AUTOMATIC PRONUNCIATION CLUSTERING USING A WORLD ENGLISH ARCHIVE AND PRONUNCIATION STRUCTURE ANALYSIS

H.-P. Shen[1,2], N. Minematsu[2], T. Makino[3], S. H. Weinberger[4], T. Pongkittiphan[2], C.-H. Wu[1]

[1] National Cheng Kung University, Tainan, Taiwan, [2] The University of Tokyo, Tokyo, Japan
[3] Chuo University, Tokyo, Japan, [4] George Mason University, Virginia, USA
[2]{happy,mine,teeraphon}@gavo.t.u-tokyo.ac.jp, [3]mackinaw@tamacc.chuo-u.ac.jp,
[4]weinberg@gmu.edu, [1]chwu@csie.ncku.edu.tw

ABSTRACT

English is the only language available for global communication. Due to the influence of speakers' mother tongue, however, those from different regions inevitably have different accents in their pronunciation of English. The ultimate goal of our project is creating a global pronunciation map of World Englishes on an individual basis, for speakers to use to locate similar English pronunciations. If the speaker is a learner, he can also know how his pronunciation compares to other varieties. Creating the map mathematically requires a matrix of pronunciation distances among all the speakers considered. This paper investigates invariant pronunciation structure analysis and Support Vector Regression (SVR) to predict the inter-speaker pronunciation distances. In experiments, the Speech Accent Archive (SAA), which contains speech data of worldwide accented English, is used as training and testing samples. IPA narrow transcriptions in the archive are used to prepare reference pronunciation distances, which are then predicted based on structural analysis and SVR, not with IPA transcriptions. Correlation between the reference distances and the predicted distances is calculated. Experimental results show very promising results and our proposed method outperforms by far a baseline system developed using an HMM-based phoneme recognizer.

Index Terms — World Englishes, speaker-based pronunciation clustering, pronunciation structure analysis, *f*-divergence, support vector regression

1. INTRODUCTION

English is the only language available for global communication. In many schools, native pronunciation of English is 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 [1,2,3,4] and they regard US and UK pronunciations just as two major examples of accented English. Diversity of World Englishes is found in various aspects of speech acts such as dialogue, syntax, pragmatics, lexical choice, pronunciation etc. Among these kinds of diversity, this paper focuses on pronunciation. If one takes the concept of World Englishes as it is, he can claim that every kind of accented English is equally correct and equally incorrect. In this situation, there will be a great interest in how one type of pronunciation is *different* from another, not in how that type of pronunciation is *incorrect* compared to US or UK pronunciation. As shown in [5], the intelligibility of spoken English heavily depends on the nature of the listeners as well as that of the speaker and the spoken content, and foreign accented English can indeed be more intelligible than native English. Generally speaking, speech intelligibility tends to be enhanced among speakers of similarly accented pronunciation.

The ultimate goal of our project is creating a global map of World Englishes on an individual basis for each of the speakers to know how his pronunciation is located in the diversity of English pronunciations. If the speaker is a learner, he can then find easy-to-communicate English conversation partners, who will have a similar kind of pronunciation. If he is too distant from many of other varieties, however, he will have to correct his pronunciation to achieve smoother communication with these others.

To the best of our knowledge, our project is the first trial to cluster World English pronunciations automatically and even on an individual basis. For this project, however, we have two major problems. One is collecting data and labeling them, and the other is creating a good algorithm of drawing the global map for a huge amount of unlabeled data. In [6], some accented English corpora with good quality were introduced. However, labeling data is needed in this paper. Luckily enough, for the first problem, the fourth author has made a good effort in systematically collecting World Englishes from more than a thousand speakers from all over the world and labeling them. This corpus is called the Speech Accent Archive (SAA) [7], which provides speech samples of a common elicitation paragraph and their narrow IPA transcriptions. To solve the second problem, we propose a method of clustering speakers only in terms of pronunciation differences. Clustering items can be performed by calculating a distance matrix among all of them.

The technical challenge here is how to calculate the pronunciation distance between any pair of speakers in the archive, where irrelevant factors involved in the archive, such as differences in age, gender, microphone, channel, background noise, etc have to be ignored adequately. To this end, we use pronunciation structure analysis for feature extraction and we also use support vector regression for distance prediction. The invariant structure analysis was proposed in [8,9] inspired by Jakobson's structural phonology [10] and it can extract invariant and robust features. The structural features were already introduced to various tasks such as pronunciation scoring [11,12], pronunciation error detection [13], language learners clustering [14], dialect analysis [15], and automatic speech recognition [16,17,18].

