Structure-based Prediction of English Pronunciation Distances and Its Analytical Investigation

Shun Kasahara
Graduate School of Engineering
The University of Tokyo, Japan
Email: kasahara@gavo.t.u-tokyo.ac.jp

Nobuaki Minematsu
Graduate School of Engineering
The University of Tokyo, Japan
Email: mine@gavo.t.u-tokyo.ac.jp

HanPing Shen
Department of Comp. Sci. and Info. Eng.,
National Cheng Kung University, Taiwan
Email: hanpinsheen@gmail.com

Daisuke Saito
Graduate School of Interdisciplinary Info. Studies
The University of Tokyo, Japan
Email: dsk_saito@gavo.t.u-tokyo.ac.jp

Keikichi Hirose
Graduate School of Info. Sci. and Tech.
The University of Tokyo, Japan
Email: hirose@gavo.t.u-tokyo.ac.jp

***Abstract*—English is the only language available for international communication and is used by approximately 1.5 billions of speakers. It is also known to have a large diversity of pronunciation partly due to the influence of the speakers' mother tongue, called accents. Our project aims at creating a global and individual-basis map of English pronunciations to be used in teaching and learning World Englishes (WE) as well as research studies of WE [1], [2]. Creating the map mathematically requires a distance matrix in terms of pronunciation differences among all the speakers considered, and technically requires a method of predicting the pronunciation distance between any pair of the speakers. Our previous but very recent study [3] combined invariant pronunciation structure analysis [4], [5], [6], [7] and Support Vector Regression (SVR) effectively to predict the inter-speaker pronunciation distances. In [3], very high correlation of 0.903 was observed between reference IPA-based pronunciation distances and the distances predicted by our proposed method. In this paper, after explaining our proposed method, some new results of analytical investigation of the method are described.**

I. Introduction

In many English classes, native pronunciation of English is often presented as a reference, which students try to imitate. It is widely accepted, however, that native-like pronunciation is not always needed for smooth communication. Due to the influence of the students' mother tongue, those from different regions inevitably have different accents in their pronunciation of English. Recently, more and more teachers accept the concept of World Englishes [1], [2] and they regard US and UK pronunciations just as two major examples of accented English. Diversity of WE can be found in various aspects such as dialogue, syntax, pragmatics, lexical choice, spelling, pronunciation, etc. Among these kinds of diversity, this paper focuses on pronunciation. If one takes the concept of WE as it is, he/she can claim that there does not exist the standard pronunciation of English. In this situation, there will be a great interest in how one type of pronunciation compares to other varieties, not in how that type of pronunciation is incorrect compared to the one and standard pronunciation.

What is the minimal unit of the pronunciation diversity? Is it country, region, prefecture, city, town, or village? Accent diversity of English pronunciation is considered to be due to diversity of the language background of individual speakers. For example, the following factors can affect one's pronunciation of English: the mother tongue of the speaker, that of his/her parents, that of his/her friends, that of the English teachers who taught English to him/her, the places where he/she was born and brought up, etc. This thinking leads us easily to the answer of the above question on the minimal unit. It should be individual and WE can have approximately 1.5 billion kinds of different pronunciations. The ultimate goal of our project is creating a global map of WE on an individual basis for each of the speakers to know where and how his/her pronunciation is located in the diversity of English pronunciations. If the speaker is a learner, he/she can then find easier-to-communicate English conversation partners, who are supposed to have a similar kind of pronunciation. If he/she is too distant from many of other varieties, however, he/she may have to correct the pronunciation for the first time to achieve smoother communication with these others.

