Sub-structure-based Estimation of Pronunciation Proficiency and Classification of Learners

Masayuki Suzuki, Nobuaki Minematsu, Dean Luo, and Keikichi Hirose

The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{suzuki,mine,dean,hirose}@gavo.t.u-tokyo.ac.jp

***Abstract*—Automatic estimation of pronunciation proficiency has its specific difficulty. Adequacy in controlling the vocal organs can be estimated from spectral envelopes of input utterances but the envelope patterns are also affected easily by different speakers. To develop a pedagogically sound method for automatic estimation, the envelope changes caused by linguistic factors and those by extra-linguistic factors should be properly separated. For this aim, in our previous study [1], we proposed a mathematically-guaranteed and linguistically-valid speaker-invariant representation of pronunciation, called speech structure. After the proposal, we have examined that representation also for ASR [2], [3], [4] and, through these works, we have learned better how to apply speech structures to various tasks. In this paper, we focus on a proficiency estimation experiment done in [1] and, based on our recently proposed techniques for the structures, we carry out that experiment again but under new and different conditions. Here, we use smaller units of structural analysis, speaker-invariant sub-structures, and relative structural distances between a learner and a teacher. Results show that correlations between human and machine rating are improved and also show extremely higher robustness to speaker differences compared to widely used GOP scores. Further, we also demonstrate that the proposed representation can classify learners purely based on their pronunciation proficiency, not affected by their age and gender.**

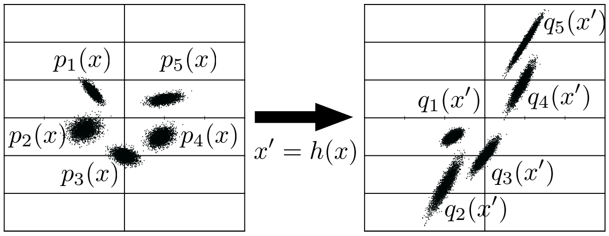

Fig. 1. Speaker-invariant system of language sounds

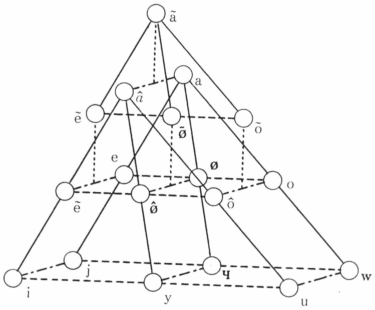

Fig. 2. Jakobson's invariant system of the French vowels

I. Introduction

How to enable computers to distinguish the spectral envelope changes caused by pronunciation improvement within a learner from the changes caused by different speakers? A good candidate answer was proposed to this question by regarding the pronunciation not as a mere set of language sounds but as a system organized by the sounds [1]. In other words, for pronunciation proficiency estimation, a focus was put not on each segment of an utterance independently but on the relationships among the segments of that utterance.

Language sounds of interest are organized into a system, i.e. a speaker-invariant sound shape [5], shown conceptually in Figure 1. The definition of the system is given by a distance matrix among these sounds because, geometrically speaking, a distance matrix can fix its own shape uniquely. In voice transformation studies, speaker difference is usually modeled as space mapping, $x'{=}h(x)$. This indicates that, if sound-to-sound distance is calculated using transform-invariant measure, the distance matrix or the speech structure becomes speaker-invariant. In Figure 1, every sound is characterized as distribution and sound-to-sound distance is measured using Bhattacharyya distance (BD) because BD is invariant with any kind of invertible transform [6]. As is well-known in ASR, vocal tract length difference and microphone difference are well modeled globally as $c'{=}Ac$ and $c'{=}c{+}b$ in the cepstrum domain, respectively [7], [8].

Acoustic assessment of each sound in an utterance can be viewed as *phonetic* assessment and that of the entire system of the sounds can be regarded as *phonological* assessment. In classical phonology, Jakobson proposed a theory of acoustic and relational invariance, called distinctive feature theory. In [5], he repeatedly emphasizes the importance of relational and systemic invariance among speech sounds and also denies the absolute invariance strongly. Figure 2 shows his speaker-invariant system of the French vowels and semi-vowels.

