Automatic and individual-basis clustering of World Englishes pronunciations using the Speech Accent Archive

MINEMATSU, Nobuaki
Graduate School of Engineering, The University of Tokyo

1. Introduction

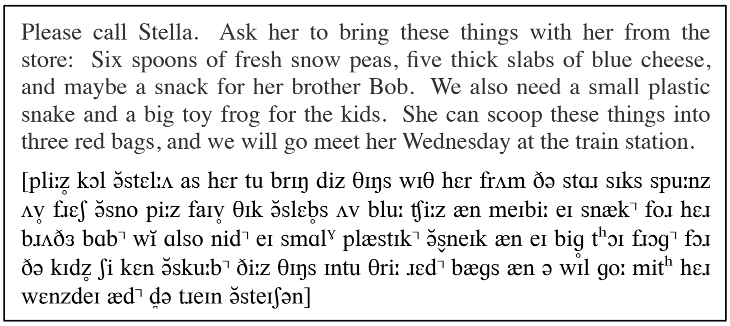
English is the only language available for global communication and its form is altered easily and variously by the mother tongue of its users. In this study, a focus is put on pronunciation diversity, called accents. The ultimate goal of this study is to create a global and individual-basis map of pronunciations of World Englishes (WE). In this paper, automatic clustering of English users only in terms of pronunciation is investigated. Clustering *N* items generally requires a method of measuring the distance between any pair of them. For that, we combine pronunciation structure analysis [1,2] and support vector regression [3] to predict the accent distance between two speakers. For regression, IPA-based reference distances are adopted for training and testing our predictor. Experiments are done in two modes of speaker-pair-open and speaker-open. Correlations between reference distances and predicted ones are 0.903 and 0.547, respectively. For the latter mode, technical improvements are still needed. Further in this paper, a tentative method for visualizing the obtained distances is also shown.

2. The minimal unit of accent diversity

What is the minimal unit of accent diversity? Is it country, region, prefecture, city, town, or village? Accent diversity is considered to be due to diversity of the language background and/or the learning background of individual speakers. This thinking leads us to the answer to the above question and it should be individual. We can say that WE have about 1.5 billion different kinds of pronunciations. These days, a huge number of people always take a high-quality microphone with themselves, called smart phone, and we can find several dialect studies [4] which use this infrastructure to collect a huge amount of data from speakers. We consider that data collection from all the users of English is not impossible. For example, [5] started collecting readings of a common and carefully designed paragraph from international users of English and this study uses a part of that.

3. The Speech Accent Archive (SAA)

The corpus is composed of read speech samples of more than 1,800 speakers and their corresponding IPA narrow transcripts. The speakers 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 [6]. These IPA transcripts are used in training and testing our predictor. The reference accent distance between a speaker pair is obtained by calculating the Levenshtein distance using the Dynamic Time Warping (DTW) algorithm. The DTW itself is an automatic procedure but it requires a huge amount of phoneticians' manual labor. In this work, we attempt to replace "human annotation + DTW" with "automatic and acoustic analysis + automatic prediction of distances". The technical challenge is how to predict the accent distance. If we measure the *acoustic* distance between readings of two speakers, the obtained distance will be influenced by difference in age and gender of the speakers. We have to create a technology that can measure the distance only in terms of pronunciation. The distance based on IPA and DTW is good as reference because phoneticians can ignore the above extra-linguistic differences and IPA transcripts do not show any attribute of age and gender.



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.

[pliːz̥ kɔl ə̆stɛlːʌ as hɛr tu brɪŋ diz θɪŋs wɪθ hɛr frʌm ðə stɑɹ sɪks spuːnz ʌv̥ fɹɛʃ ə̆sno piːz faɪv̥ θɪk ə̆slɛb̥s ʌv bluː ʧiːz æn meɪbiː eɪ snæk˺ foɹ hɛɹ bɹʌðɜ bɑb˺ wĭ ɑlso nid˺ eɪ smɑlˠ plæstɪk˺ ə̆s̬neɪk æn eɪ big̥ tʰɔɪ fɹɔg˺ fɔɹ ðə kɪdz̥ ʃi kɛn ə̆skuːb˺ ðiːz θɪŋs ɪntu θriː ɹɛd˺ bægs æn ə wɪ̊l goː mitʰ hɛɹ wɛnzdeɪ æd˺ d̪ə tɹeɪn ə̆steɪʃən]