2. SPEECH ACCENT ARCHIVE

The corpus is composed of read speech samples of more than 1,700 speakers and their corresponding IPA narrow transcriptions. The speakers are from different countries around the world and they read a common elicitation paragraph, shown in Fig. 1, where an example of IPA transcription is also presented. The paragraph contains 69 words and can be divided into 221 phonemes using the CMU dictionary as reference [19]. The IPA transcriptions will be used to prepare reference inter-speaker pronunciation distances as label, which will be adopted as target of prediction using SVR in our study. This is because IPA transcription is done through phoneticians' ignorance of non-linguistic and acoustic variations found in utterances such as differences in age, gender, channel, etc. It should be noted that the recording condition in the corpus varies from sample to sample because the audio data were collected under many different situations. 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 us only "controlled" diversity. In [20], however, English sentences read by 200 Japanese university students showed a very large pronunciation diversity and [21] showed that the intelligibility of the individual utterances to American listeners covered a very wide range. Considering these facts, we considered that read speech samples can still show well how diverse World English pronunciations are.

It is well-known that pronunciation diversity is found in both segmental and prosodic aspects. In this study, however, we will prepare reference pronunciation distances by using IPA transcriptions, which means that prosodic diversity will be ignored. We do not claim that the prosodic diversity is minor but, as will be shown in this paper, clustering of English users only based on the segmental aspect seems able to show validly how diverse World Englishes are in terms of pronunciation. Preparation of reference distances with prosodic variation considered will be a future work.

"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."

[pʰl̥iːz kʰɑːl stɛlʌ æsk hə˞ ɾə bɹĩŋ ðiːz θĩŋz wɪθ heɪ fɪʌ̃m ðə stoːɹ sɪks spɵ̃ːnz əɣ̞ fɪɛʃ snoʊ pʰiːʐ faːɣ̞ θɪkˀ slæːbz ə blɥ̩ː ʃĩːʐ ɛ̃n meɪbi ə snæ̀kˀ foɪ hə˞ bɪʌðə bɑb wi ɑlso nɨːd ə smɑːl pʰl̥æstɪk sneɪkˀ æ̃n ə bɪgˀ tʰɔɪ fɪog fõɹ ðə kʰiɪdʐ ʃi kʰĩn skɥ̩ːp ðiːʐ θĩŋz̃ ĩntʰɥ̩ θɹi ɹɛ̨d bæːgz æ̃n wi wɪl miːt heɪ wɛ̃ntsdi æt ðə tɹeĩn steɪʃĩn]

Fig. 1 The elicitation paragraph used in the SAA and an example of narrow IPA transcription

In 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 as candidate files and then, 515 speakers' files were obtained. Some of these files were found to include a very high level of background noise and many pauses, and we manually removed them. Finally, 381 speakers' data were obtained and used here.

3. REFERENCE INTER-SPEAKER PRONUNCIATION DISTANCE

In this study, a pronunciation distance predictor based on pronunciation structure analysis is constructed. To this end, we have to prepare reference inter-speaker distances in the speech data, which can be used to train the distance predictor and verify the predicted distances. In this paper, the reference pronunciation distance between two speakers is calculated through comparing their individual IPA transcriptions using dynamic time warping (DTW). Since all the transcriptions contain exactly the same number of words, word-level alignment is easy and we only have to deal with phone-level insertions, deletions, and substitutions between a word and its counterpart in a transcription pair.

The process of estimating reference inter-speaker distances can be divided into two steps. Since DTW-based alignment of two IPA transcriptions needs a distance matrix among all the existing IPA phones in the archive, we prepared the distance matrix in the first step. We calculated frequency of each kind of the IPA phones, many of which were with a diacritical mark, and extracted the IPA phones that covered 95% of all the phone instances in the archive. The number of the kinds of the extracted phones with/without a diacritical mark was 153. One phonetician, the third author, was asked to pronounce each of these phones twenty times. Here, he was requested to pay attention to diacritical difference within the same IPA phone. In the recording, the phonetician pronounced each vowel twenty times. For consonants, a consonant was succeeded and preceded at the same time by vowel [a]. For example, to

collect samples of phone [p], the phonetician spoke [apa] twenty times. In this way, every consonant was recorded.

