Improved Structure-based Automatic Estimation of Pronunciation Proficiency

Masayuki Suzuki, Luo Dean, Nobuaki Minematsu, Keikichi Hirose

The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, 113-8656, Japan

{suzuki,dean,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract

Automatic estimation of pronunciation proficiency has its specific difficulty. Adequacy in controlling the vocal organs is often estimated from spectral envelopes of input utterances but the envelope patterns are also affected by alternating speakers. To develop a good and stable method for automatic estimation, the envelope changes caused by linguistic factors and those by extra-linguistic factors should be properly separated. In our previous study [1], to this end, we proposed a mathematically-guaranteed and linguistically-valid speaker-invariant representation of pronunciation, called speech structure. After the proposal, we have tested that representation also for ASR [2, 3, 4] and, through these works, we have learned better how to apply speech structures for various tasks. In this paper, we focus on a proficiency estimation experiment done in [1] and, using the recently developed techniques for the structures, we carry out that experiment again but under different conditions. Here, we use a smaller unit of structural analysis, speaker-invariant substructures, and relative structural distances between a learner and a teacher. Results show higher correlation between human and machine rating and also show extremely higher robustness to speaker differences compared to widely used GOP scores.

1. Introduction

How to separate the spectral envelope changes caused by pronunciation improvement within a learner and those caused by alternation of learners? A good candidate answer was given 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 or a speaker-invariant sound shape, 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 [5]. As is well-known in ASR, vocal tract length difference and microphone difference is well modeled globally as $c'=Ac$ and $c'=c+b$ in the cepstrum domain, respectively [6].

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 [7],

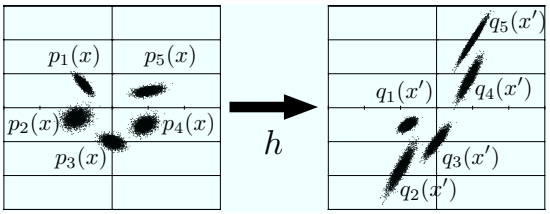

Figure 1: *Speaker-invariant system of language sounds*

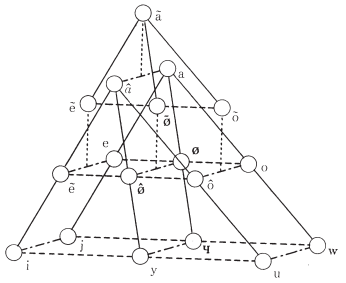

Figure 2: *Jakobson's invariant system of the French vowels*

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, 8, 9]. In addition, we've also done a series of studies of structure-based ASR systems [2, 3, 4]. In this paper, a proficiency estimation experiment, which was done in [1], is carried out again but in new experimental conditions. Here, the techniques we've developed for the structure-based ASR are applied and more accurate and robust estimation is highly expected.

2. Structure analysis for ASR and CALL

In the structure-based ASR studies [2, 3, 4], to form a BD-based distance from an input 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 automatically converted into an HMM. Once two structures are formed from two different utterances, 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. It approximates well the minimum summation of the distances between corresponding two distributions after shifting and rotating a structure so that

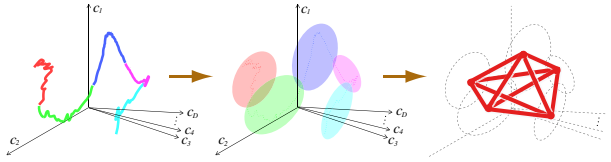

Figure 3: *An utterance structure composed only of BDs*

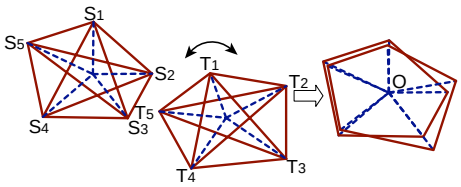

Figure 4: *Structure comparison through shift & rotation*

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 [10]. This indicates that the structure-based ASR gives matching scores after global adaptation without explicit adaptation. This is why the structure-based ASR is extremely robust to microphone and speaker differences [2, 3, 4].