In this paper, we used the Speech Accent Archive (SAA) [8], which provides speech samples of a common elicitation paragraph read by more than 1,800 speakers from all over the world. The SAA also provides IPA-based narrow transcripts of all the samples, which were used for training a pronunciation distance predictor [3]. To calculate the pronunciation distance between two speakers of the SAA, [9], [10] proposed a method of comparing two IPA transcripts using a modified version of the Levenshtein distance. Although it was shown that the calculated distances had reasonable correlation with the pronunciation distances perceived by humans, [9], [10] cannot handle unlabeled data, i.e., raw speech. Very recently, we proposed a method of predicting the pronunciation distance only using spoken paragraphs of the SAA [3]. The technical challenge is how to make prediction independent of irrelevant but inevitably involved factors such as differences in age, gender, microphone, channel, background noise, etc. To this end, we used invariant pronunciation structure analysis [4], [5], [6], [7] for feature extraction and SVR for prediction. In training the predictor, reference distances had to be prepared as ground truth. In [3], IPA-based phonetic distances calculated through string-based DTW of two IPA transcripts were used. The correlation between the reference distances and the predicted

978-1-4799-4808-6 /14/$31.00 ©2014 IEEE

Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

[pliːz̩ kəl ˈstɛlːʌ as hɛr tu brɪŋ diz θɪŋs wɪθ hɛr frʌm ðə stɑɪ sɪks spuːnz ʌɣ̩ fɪɛʃ ˈsno piːz fɑɪɣ̩ θɪk ˈsleɮs ʌv bluː ʧiːz æn meɪbiː eɪ snæk˺ foɪ hɛɪ bɹʌðɜ bɑb˺ wĭ ɑlso nid˺ eɪ smɑlˠ plæstɪk˺ ˈs̪neɪk æn eɪ biġ tʰɔɪ fɹɔg˺ fɔɪ ðə kɪdz̩ ʃi kɛn ˈskuːb˺ ðiːz θɪŋs ɪntu θriː ɹɛd˺ bægs æn ə wĭl goː mitʰ hɛɪ wɛnzdeɪ æd˺ d̪ə tɹeɪm ˈstɛɪʃən]

Fig. 1. The SAA paragraph and an example of IPA transcription

ones was 0.903 and in this paper, after explaining the proposed method, several new results on its analytical investigation are described and discussed.

II. The Speech Accent Archive

The corpus is composed of read speech samples of more than 1,800 speakers and their corresponding IPA narrow transcripts. The speakers are from all over the world and they read the common elicitation paragraph, shown in Figure 1, where an example of IPA transcription is also presented. The paragraph contains 69 words and can be divided into 221 phoneme instances using the CMU dictionary as reference [11]. The IPA transcripts were used in [3] to prepare reference inter-speaker pronunciation distances, which were adopted as target of prediction using SVR. This is because IPA transcription is done through phoneticians' ignorance of non-linguistic and acoustic variations involved in utterances such as differences in age, gender, etc. It should be noted that the recording condition in the corpus varies from sample to sample because data collection was done voluntarily by those who had interest in joining the SAA project. To create a suitable map automatically, these non-linguistic variations have to be cancelled adequately.

Use of read speech for clustering is considered to reduce pronunciation diversity because read speech may show only "controlled" diversity. In [12], however, English sentences read by 200 Japanese university students showed a very large pronunciation diversity and in [13], a large listening test by Americans showed that the intelligibility of the individual utterances covered a very wide range. Following these facts, we considered that read speech samples can still show well how diverse World Englishes pronunciations are.

It is known that pronunciation diversity is found in both the segmental and prosodic aspects. In [3], we prepared reference distances by using IPA transcripts only, meaning that prosodic diversity was lost. We do not claim that the prosodic diversity is minor but, as shown in [14], clustering only based on the segmental aspect seems able to show validly how diverse World Englishes are in terms of pronunciation. Reference distances with prosodic features will be treated as future work.

In [3] and this study, only the data with no word-level insertion or deletion were used. The audio files that had exactly 69 words were automatically detected. Some of them were found to include a very high level of background noise and/or many pauses, and we manually removed them. Finally, we used 370 speakers' data only but the number of speaker pairs was still large and it was 68,265 (=370×369 / 2).

