Pronunciation Assessment Based upon the Phonological Distortions Observed in Language Learners' Utterances

Nobuaki MINEMATSU

Graduate School of Information Science and Technology, University of Tokyo
mine@gavo.t.u-tokyo.ac.jp

Abstract

Speech representation provided by acoustic phonetics, spectrogram, is very noisy representation in that it shows every acoustic aspect of speech. Age, gender, size, shape, microphone, room and line are completely irrelevant to speech recognition, pronunciation assessment, and so on. But the spectrogram is affected easily by these factors. This is the very essential reason why speech systems are sometimes unreliable and the author supposes that the education should not endure this inevitable characteristics. The author proposed a novel method of *acoustic* representation of speech where no dimensions of the above factors exist. The method was derived by implementing structural phonology on physics. This paper examines whether the new representation of speech can provide a good tool of pronunciation assessment. Results of the experiments with good and intentionally-bad pronunciations of a single speaker showed that all the students are *acoustically* located between the two pronunciations, indicating that all 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 and is extremely reliable and effective in CALL.

1. Introduction

Pronunciation training should be based upon articulatory phonetics because a speech sound is produced adequately only by the correct articulation. But it is very difficult and expensive to measure or estimate movements of the articulators of *students*, and then training on articulatory phonetics requires specialists. It is true that all the language teachers cannot be specialists. As an alternative to articulatory phonetics, it has been investigated whether speech representation of acoustic phonetics, spectrogram, can be a good tool for pronunciation assessment. In a meaning, the spectrogram can show well how good the pronunciation is because teachers can judge goodness of the segmental aspect of the pronunciation by *hearing* the spectrogram. In another meaning, however, the spectrogram cannot show well goodness of the pronunciation because teachers cannot judge it by *looking at* the spectrogram. Teachers expected speech engineering, i.e., computers, to judge it well by looking at the spectrogram, and then CALL systems were developed with speech recognition technologies. The question is whether computers can do reliable and pedagogically-sound enough judgment. In the beginning, CALL systems were accepted to teachers and students because students could have virtual teachers anytime and anywhere with multimedia attractions. But recently, some papers report unreliability and instability of the systems[1]. Native speakers are sometimes judged to be worse than students, for example. Strictly speaking, the spectrogram is a noisy representation of speech in that it shows every acoustic aspect of speech. 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. By collecting a large amount of data, speaker-independent models are built. But they often require speaker adaptation techniques, which implies that the speaker-independent models are not really speaker-independent. Collection of more data, i.e., quantitative solution, seems not to work pedagogically-sound enough.

A novel and qualitative solution was proposed by the author. Deletion of the non-linguistic information was done not by collecting data but by deleting all the dimensions mathematically to represent the irrelevant things from speech[2]. The obtained acoustic representation of speech is regarded as physically-implemented phonology because only the interrelations of speech events are focused. The following section briefly introduces how to implement 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 effective enough for pronunciation assessment.

2. Physical implementation of phonology

2.1. Acoustic modeling of the non-linguistic information

In order to delete the non-linguistic information 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 because it is not inevitable. Students can turn off a TV set before learning English. 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. GMM speaker modeling 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 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 spectrum, where formant shifts are well approximated. According to [3], any monotonously continuous frequency warping of the log spectrum is mathematically 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 phonetics to phonology

In phonology, the non-linguistic information is ignored in researchers' brain and speech sounds are represented as abstract entities named phonemes. Phonology is a science to clarify a phonemic system hidden in a language. Inspired by Saussure's structuralism, Jakobson, Halle, and others have discussed structure of the phonemes embedded in a language with distinctive features[4] and drew a tree diagram of the phonemes. Classification of the phonemes is done so that a set of phonemes under every node of the tree comprise a natural class. In phonology, the structure is extracted in a top-down way based upon researchers' knowledge on the language. In this work, the structure is determined in a bottom-up way where not knowledge but distance measure between two elements is required. An n-point structure is represented uniquely by distance matrix among the n points. Viewing n elements as structure means that the elements are observed only relatively and 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 non-linguistic information in the ignored information, the resulting structure will be the desired acoustic representation.

2.3. Implementation of phonology on physics

Phonology claims that the structure is universal with regard to all the kinds of non-linguistic information, which is mathematically translated that an n-point structure (distance matrix) is invariant with any affine transformation. This looks impossible, which can become possible 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.

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

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.

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

μ_x and Σ_x are the average vector and the variance-covariance matrix of $d_x(c)$, respectively. μ_{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.

$$\begin{aligned} &BD(A\mu_x + b, A\Sigma_x A^T, A\mu_y + b, A\Sigma_y A^T) \\ &= BD(\mu_x, \Sigma_x, \mu_y, \Sigma_y) \end{aligned} \qquad (3)$$

These facts mean that BD between any two of the n distributions (phonemes) 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 cannot be changed at all. Now, the desired acoustic representation is gracefully derived.

