Training of Pronunciation as Learning of the Sound System Embedded in the Target Language

Nobuaki Minematsu

Graduate School of Engineering, The University of Tokyo
mine@gavo.t.u-tokyo.ac.jp

Abstract

The current speech recognition technology consists of acoustic models, language models, a pronunciation dictionary, and a decoder. Computer-aided pronunciation training systems often use the acoustic matching module to compare a learner's utterance to its corresponding template stored in the systems. The template is usually calculated through collecting native utterances of that phrase and estimating its acoustic distribution. On this framework, a learner's utterance is acoustically compared to its distribution and the notorious mismatch problem happens more or less inevitably. The author claims that this framework has a serious problem because learners don't have to imitate the utterances (voices) of a teacher acoustically. What in a teacher's utterances should be physically imitated by learners? This question was answered by the author through deriving mathematically and linguistically a speaker-invariant sound pattern embedded in an utterance [1, 2]. Based on this answer, a novel technique was realized for CALL system development [3, 4, 5]. This paper describes the answer and the new technique proposed in the previous studies and shows that the new technique can also provide a very motivating interface with learners.

1. Introduction

Developmental psychology tells that infants acquire spoken language through imitating the utterances of their parents, called vocal imitation [6]. But no infants try to imitate the voices. As they have little phonemic awareness [7, 8], they cannot identify a sound as phoneme although they can discriminate two different sounds very well. They cannot decode the speech into sequence of phonemes or convert the phonemes into sounds. In this situation, what in a father's speech is acoustically imitated by infants? Researchers claim that they firstly learn the holistic sound pattern of the word [7, 8], called word Gestalt. Then, what is the acoustic definition of that word Gestalt? If it carries speaker information, many infants must try to produce their fathers' voices. This consideration indicates that the word Gestalt has to be speaker-invariant. But what is that acoustically? I asked this question to many researchers in some conferences on infant study [9] but no researchers gave me a definite answer. If the word Gestalt could be defined acoustically, I'm wondering whether it might be something like the linguistic skeleton.

No infants imitate the voices but myna birds imitate sounds of cars, doors, and animals as well as human voices. Hearing a very adept myna bird say something, one can guess its keeper [10]. Hearing a very good child say something, however, it is impossible to guess its keeper. If one trains a myna bird to be a better imitator, the bird's voice and the target sound will be acoustically and directly compared and, to reduce the difference, another training method will be examined. Most of the CALL systems acoustically compare an input utterance to the distribution estimated from its corresponding native utterances. This fact claims that the systems assume that learners are myna birds to the distributions calculated from native utterances. Is this assumption correct and pedagogically-sound enough?

Figure 1: The tallest man (7.9ft) and the shortest adult (2.4ft)

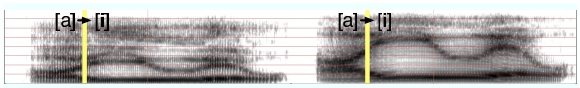

Figure 2: /aiueo/s produced by a tall speaker and a short speaker

Figure 1 shows a picture of the tallest man and the shortest adult in the world. Figure 2 shows two /aiueo/s generated by a pseudo-speaker as tall as the tallest man and a pseudo-speaker as short as the shortest adult. It is very easy to perceive that the two utterances carry an equivalent linguistic content and it is the case even for young children. Considering their weak phonemic awareness, the equivalence perception is not based on string-based (symbol-based) comparison between the utterances but based on Gestalt perception. It should be noted that the Gestalt-based equivalence perception does not seem to require sound-to-phoneme (symbol) conversion. But the question here is what is the acoustic definition of the speaker-invariant patterns embedded in the two utterances in Figure 2.

The question I'm addressing is one of the classical but still unsolved questions in speech science, which is variability of speech acoustics and invariance of speech perception [11]. By considering a very good similarity between language and music, I proposed a novel answer [1, 2]. Some people can identify individual tones in a given melody as syllable names. This identification is done independently of the key of that melody. If a melody is transcribed as sequence of syllable names such as So-Mi-So-Do, then, its transposed version into any key is also transcribed as So-Mi-So-Do. It is interesting that the listeners perceive the internal and silent voice of "Mi" for acoustically different tones, for example. This key-invariant perception is known to owe much to relative pitch perception. In [1, 2], relative *timbre* perception was hypothesized and examined. Based on this hypothesis, the speaker-invariant pattern was mathematically derived and used for CALL application [3, 4, 5].