III. Reference Inter-speaker Pronunciation Distances

To train a pronunciation distance predictor, reference inter-speaker distances were needed, which were also used to evaluate the trained predictor. Following [10], the reference distance between two speakers was calculated through DTW of their IPA transcripts. Since all the transcripts contain exactly the same number of words, word-level alignment was easy and we only had to treat phone-level insertions, deletions, and substitutions between a word and its counterpart.

Since DTW-based alignment of two IPA transcripts needed the distance matrix among all the existing IPA phones in the SAA, we prepared it in the following way. The most frequent 153 kinds of phones were extracted from the SAA, which covered 95% of all the phone instances and we asked an expert phonetician to pronounce each of the 153 phones twenty times. Using the recorded data, a speaker-dependent three-state HMM was built for each phone, where each state contained a Gaussian distribution. Then, for each phone pair, the phone-to-phone distance was defined as the average of three square roots of the state-to-state Bhattacharyya distance. We note that, since the HMMs were speaker-dependent, all the distances were calculated in a matched condition. The other 5% of the phones were all with a diacritical mark. For each of them, we substituted the HMM of the same phone with no mark.

Using the distance matrix among all the kinds of phones in the SAA, word-based DTW was conducted to compare a word and its counterpart in IPA transcripts. The accumulated distance was normalized by the number of phones in the word pair and the normalized distances were summed for all the 69 words. This final distance was used as reference pronunciation distance. Detailed explanation of our string-based DTW, such as configuration of local paths and penalty scores, is found in [14] as well as a result of bottom-up clustering of a part of the SAA speakers using IPA-based and string-based DTW.

Although the DTW-based distances were adopted as reference distances in [3] and this study, we do not claim at all that the above method is the only method of calculating phonetic pronunciation distance between two speakers. The definition of reference distances should be dependent on how the resulting speaker-based pronunciation distance matrix is used for education and research, and we can say that different purposes may require different definitions of reference distances. We consider that our proposed method can be applied to other definitions.

IV. Two Baseline Systems

For comparison, we built two baseline systems, which corresponds directly to two automated versions of the reference distance calculation procedure described in Section III.

The calculation procedure is composed of two steps: 1) IPA manual transcription and 2) DTW alignment for distance calculation. Here, the first process was replaced with automatic recognition of phonemes in input utterances[1]. We used a phoneme recognizer of American English (AE). Using all the utterances of the 370 speakers as training data, monophone HMMs were constructed with the WSJ-based HMMs [15]

[1]As far as we know, there does not exist an automatic recognizer of IPA phones with a diacritical mark.

