Pronunciation Diagnosis
— What to correct at first in YOUR case? —

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

Native-sounding vs. intelligible. This has been a controversial issue for a long time in language learning and many teachers claim that intelligible pronunciation should be the goal. What is the physical definition of intelligibility? The current work shows a very good candidate answer to this question. The first author proposed a new paradigm of observing speech acoustics based upon structural phonology, where all the kinds of speech events are viewed as an entire structure and this structure was shown to be mathematically invariant with any static non-linguistic features such as age, gender, size, shape, microphone, room, line, and so on. This acoustic structure is purely linguistic and the phoneme-level structure is regarded as the pronunciation structure of individual students. This structure is matched with another linguistic structure, the lexical structure of the target language, and degree of compatibility between the two different levels of structures is calculated, which is defined as the intelligibility in this work. To increase the intelligibility, different instructions should be prepared for different students because no two students are the same. The phonological structure can be divided into some sub-structures. By evaluating which sub-structure causes the largest damage when communicating in the target language with the student's phonological structure, the instruction is automatically generated on what to correct at first in his/her case.

Keywords: *Intelligibility, structural phonology, pronunciation diagnosis, lexical density, phonological distortion, non-linguistic features*

1 Introduction

There exist many kinds of English pronunciations socially accepted as intelligible all over the world, although some of them are clearly different from the native pronunciation. Many teachers claim that the intelligible pronunciation should be the goal of pronunciation training because pronunciation is just a tool for smooth speech communication. But it is very difficult to define the intelligible pronunciation physically because the intelligibility depends upon listeners. Especially in the case of non-native listeners, it is highly expected that different mother tongues will define different intelligible pronunciations. Against this difficulty, some bold attempts were made to discuss intelligible pronunciation (Bernstein 2003; Minematsu 2003), where non-native utterances were directly presented to listeners who were asked to repeat or type what they heard. A large number of miscommunications were observed and, based upon these, intelligible pronunciation was discussed. According to Minematsu (2003), acoustic and linguistic analysis of the facts implied that the most influential factor on the intelligibility is speech rhythm involved in an utterance. In both works, the listeners were all Americans, which are just one candidate of the listeners, and this approach may have to continue until everybody on Earth joins the experiment if different listeners are considered to define different intelligible pronunciations.

In this paper, another approach is taken, where the intelligibility is defined quantitatively with little attention to the listeners. The first author proposed a new paradigm of observing speech acoustics based

upon structural phonology (Minematsu 2004a; Minematsu 2004b; Minematsu 2005). 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. In short, structural phonology was implemented in physics, and the structure is purely acoustic and linguistic at the same time.

How do we define the intelligibility quantitatively with little attention to the listeners? It is true that a student will communicate with many different non-native listeners and, in this sense, the intelligibility may have to be defined based upon the listeners. However, it is true that the student is learning English of a single specific accent, i.e., British, American, Canadian, Australian English, or another. As is mentioned above, the pronunciations of individual students are acoustically and linguistically modeled as structures, which is similar to Halle's phoneme tree diagram (Halle 1959) or Jakobson's geometrical structure of phonemes (Jakobson 1975). It is also possible to extract the lexical structures from vocabulary of the individual Englishes. The pronunciation structure is determined by fixing a student and the lexical structure is determined by fixing an English. If compatibility between the two different levels of structures is measured, it will be another definition of intelligibility. It is desired to measure the compatibility based upon some cognitive models because speaking is always intended for a human listener.

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 deletion is implemented. In speech recognition, distortions caused by 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 a 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 the speaker individuality is also regarded as the multiplicative distortion. If a speech event is represented by the 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 as 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 the 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 the researchers' brain and speech sounds are represented as abstract entities named phonemes. Phonology is a speech science to clarify a phonological system 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