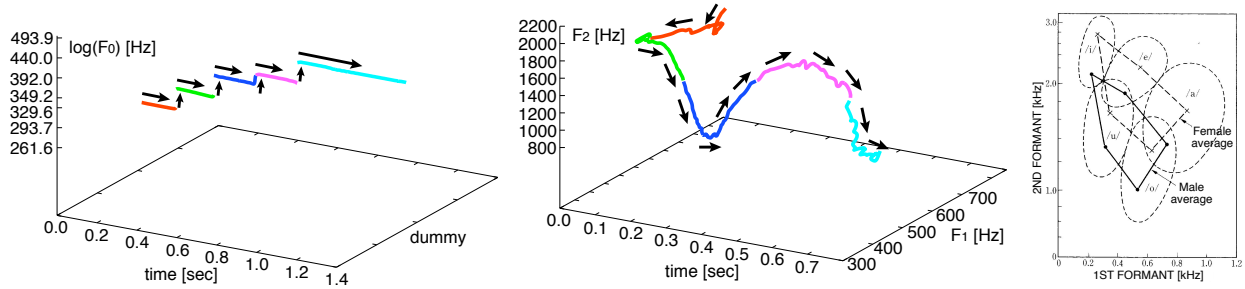

Figure 3: Dynamic changes of pitch in CDEFG and those of timbre in /aiueo/ with the Japanese vowel chart

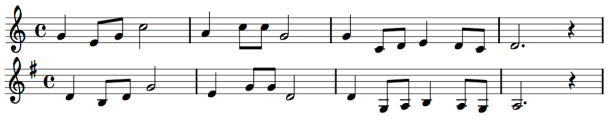

Figure 4: A musical piece and its transposed version

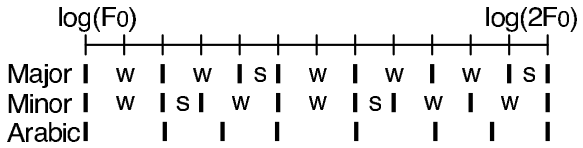

Figure 5: Three musical scales of Major, Minor, and Arabic

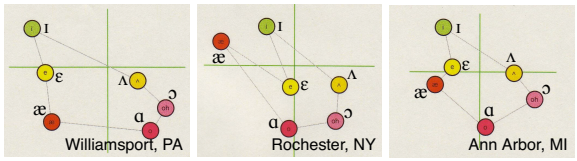

Figure 6: Accented pronunciations of American English [14]

2. Speaker-invariant representation

2.1. Key-invariant representation of melody

A musical piece (C-major) and its transposed version (G-major) are shown in Figure 4. Hearing them, it is usually easy to perceive that they carry an equivalent melodic content although they are acoustically different. People with relative pitch (RP) who can verbalize a melody contour perceive the same string of syllable names, i.e. So-Mo-So-Do La-Do-Do-So, in the above two pieces. People with absolute pitch (AP) can easily transcribe the first piece, using pitch names, as G-E-G-C A-C-C-G and the second one as D-B-D-G E-G-G-D. If they have very strong AP, they come to have some difficulty in perceiving the equivalence [12]. They have to transform the first symbol sequence into the second consciously. A large number of people cannot transcribe a melody. They have RP but cannot verbalize. It is true, however, that they can easily perceive the equivalence between the two pieces even *with no tonal symbols*. In psychology, what they perceive is regarded as melody Gestalt.

As described above, developmental psychology claims that young children perceive the holistic and equivalent sound pattern embedded in the two utterances in Figure 2, called word Gestalt. It also claims that this perception requires no phonemic or phonetic symbols. In [1, 2], the Gestalt was mathematically introduced and its experimental validity was examined.

Melody consists of a dynamically changing pattern of pitch. Figure 3 shows a pitch contour of CDEFG. Between this and its transposed version, I can derive easily the acoustic definition of the key-invariant tonal pattern. If a melody is represented as a sequence of local and relative pitch changes, namely, $\Delta F0_t$ ($=F0_t - F0_{t-1}$), the sequence becomes key-invariant. In Western music, an octave is divided into 12 semi-tone intervals and a musical scale is composed of 8 tones, which have 5 whole-tone intervals (Ws) and 2 semi-tone ones (Ss). It is important that the tones' relative arrangement is invariant with key. This is why a $\Delta F0_t$ sequence is key-invariant. Further, RP people who can carry out key-invariant transcription of a melody use this invariant arrangement [13]. Figure 5 shows two well-known musical scales, Major and Minor, and a very local scale, Arabic. If a Western melody is played in the Arabic scale, RP people are expected to have difficulty in transcribing the melody as a syllable name sequence. This expectation is very valid when they are unfamiliar with the Arabic tonal arrangement.