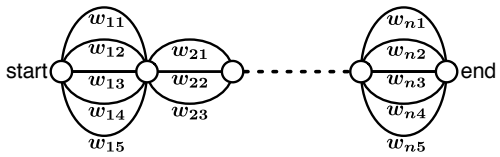

Fig. 2. An example of word-based network grammar

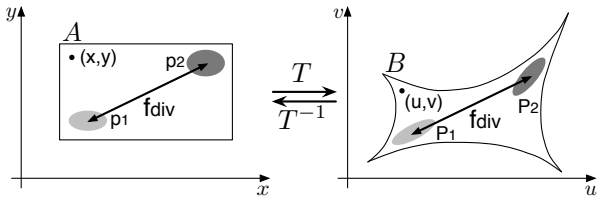

Fig. 3. Transform-invariance of f-divergence

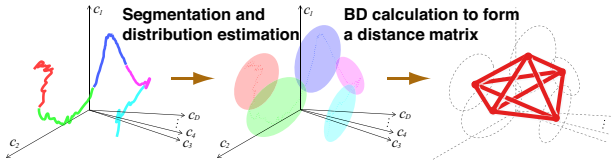

Fig. 4. Utterance structure composed only of BDs

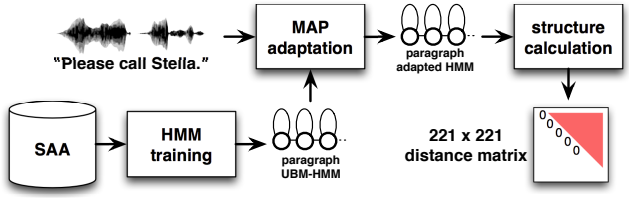

Fig. 5. Procedure to calculate the pronunciation structure

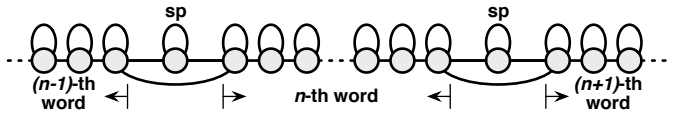

Fig. 6. Explicit modeling of inter-word short pauses

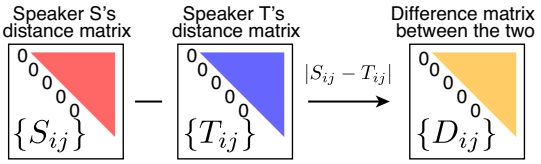

Fig. 7. Difference matrix derived from two speakers' matrices

adopted as initial model. For this training, each IPA transcript was converted into its AE phoneme transcript. This conversion was done by preparing a phone-to-phoneme mapping table with special attention paid to conversion from two consecutive IPA vowels to an AE diphthong.

Since IPA transcription is based on phones and the HMMs were trained based on phonemes, even if we could have a perfect phoneme recognizer, generated transcripts have to be phonemic versions of IPA transcripts. Phone to phoneme conversion is an abstraction process and some detailed phonetic information has to be lost inevitably. Our first baseline system used a perfect but imaginary phoneme recognizer and the pronunciation distance was calculated by comparing two phonemic transcripts based on DTW. Here, the phoneme-to-phoneme distance matrix was needed and prepared by using the WSJ-based HMMs. Our second system used a real phoneme recognizer with word-based network grammar that covered the entire pronunciation diversity found in the 370 speakers. Figure 2 shows an example of the network grammar of an n-word sentence, where w_{ij} denotes the i-th word spoken with the j-th pronunciation. In this system, the generated phoneme transcripts often included recognition errors.

The correlation between the IPA-based inter-speaker reference distances and the phoneme-based distances obtained from the first system was 0.829, meaning that information loss existed to some degree. On the other hand, the phoneme recognition accuracy of the second system was 73.5% and the correlation was found to be so low as 0.458. This clearly indicates that recognition errors are very fatal in this task.

V. Structure-based Prediction of Pronunciation Distances

As told in Section I, we need a robust method to predict the pronunciation distance. The second author proposed a unique method of representing speech, called speech structure, and proved that acoustic and non-linguistic variations involved in speech can become effectively unseen in the representation [4], [5], [6], [7]. This proposal was done being inspired by Jakobson's structural phonology [16] and infants' sensitivity to distributional properties of sounds [17], [18]. The speech structure is invariant against any kind of continuous and convertible transform and this invariance is due to the transform-invariance of f-divergence (See Figure 3) [6], which is calculated as

$$f_{\mathrm{div}}(p_1, p_2) = \int_{\mathcal{X}} p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x}, \qquad (1)$$

where $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ are density functions of two distributions on measurable space $\mathcal{X}$. $g(t)$ is a convex function for $t > 0$. If we take $\sqrt{t}$ as $g(t)$, $-\log(f_{\mathrm{div}})$ becomes the Bhattacharyya distance (BD).

Figure 4 shows the procedure of representing an utterance only by BDs. The utterance is a sequence of vectors and it is converted into a sequence of distributions. Here, any speech event is characterized as distribution. Then, the BD is calculated between any distribution pair and the resulting BD-based distance matrix is an invariant speech structure. When this representation is applied to pronunciation analysis, the matrix is called pronunciation structure [5], [7]. Pronunciation structure is interesting because it captures only local and distant acoustic contrasts and discards absolute acoustic features at all. Figure 5 shows the detailed procedure to calculate the pronunciation structure, where the Universal Background Model (UBM) was trained as paragraph-based HMM by using all the 370 speakers. Here, by arranging the WSJ-based mohophone HMMs [15] following the SAA phonemic transcript, the initial paragraph-based HMM was prepared. Since most of the speakers of SAA are non-native speakers of English, pauses are sometimes inserted at word boundaries where native speakers