Using the wav files and their IPA transcriptions, a speaker-dependent three-state HMM was constructed for each phone, where each state contained a Gaussian distribution. After training the HMMs for all the phones, the Bhattacharyya distance was calculated between two corresponding states of each phone pair. By averaging the three state-to-state distances, we could finally define the acoustic distance between any phone pair. We note here that, since the HMMs were speaker-dependent, all the models were built in the same and matched condition.

The other 5% phones, which were not pronounced by the phonetician, were all with a diacritical mark. So, for these phones, we substituted the HMMs of the same phones with no diacritical mark. Using these HMMs, the inter-phone distance information among all the existing kinds of phones in the archive can be estimated. Due to limit of space, we do not visualize the 153x153 phone-based distance matrix in this paper, but by converting it to a tree diagram, we confirmed that we obtained a phonetically valid distance matrix. This was used as local distance or penalty in the next step to estimate the inter-speaker distance through DTW alignment between any two transcriptions.

In the next step, DTW was conducted to compare two IPA transcriptions in a word-by-word manner by using the phone-to-phone distance matrix. The resulting distance between two speakers will be used as reference inter-speaker distance. Because all the used files contained exactly 69 words, word-level alignment was easy and we could focus only on phone-level differences in each word pair between two IPA transcriptions. The local and allowable path of the DTW used in this section is shown as Fig. 2.

P1, *P2* and *P3* are allowable paths of insertion, match and deletion. Path selection is done based on equation 1.

$$DTW[m, n] := minimum(DTW[m-1,n] + phone\_dist[m,n], DTW[m-1,n-1] + 2*phone\_dist[m,n], DTW[m,n-1] + phone\_dist[m,n]) \quad (1)$$

DTW[m, n] is the current accumulated cost at position *(m,n)* and *phone_dist[m,n]* is a distance between the phone of time *m* and the phone of time *n*. Out of P1, P2, and P3, the path of which the accumulated cost at *(m,n)* is the minimum is selected. For DTW, phone-to-phone distances were used as penalty and we obtained a distortion score for each word pair between the two transcriptions. After normalizing this score by the number of phones found in the word pair, the score was summed for all the 69 words existing in the two transcriptions. This final score will be used as reference inter-speaker distance, namely, in training our predictor and in verifying the predicted distances.

After obtaining the inter-speaker distances, all the speakers can be clustered using Ward's method, one of the hierarchical clustering methods. Pronunciation can be affected by their mother tongue in different ways and to different degrees. In Fig.3, the clustering result of 18 selected speakers is shown. We picked up German speakers from the archive who were born in Germany, the number of whom was 9. Then, 9 native American English speakers were randomly selected. "EN" and "GE" denote American and German, respectively. The numbers succeeding "EN" or "GE" in the figure are speaker IDs. From Fig. 3, it can be seen that the all American speakers are clustered into one sub-tree and eight German speakers are clustered into the other sub-tree. Although GE16 is clustered into the same sub-tree with American speakers, by inspecting his biography included in the SAA, it is found that he had lived in USA for 4 years. It seems that his pronunciation has been reasonably affected by and adapted to American accent. On the other hand, most of the other German speakers had lived in America less than 1 year. We consider that this result indicates that the estimated inter-speaker distances are valid enough.

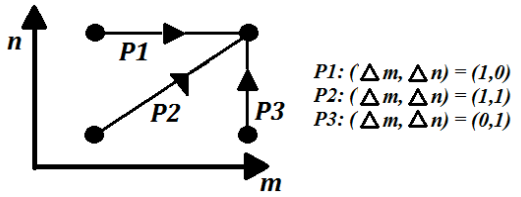

Fig. 2 Allowable paths of the DTW

4. BASELINE SYSTEM

For comparison, we built a baseline system, which corresponds directly to an automated version of the inter-speaker distance calculation procedure described in section 3. As mentioned above, the procedure is composed of two steps: 1) IPA manual transcription and 2) DTW alignment for distance calculation. In the baseline system, the process of 1) is replaced with automatic recognition of phonemes in input utterances[1]. Here, monophone HMMs were obtained through ML-based training using the WSJ-based monophone HMMs [22] as initial model and all the utterances of the 381 SAA speakers as training samples. For this training, each IPA transcription was converted into American phoneme transcription. 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 American diphthong.