2.2. Speaker-invariant representation of speech

If only resonant linguistic sounds are considered, speech is a dynamically changing pattern of timbre. Figure 3 also shows a timbre contour of /aiueo/. Can the timbre contour be speaker-invariant like melody? Transposition of melody translates the pitch contour but the shape of the contour is not altered. The Japanese F_1/F_2-based vowel chart is also shown in Figure 3. It seems that the male vowel system is translated to fit to the female version in which the vowel arrangement is not changed. It is regarded as multi-dimensional transposition. As shown in Figure 5, a different tonal arrangement gives us a different color of music. Similarly, as shown in Figure 6, a different timbre arrangement gives us a different color of language, namely, regional accents. In both cases, not the sounds themselves but the sound system may play a critical role in invariant perception.

If this vowel system invariance is always satisfied with any kind of the non-linguistic factors, the timbre contour or trajectory can be considered as speaker-invariant representation, i.e. word Gestalt. As explained shortly, however, this simple derivation does not work at all mathematically and experimentally.

2.3. Directional dependence of cepstrum on speakers

Difference of the vocal tract length changes formant frequencies. If it becomes shorter or longer, they will become higher or lower, respectively. This change is often modeled as frequency warping of a spectrum envelope in the spectral domain and as multiplying matrix A ($=\{a_{ij}\}$) in the cepstral domain [15].

$$a_{ij} = \frac{1}{(j-1)!} \sum_{m=m_0}^{j} \binom{j}{m} \frac{(m+i-1)!}{(m+i-j)!} (-1)^{(m+i-j)} \alpha^{(2m+i-j)},$$

where $|\alpha| \leq 1.0$, $m_0 = \max(0, j-i)$, and

$$\binom{j}{m} = \begin{cases} {}_jC_m & (j \geq m) \\ 0 & (j < m). \end{cases}$$

Using $c' = Ac$, it is possible to convert a speech sample of a male adult into that of a boy. I carried out a geometrical analysis of

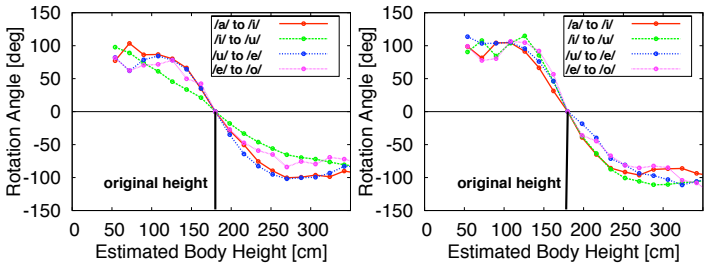

Figure 7: Rotation of cepstrum vectors and their Δs

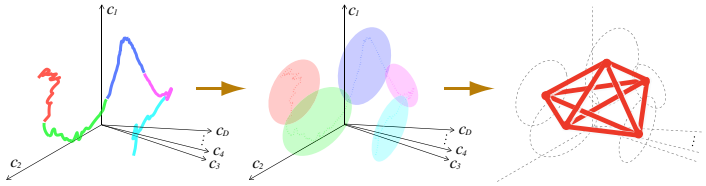

Figure 8: Distribution-based structuralization

this matrix and found that A has a very strong function in rotating cepstrum vectors although A is not a complete rotation matrix [16]. In other words, A approximately satisfies the conditions of rotation matrix; $A^tA{=}AA^t{=}I$ and $|A| = 1$.

Figure 2 shows two speech samples of /aiueo/; the original (male adult) and its warped version (boy) using A. Figure 7 shows some results of analyzing the relation between rotation angles and the degree of body height change through warping. Using different values of α, an /aiuoe/ utterance of a male adult was warped into that of speakers of different heights. The original height was 167 cm and the height was changed into 50 cm to 350 cm. From these utterances, four fixed points were detected; the central transition points of /a/ to /i/, /i/ to /u/, /u/ to /e/, and /e/ to /o/. The /a/ to /i/ transition points are shown in Figure 2. I can say that cepstrum rotation and Δ cepstrum rotation are very similar and that a father's direction and his son's direction are almost orthogonal. Based on this mathematical and experimental fact, I claim that the timbre contour or trajectory is not a good answer to the question of "What is the acoustic definition of speaker-invariant word Gestalt?" It should be noted that pitch rotation is impossible because pitch is one-dimensional. Since timbre is multi-dimensional, its rotation is possible.