In the structure-based CALL studies [1, 8, 9], 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 [8, 9], a vowel structure is extracted from eleven word utterances containing the eleven American English monophthongs. In [1], through structural comparison between each student of a Japanese-English database [11] 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 [8, 9], $D_1(S,T)$ is decomposed into vowel pairs and, through pairwise structural analysis, diagnostic instructions on which vowel to correct at first are provided for each student.

3. Proficiency estimation of Japanese learners reading English sentences

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

Experimental discussions of the structure-based ASR 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 us the best performance. In other words, after estimating usual HMMs from an utterance, its speech structure should be formed by using states of these HMMs. The finer structures are expected to improve the performance also for CALL.

The use of speech structures enabled us to introduce a new normalization technique, that is normalization of the magnitude

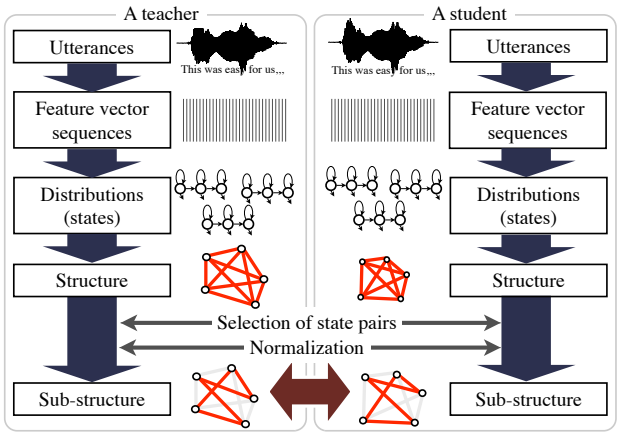

Figure 5: *Sub-structure extraction for a student and a teacher*

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 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 two 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 highly effective to improve the ASR 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 input utterance was matched with template structure patterns, which were *statistical* structure patterns trained with several training speakers. Use of the statistical patterns can calculate matching scores by taking parameter variances into account. In the case of one-to-one comparison in Equation (1), however, this is impossible and accidentally large values of $|S_{ij} - T_{ij}|$ can dominate pronunciation estimation. To avoid this defect, the following formula is tested experimentally in this paper.

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

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.

3.2. 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 [11]. 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

Table 1: *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

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 two read all the sentences. In this paper, a male speaker (M08) of the two is used as a teacher commonly for all the 200 students.

3.3. Structure-based analysis and GOP-based analysis

Table 1 shows the acoustic analysis conditions and the number of AE monophones is 43. From the students, 200 sets of mohophone HMMs are trained. From the single teacher, eight sets of HMMs are trained, each corresponding to a sentence set in ERJ. From all the sets of HMMs, 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 determined to maximize the correlation between machine rating and human rating. 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. Following the obtained optimal definition of sub-structures, those of the 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 probabilities of the intended phonemes given input utterances.

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

where T is the total 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 [12] 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 speakers of American English.