Since IPA transcription is based on phones and HMMs are trained based on phonemes, even if we could have a perfect phoneme recognizer, the generated transcriptions have to be phonemic versions of IPA transcriptions. Phone to

[1]As far as we know, there does not exist an automatic recognizer of IPA phones with a diacritical mark. Then, we used a phoneme recognizer of American English instead in this study.

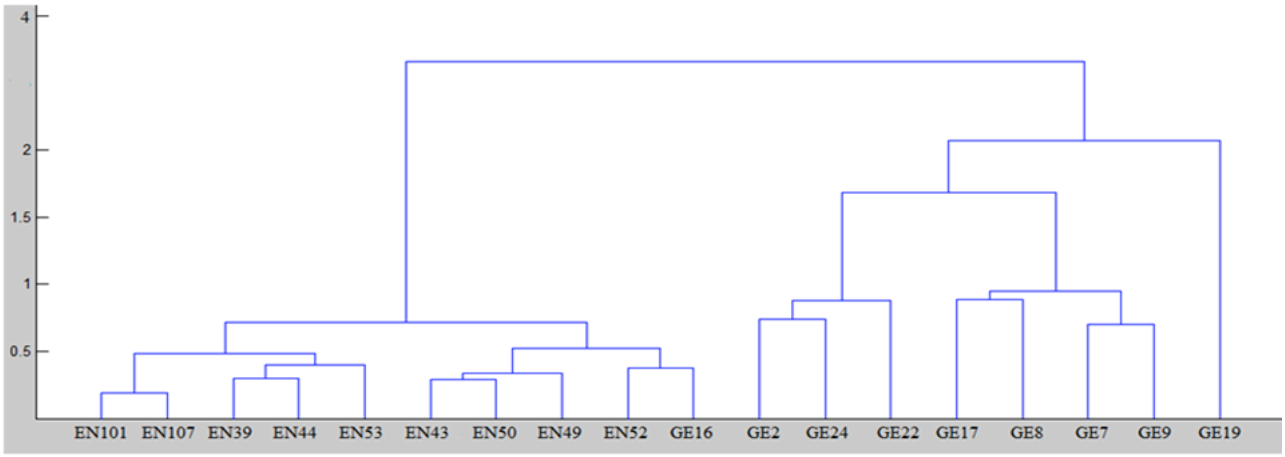

Fig.3 The clustering result of 18 selected speakers

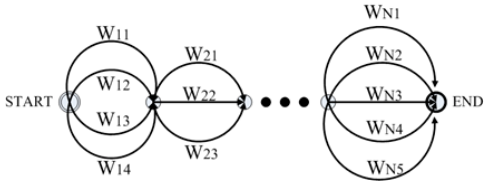

Fig.4 An example of word-based grammar

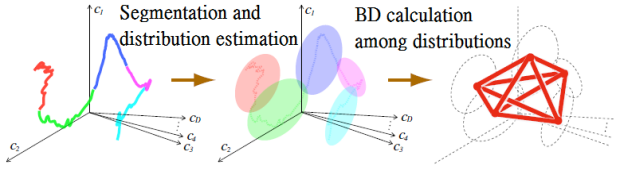

Fig. 5 Procedure of representing an utterance only by BD

phoneme conversion is an abstraction process and some detailed phonetic information will be lost inevitably. To evaluate this abstraction process quantitatively, we calculated correlation between the inter-speaker distances obtained in section 3 and those obtained by using perfect phoneme recognition results and DTW. The perfect results are the phone-to-phoneme conversion results explained above. Here, DTW alignment between any two phoneme transcriptions was done by using a phoneme-to-phoneme distance matrix, which was obtained from the same monophone HMMs as above. The correlation was found to be 0.882, meaning that information loss exists to some degrees.