2.4. Robust and structural invariance in speech

If matrix A is a complete rotation matrix, speaker-invariant features can be obtained as follows. A speech stream is converted into a sequence of N cepstrum vectors. If every distance is calculated between any pair of the N cepstrums, which provides an $N{\times}N$ distance matrix, the matrix is invariant. In the cepstral domain, a difference in microphones or lines is represented as addition of another static vector b, $c'{=}c{+}b$. And it is very clear that the matrix is also invariant with any kind of b. It seems that the distance matrix can be a good candidate to the acoustic definition of word Gestalt but I have to note that matrix A is not a complete rotation matrix. I have to say that the *point-based* distance matrix is easily modified by a difference in speakers.

Is there a good method to make the distance matrix invariant? The answer is to calculate the matrix as *distribution-based* matrix. Figure 8 shows a timbre contour in a cepstrum space. The contour is converted into a sequence of distributions, from which a distance matrix is extracted. It should be noted that distance is calculated also from *temporally distant* events. I can guarantee mathematically that this *holistic* matrix is invariant with any kind of 1-to-1 linear or non-linear transform [17].

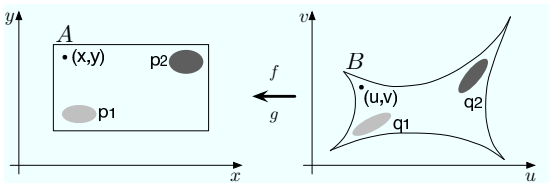

Figure 9: Linear or non-linear mapping between two spaces

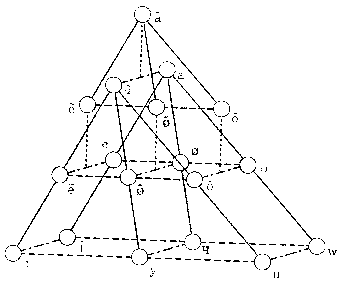

Figure 10: Jakobson's structure of the French vowels [18]

In Figure 9, there are two spaces, one of which is mapped into the other by a linear or non-linear transform. Point (x, y) in space A is mapped uniquely on (u, v) in B, where $x{=}f(u, v)$ and $y{=}g(u, v)$. Bhattacharyya distance between two distributions is invariant with any kind of linear or non-linear transform.

$$BD(p_1, p_2) = -\log \iint \sqrt{p_1(x, y)p_2(x, y)}dxdy$$

$$= -\log \iint \sqrt{p_1(f(u, v), g(u, v))|J| \cdot p_2(f(u, v), g(u, v))|J|}dudv$$

$$= -\log \iint \sqrt{q_1(u, v)q_2(u, v)}dudv = BD(q_1, q_2),$$

where $J{=}J(u, v)$, Jacobian. The distribution-based distance matrix is robustly invariant. Since a distance matrix determines its own shape of the geometrical structure, I can call this matrix-based representation a structural representation of speech.

2.5. Linguistic, psychological, and technical verifications

Phonetics discusses the absolute and local values of language sounds, i.e. "elements first!". Phonology discusses their relative and holistic values, i.e. "system first!". The proposed structural representation is regarded as mathematical and physical implementation of Jakobson's structural phonology (See Figure 10). In his case, he mentally ignored non-linguistic factors in speech and built a speaker-invariant structure of a language. In my case, however, I removed the factors purely mathematically and extracted a speaker-invariant structure from an utterance.

A previous study of speech perception [19] showed that isolated vowel sounds generated by pseudo-giants and pseudo-fairies like the two male adults in Figure 1 were very difficult to identify. This is because their vowels are out of the range of the vowels of normal-sized speakers, shown in Figure 3. When they produced meaningless sequences of morae, which contained timbre dynamics, however, subjects became reasonably able to identify the vowels in the utterances [19]. It should be noted that the utterances contained no semantic and syntactic information. This performance is similar to that of RP people who can verbalize a melody as syllable name sequence. They cannot identify an isolated tone at all but can transcribe a melody. Their transcription is based on the key-invariant tonal arrangement, not based on tone-to-symbol conversion. Utterances can be transcribed not based on sound-to-symbol conversion.