We consider that the BD-based distance matrix is a mathematical realization of Jakobson's claim and that pronunciation assessment should be done not by evaluating individual sounds in a learner's pronunciation independently but by examining whether an adequate sound system underlies a learner's pronunciation of the target language. Based on this philosophy, we've already conducted a series of studies of structure-based CALL systems [1], [9], [10]. In addition, we've also done a series of studies of structure-based ASR [2], [3], [4]. In this paper, a proficiency estimation experiment, which was

978-1-4244-5479-2/09/$26.00 © 2009 IEEE

done in [1], is carried out again but under new and different experimental conditions. Here, the new techniques which we have developed for the ASR are utilized effectively for CALL and more accurate estimation is highly expected.

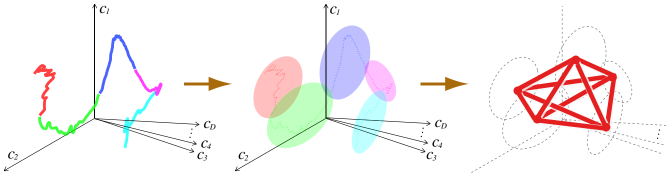

Fig. 3. An utterance structure composed only of BDs

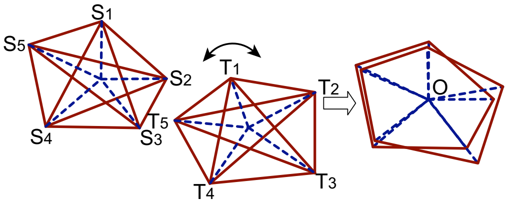

Fig. 4. Structure comparison through shift & rotation

II. Structure analysis for ASR and CALL

In the structure-based ASR studies [2], [3], [4], to form a BD-based distance matrix from an utterance, the utterance, i.e. a cepstrum vector sequence, is converted to a distribution sequence (See Figure 3). This preprocessing is implemented as MAP-based training of an HMM and an utterance is converted into an HMM. Once two utterances are converted into two structures, how to match them? In the current implementation of the structure-based ASR, two structures have to have the same number of distributions. The matching score is simply calculated in the following formula, which well approximates the minimum summation of the distances between two corresponding distributions after shifting and rotating a structure so that the two structures are overlapped the best (See Figure 4).

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

where S and T are two distance matrices whose elements are calculated as $\sqrt{\mathrm{BD}}$. M is the number of distributions. In the cepstrum domain, shift and rotation of a structure correspond to cancellation of differences in microphone and in vocal tract length, respectively [11]. This means that, without explicit adaptation, the structure-based ASR gives matching scores after global adaptation. This is why the structure-based ASR is extremely robust to extra-linguistic differences [2], [3], [4].

In the structure-based CALL studies [1], [9], [10], a student's structure S and a teacher's structure T are extracted from their plural utterances. In [1], from about 60 sentence utterances, a structure of the entire phonemes is formed for a student while, in [9], [10], a vowel structure is extracted from eleven word utterances, which contain the eleven American English monophthongs. In [1], through structural comparison between each student in a Japanese-English database [12] and a specific teacher, pronunciation proficiency is automatically estimated. The obtained scores are compared to the proficiency scores given by five native teachers of American English and high correlation is found. In [9], [10], $D_1(S,T)$ is decomposed into vowel pairs and, through pairwise structural analysis, student-dependent and diagnostic instructions on which vowel to correct at first are provided for each student.

III. Proficiency estimation of Japanese learners reading English sentences

A. What we have developed for the structure-based ASR

The structure-based ASR experiments [2], [3], [4] enabled us to apply the structures in a more proper way to various tasks. In this paper, we examine the following three techniques.

As shown in Figure 3, a speech structure is a BD-based distance matrix among speech events, namely, distributions. In [1], phonemes were used as units of estimating distributions and forming their structure. In [2], [3], [4], however, we found that a phoneme-based distance matrix is too coarse to obtain a good performance for ASR. Three to five distributions per phoneme gave the best performance, meaning that, after estimating usual HMMs from an utterance, its speech structure should be formed by using states of those HMMs. The finer structures are expected to improve the CALL performance.

