Speaker-based Pronunciation Clustering using World Englishes and Pronunciation Structure*

☆H.-P. Shen[1,2], N. Minematsu[2], S. H. Weinberger[3], T. Makino[4], T. Pongkittiphan[2] , C.-H. Wu[1]
(National Cheng Kung Univ.[1], The Univ. of Tokyo[2], George Mason Univ.[3] Chuo Univ. [4])

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] and they regard US and UK pronunciations as just two major examples of accented English. If one takes the philosophy of World Englishes as it is, we can claim that every kind of accented English is equally correct and equally incorrect. In this situation, there is a great interest in how one type of pronunciation is different from another. As shown in [3], the intelligibility of spoken English heavily depends on the nature of the listeners, and foreign accented English can indeed be more intelligible than native English. Generally speaking, intelligability tends to be enhanced among speakers of similarly accented pronunciation.

The ultimate goal of our project is creating a global map of World and *individual* Englishes, for speakers to use to locate similar Englishes. The speaker can then find the best English conversation partner. A learner can also know how his pronunciation geographically compares to other varieties. If he is too distant from these other varieties, he may have to correct his pronunciation *for the first time* to achieve smoother communication with these others. For this project, 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. Luckily enough, for the first problem, the third author has made a good effort in systematically collecting World Englishes from more than a thousand speakers from all over the world. This corpus is called Speech Accent Archive [4]. To solve the second problem in this paper, we propose a method of clustering speakers only in terms of their pronunciation. Clustering of items can be done by calculating a distance matrix among them. The technical challenge here is how to calculate the pronunciation distance between any pair of the speakers in the archive, where irrelevant factors involved in the data, such as differences in age, gender, microphone, channel, background noise, etc have to be ignored. For that, we use a pronunciation structure paradigm [5,6] with support vector regression (SVR). Our experiments demonstrate very promising results.

2 Speech Accent Archive

The corpus is composed of read speech samples of 1,716 speakers and their corresponding IPA transcriptions. The speakers are from different countries around the world and they read a common elicitation paragraph, shown in Fig. 1. It contains 69 words and can be divided into 221 phonemes based on the CMU dictionary [7]. Each sample has its detailed IPA transcription, which is provided by trained phoneticians, and an example is also shown in Fig. 1. The transcriptions can be used to prepare a reference for inter-speaker distances, which will be adopted as a target of prediction using support vector regression in our study.

The recording condition varies among samples because the audio data were collected under many different situations. To create a suitable

---

"話者を単位とした世界英語発音クラスタリング", 沈涵平[1,2], 峯松信明[2], スティーブン・ワインバーガー[3], ポンキッティパン・ティーラポン[2], 牧野武彦[4], 呉宗憲[1] (成功大[1], 東京大[2], ジョージ・メイソン大[3], 中央大[4])

“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.”

[pʰl̥i:z kʰɑ:l stɛlʌ æsk hɚ ɾə bl̃iŋ ði:z θl̃iŋz wɪθ hɛɪ fɪl̃m ðə sto:ɹ sɪks spl̃u̟:nz ə̯ fɪɛʃ snoʊ pʰi:ɻ̟ fa:ɣ̟ θɪkˀ slæ̟:bz ə blu̟: ɡl̃i:ɻ̟ l̃ɛn meɪbi ə snæ̟kˀ foɹ hɚ bɹʌðə bɑb wi ɑlso ni̟:d ə smɑ:l pʰl̥æstɪk sneɪkˀ æ̃n ə bɪgˀ tʰɔɪ fɪog fõɹ ðə kʰiɪdʑ̟ ʃi kʰl̃ĩn sku̟:p ði:ɻ̟ θl̃iŋz̃ l̃ɪntʰu̟ θɹi ɹɻ̟d bæ̟:gz æ̃n wi wɪl mi:t hɛɪ wɛ̃ntsdi æt ðə tɹɛ̃ĩn steɪʃĩn]

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

map, these acoustic variations including age- and gender-variation have to be cancelled well because these are irrelevant to the pronunciation structure analysis.

In this study, only the data with no word-level insertion or deletion were used. The audio files with exactly 69 words were selected as candidate files and 515 speakers’ files were obtained. Some of these files were found to include a high level of background noise, and we manually removed them. At the end of the day, 381 speakers’ data were used.