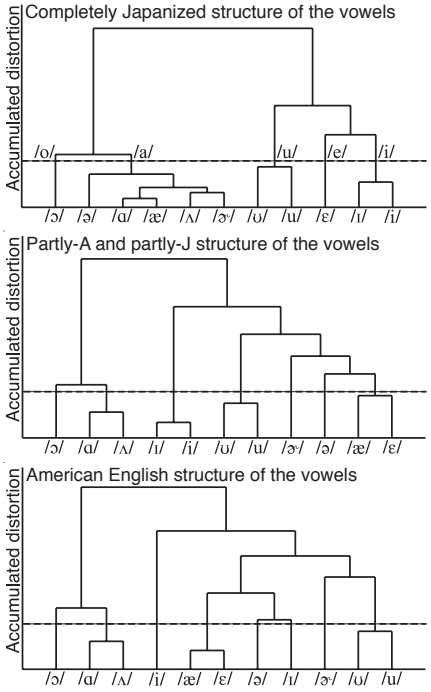

Figure 11: From the Japanized structure to the American structure

The speaker-invariant structural representation of speech was applied in isolated word recognition [20, 21], where a word was defined artificially as a sequence of five Japanese vowels such as /eoiau/. The recognition performance of the proposed method could not drop with the change of body size. But the performance of word-HMMs had to drop drastically. For example, a structure-based recognizer showed 80% with extremely small speakers but an HMM-based recognizer showed 1%. The former recognizer cannot identify an isolated sound at all.

3. Use of the structure for CALL

Infants and learners don't have to imitate the voices of a parent and a teacher, respectively. What they have to imitate is the speaker-invariant sound system of the target language, embedded in utterances. Based on this principle, a structure-based vowel training system was built [3, 4, 5]. Here, a vowel structure was extracted from a set of words which contained a full set of English monophthongs such as beat, bit, bed, bat, etc.

3.1. Development of the vowel system and its visualization

Various pronunciations of the vowels were simulated by a Japanese speaker who can speak American English (AE) well. Each of the 11 AE vowels was recorded only once as /bVt/ and each of the 5 Japanese ones was done five times as /bVto/. Using the vowel segments of these data, various vowel structures were calculated. For example, the completely Japanized English structure can be obtained by substituting Japanese /a/ sounds for /ʌ, æ, ɑ, ɘ, ɚ/ and the other Japanese vowels properly for the other AE vowels. Partly-American and partly-Japanese vowel structures can be generated by changing the substitution pattern. Figure 11 shows the completely Japanized structure, a partly-American and partly-Japanese structure, and the AE structure. Here, hierarchical clustering was used to visualize the vowel structures (distance matrices). The second tree was obtained from the first one by correcting /ʌ, æ, ɑ, ɘ, ɚ/.

Table 1: Vowel substitution table

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

Table 2: 8 patterns of the vowel substitution

	ɑ	æ	ʌ	ɘ	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.

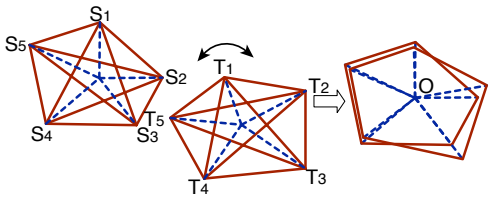

Figure 12: Distance calculation after shift and rotation

3.2. Classification of learners based on their pronunciations

A learner was visualized as tree diagram, which was generated by a full set of Bhattacharyya distances between any two of the vowels. If distance measure between two vowel matrices, i.e, two learners, is adequately defined, then, it will be possible to calculate a full set of distances between any two of the learners. This means that the learners can be classified purely based on their vowel structures, with no undesired effects from age, gender, speaker, microphone, etc. This section shows that, with the proposed technique, the learner classification worked very well. In the experiment, various vowel structures were used and they were obtained from twelve Japanese returnees from US.

3.2.1. Speech material used in the experiment

Six male and six female high school or university students, who were returnees from US, joined the recording. The 11 AE vowels and the 5 Japanese vowels were recorded once as /bVt/ and five times as /bVto/, respectively. This is because five different American vowels, at most, was replaced by a Japanese vowel.

Considering Japanese habits of producing AE vowels, the substitution table was prepared, shown as Table 1. Using this table, 8 patterns of the vowel substitution were obtained and listed in Table 2. P1 and P8 correspond to the completely Japanized English and the good American English pronunciations, respectively. P2 to P7 are half-Japanese and half-American pronunciations. 8 different vowel structures were prepared per speaker and 96 vowel structures all together. The aim of the experiment was to examine whether the 96 structures could be classified based on the vowel structures, not based on gender or speaker.