The use of speech structures lead to a new normalization technique, that is normalization of the magnitude of articulatory efforts. The size of a structure is highly correlated with how articulate a speaker's phonation is and the performance of ASR should not be affected by this. In [2], [3], [4], the size-normalized structures improved the ASR performance and, in this paper, this technique is tentatively examined.

In CALL, a structure of an utterance and another structure of another utterance are compared based on Equation (1). For ASR, two utterances of different words should be modeled discriminatively. In [2], [3], [4], features observed commonly in different words were removed and not used to form their structures. PCA, LDA, and feature selection were examined and we found that parameter (dimension) reduction was effective to improve the performance. In this paper, adequate selection of distribution pairs is also investigated to find the optimum sub-structures for estimating pronunciation proficiency and emphasizing differences between good and bad learners.

In addition to these three techniques, we examine another new technique, that is normalization of local and structural differences. In [2], [3], [4], a speech structure formed from an utterance was matched with template structure patterns, which were *statistical* structure patterns trained with some training speakers. Use of the statistical patterns can calculate matching scores by taking parameter variances into account. In the case of comparison between a student and a teacher using Equation (1), however, this is impossible. Then, accidentally large values of $|S_{ij}-T_{ij}|$ can dominate pronunciation estimation. To avoid this defect, the following formula is tested.

$$D_2(S,T) = \sqrt{\frac{1}{M}\sum_{i<j}\left\{\frac{S_{ij}-T_{ij}}{\frac{1}{2}(S_{ij}+T_{ij})}\right\}^2}. \qquad (2)$$

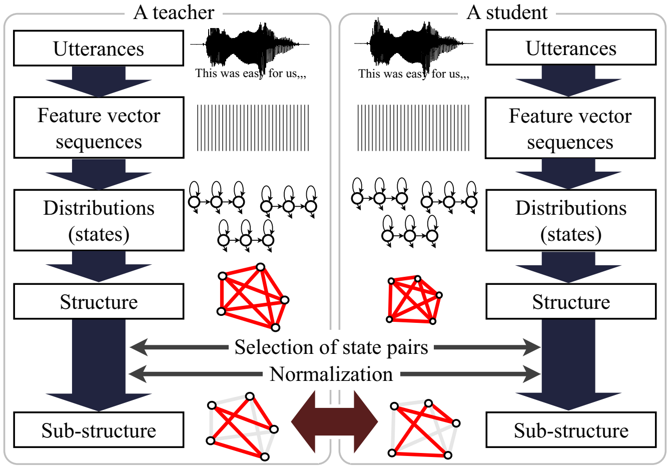

Fig. 5. Sub-structure extraction for a student and a teacher

TABLE I
CONDITION FOR THE ACOUSTIC ANALYSIS

sampling	16bit / 16kHz
windows	25ms length and 10ms shift
training data	about 75 sentences per speaker
parameters	MFCC + Δ + ΔPower (25dim.)
HMMs	speaker-dependent, context-independent, and 1-mixture monophones with diagonal matrix
topology	5 states and 3 distributions per HMM
monophones	aa,ae,ah,ao,aw,ax,axr,ay,b,ch,d,dh,eh,er,ey, f,g,hh,ih,iy,j,jh,k,l,m,n,ng,ow,oy,p,r,s,sh,t,th, uh,uw,v,w,y,z,zh,sil

Figure 5 shows the procedure of extracting state-based sub-structures from two corpuses of a student and a teacher. First, a set of speaker-dependent HMMs are trained, where each state corresponds to an event (distribution). Then, a BD-based distance matrix is formed. Next, by selecting an appropriate subset of state pairs, a sub-structure is formed. This procedure is conducted for a teacher and a learner and their sub-structures are compared to estimate the proficiency of that learner.