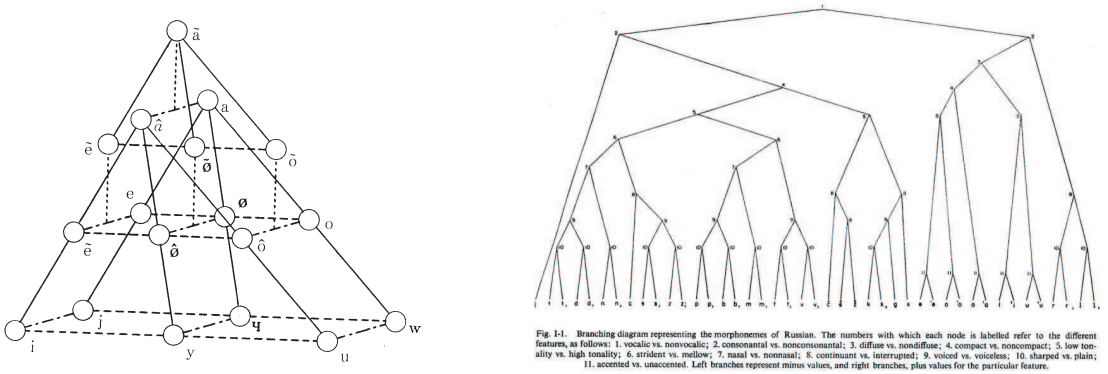

Figure 1: Jakobson's geometrical structure of some French phonemes (left) and Halle's tree diagram of the Russian phonemes (right)

(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 in 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 a mathematical and physical interpretation of 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 as an n-point structure (distance matrix) that is invariant with any affine transformation. This looks impossible because affine transformation is a transformation that 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 dis-

tributions follow the Gaussian distribution, the following is obtained.

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

where μ_x and Σ_x are the average vector and the variance-covariance matrix of $d_x(c)$, respectively, and μ_{xy} is $\mu_x - \mu_y$. Although an 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.

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

These facts mean that BD between any two distributions is not changed by any 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, are mathematically regarded as a 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 ERJ speech database

ERJ (English Read by Japanese) database (Minematsu 2004c) was used, which contains English read by 202 Japanese, Japanese English (JE), and 20 General American speakers (GA). The individual students have pronunciation scores rated by 5 American teachers of English. Table 1 shows the acoustic analysis

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

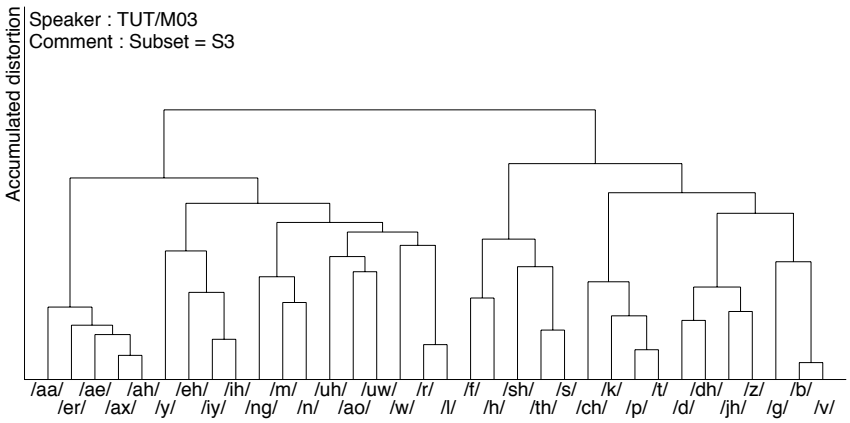

Figure 2: *A structurally represented poor Japanese student*

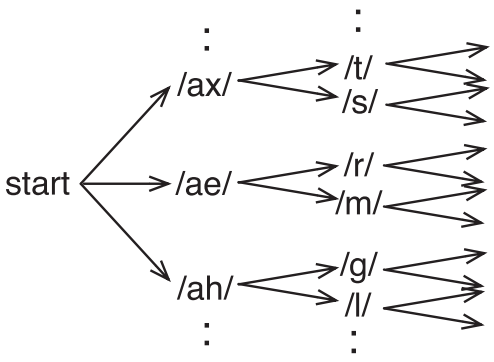

Figure 3: *An example of the tree structured lexicon*

conditions. Phoneme-to-phoneme distance is calculated as the average distance over the three state-to-state BDs between two phoneme HMMs. Figure 2 shows a tree example of a Japanese student extracted from his HMMs and the well-known Japanese habits are clearly seen. Confusions of /r/&/l/, /s/&/th/, /z/&/dh/, /f/&/h/, /iy/&/ih/, /v/&/b/, etc are found. Mid and low vowels of English are located close to each other because there is the only one mid and low vowel in Japanese. Schwa is close to the above vowels because Japanese often produce the mid and low Japanese vowel for schwa. Remarkably high performance of discarding the non-linguistic information was experimentally verified in (Minematsu 2004a; Minematsu 2004b; Minematsu 2005; Minematsu 2004d) , and interested readers should refer to them.

4 Estimation of the intelligibility

In this section, compatibility between the pronunciation structure and the lexical structure is introduced based upon the Cohort Model, one of the isolated word perception models.