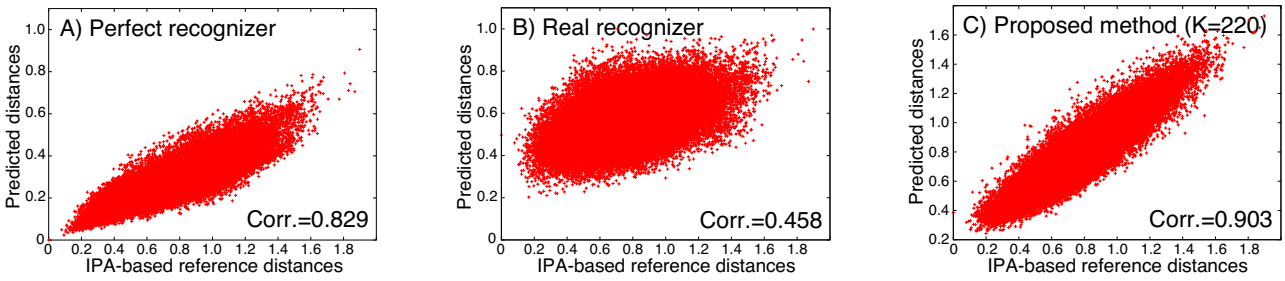

Fig. 8. Correlations between the IPA-based reference pronunciation distances and the predicted distances

do not. As shown in Figure 6, a sp model was inserted in the initial paragraph HMM at every word boundary and the resulting HMM was retrained by using all the 370 speakers' data. The number of states of the HMM is 3×221 + N, where 221 is the number of phoneme instances of the SAA paragraph and N is the number of word boundaries. The UBM was adjusted to each speaker separately with MAP adaptation. After adaptation, sp states were removed from the HMM. By calculating the averaged BD between every pair of phoneme instance HMMs in the paragraph HMM, a 221×221 phoneme-based distance matrix was obtained for each speaker. The i-th phoneme instance HMM in a paragraph HMM is the three-state HMM spanning from the $(3i-2)$-th state to the $3i$-th state of that paragraph HMM. As shown in Figure 7, from the two distance matrices of speakers S and T, we derived another matrix D to represent the differences between them.

$$D_{ij} = |S_{ij} - T_{ij}|\,, \text{where } i < j. \tag{2}$$

$\{D_{ij}\}$ were used as input features to SVR to predict the pronunciation distance. The total number of the features was 24,310 and the ϵ-SVR in LIBSVM [19] was used with the radial basis function kernel of $K(x_1, x_2) = \exp(-\gamma|x_1 - x_2|^2)$.

Acoustic features used for training the paragraph-based UBM-HMM and adapting it were MFCC-based features; MFCC + ΔMFCC. For pronunciation analysis, BD was calculated by using MFCC features only. 68,265 speaker pairs of the SAA were sorted by the order of IPA-based reference distances and they were divided into two groups of even-numbered pairs and odd-numbered pairs. For training a predictor and testing it, 2-fold cross-validation was done using these two groups.

Figure 8 shows three correlations of A) perfect phoneme recognizer, B) real phoneme recognizer, and C) our proposed method [3]. A large improvement is achieved and C) is higher than A). In the next section, some new results of analytical investigation of the proposed method are described. Here, feature selection and addition of other features are tested.