B. The speech database used in the experiment

ERJ (English Read by Japanese) corpus is used in our experiments, which contains eight sets of read sentence utterances [12]. Each set is composed of about 75 sentences and they are read by about 25 university students, among whom about a half are male and the other are female. Those sentences are a part of the TIMIT sentences and students of different sets read different sentences. The eight sets cover the TIMIT sentences completely. Proficiency scores are also provided for all the students, which were manually given by five native teachers of American English with good knowledge of phonetics and Japanese English. In addition to speech and label data of Japanese English, in the corpus, the utterances of the same sentences read by 20 native speakers of General American English (GA) are also included. 18 of them read a half of the entire sentences and the remaining two read all the sentences. In structural analysis, only a male speaker (M08) of the two is used as a reference teacher commonly for all the 200 students.

C. Structure-based analysis and GOP-based analysis

Table I shows the acoustic analysis conditions and the number of AE monophones is 43. From ERJ, 200 sets of speaker-dependent monophone HMMs are trained from the individual students. From the teacher (M08), eight sets of HMMs are trained, each corresponding to a sentence set in ERJ. Eventually, 208 129×129(=43×3) BD-based distance matrices are formed in total. Using the students of all the sets but set-6, the optimal definition of state-based sub-structures is estimated. Selection of state pairs is incrementally and greedily determined so as to maximize the correlation between machine rating scores and human scores. Here, $-D_1$ or $-D_2$ is used as machine scores and they are calculated by matching a student's sub-structure and the sub-structure of the corresponding sentence set of the teacher. By following the obtained optimal definition of sub-structures, those of 26 students of set-6 are used as open data and compared to the teacher's sub-structure. Then, correlation between machine and human is calculated.

For comparison, the pronunciation proficiency is estimated as GOP (Goodness Of Pronunciation) scores, i.e. posterior probability of the intended phonemes given input utterances.

$$\begin{aligned} & GOP(o_1, ..., o_T, p_1, ..., p_N) \\ & = P(p_1, ..., p_N | o_1, ..., o_T) \\ & \approx \frac{1}{N} \sum_{i=1}^{N} \frac{1}{D_{p_i}} \log \left\{ \frac{P(o^{p_i}|p_i)}{\max_{q \in Q} P(o^{p_i}|q)} \right\}, \end{aligned} \tag{3}$$

where T is the length of given observation sequences and N is the number of the intended phonemes. o^{p_i} is the speech segment obtained for p_i through forced alignment and D_{p_i} is its duration. $\{o^{p_1},...,o^{p_N}\}$ correspond to $\{o_1,...,o_T\}$. Q is the inventory of phonemes. The GOP was originally proposed in [13] and is widely accepted as pronunciation proficiency. Since GOP is probability ratio, it internally has a function of canceling acoustic mismatch between teachers' HMMs and a learner's utterance. In this paper, nine sets of HMMs are prepared to calculate the GOP. Eight sets are from eight sentence sets of the common teacher (M08). The other set is trained with all the utterances of the 20 native teachers.

D. Results of pronunciation proficiency estimation

Results of proficiency estimation by phoneme-based structure analysis are shown in Figure 6. X-axis represents the number of selected phoneme pairs. The maximum is ${}_{43}C_2$=903. Colors indicate differences in normalization methods. The red curve is obtained using relative differences of Equation (2) and the green curve is drawn by normalizing the size of sub-structures. The blue curve indicates no normalization.

When we used finer units of structure analysis, state-based structure analysis, as we expected, higher correlations were obtained, shown in Figure 7. Here, X-axis is the number of selected state pairs and its maximum is ${}_{43\times3}C_2$=8,256. Similarly to Figure 6, colors indicate differences in normalization.

Looking at both the figures, we can find easily that feature selection works effectively to improve the performance and