Figure 1: The SAA elicitation paragraph and an example of IPA transcription

4. Reference inter-speaker accent distance

DTW is done between any pair of the transcripts. Here, it has to be assumed that speakers read 69 words with no word-level insertion or deletion. However, a large number of speakers in the SAA did inserted or deleted words. We removed those speakers and found that the number of the remaining speakers was 370, which is small but the number of speaker pairs is very large (370x369/2=68,265). In this work, a speaker pair, not a speaker, is a sample for predicting the distance.

Between two transcripts, word-level alignment is easy and we only had to treat phone-level insertions, deletions, and substitutions between a word and its counterpart. Since DTW-based alignment needs the distance matrix among all the existing IPA phones in the SAA, we prepared it in the following way. We found that the most frequent 153 kinds of phones in the SAA can cover 95% of all the phone instances. Then, we asked an expert phonetician to pronounce each of them twenty times, which were recorded. Using the recorded data, a speaker-dependent three-state acoustic model (Hidden Markov Model, HMM) was built for each phone, where each state was modeled as Gaussian distribution. For each phone pair, the phone-to-phone distance was defined as the average of three state-to-state Bhattacharyya distances. The other 5% of the phones were all with a diacritical mark. For each of them, we substituted the HMM of its base phone.

5. Pronunciation structure analysis

As is explained in Section 3, the acoustic distance between speakers is not the accent distance. The latter can be measured, for example, by removing speaker and age components in acoustic streams of given utterances. What remains can be considered as pronunciation skeleton and the skeletons of two speakers should be compared for distance prediction. In this work, we used pronunciation structure analysis [1,2], where only sound contrasts, not sound instances, were extracted, where the contrasts are carefully calculated so that they become independent of age and gender. After the structure analysis, a given utterance is represented as a distance matrix of the sounds observed in that utterance. Generally and geometrically speaking, an *N*x*N* distance matrix can determine uniquely its geometrical shape formed by the *N* items (points). Figure 2 is a conceptual illustration of the procedure of extracting the pronunciation structure from a given utterance. If readers want to know more of technical details, please refer to [1,2]. Figure 3 also shows formant-based vowel structures (distributions) of three US dialects [7]. It is clear that the geometrical shape of the vowels is strongly dependent on dialects.

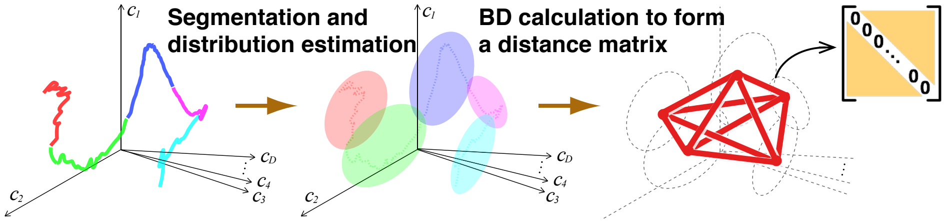

c_1
Segmentation and distribution estimation
c_1
BD calculation to form a distance matrix
c_2
c_D
c_4
c_3
c_2
c_D
c_4
c_3
0
0
0...
0
0

Figure 2: Procedure to extract the pronunciation structure from a given utterance

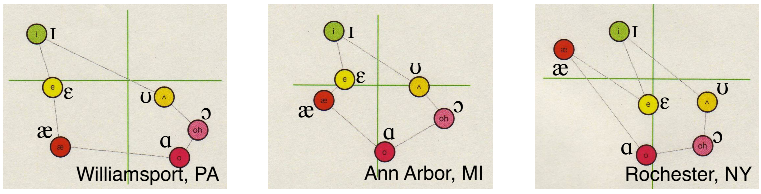
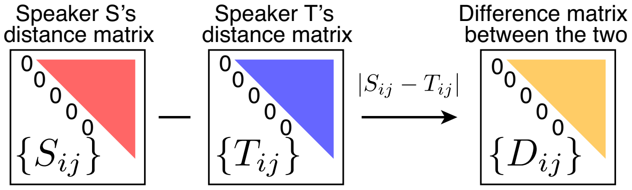