3. Distance measure between two structures

3.1. Speech database used in the analysis

ERJ (English Read by Japanese) database[5] 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[1] and phoneme-to-phoneme distance is defined as average distance over the three state-to-state BDs between two phonemes.

3.2. What's possible with the new representation?

With the proposed representation, the pronunciation of a student is modeled as a structure composed of the n phonemes and this structure is visualized by a tree diagram, for example. Figure 1 shows a tree example of a poor Japanese student.

It should be noted that the representation contains only the acoustic interrelations of speech events with no absolute acoustic properties of the individual events. If all the phones are modeled with this method, the entire model cannot recognize even 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? In the following discussions, it is shown that the interrelational model of all the speech events can do a very good job.

3.3. Distance measure between two structures

If an M-point structure, P, exists in Euclidean space, the following equation is true, where P_G is a gravity center of $\{P_i\}$.

$$\sqrt{\frac{1}{M^2} \sum_{i<j} \overline{P_i P_j}^2} = \sqrt{\frac{1}{M} \sum_i \overline{P_i P_G}^2} \qquad (4)$$

If BD is used for Euclid distance, the above equation is not satisfied. But $\sqrt{\text{BD}}$ satisfies the equation approximately. Figure 2

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

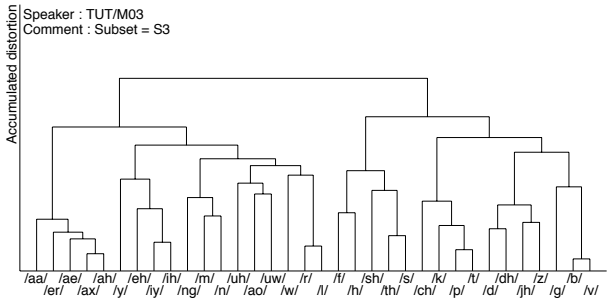

Figure 1: *A structurally represented poor Japanese student*

[1] 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.

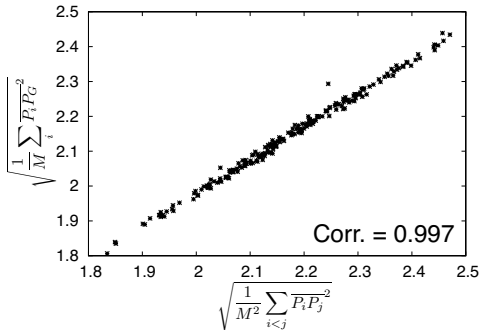

Figure 2: $\sqrt{BD}$ *approximately satisfies Equation (4).*

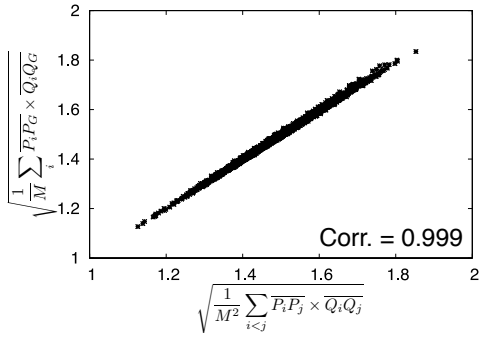

Figure 3: $\sqrt{BD}$ *approximately satisfies Equation (5).*

shows values of the left and the right terms of the above equation calculated 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.

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

Figure 3 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.

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

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 4) 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.

4. Automatic scoring of the proficiency

4.1. Preliminary discussions of the structural comparison

Figures 5 and 6 show the structural distortion and the positional distortion, which is defined as the averaged cepstrum distance