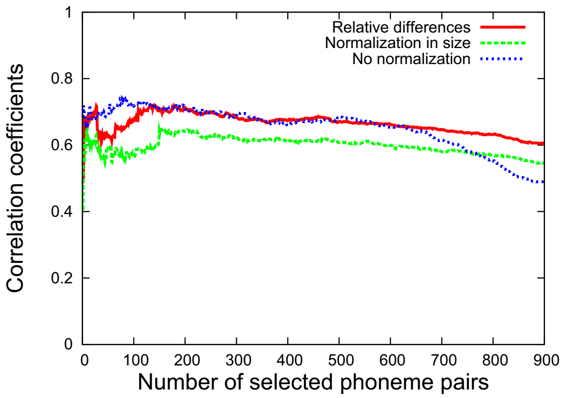

Fig. 6. Correlations with phoneme-based structure analysis

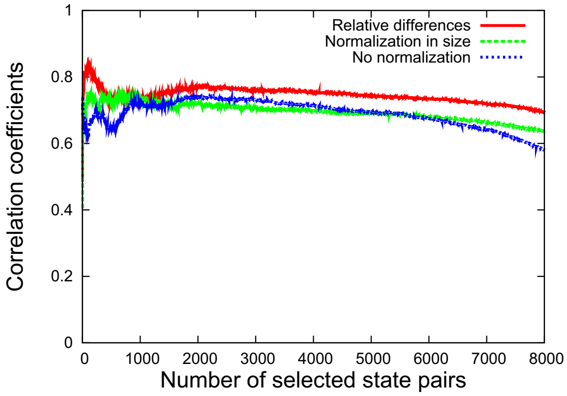

Fig. 7. Correlations with state-based structure analysis

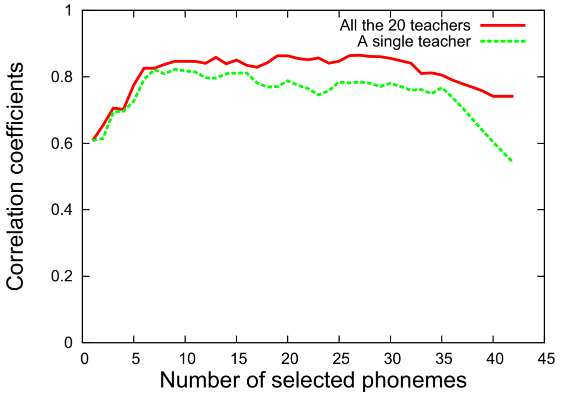

Fig. 8. Correlations with GOP analysis

that finer units of structure analysis, i.e. state-based sub-structures, are also beneficial. Checking each of them, we can find that the effect of normalization somewhat differs between them. In the phoneme-based structure analysis, the size-based normalization works poorly and, in the state-based structure analysis, the effect of Equation (2) is significant. In Figure 7, the highest correlation (0.84) is obtained with Equation (2) in the case of 86 selected state pairs. Although this number is small, the 172(=86×2) states cover 41 phonemes out of 43.

Figure 8 shows the results of estimating the GOP scores for two cases. One is using the HMMs of the common teacher and the other is using those of all the 20 AE native teachers.

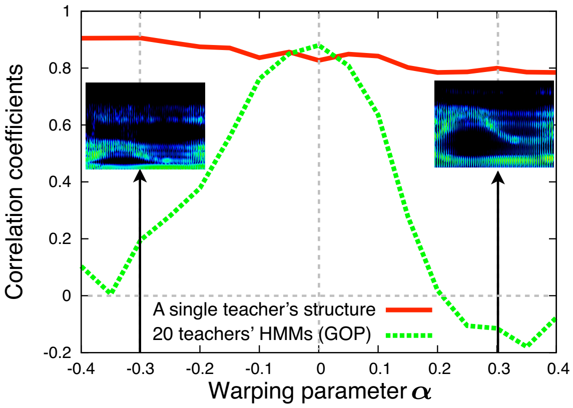

Fig. 9. Correlations with warped utterances

As in structure analysis, we carried out incremental phoneme selection to realize discriminative comparison. This selection is also effective here and the highest correlation (0.87) is found at the number of 27. The performance difference between two sets of HMMs can be interpreted as follows. Although GOP has an internal function of mismatch cancelation, this function works when forced alignment performs well. In some cases, this is not the case. Then, the GOP scores of the common teacher shows less correlations than those of all the teachers.

IV. Robustness of the Proposed Method with Respect to Speaker Differences

A. *Urgent requirement for extremely robust technologies*

The Japanese government decided to introduce lessons for oral English communication to every primary school from 2011. But it is true that we don't have a sufficient number of English teachers. The government expects class teachers, many of whom did not receive a good education for teaching English, to play an important role in the lessons. In this situation, we consider that some technical solutions will be introduced to classrooms. Automatic estimation of pronunciation proficiency is one of the key technologies and it requires high robustness [15] because the pronunciations of adult teachers and those of young children have to be treated properly at the same time.