4.1 Cohort Model of word perception

The original Cohort Model characterizes a human process of perceiving an isolated word as a simple left-to-right process (Marslen-Wilson1980). When the initial phoneme of a word input is perceived, a set of words starting with the phoneme are activated in the brain. The number of activated words is reduced by the subsequent input of phonemes and finally reaches one, which means the end of word perception. Cohort means a set of the activated words in the brain. It is clear that Cohort Model assumes a tree structured lexicon in the brain, which is shown in Figure 3. As is mentioned in Section 2.2, phonology clarifies a phonological structure hidden in the entire set of phonemes or in sequences of phonemes. The pronunciation structure discussed in the previous section corresponds to the former and is determined by fixing a student. The tree structured lexicon corresponds to the latter and is determined by fixing a target language. The intelligibility is defined as compatibility calculated between the two different levels of *phonological* structures based upon the Cohort Model. An algorithm for the calculation is shown below.

The Cohort Model is often discussed with phonemes as its basic acoustic units. In this work, however, syllables are used as the basic units for cohort development. This is because an acoustic unit of speech production in English is said to be a syllable.

4.2 Estimation of the intelligibility as cohort size

Clearly seen in Figure 2, many phonemic confusions occur in Japanese English. This is natural because Japanese has only 25 phonemes and English has more than 40. If Japanese students use their own sounds only, 1-to-N mapping is inevitable. With the phonemic confusions, different words get acoustically closer and the acoustic lexical density is increased. In this work, larger lexical density is interpreted

as less intelligibility. The compatibility between a student's pronunciation structure and the target language's lexical structure is defined as the cohort size calculated from the two structures. The smaller the cohort size is, the higher the compatibility and the intelligibility are.

The cohort activated only with the initial syllable input was focused upon. A 20K-sized lexicon in WSJ database was used as vocabulary and each entry of the lexicon has a unigram score. The phoneme sequence of each entry is obtained from the PRONLEX dictionary. Each word (each phoneme sequence) was converted into a syllable sequence by tsylb software. Speech samples of some students in ERJ did not include a part of diphthongs and, in this case, HMMs for these phonemes could not be trained (See Table 1). Then, the words starting with a syllable including a diphthong were ignored. The number of the remaining words was about 18K. It should be noted that the vocabulary includes different words whose baseforms are identical, such as walk, walked, and walking. Syllabification of the words showed that approximately 3,200 different kinds of syllables were found as word-initial syllables.

For each of the different word-initial syllables s_i, the number of words starting with s_i or with a syllable acoustically close to s_i was calculated as $CS_0(s_i)$. Distance between two syllables was calculated by DP matching between two sequences of phoneme HMMs (syllables) and the calculation required only the phoneme-to-phoneme distance matrix. The syllables acoustically close to s_i were defined as the syllables distant from s_i by less than threshold θ. Thus, $CS_0(s_i)$ was actually obtained as $CS_0(s_i, \theta)$, using thesize of the cohort activated by the initial syllable of word w_j, which was calculated as $CS_1(w_j, \theta)$=$CS_0(s^1(w_j), \theta)$, where $s^1(w_j)$ is the initial syllable of w_j. Finally, the expected cohort size $ECS(\theta)$ over the entire vocabulary was obtained by the following equation.

$$ECS(\theta) = \sum_j p(w_j) CS_1(w_j, \theta), \tag{4}$$

where $p(w_j)$ is a normalized uni-gram probability satisfying a condition of $\sum_j p(w_j)$=1.0 over the words selected by deleting those starting with a syllable including a diphthong.

4.3 Results and discussions

Japanese students and GA speakers who read sentence set 6 were used in the experiment. The pronunciation structure somewhat depends upon the sentences read and set 6 was adopted because it covered a wide range of the proficiency with a rather even distribution. In Figure 4, the number of speakers of the individual ranges of proficiency is also listed. The numbers of Japanese and Americans are 26 and 4, respectively. Proficiency scores of the Americans were assumed to be 5.0, which is the full score.

Figure 4 shows relations between the ECS and threshold θ for all the speakers, where the best three and the poorest three students are indicated by showing their pronunciation scores assigned by teachers. It is clearly indicated that words produced by the poorest students are very confusing and those by the best students are very distinct. This result shows good validity of the definition of the intelligibility adopted in this work.

Figure 5 shows the correlation between the ECS and the pronunciation scores. Rather good correlation is found between the two. The ECS values are those at $\theta = 0.35$ in Figure 4 and the pronunciation scores were obtained by asking the teachers to rate the individual students with regard to the segmental aspect of the pronunciation. In the figure, the four Americans are explicitly indicated. Rather good correlation denotes high validity of the proposed method to estimate the intelligibility.