Figure 3: Vowel distribution patterns of three US dialects

6. Experiments of predicting the accent distances

All the available speakers from the SAA were divided into training data and testing data. A predictor was trained based on machine learning by using the training data and it was assessed by using the testing data. Two schemes of speaker division were examined. In one scheme, all the speaker pairs were sorted by referring to their IPA-based reference distances and even-numbered speaker pairs were adopted as training and the others were as testing, i.e., speaker-pair-open mode. In this mode, every speaker can appear in either of the two data sets. In the other scheme, a fifth of the speakers were selected as testing speakers and the others as training data. By changing the training speakers, a similar experiment can be repeated five times. This is a speaker-open mode, where any single speaker cannot appear simultaneously in both data sets.

In each mode, every speaker was represented as pronunciation structure (distance matrix). The SAA paragraph consists of 69 words, corresponding to 221 phonemes. Then, the pronunciation structure was formed as phoneme-based distance matrix and the number of elements in the matrix was 221x220/2 = 24,310. For each of the training speaker pairs, the *difference* matrix was derived from their distance matrices (See figure 4). Since any element in the difference matrix is considered to contribute to prediction of the IPA-based reference distance of the two speakers, all the 24,310 elements in the difference matrix were used for prediction, where Support Vector Regression (SVR) was used as prediction model.



Speaker S's distance matrix
0
0
0
0
0
$\{S_{ij}\}$
Speaker T's distance matrix
0
0
0
0
0
$\{T_{ij}\}$
$|S_{ij} - T_{ij}|$
Difference matrix between the two
0
0
0
0
0
$\{D_{ij}\}$

Figure 4: Derivation of the difference matrix from two pronunciation matrices

Experiments in the speaker-pair-open mode showed a very high correlation between reference and prediction, which is 0.903 (See figure 5). However, in the speaker-open mode, it is 0.547, which we have to admit to be very low. It is easy to understand that difficulty of prediction is higher in the latter mode because no data is common between training and testing conditions. When we consider the prediction mechanism of SVR carefully, we can derive practical interpretation of the conditional difference between the speaker-pair-open and speaker-open modes, which is illustrated as Figure 6. Suppose that N speakers are given as training data, the task of the former mode is predicting the distances from a new speaker to the training N speakers. However, that of the latter mode is predicting the distances between all the speaker pairs of M new speakers.

While the prediction performance in the latter mode is not high enough to be used in building real applications, we can say that applicability of prediction in the former mode is not low. For example, suppose that we can obtain spoken SAA paragraphs from all the TED speakers and their IPA transcripts, an SVR-based predictor can be trained using those data. Then, if a spoken paragraph is given from a new student, where his/her IPA transcript is not available, the distances from the student to all the TED speakers can be predicted accurately.