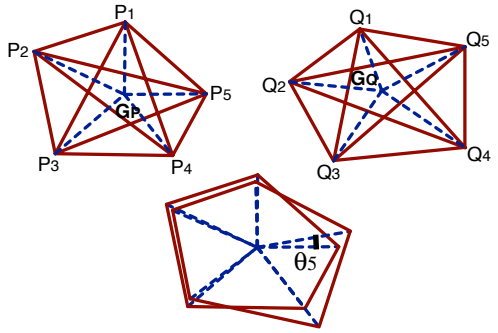

Figure 4: *Two structures and their shift & rotation for fitting*

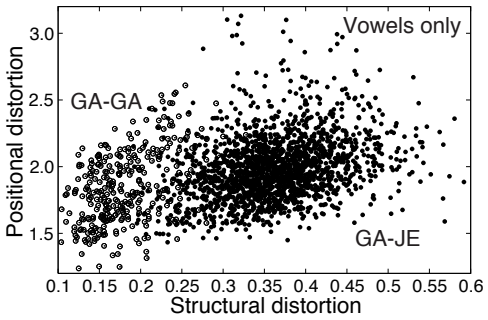

Figure 5: *Structural and positional distortions for vowels*

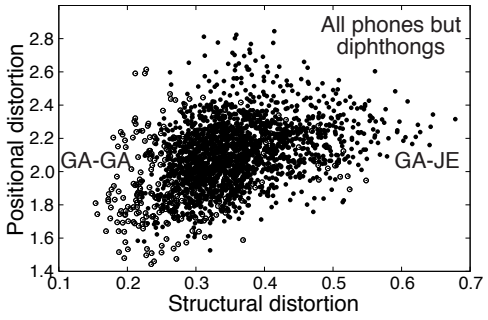

Figure 6: *Structural and positional distortions for all the phones*

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 speaker (GA-JE). In Figure 5, only the vowels are used and, in Figure 6, all the phones 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 author considers 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 highly expected among the three states. The other is 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, adequate selection of the phone pairs are done.

4.2. Automatic scoring of the pronunciation proficiency

Student i in ERJ has his/her pronunciation score s_i rated by 5 American teachers ($1 \le s_i \le 5$). Then, the phone pair selection was done so that correlation between $5 - s_i$ and the struc-

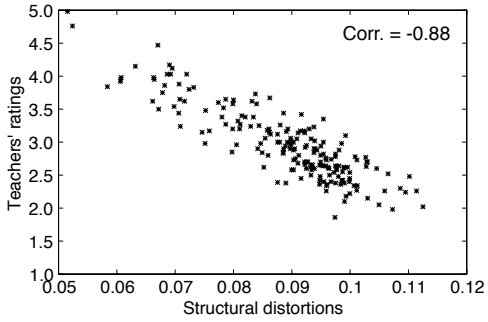

Figure 7: *Proficiency assessment with the structural distortion*

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

tural distortion between student i's structure and the teacher's one should be maximized. The number of the selected phone pairs is 52. Figure 7 shows results of automatic scoring of the pronunciation proficiency based upon the structural distortion. Good correlation is obtained between the two quantities.

5. Two different pronunciations of a speaker

An interesting experiment was carried out with two different pronunciations of a single speaker. The 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 very well how to control muscles around the mouth and the belly and how to control air flow, which was perceived by the author as rather different from the Japanese way of control. Four pronunciations were prepared, shown in Table 2. Two are a male (M) and a female (F) Americans. The other two are the author's normal pronunciation (A) and his intentionally Japanized pronunciation (B). Speaker-dependent HMMs were built for (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 $P(o|M)$.
- With the posteriori probability score of $P(M|o)$.
- With the proposed structural distortion score.

$P(M|o)$ is often used in CALL systems to normalize differences in compatibility between an input speaker and the acoustic models. If the author's normal pronunciation (A) should be pedagogically judged to be closer to (F) than to (B), the author 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 8 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 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

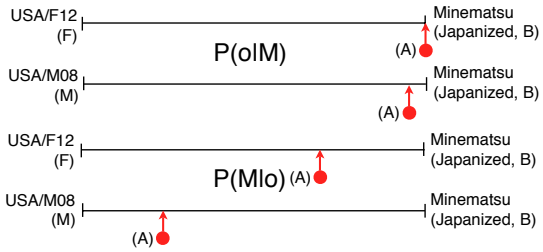

Figure 8: *Proficiency rating with $P(o|M)$ and $P(M|o)$*

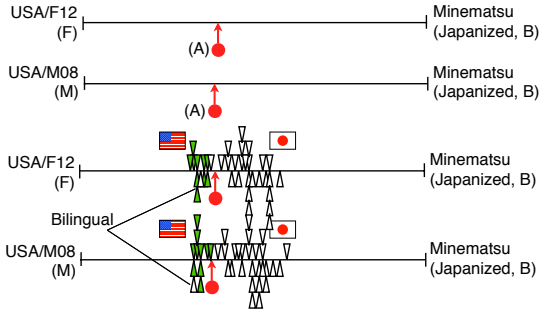

Figure 9: *Proficiency rating with the structural distortion*

why the conventional CALL systems are sometimes criticized. Figure 9 shows results with the structural distortion. The other Japanese and Americans (set 6 in ERJ) are also plotted. White and green 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 a 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 author believes that the education should be supported only by the reliable and stable technology.

6. Conclusions

This paper investigates whether the phonological representation of speech, recently proposed by the author, can realize good and reliable assessment of the pronunciation. Although it is impossible to recognize or synthesize a single phone based upon the proposed representation, it can assess the pronunciation proficiency accurately and its reliability and stability is remarkably high. The author is further trying to increase the reliability and examining the method with children's voices.

7. References

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

[2] N. Minematsu, "Yet another acoustic representation of speech sounds," Proc. ICASSP, pp.585–588 (2004)

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

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

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