3.2.2. Matrix-to-matrix distance measure

Suppose that two vowel structures, S and T, are given as two distance matrices. Then, structure-to-structure distance is ob-

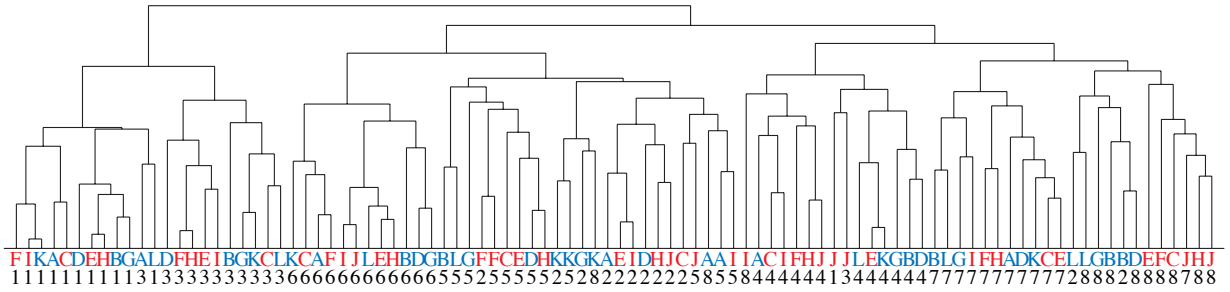

Figure 13: Classification of the 96 vowel structures based on the *contrast-based* comparison (D_1)

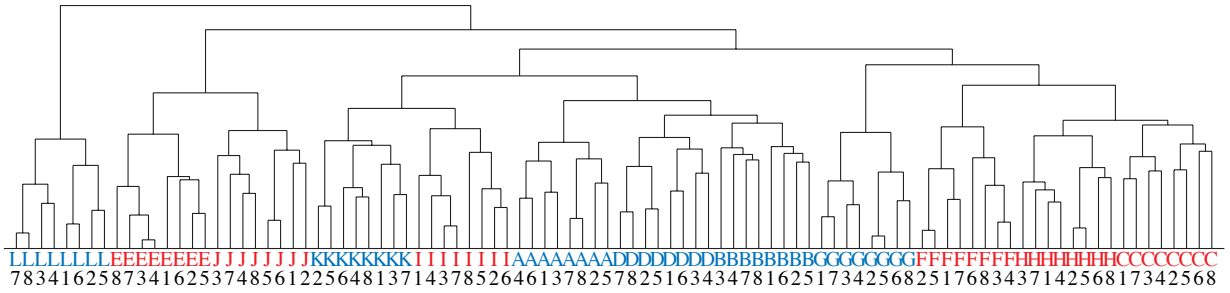

Figure 14: Classification of the 96 vowel structures based on the *substance-based* comparison (D_2)

tained after shifting and rotating a structure so that the two can be overlapped the best, shown in Figure 12. The distance is calculated as the minimum of the total distance between the corresponding two points after shift and rotation. In [3], it was shown that the minimum distance, D_1, can be approximately calculated as euclidean distance between the two distance matrices, where the upper-triangle elements form a vector;

$$D_1(S,T) = \sqrt{\frac{1}{M}\sum_{i<j}(S_{ij} - T_{ij})^2}, \quad (1)$$

where S_{ij} is (i,j) element of matrix S and M is the number of the vowels. D_1 can be regarded as summation of differences of vowel contrasts between the two. For example, distance between /ʌ/ and /ɛ/ is compared between the two structures. In the conventional acoustic matching framework such as DTW and HMM, however, vowel substance /ʌ/ of a learner and that of another was directly compared. In this framework, distance between two vowel structures, D_2, is formulated as follows.

$$D_2(S,T) = \sqrt{\frac{1}{M}\sum_{i} BD(v_i^S, v_i^T)}, \quad (2)$$

where v_i^S is vowel i of structure S.

3.2.3. *Results of classifying the 96 simulated learners*

Figures 13 and 14 show the results of classifying the 96 vowel structures in two different ways. Numbers and alphabets represent the vowel patterns (1 to 8) and the speakers (A to L), respectively. It is clearly shown that contrast-based comparison led to pronunciation classification and substance-based comparison led to speaker classification. In [5], another tree was built manually by a phonetician and a good similarity between the manual tree and the automatic tree of Figure 13 was verified.