What about a real phoneme recognizer? By using the phone-to-phoneme conversion results above, we can build word-based network grammar which can cover all the pronunciation diversity found in the 381 speakers. Fig. 4 shows an example of word-based network grammar. In this figure, W_{ij} denotes the i-th word and the j-th possible pronunciation. Using this grammar, each utterance can be converted into a phoneme sequence automatically. It should be noted that the monophone HMMs and the network grammar were built in a speaker-closed manner. The phoneme recognition accuracy was 73.36%. Considering a recent study on pronunciation error detection [23], this performance is very reasonable. However, the correlation between the IPA-based reference inter-speaker distances and the inter-speaker distances using automatically generated phonemic transcriptions and DTW was found to be so low as 0.313. This clearly indicates that phoneme recognition errors are very influential to inter-speaker distance calculation and real phoneme recognizers are not working well for this task.

5. INVARIANT PRONUNCIATION STRUCTURE

As described in section 1, we have to use a very robust method to estimate the pronunciation distance. Minematsu et al. proposed a new method of representing speech, called speech structure, and proved that the acoustic variations, corresponding to any linear transformation in the cepstrum domain, can be effectively unseen in the representation [9]. This invariance is due to the invariance of the Bhattacharyya distance (BD), which is calculated using equation 2 and is proved to be invariant with any linear transform.

$$D_B = \frac{1}{8}(M_1 - M_2)^T P^{-1}(M_1 - M_2) + \frac{1}{2}\ln\left(\frac{\det P}{\sqrt{\det P_1 \det P_2}}\right), \quad (2)$$

where M_1, M_2 are mean vectors and P_1, P_2 are covariance matrices of two Gaussian distributions. $P = (P_1+P_2)/2$.

Fig. 5 shows the procedure of representing an input utterance only by BD. The utterance in a cepstrum space is a sequence of vectors and it is converted into a sequence of distributions through automatic segmentation. Here, any speech event is characterized as distribution. The BD is calculated from any pair of distributions and the resulting full set of the BDs forms an invariant distance matrix. This ma-

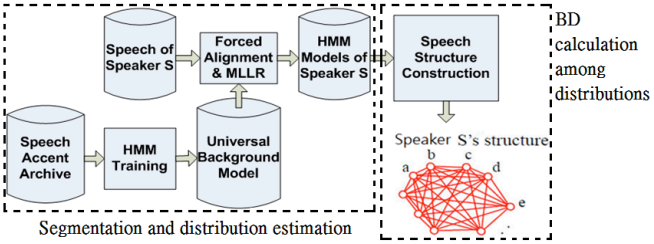

Fig.6 Speaker-independent pronunciation structure

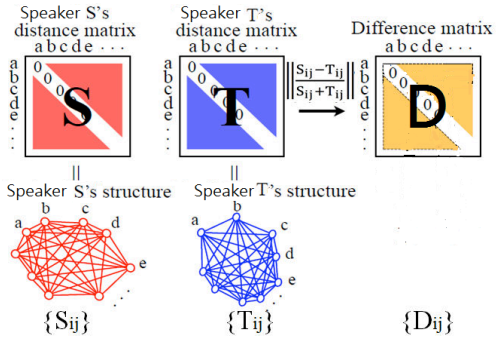

Fig.7 Inter-speaker structure difference [12]

trix-based representation of an utterance is called pronunciation structure [9]. The structure only represents the local and global contrastive aspects of a given utterance, which is theoretically similar to Jakobson's structural phonology [10]. By calculating the BD of every pair of sound units in the elicitation paragraph read by a speaker, the pronunciation structure specific to that speaker can be obtained. Thus, the structural differences between two speakers can be used as features to predict the inter-speaker pronunciation distance.

Fig. 6 shows the procedure to construct a pronunciation structure much more in detail. We firstly trained a paragraph-based universal background HMM using all the data available. 24-dimensional MFCCs (MFCC +Δ MFCC) were used to train the HMM. Here, the paragraph was converted into its phoneme sequence by using the canonical pronunciation of each word found in the CMU dictionary. The number of the states in the background HMM was 3M, where M is the number of phonemes in the paragraph. To construct a specific speaker's HMM, forced alignment of that speaker's utterance was done to obtain state boundaries and MLLR adaptation was done to adapt the background model to that speaker. In MLLR adaptation, the number of regression classes used is 32. In the adapted model, each state contains one Gaussian. Finally, the BD is calculated between a state and another in the adapted HMM. By assuming that three consecutive states form a phoneme-like unit, the averaged BD distance ($d_{p_i p_j}$) was calculated between a unit p_i and another unit p_j in equation 3.