B. *Robustness of the structure and the GOP*

Figure 9 shows the results of proficiency estimation using the sub-structures (the common teacher) and the GOP (all the 20 teachers). In this case, by using frequency warping techniques, all the input utterances of set-6 were transformed as if they had been generated by speakers of various vocal tract lengths. X-axis means warping parameter α [7], [8] and, with α=-0.4/$+0.4$, the vocal tract length is doubled/halved, respectively. In the figure, two speech segments which are obtained by transforming a speech segment with α=$+0.3$ and -0.3 are shown visually. Frequency warping resulted in a drastic acoustic modification. In spite of this large change, Figure 9 shows the extreme robustness of the sub-structures but it also shows the extreme weakness of the GOP. We can say definitely that even a single teacher's sub-structure can be used directly and effectively for any student of any size.

C. Learning to pronounce or learning to impersonate

Basically speaking, GOP is a posterior probability and it has an internal function of canceling acoustic mismatch between HMMs and learners. But this function only works when forced alignment (numerator of Equation (3)) and continuous phoneme recognition (denominator of Equation (3)) perform properly. With a large acoustic mismatch, however, the two processes inevitably fail. To circumvent this, teachers' models (HMMs) are often adapted to learners or the models are trained from many teachers who have similar voice quality to that of learners. In this case, however, another critical problem appears though it is not technical but theoretical. The requirement of no acoustic mismatch in voice quality between learners and teachers leads us to consider that the pronunciation assessment based on the current ASR framework such as GOP identifies learning to pronounce as learning to impersonate [14] and it quantifies how well learners can impersonate the model speakers. In other words, the conventional pronunciation assessment framework is just application of the impersonation assessment technology by preparing no-mismatch conditions in advance.

But learning to pronounce is not learning to impersonate at all. No male student tries to produce female voices when asked to repeat what a female teacher said. No young child tries to produce deep voices to repeat what a tall male teacher said. They are not parrots but we have to wonder whether the conventional framework assumes students as parrots [14]. Rational teachers may be unwilling to use the products based on this framework. As Jakobson claimed, however, students extract a speaker-invariant sound system underling a given utternace and try to reproduce that system orally. But inevitable differences in size and shape of the vocal organs between a learner and a teacher have to cause acoustic differences between them.

V. Structure-based classification of learners

A. Learner-based distance matrix

Using $D_1(L_a, L_b)$ or $D_2(L_a, L_b)$, where L_x stands for a pronunciation sub-structure of learner x, it is possible to calculate a distance matrix among all the learners. This learner-based distance matrix enables bottom-up learner classification. Considering the results of structure-based estimation of pronunciation proficiency (See Figure 9), the learner classification based on pronunciation (sub-)structures will be a classification purely based on pronunciation not based on age and gender. To verify this through comparison, we prepare another criterion for calculating a difference between two speakers S and T.

$$D_3(S,T) = \sqrt{\frac{1}{M}\sum_i BD(s_i^S, s_i^T)}, \qquad (4)$$

where i is a state index of the 86 selected state pairs and s_i^T is a distribution of state i and speaker T. BD means Bhattacharyya distance and M is the number of physically different states in the selected 86 state pairs (=90, which is out of 129=43×3).

Although both of $D_1(S,T)$ and $D_2(S,T)$ compare timber *contrasts* between two speakers of S and T, $D_3(S,T)$ focuses on timber *substances* and, using them directly and absolutely, two speakers are compared. The former scheme corresponds to contrast-based (structure-based) acoustic matching [2], [3], [4] and the latter scheme corresponds to substance-based acoustic matching used conventionally in DTW and HMM.

26 students of set-6 were used with multiple values of α. Here, -0.3 (very tall), 0.0, $+0.3$ (very short) were used and 78 (=26×3) students of different sizes were virtually created in total. For bottom-up clustering, we used Ward's method.

