognition technology," Proc. Int. Conf. Spoken Language Processing(ICSLP), pp.1739–1742 (1998)

A. Neri *et al.*, "Automatic speech recognition for second language learning: how and why it actually works," Proc. Int. Congress on Phonetic Science(ICPhS), pp.1157–1160 (2003)

N. Minematsu, "Yet another acoustic representation of speech sounds," Proc. Int. Conf. Acoustics, Speech, and Signal Processing(ICASSP), pp.585–588 (2004a)

N. Minematsu, S. Asakawa, and K. Hirose, "The acoustic universal structure in speech and its correlation to para-linguistic information in speech," Proc. Int. Workshop on Man-Machine Symbiotic Systems(IWMMS), pp.69-79 (2004b)

N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," Proc. Int. Conf. Acousitcs, Speech, and Signal Processing(ICASSP) (2005, accepted)

M. Pitz and H. Ney, "Vocal tract normalization as linear transformation of MFCC," Proc. EUROSPEECH, pp.1445–1448 (2003)

F. Saussure, "Cours de linguistique general," publie par Charles Bally et Albert Schehaye avec la collaboration de Albert Riedlinge, Lausanne et Paris, Payot (1916)

R. Jakobson, G. Fant, and M. Halle, "Preliminaries to speech analysis: the distinctive features and their correlates," MIT Press, Cambridge (1952)

R. Jakobson and M. Halle, "Fundamentals of language," The Hague: Mouton (1975)

M. Halle, "The sound patterns of Russian," The Hague: Mouton (1959)

A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," Bulletin of Calcutta Maths Society, vol.35, pp.99–110 (1943)

T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," IEEE transaction on communication technology, vol.15, no.1, pp.52–60 (1967)

N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "Development of English speech database read by Japanese to support CALL research," Proc. Int. Conf. Acoustics(ICA), pp.557–560 (2004c)