$$d_{p_i p_j} = \sqrt{\frac{BD(p_i^1, p_j^1) + BD(p_i^2, p_j^2) + BD(p_i^3, p_j^3)}{3}}. \quad (3)$$

p^1, p^2 and p^3 are the first, second and third states of the phoneme-like unit p. All the distances $\{d_{p_i p_j}\}$ are used together to derive the pronunciation structure. The distance matrix S_{matrix} of speaker S can be represented as follows

$$S_{matrix} = \begin{bmatrix} 0 & d_{p_1p_2} & . & . & . & d_{p_1p_N} \\ d_{p_2p_1} & 0 & . & . & . & . \\ . & . & 0 & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & 0 & d_{p_{N-1}p_N} \\ d_{p_Np_1} & . & . & . & d_{p_Np_{N-1}} & 0 \end{bmatrix}, \quad (4)$$

This matrix is reasonably symmetric and only the elements found in the upper triangle are used to form the pronunciation structure of a specific speaker.

For two given pronunciation structures (two distance matrices) from speakers S and T, a difference matrix between the two is calculated by equation 5 (D in Fig. 7).

$$D_{ij}(S,T) = \left| \frac{S_{ij} - T_{ij}}{S_{ij} + T_{ij}} \right|, \quad \text{where i < j.} \quad (5)$$

S_{ij} and T_{ij} are (i,j) elements in S and T. Since S_{ij} and T_{ij} are invariant features, D_{ij} also becomes an invariant and robust feature. For speaker-based clustering of World Englishes, we use D_{ij} as a feature in support vector regression.

6. SVR TO PREDICT PRONUNCIATION DISTANCES AMONG SPEAKERS

Using the IPA-based reference distance between any two speakers as target and using the upper triangle elements of the difference matrix D between them as input attribute, we trained a model of support vector regression (SVR). In this paper, LIBSVM [24] was adopted to train the SVR. Here, the epsilon-SVR was used. The kernel type is a radial basis function: *exp(-gamma * |x1-x2|^2)*.

For this experiment, we divided the elicitation paragraph into 9 sentences. Therefore 9 pronunciation structure matrices were obtained, one for each sentence. From all of them, a set of 2,804 unit-to-unit distances were obtained for each speaker. Then, between any two speakers, 9 difference matrices can be obtained, which also have 2,804 elements.

For performance evaluation, the correlation between the IPA-based reference distances and the predicted distances was calculated. We divided all the speaker pairs into 2 sets based on the reference distances and performed a 2-fold cross-validation, where a set was used to train SVR and the other set was used for testing. The correlations found in both test sets were 0.808 and 0.812. The average correlation was 0.810. Fig. 8 shows the prediction results of both sets simultaneously. It is clearly shown that our system outperforms by far the speaker-closed baseline system (corr. = 0.313) and the performance of our system can be said to be close to

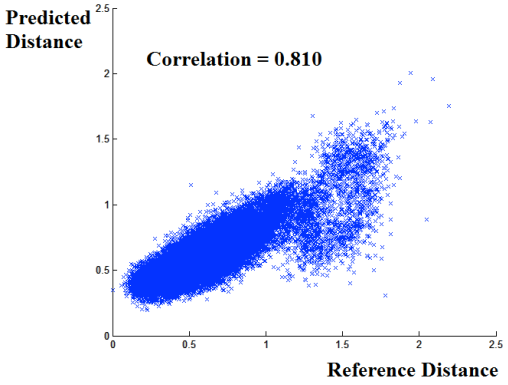

Fig.8 Correlation of the predicted distances and the reference inter-speaker distances

that of an imaginary perfect phoneme recognizer (corr. = 0.882), although there still exists a certain performance gap. In Fig. 8, a large number of dots are found closer to the diagonal line, but not a small number of dots are found off the line. Currently, we're investigating these data. We also consider that our system can become more comparable to the perfect recognizer by tuning input features and regression methods. For features, we can use Multiple Stream Structuralization (MSS) [9] and, as discussed in [12], use of absolute features in addition to contrast (relational) features will also be effective to improve the performance. For regression, we're interested in applying kNN-SVR [25] to our task.