3 Inter-speaker Pronunciation Distance

A pronunciation distance predictor based on pronunciation structure was constructed. To this end, we prepared 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. Since all the transcriptions contain exactly the same number of words, word-level alignment between transcriptions is easy and we only have to deal with phone-level insertions, deletions, and substitutions between a word and its counterpart. It should be noted that diacritical marks in our IPA transcriptions were ignored because we wanted to focus mainly on phone-level differences between the two transcriptions. DTW-like comparison between a word and its counterpart gives us a penalty score depending on what kind of phone-level changes are found between the two. For insertion and deletion, a high penalty is given because they change syllable structure**.** For substitution, a lower penalty is assigned. By accumulating these phone-level penalties, a word-based penalty score was obtained. By accumulating these word-level penalties, we get the paragraph-based penalty between two speakers. Furthermore, in [4], some phonological generalization rules, which were defined by phoneticians, are used to describe each speaker’s pronunciation. These rules represent commonly observed substitution patterns. As they are very common, the lowest penalties, were assigned to them. The phonological generalization rules, penalties and their corresponding examples are shown in Table 1. By normalizing the penalty using the averaged number of phones over the two transcriptions, the final (and normalized) penalty score was obtained. This was defined as the inter-speaker pronunciation distance in this study.

Table 1 The used phonological generalization rules and penalties in calculating reference inter-speaker distances

Phonological generalization rules	Examples	Penalty (Distance increasing)
Final obstruent devoicing & Consonant voicing	b ⇔ p t ⇔ d	+1
Stop (plosive) =>Fricative	p => ɸ b => β	+1
Interdental fricative change	θ => t θ => d	+1
Alveolar approximant change	ɹ => r ɹ => X	+1
w => fricative	w => v	+1
h => velar fricative	h => x h => ɣ	+1
S -> s	ʃ => s	+1
Syllable structure change	Vowel insertion, Consonant deletion, Consonant insertion	+5
Phone-level substitution	Other substitutions	+3

Although the final and normalized scores are used as reference for inter-speaker distances in the following sections, we do not claim at all that the above procedure of calculating scores is the best and only procedure for our purpose. Our definition of penalty scores is very heuristic and what we want to claim here is that our proposed

"prediction" algorithm is expected to work independently of the definition of penalties.

4 Invariant Pronunciation Structure

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 completely unseen in the representation [5]. This invariance is attributed to the invariance of Bhattacharyya distance (BD), which is calculated using equation 1 and is proved to be invariant with any linear transform.

$$D_B = \frac{1}{8}(\mu_1-\mu_2)^T P^{-1}(\mu_1-\mu_2)+\frac{1}{2}\ln\left(\frac{\det\Sigma}{\sqrt{\det\Sigma_1 \det\Sigma_2}}\right) \quad (1)$$

where μ_1, μ_2 are mean vectors and Σ_1, Σ_2 are covariance matrices of two Gaussian distributions. $\Sigma = (\Sigma_1+\Sigma_2)/2$.

By calculating the BD of every pair of sound units in the elicitation paragraph read by a specific speaker, the unique distance matrix with respect to that speaker can be obtained. This sound structure is called the pronunciation structure in this paper and is shown in Fig. 2. The structure only represents the local and global contrastive aspects of a given utterance, which is theoretically similar to Jakobson's structural phonology [8]. [7] showed experimentally that the invariant pronunciation structure is useful to group dialects into clusters. Thus, the differences of the structures between two speakers can be used as features to estimate inter-speaker pronunciation distances.

To construct a specific speaker's pronunciation structure, we first trained a paragraph-based universal background hidden Makov model (HMM) using all the data available. Here, the paragraph was converted into a phoneme sequence using the CMU dictionary and, by using this as a reference phoneme sequence, a paragraph-based HMM was trained using all the data. Then, for each speaker, forced alignment of that speaker's utterance was done to obtain phoneme boundaries. Then, MLLR adaptation was done to adapt the universal model to that speaker. Using this adapted model, a phonemic segment was characterized as three Gaussians. Finally, three BDs are calculated