VI. Analytical Investigation of Our Proposal

A. Feature selection based on locality of speech contrasts

In [3], from a difference matrix of $\{D_{ij}\}$, all the elements were used. As shown in Figure 7, D_{ij} is a difference between two speech contrasts or edges of S_{ij} and T_{ij}. Some contrasts are local contrasts but others are distant contrasts. For example, the last phoneme of the SAA paragraph is distant from the first one by 220 phonemes, which is the most distant speech contrast in $\{S_{ij}\}$ and $\{T_{ij}\}$. To investigate whether local contrasts

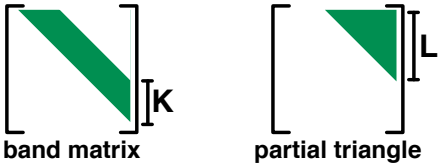

Fig. 9. Feature selection based on locality of speech contrasts

contribute better than distant contrasts, feature selection was done based on locality of speech contrasts. Figure 9 shows two methods compared in this section. The first one is use of only local contrasts in a band matrix, where K is the width of the band and it varies from 1 to 220. The last one is use of only distant contrasts in a partial triangle, where L also varies from 1 to 220. Use of the full triangle gives us 24,310 features and the number of features in the two cases depends on K and L.

B. Feature selection based on phonetic attributes

Transform-invariance of f-divergence requires an assumption that an entire space has to be mapped to another by a single transformation. This assumption is not always valid because, for example, MLLR-based HMM adaptation often needs multiple transform matrices for multiple phonetic classes. In this section, the 221 phoneme instances in the SAA paragraph were divided into four phonetic classes of A) vowels, B) resonant consonants, which are all voiced, C) other voiced consonants, and D) the rest (unvoiced consonants). Two cases were examined. In one case, two distance matrices were calculated for two groups of A)+B)+C) and D) and in the other, four matrices were calculated for four groups of A), B), C), and D). In the former case, the number of features were 14,752 and in the latter, it was 6,593.

C. Addition of absolute features

We investigated effectiveness of using acoustic distances obtained by direct and absolute comparison. Since the paragraph HMM is a sequence of 221 phoneme instance HMMs, we can get new 221 phoneme-based BDs between a speaker's paragraph HMM and another speaker's. These 221 BDs can be used as additional features to our original feature set.

D. Results and discussion

All the results of the above investigations are shown in Figure 10, where the correlations that were obtained in the individual experiments are plotted. If we compare use of band

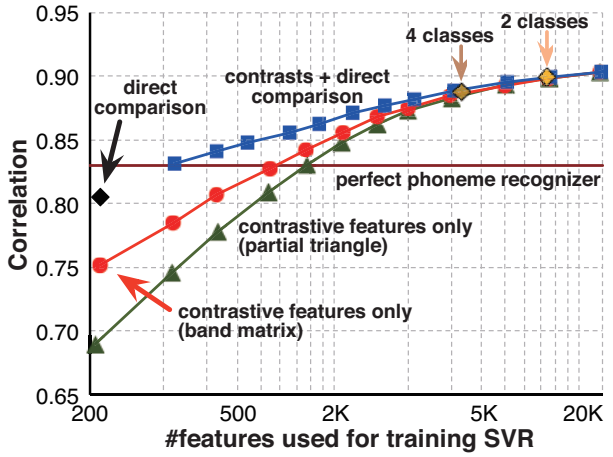

Fig. 10. Correlations obtained by analytical investigations

matrices and that of partial triangles, we can say that local contrasts are more effective than distant contrasts when the number of features is small but when it is large enough, they show very similar performances. Use of phonetic knowledge for feature selection (feature grouping) also seems ineffective when the number of features is large enough.

As for absolute comparison, when the 221 phoneme-based BDs are only used in SVR, the correlation was 0.805, which is much higher than the performance of K=1 (#features=220). However, it was also found that absolute features were ineffective when the number of contrastive features is large enough.