B. Results of learner classification

Figures 10 and 11 show the results of learner classification using D_2 and D_3, respectively. Alphabets are student IDs. $\overline{\text{X}}$ and $\underline{\text{X}}$ stand for taller and shorter versions of X. Color represents gender and numbers below the student IDs are pronunciation proficiency scores rated manually by teachers.

We can find a remarkably clear difference between the two figures. While *contrast*-based comparison (D_2) results in classifying the students purely according to their pronunciations, not affected by size and gender, *substance*-based comparison (D_3) leads to complete classification based on size and gender.

Linguistically speaking, Figure 10 corresponds to dialect-based speaker classification because systemic variation in the phonemes, especially the vowels, characterizes dialects [16]. Technically speaking, Figure 10 indicates the possibility of classifying all the individuals on earth based on their English pronunciations. Further, for a specific learner, Googling all the individuals can search for the one with an extremely similar accent, who surely assesses the learner's pronunciation as the most intelligible because these two are the closest dialectally. We already started classifying the Chinese speakers based on their native dialects, sub-dialects, and sub-sub-dialects [17].

VI. Discussions

As described in Section I, the structural representation of speech or pronunciation was originally proposed to remove extra-linguistic factors from speech acoustics [1] and to model only the linguistic aspect of utterances. In Section IV-B, we demonstrated the effectiveness of using pronunciation structures to estimate pronunciation proficiency by ignoring extra-linguistic variations. In Section V-B, a similar and good effect was obtained again to classify learners not affected by their age and gender. In contrast, we claimed that GOP-based estimation of goodness of pronunciation should be regarded as estimation of goodness of impersonation. Further, we also showed that, if speech sounds are compared directly among speakers, what we obtain is classification of speakers, not pronunciations.

What is a difference between the proposed framework and the conventional one? The answer is what to model in speech. In the former, speech (timbre) contrasts are modeled and, in the latter, speech substances are modeled. With speech contrasts, we can organize them into a linguistic sound system. We consider that Jakobson and others focused on this speaker-invariant sound system [5]. In [18], we can find two different definitions of the phoneme. *1) A phoneme is a class of sounds that are phonetically (acoustically and/or articulatorily) similar and show certain characteristic patterns of distribution*

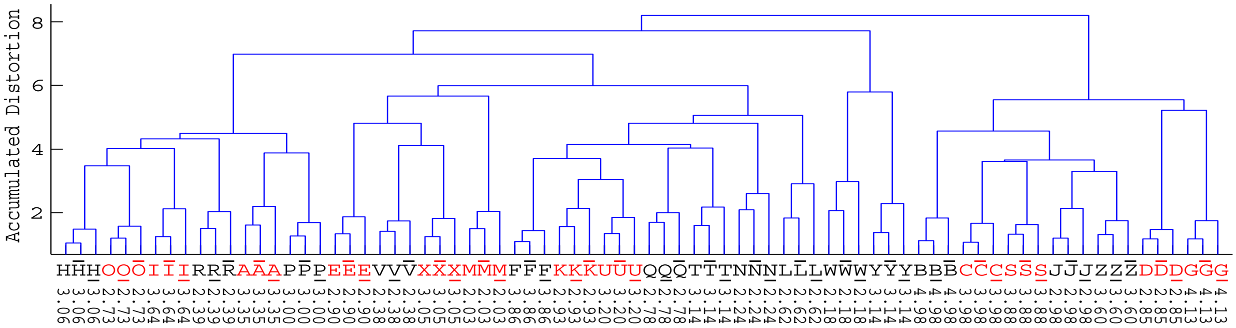

Fig. 10. Classification of the 78 virtual students based on the *contrast-based* comparison (D_2)