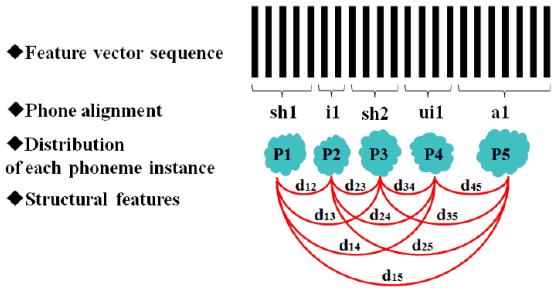

Fig.2 Extraction of structural features

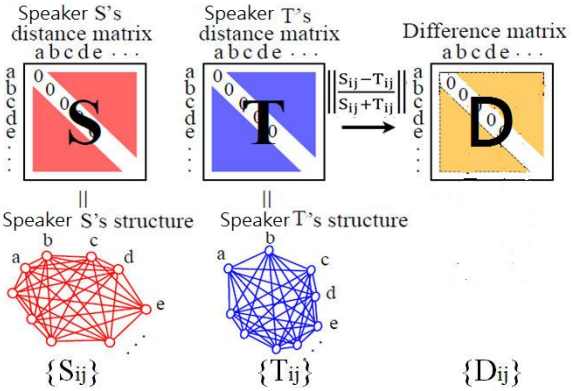

Fig.3 Inter-speaker structure difference

between two sequences of three Gaussians. They are averaged to give us the final BD score to derive the pronunciation structure. This process is shown in Fig. 2.

For two given pronunciation structures (two distance matrices) from speakers S and T, a difference matrix between the two is calculated by equation 2, which is shown as D in Fig. 3.

$$D_{ij}(S,T) = \left|\left(\frac{S_{ij}-T_{ij}}{S_{ij}+T_{ij}}\right)\right|, \quad \text{where i < j} \quad (2)$$

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.

5 SVR to Predict Pronunciation Distances among Speakers

Using the reference distance between any two speakers and the difference matrix D between them, we trained a support vector regression (SVR). Here, the SVR is expected to predict the reference distance using D features as input. In this paper, the epsilon-SVR is used. The kernel type is radial basis function: *exp(-gamma * |x1-x2|^2)*

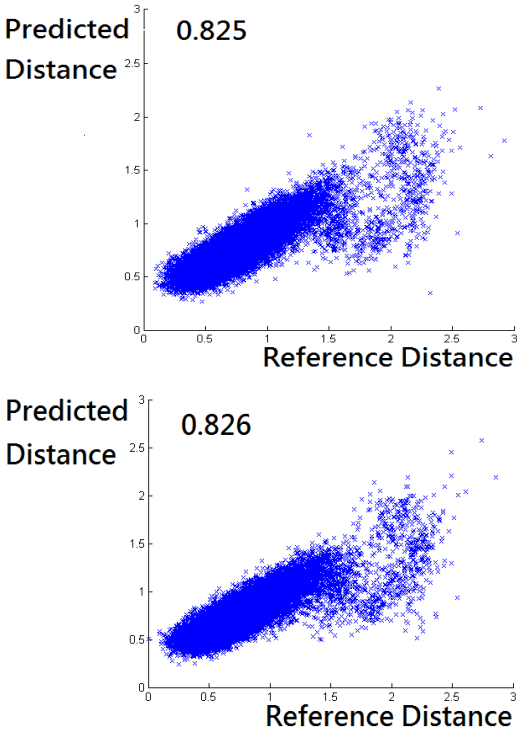

Fig.4 Correlation results of the two testing sets

We divided the elicitation paragraph into 9 sentences. Therefore 9 pronunciation structures were obtained, one for each sentence. From all of the 9 structures, a set of 2,806 phone distances were obtained for each speaker. An inter-speaker distance vector between two speakers was also represented as a set of 2,806 values. In speech structure construction, conventional 24-dimensional MFCCs (MFCC +Δ MFCC) were used to train the HMMs.

For the performance evaluation, correlation between the reference distances and the predicted distances was used. We divided all the speaker pairs into 2 sets based on the reference distances and performed a 2-fold cross-validation (1 set was used to train SVR and the other set was used for testing). The correlation results of the first set and second set were 0.825 and 0.826, respectively. Fig. 4 shows the correlation results of these two sets.

6 Conclusions

With the ultimate aim of drawing the global map of World and individual Englishes, 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. In future work, we are planning to further define the list of penalties, which may be obtained by acoustic analysis of every phone pair spoken by a single speaker. Moreover, a more extensive collection of data is planned 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.

References

[1] D. Crystal, *English as a global language*, Cambridge University Press, New York, 1995.

[2] J. Jenkins, *The phonology of English as an international language*, Oxford University Press, 2000.

[3] M. Pinet *et al*. “Second-language experience and speech-in-noise recognition: the role of L2 experience in the talker-listener accent interaction”, Proc. SLaTE, CD-ROM, 2010

[4] S. H. Weinberger, *Speech Accent Archive*, George Mason University http://accent.gmu.edu

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

[6] X. Ma, *et al*. i, *Dialect-based speaker classification using speaker invariant dialect features*, in Proc. Int. Symposium on Chinese Spoken Language Processing, pp.171-176, 2010

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

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