It should be noted that the proposed algorithm is implemented without any acoustic matching between a student and a teacher. The student's pronunciation is matched with the target language itself. The pronunciation structure can be said to be purely acoustic and linguistic at the same time. Then, the structure is matched with another level of linguistic structure, which is the lexical structure of the target language. In this sense, the proposed algorithm enables the linguistic matching between a student and the target language.

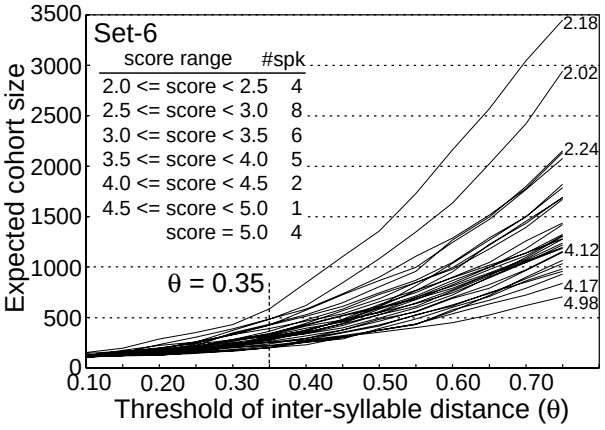

Figure 4: *ECS as function of θ*

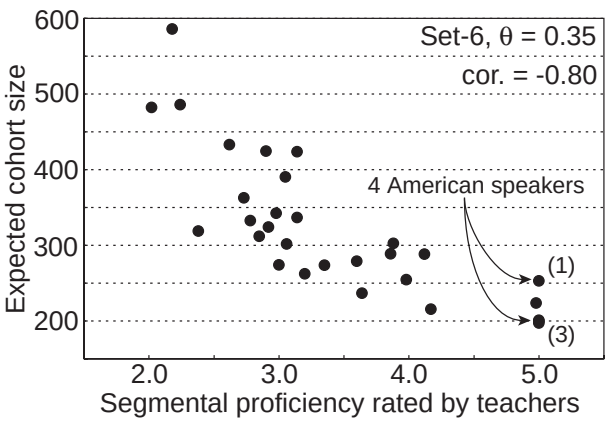

Figure 5: *Proficiency rating without any acoustic matching*

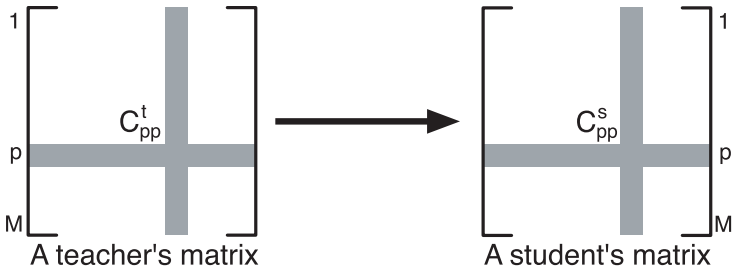

Figure 6: *Replacement of a sub-structure*

5 Effective and efficient instructions optimized for the individual students

5.1 Exchange of sub-structures between two speakers

The pronunciation structure is extracted so that all the static non-linguistic features are discarded from speech. This characteristics enables an interesting operation; exchange of sub-structures between two speakers. If a sub-structure in a student is replaced with its corresponding one in a teacher, the student will have a better pronunciation structure (See Figure 6). This operation is meaningless if the other acoustic representation of speech, spectrogram, is used. If a portion of the spectrogram of a speaker is replaced with its corresponding spectrogram of another, all the acoustic features are changed, such as age, gender, size, shape, microphone, room, and so on. The resulting spectrogram will be completely confused.

5.2 Instructions optimized for individual students

Replacement of which sub-structure minimizes the cohort size? The answer to this question will provide the pedagogical instruction optimized for the student. In the current work, a sub-structure of phoneme p is defined as the following set of elements in the distance matrix; $\{c_{pj}\}$ and $\{c_{jp}\}$ $(1 \leq i, j \leq N)$. Here, c_{ij} is an element of the matrix. If replacement of the sub-structure of phoneme p_0 minimizes the cohort size, it means that the student should correct the articulation of phoneme p_0 among others.