Although all the results obtained in the experiments show that contrastive or structural features are very useful when a large number of them can be used for prediction. However, this effectiveness may have been attributed to the experimental condition adopted in this paper. In the task of pronunciation distance prediction between a speaker pair, differential features are extracted from the speaker pair and used as input to SVR. In this paper, $\{D_{ij}\}$ are used and, as shown in Section V, 2-fold cross-validation was done. However, this cross-validation was designed in terms of speaker pairs, not speakers. Because differential features of two speakers are used as input to SVR, when the number of speakers used in the experiments is N, the intrinsic diversity of differential features will be estimated as $O(N^2)$. In the evaluation experiment in Section V, when one of the testing data is $\{D_{ij}\}$ of speakers A and B, the training data include samples of A-to-$\{x_n\}$ and B-to-$\{y_n\}$, where $x_n \neq$B and $y_n \neq$A. In SVR, the inner product of an input sample and each of the training samples is calculated in a very high dimensional space. Values of inner product can be regarded as similarity scores and regression or prediction is done by using these scores as weight. In the above case of speakers of A and B, the prediction performance is supposed to be influenced by whether $\{x_n\}$ include a speaker who is close to B or whether $\{y_n\}$ include a speaker who is close to A in the training data. Considering these facts, the experimental condition adopted in this paper may have reduced the intrinsic diversity of input features to $O(N)$ from $O(N^2)$. In the near future, we will run experiments based on speaker-based n-fold cross-validation to clarify this point.

VII. CONCLUSIONS

This paper firstly explained our recently proposed method of predicting pronunciation distances between any speaker pair who read a common paragraph. This method is based on combining pronunciation structure analysis for feature extraction and support vector regression for prediction. Then in this paper, some analytical investigations were done to understand how the proposed method works better. Although high effectiveness of contrastive (structural) features was shown through the investigations, a possible problem in the adopted experimental condition was also indicated. In the future, as well as clarifying this point, we will collect samples of World Englishes more intensively and extensively to approach our ultimate goal.

REFERENCES

[1] B. Kachru *et al.*, *The handbook of World Englishes*, Wiley-Blackwell, 2009.

[2] J. Jenkins, *World Englishes: a resource book for students*, Routledge, 2009.

[3] S. Kasahara, *et al.*, "Improved and robust prediction of pronunciation distance for individual-basis clustering of world Englishes pronunciation," *Proc. ICASSP*, 2014 (to appear).

[4] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," Proc. ICASSP, 889–892, 2005.

[5] N. Minematsu, *et al.*, "Speech structure and its application to robust speech processing", *Journal of New Generation Computing*, 28, 3, 299–319, 2010.

[6] Y. Qiao, *et al.*, "A study on invariance of f-divergence and its application to speech recognition", *IEEE Trans. on Signal Processing*, 58, 7, 3884–3890, 2010.

[7] M. Suzuki, *et al.*, "Integration of multilayer regression with structure-based pronunciation assessment," *Proc. INTERSPEECH*, 586–589, 2010.

[8] Speech Accent Archive, http://accent.gmu.edu.

[9] M. Wieling *et al.*, "A cognitively grounded measure of pronunciation distance," *PLoS ONE*, DOI: 10.1371/journal.pone.0075734, 2014.

[10] M. Wieling *et al.*, "Automatically measuring the strength of foreign accents in English," http://urd.let.rug.nl/nerbonne/papers/WielingEtAl-Accents-Validating-2013-final1.pdf

[11] The CMU pronunciation dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict

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

[13] N. Minematsu, *et al.*, "Measurement of objective intelligibility of Japanese accented English using ERJ (English Read by Japa-nese) database," *Proc. INTERSPEECH*, 1481–1484, 2011.

[14] H.-P. Shen, *et al.* "Speaker-based accented English clustering using a world English archive", *Proc. SLaTE*, CD-ROM, 2013.

[15] HTK Wall Street Journal Training Recipe, http://www.keithv.com/software/htk/

[16] R. Jakobson *et al.*, *The sound shape of language*, Mouton de Gruyter, 2002.

[17] J. Maye *et al.*, "Infant sensitivity to distributional information can affect phonetic discrimination," *Cognition*, 82, B101–B111, 2002.

[18] J. F. Werker *et al.*, "Infant-directed speech supports phonetic category learning in English and Japanese," *Cognition*, 103, 147–162, 2007.

[19] C.-C. Chang *et al.*, LIBSVM, a library for support vector machine, 2001.