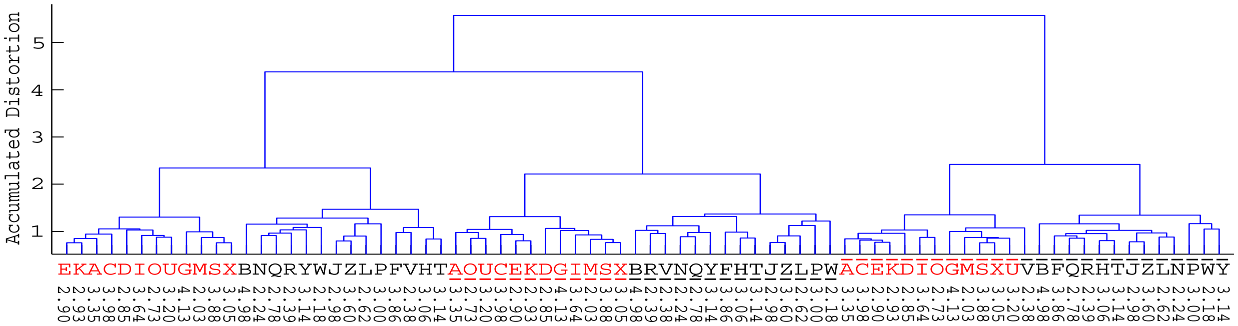

Fig. 11. Classification of the 78 virtual students based on the *substance-based* comparison (D_3)

in the language or dialect under consideration. This is the absolute definition and the conventional framework is based on this definition. *2) A phoneme is one element in the sound system of a language having a characteristic set of interrelations with each of the other elements in that system.* This is the relational and contrastive definition and our proposed framework is built on it. We already applied the contrast-based structural representation of speech to robust ASR [2], [3], [4] and speaker-independent speech recognition was implemented by using only a small number of training speakers and not using speaker adaptation techniques explicitly.

VII. Conclusions

In this paper, we carried out experiments to estimate pronunciation proficiency from learners' utterances. The experiments had been originally done in our previous work and, in the current work, we introduced recently developed techniques for structure-based ASR. The results of the experiments showed the effectiveness of sub-structures and they were also shown to be useful in classifying learners based on their pronunciations. Through comparison between the proposed framework and the conventional one, we pointed out that the conventional framework inappropriately assumes pronunciation learning as impersonation learning. We consider that this is attributed to inappropriate modeling of speech. Although learners don't imitate the voices of teachers acoustically, the conventional framework builds acoustic models of their voices and uses them to estimate the proficiency. What learners ignore should be discarded when building pronunciation models of teachers. What should be used and what should be ignored in speech? As Jakobson claimed, in this work, we focused on the speaker-invariant sound system underlying teachers' utterances.

References

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

[2] Y. Qiao *et al.,* "Random discriminant structure analysis for continous Japanese vowel recognition," *Proc. ASRU*, pp.576–581, 2007.

[3] S. Asakawa *et al.,* "Multi-stream parameterization for structural speech recognition," *Proc. ICASSP*, pp.4097–4100, 2008.

[4] N. Minematsu *et al.,* "Implementation of robust speech recognition by simulating infants' speech perception based on the invariant sound shape embedded in utterances," *Proc. SPECOM*, pp.35–40, 2009.

[5] R. Jakobson *et al.*, *The sound shape of language,* Mouton De Gruyter, 1987

[6] Y. Qiao *et al.,* "f-divergence is a generalized invariant measure between distributions," *Proc. INTERSPEECH*, pp.1349–1452, 2008.

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

[8] T. Emori *et al.,* "Rapid vocal tract length normalization using maximum likelihood estimation," *Proc. EUROSPEECH*, pp.1649-1652, 2001.

[9] N. Minematsu *et al.,* "Structural representation of the pronunciation and its use for CALL," *Proc. SLT*, pp.126–129, 2006.

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

[11] D. Saito *et al.,* "Directional dependency of cepstrum on vocal tract length," *Proc. ICASSP*, pp.4485–4488, 2008.

[12] N. Minematsu, *et al.*, "Development of English speech database read by Japanese to support CALL research," *Proc. ICA*, pp.577–560, 2004.

[13] S. M. Witt *et al.*, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, 30, pp.95–108, 2000.

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

[15] M. Russell *et al.*, "Challenges for computer recognition of children's speech," *Proc. SLaTE*, CD-ROM, 2007.

[16] W. Labov *et al.*, *Atlas of North American English,* Mouton and Gruyter, 2005

[17] X. Ma *et al.*, "Structural analysis of dialects, sub-dialects, and sub-sub-dialects of Chinese," *Proc. INTERSPEECH*, pp.2219–2222, 2009.

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