7. CONCLUSIONS

With the ultimate aim of drawing the global map of World Englishes on an individual basis, this paper investigated invariant pronunciation structure and SVR to predict inter-speaker pronunciation distances for new speaker pairs. The speech accent archive, containing data from worldwide accented English speech, was used as training and testing samples. Evaluation experiments showed very promising results. The correlation between the IPA-based reference inter-speaker distances and the predicted inter-speaker distances obtained using the proposed method was 0.810, which is absolutely higher than the correlation obtained by the baseline system using a phoneme recognizer. In future work, we are planning to make the proposed predictor more comparable to the perfect phoneme recognizer and collect a more data using smart phones and social network infrastructure such as crowdsourcing. Pedagogical application of the World and individual English map will also be considered in collaboration with language teachers.

8. REFERENCES

[1] D. Crystal, *English as a global language*, Cambridge University Press, New York, 1995.
[2] J. Jenkins, *World Englishes: a resource book for students*, Routledge, 2009.
[3] B. Kachru, et al., *The handbook of World Englishes*, Wiley-Blackwell, 2009.
[4] A. Kirkpatrick, *The Rougledge handbook of World Englishes*, Routledge, 2012.
[5] M. Pinet, et al, "Second-language experience and speech-in-noise recognition: the role of L2 experience in the talker-listener accent interaction", *in Proc. of SLaTE*, CD-ROM, 2010.
[6] A. Hanani, et al, "Human and computer recognition of regional accents and ethnic groups from British English speech", Computer Speech & Language, vol. 27, Issue 1, pp. 59–74, 2013
[7] S. H. Weinberger, Speech Accent Archive, George Mason University, http://accent.gmu.edu ·
[8] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," *Proc. ICASSP*, pp.889-892, 2005.
[9] N. Minematsu, et al., "Speech structure and its application to robust speech processing", *Journal of New Generation Computing*, 28, 3, pp. 299-319, 2010.
[10] R. Jakobson and L. R. Waugh, *Sound shape of language*, Branch Line, 1979.
[11] M. Suzuki, et al., "Sub-structure-based estimation of pronunciation proficiency and classification of learners," *Proc. ASRU*, pp.574-579, 2009.
[12] M. Suzuki, et al., "Integration of multilayer regression with structure-based pronunciation assessment," Proc. INTERSPEECH, pp.586-589, 2010.
[13] T. Zhao, et al., "Automatic Chinese pronunciation error detection using SVM with structural features," *Proc. Spoken Language Technology*, pp.473-476, 2012.
[14] N. Minematsu, et al., "Structural representation of the pronunciation and its use for clustering Japanese learners of English," *Proc. SLaTE*, CD-ROM, 2007.
[15] X. Ma, et al. , "Dialect-based speaker classification using speaker invariant dialect features", in *Proc. of Int. Symposium on Chinese Spoken Language Processing*, pp.171-176, 2010.
[16] Y. Qiao, et al., "A study of Hidden Structure Model and its application of labeling sequences," *Proc. ASRU*, pp.118-123, 2009.
[17] Y. Qiao and N. Minematsu, "A study on invariance of f-divergence and its application to speech recognition," *IEEE Trans. on Signal Processing*, vol.58, no.7, pp.3884-3890, 2010.
[18] M. Suzuki, et al., "Discriminative reranking for LVCSR leveraging invariant structure," *Proc. INTERSPEECH*, CD-ROM, 2012.
[19] The CMU pronunciation dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[20] N. Minematsu, et al., "Development of English speech database read by Japanese to support CALL research," *Proc. ICA*, pp.557-560, 2004.
[21] N. Minematsu, et al., "Measurement of objective intelligibility of Japanese accented English using ERJ (English Read by Japanese) database," *Proc. INTERSPEECH*, pp.1481-1484, 2011.
[22] HTK Wall Street Journal Training Recipe http://www.keithv.com/software/htk/
[23] Y.B. Wang, "Improved Approaches of Modeling and Detecting Error Patterns with Empirical Analysis for Computer-Aided Pronunciation Training," *Proc. ICASSP*, pp.5049-5052, 2012.
[24] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, 2001.
[25] W.-L. Chao et al., "Facial age estimation based on label-sensitive learning and age-specific local regression", *In Proc. of ICASSP*, pp.1941-1944, 2012.