3.3. Which vowel to correct at first in your case?

3.3.1. *The vowel generating the largest structural distortion*

Matrix-to-matrix distance was derived as D_1 which indicates the total distortion between the two structures. It can be decomposed into components of the individual vowels. The *local* structural distortion caused by vowel v, $d(v)$, was defined as

$$d(v) = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(S_{vj} - T_{vj})^2}. \quad (3)$$

S and T correspond to a learner matrix and a teacher's one. The vowel giving the largest $d(v)$ should be corrected at first.

3.3.2. *Estimation of the order of vowel correction*

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 vowel structure for each pattern. P8 is regarded as distance matrix of a teacher and there were 7 learners, one of which had the complete Japanese accent (P1) and the others had partly Japanese accented pronunciations. Using Equation 3, the order of vowel correction was estimated for each learner. It was examined whether the replaced AE vowels (see Table 2) were ranked as higher or not.

The estimated orders for P1 to P6 are shown in Figure 15. Bars represent $d(v)$ and gray bars mean that of the replaced vowels. It can be said that the replaced vowels are ranked higher with some exceptions. In several figures, the replaced vowels of /ʊ, u, i/ are ranked lower than unreplaced vowels. This result is considered reasonable because American vowels of /ʊ, u, i/ are known to closer to Japanese vowels of /u, u:, i:/.

4. Use of the structure to motivate learners

4.1. Selection of favorite teachers

The speaker-invariant structural representation of speech is used to build a very motivating user-interface for CALL systems. With the proposed technique, a learner's pronunciation can be compared to a specific teacher's pronunciation. For example, the pronunciation of the tallest man can be compared to that of the shortest adult without any mismatch. Figure 16 shows a window of teacher selection, where learners can select their favorite teachers. In the current demonstration system, the pronunciations of world-famous phoneticians are stored. If those of movie stars are available, learners can select them if they want.

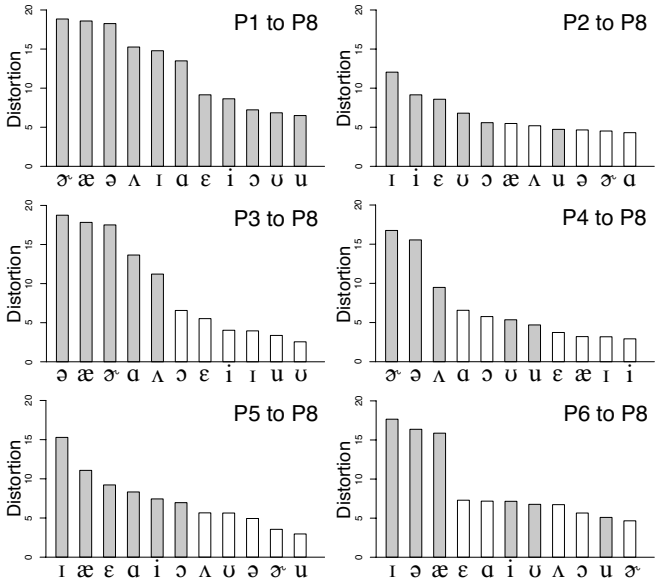

Figure 15: The estimated order of vowel correction

The Kashiwa campus of The University of Tokyo holds an open-campus activity once a year. Every lab. shows demonstrations and my lab. carries out "Pronunciation Clinic" (PC) every year. For the last three years, the total number of the visitors to the open-campus were around 3,800, 4,500, and 2,700. The small number of the latest year was due to rains. In contrast, the total number of people who joined PC was about 200, 250, and 250. Posters of PC with Figure 16 were put on walls. I believe that the interface successfully attracted interests of the visitors.

4.2. Overview of all the learners' development in a class

Figure 17 shows the changes of 18 learners before and after a 1-week training. These illustrations are obtained with multi-dimensional scaling and five teachers are also plotted. The learners' efforts are clearly visualized and I found that this kind of image overviewing the class motivated the learners very well.

5. Conclusions

This paper describes the theoretical background of the structural representation of speech and shows its application to CALL. The new representation claims that infants and learners should acquire the speaker-invariant sound system embedded in utterances of parents and teachers. With the new representation, some new techniques and new interfaces are created. I hope that the new representation helps learners improving their pronunciation of the target language pleasantly and successfully.