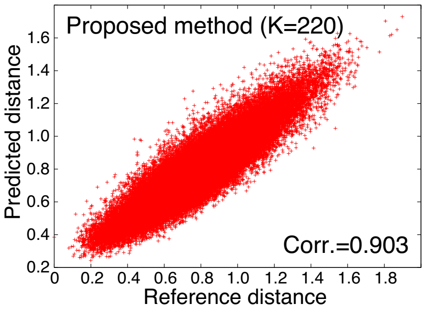


Figure 5: Correlation in the speaker-pair-open mode

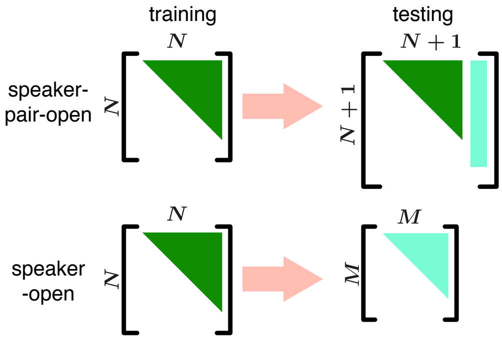


Figure 6: Conditional difference between the two modes

7. Visualization of a speaker-to-speaker distance matrix

Two methods are well-known to visualize a distance matrix among N items: dendrogram and Multi-Dimensional Scaling (MDS). Although the latter is better to represent how the N items are distributed in their feature space, it has an inevitable and big problem, called stress. The MDS projects the original geometrical shape in the original higher dimensional space (original distance matrix) into a two-dimensional geometrical shape (projected distance matrix). The latter matrix generally differs from its original one and this difference is called stress. Figure 7 shows a result of applying the MDS directly to a distance matrix among students. It seems to show well how they are distributed in terms of pronunciation but we can say, although stress is small in some parts of the result, it may be very large in other parts. It should be noted that students cannot know which part includes larger stress or distortion. Pedagogically, this is a very critical problem.

The MDS attempts to visualize a distance matrix wholly (See figure 7) but a student is easily expected to pay much more attention to how he/she is different from the others and much less to relations among the others. This indicates that, to student i, the i-th row in the matrix is very important. If the i-th row only is used for visualization, then, stress-free visualization is possible [8]. Based on this strategy, we tentatively realized a method of stress-free visualization (See figure 8), where student i is put at the center and the radius indicates how the pronunciation of student i is different from that of another. Other students of the same gender are plotted on the upper hemisphere and the angle is the age of speakers.

In the proposed visualization, not only pronunciation differences but also differences of age and gender are also used for visualization.

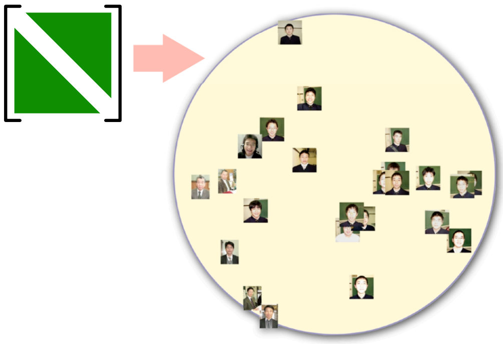


Figure 7: MDS-based visualization

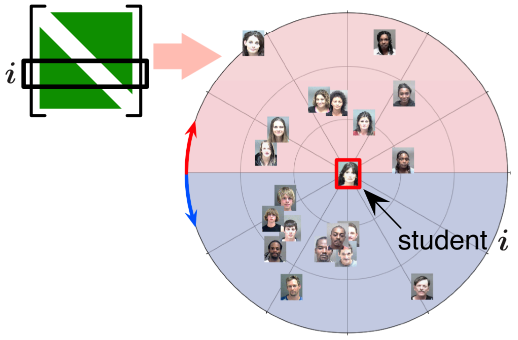


Figure 8: Stress-free visualization

In the future, we're planning to collect SAA spoken paragraphs from TED talkers. If their talks and pronunciations are plotted from a student's self-centered viewpoint, it is surely a browser of TED talks in terms of English pronunciation, especially designed for that specific student. We believe that the browser can help that student to learn WE in a very efficient and effective way.

8. Conclusions

This paper describes our recent development of a method of automatic and individual-basis clustering of English pronunciations by using the SAA and a method of stress-free visualization. Performance of accent distance prediction is very good in a speaker-pair-open mode but it is very low in a speaker-open mode. However, we consider that prediction in a speaker-pair-open mode will be able enough to be used efficiently and effectively in teaching/learning WE.

References

[1] N. Minematsu, et al, "Speech structure and its application to robust speech processing," *Journal of New Generation Computing*, 28, 3, 399-319, 2010.

[2] S. Kasahara, et al, "Structure-based prediction of English pronunciation distances and its analytical investigation," *Proc. Int. Conf. Information Science and Technology*, 331-335, 2014.

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

[4] K. Marie-José, et al., "Dialäkt App: communication dialectology to the public – crowdsourcing dialects from the public", *Trends in Phonetics and Phonology in German-speaking Europe* (in press)

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

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

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

[8] 川瀬他，"訛り・性別・年齢を考慮した自己視点からの世界諸英語発音の可視化"，電子情報通信学会音声研究会 SP2014-12，pp.127-132，2014.