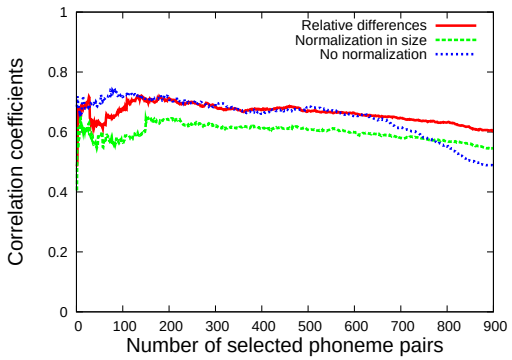

Figure 6: *Correlations with phoneme-based structure analysis*

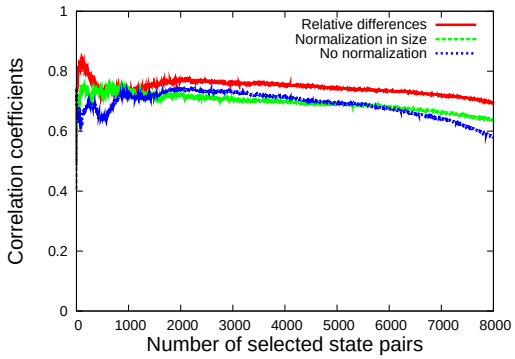

Figure 7: *Correlations with state-based structure analysis*

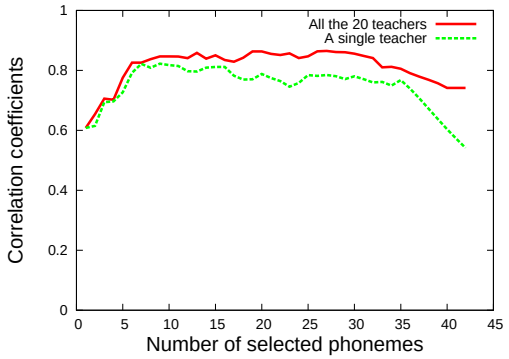

Figure 8: *Correlations with GOP analysis*

3.4. 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 and structural differences of Equation (2) and the green one 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 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

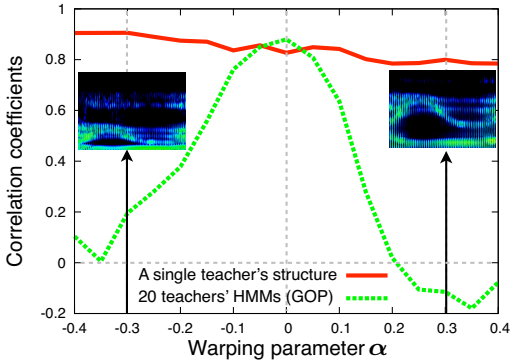

Figure 9: *Correlations with warped utterances*

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, with Equation (2), the highest correlation (0.84) is obtained in the case of 86 selected state pairs. Although this number is very 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 AE speakers. 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 only when forced alignment performs well. In some cases, this condition is less satisfied. Then, the GOP scores of the common teacher shows less correlations than those of all the teachers.

4. Robustness of the proposed method with respect to speaker differences

4.1. Urgent requirement for extremely robust technologies

The Japanese government decided to introduce lessons for oral English communication to every primary school from 2011 but 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 because the pronunciations of adult teachers and young children have to be treated properly.

4.2. Robustness of the structures and the GOP

Figure 9 shows the results of proficiency estimation using the structures (the common teacher) and the GOP (all the 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 and, with α=−0.4/+0.4, the vocal tract length is doubled/halved. Two speech segments warped from the same segment with α=+0.3 and α=−0.3 are shown. Frequency warping resulted in a drastic acoustic modification. In spite of this large change, Figure 9 shows the extreme robustness of the structures but it also shows the extreme weakness of the GOP. We can say that even a single teacher's structure can be used directly and effectively for any student of any size.

5. Discussions and conclusions

In this paper, the improvement of structure-based proficiency estimation is realized and its high robustness to speaker variability is experimentally verified. Further, the weakness of GOP is also made clear. As GOP is basically a posterior probability, it internally has a 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 avoid this, teachers' models (HMMs) are often adapted to learners. If one wants to prepare the most adequate models for a specific learner, one has to build the models trained with that learner who would pronounce the target language correctly. It is ideal that a student and his/her teacher have the same voice quality because of no mismatch.

This technical requirement leads us to consider that GOP should stand for, not Goodness Of Pronunciation, but Goodness Of imPersonation, which quantifies how well a learner can impersonate the model speaker [13]. 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 produces deep voices to repeat what a tall male teacher said. As Jakobson claimed, we consider that they 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 their utterances. Computer-Aided Language Learning (CALL) or Computer-Aided Impersonation Learning (CAIL), which is needed for classrooms? We believe that the answer is obvious.

6. 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*, 2009.

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

[6] 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.

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

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

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

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

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

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

[13] 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.