6. References

[1] N. Minematsu *et al.*, "Consideration of infants' vocal imitation through modeling speech as timbre-based melody," in *New Frontiers in Artiticial Intelligence*, LNAI4914, pp.26–39, Springer Verlag (2008)

[2] N. Minematsu, "Are learners myna birds to the averated distributions of native speakers? – a note of warning from a serious speech engineer –," *Proc. SLaTE*, CD-ROM (2007)

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

[4] N. Minematsu *et al.*, "Structural representation of the pronunciation and its use for CALL," *Int. Workshop on Spoken Language Technology*, pp.126–129 (2006)

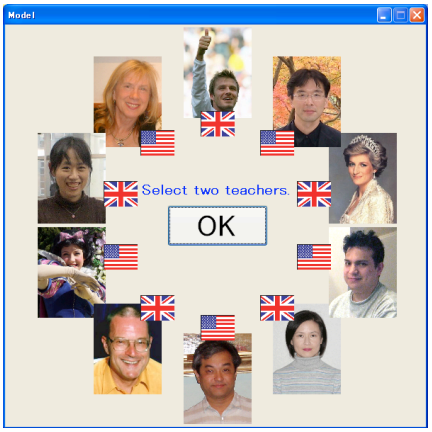

Figure 16: Select your favorite teachers!!

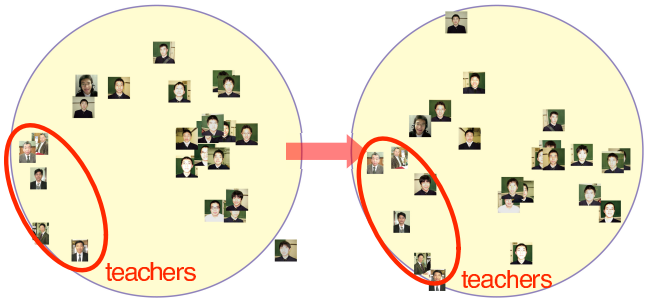

Figure 17: Overview of the changes of 18 learners in a class

[5] N. Minematsu *et al.*, "Structural representation of the pronunciation and its use for classifying Japanese learners of English," *Proc. SLaTE*, CD-ROM (2007)

[6] P. K. Kuhl *et al.*, "Infant vocalizations in response to speech: vocal imitation and developmental change," *J. Acoust. Soc. Am.*, vol.100, no.4, pp.2425–2438 (1996)

[7] S. E. Shaywitz, *Overcoming dyslexia*, Random House Inc. (2005)

[8] M. Kato, "Phonological development and its disorders," *J. Communication Disorders*, 20, 2, 84-85 (2003)

[9] N. Minematsu *et al.*, "Universal and invariant representation of speech," Proc. Int. Conf. Infant Study (2006)
http://www.gavo.t.u-tokyo.ac.jp/~mine/paper/PDF/2006/ICIS_t2006-6.pdf

[10] K. Miyamoto, *Making voices and watching voices,* Morikita Pub. (1995)

[11] K. Johnson *et al.*, *Talker variability in speech processing*, Academic Press (1997)

[12] K. Miyazaki, "How well do we understand absolute pitch?," *J. Acoust. Soc. Jpn.*, 60, 11, 682-688 (2004)

[13] T. Taniguchi, *Sounds become music in mind – introduction to music psychology –*, Kitaoji Pub. (2000)

[14] W. Labov *et al.*, *Atlas of North American English*, Walter De Gruyter (2001)

[15] M. Pitz *et al.*, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech and Audio Processing,* 13, 5, 930-944 (2005)

[16] D. Saito *et al.*, "Directional dependency of cepstrum on vocal tract length," *Proc. ICASSP*, (2008, to appear)

[17] N. Minematsu, *et al.*, "Linear and non-linear transformation invariant representation of information and its use for acoustic modeling of speech," *Proc. Spring Meeting Acoust. Soc. Jpn.*, 147-148 (2007)

[18] R. Jakobson *et al.*, *Notes on the French phonemic pattern*, Hunter (1949)

[19] Y. Hayashi *et al.*, "Comparison of perceptual characteristics of scaled vowels and words," Proc. Spring Meeting Acoust. Soc. Jpn., pp.473–474 (2007)

[20] S. Asakawa *et al.* "Multi-stream parameterization for structural speech recognition," *Proc. ICASSP* (2008, to appear)

[21] N. Minematsu *et al.* "'Static and dynamic' or 'local and holistic': that is the question," Proc. ISCA Tutorial Research Workshop on Speech Analysis and Processing for Knowledge Discovery (2008, submitted)