Using a female speaker, RYU/F06 (pronunciation score is 2.02), the cohort size reduction is done by replacing her sub-structures with a teacher's ones. Figure 7 shows the results of the cohort reduction by a single replacement. Her original cohort is more than double that of the native cohort and replacement of a sub-structure of schwa is shown to be the most effective and efficient for her to improve the intelligibility of the pronunciation. What is next if the schwa is corrected? The most effective replacement after

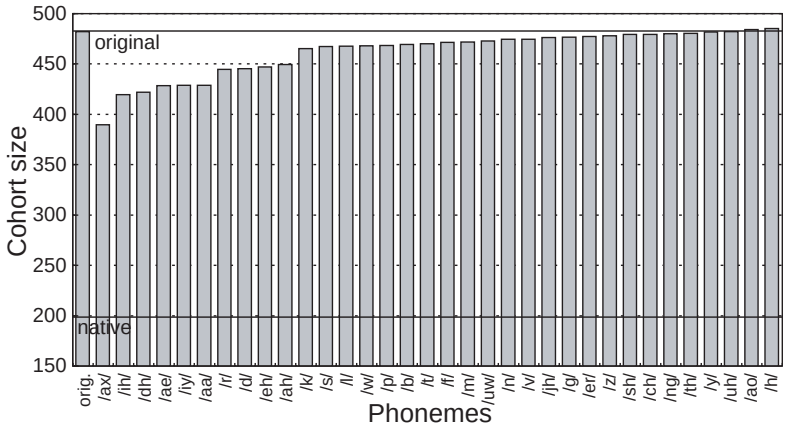

Figure 7: *Cohort size reduction by a single replacement*

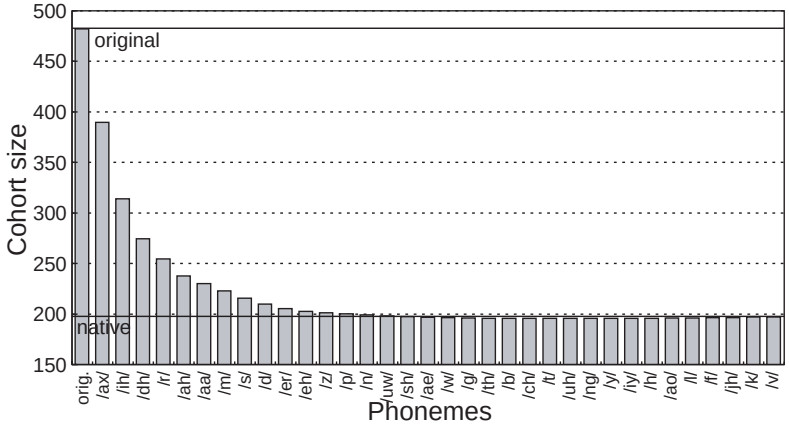

Figure 8: *Cohort size reduction by sequential replacements*

schwa's correction can be discussed in the same manner. Figure 8 shows the order of English phonemes for her to correct, and the size of the effect accumulated by the sequential corrections. It is shown in Figure 8 that the phonemes with higher priority for corrections are those which are well-known to be pronounced by Japanese to be acoustically similar to other phonemes. This result shows good validity of the proposed method for automatic generation of instructions.

6 Conclusions

This paper proposed a novel method to estimate compatibility between a student's pronunciation structure and the target language's lexical structure, which is regarded as the intelligibility of the pronunciation. The proposed algorithm does not require any acoustic matching between a student and a teacher, which means that the algorithm cannot face "mismatch problems" at all. The algorithm can directly match a student's pronunciation with the target language linguistically. The linguistic matching became possible because the first author proposed a novel method of representing speech acoustics with no dimensions to indicate the static non-linguistic features. This paper also showed that it is possible to determine the order of phonemes for individual students to correct. This determination is done by sequential replacements of sub-structures and this operation is possible only with the new speech representation. As future work, the authors are planning to verify the effectiveness of the proposed method in actual classrooms with not only university students but also young children.

References

J. Bernstein, "Objective measurement of intelligibility," Proc. Int. Congress on Phonetic Science(ICPhS), pp.1581–1584 (2003)

N. Minematsu, C. Guo, and K. Hirose, "CART-based factor analysis of intelligibility reduction in Japanese English," Proc. EUROSPEECH, pp.2069–2072 (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)

N. Minematsu, "Pronunciation assessment based upon the phonological distortions observed in language learners' utterances," Proc. Int. Conf. Spoken Language Processing(ICSLP), pp.1669-1672 (2004d)

W. D. Marslen-Wilson *et al.*, "The temporal structure of spoken language understanding," Cognition, vol.8, pp.1–71 (1980)