VISUALIZATION OF PRONUNCIATION DIVERSITY OF WORLD ENGLISHES FROM A SPEAKER'S SELF-CENTERED VIEWPOINT

Yuji Kawase, Nobuaki Minematsu, Daisuke Saito, Keikichi Hirose

The University of Tokyo, Tokyo, Japan
{kawase,mine,dsk_saito,hirose}@gavo.t.u-tokyo.ac.jp

ABSTRACT

English is the only language available for global communication and is known to have a large diversity of pronunciations due to the influence of speakers' mother tongue, called accents. Our previous studies [1, 2] made an attempt to do speaker-basis clustering of those pronunciations, where every speaker was assumed to speak with his own accent. The clustering procedure required a distance matrix only in terms of pronunciation differences among speakers and [1, 2] proposed a method to predict the pronunciation distance between any pair of the speakers. A distance matrix is often visualized on a two-dimensional plane by using the Multi-Dimensional Scaling (MDS) or drawing a dendrogram. In this study, considering learners' perceptual characteristics, a new method is proposed for visualization. When a visualization result is fed back to a learner, his main interest will be in the relations from himself to the others, not those among the others. Then, by using only a part of the distance matrix and other kinds of information such as age and gender, the proposed method can visualize multiple kinds of diversity found in acoustics of English pronunciation from a speaker's self-centered viewpoint. Unlike the conventional methods, our proposal is guaranteed to cause no distortion at all in results of visualization.

Index Terms— World Englishes, pronunciation clustering, visualization, self-centered viewpoint, difference of age and gender

1. INTRODUCTION

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 (WE) [3, 4] 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 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.

In our previous studies [1, 2], with the ultimate goal of creating a global pronunciation map of WE on an individual basis, we proposed a method of predicting the pronunciation distance or accent distance between speakers, where non-linguistic differences such as those of age and gender were well ignored. If a learner of English is on the map, he can then find easier-to-communicate English conversation practice partners, who are supposed to have a similar kind of pronunciation. On the other hand, he may want to find partners with very different accents because he wants to expose himself to different accents to improve his capability of listening and generalizing.

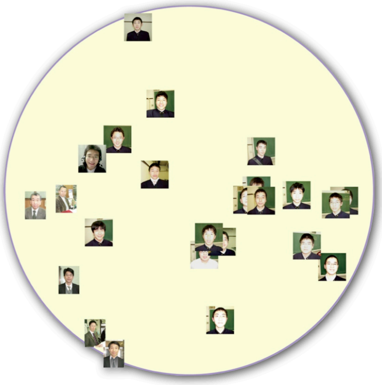

Fig. 1. The MDS chart of a pronunciation distance matrix

In [1, 2], however, a main focus was put only on measuring the pronunciation distance automatically between speakers to form their distance matrix. To create an easy-to-understand map of the speakers, an effective method of visualizing the distance matrix is required and different methods may be required to use the map for different purposes. Two well-known methods to visualize a distance matrix is drawing an MDS-based scatter chart (See Figure 1) [5] and drawing a dendrogram from the matrix (See Figure 2) [6]. Both methods try to reflect the relations among all the items in the matrix on a two-dimensional plane. If those methods are used for learners in a language class and the result is fed back to them, they will receive one and the same visualization result. It is expected, however, that different learners may pay special attention to different parts of the result. A learner's main interest will be in the relations from *himself* to others, which should be emphasized compared to the other relations. Learner-dependent visualization is needed.

It is known that learners' ability to listen to variously accented Englishes is lower than that of native speakers [7]. Using the pronunciation map, a learner will be able to find another learner with a different accent. Through conversation between the two, they will improve their ability to listen to accented Englishes. It is also interesting that a learner's listening ability is sometimes overfitted to a specific speaker, i.e., his teacher. A learner can understand easily what his teacher says but cannot understand well what other teachers say. It is known that learners' robustness of listening against differences of age and gender is lower than that of native speakers [8]. Considering this fact, extra-linguistic diversity found in pronunciation should also be considered and included in the map.

In this paper, a novel method to visualize a given pronunciation

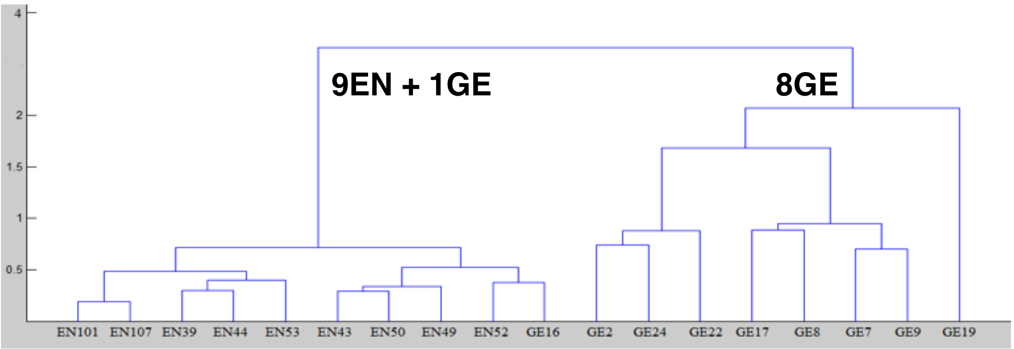

Fig. 2. Dendrogram of 9 German speakers (GE) and 9 American speakers (EN) in the Speech Accent Archive [9]

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 big tʰɔɪ fɹɔg˺ fɔɹ ðə kɪdz̥ ʃi kɛn ə̆skuːb˺ ðiːz θɪŋs ɪntu θriː ɹɛd˺ bægs æn ə wi̊l goː mitʰ hɛɹ wɛnzdeɪ æd˺ d̪ə tɹeɪn ə̆steɪʃən]

Fig. 3. The SAA paragraph and an example of transcription

distance matrix is proposed. In the method, from the matrix, only the relations from a specific speaker to others are firstly extracted. Next, we use other kinds of information about those other speakers. Here, age and gender is used so that that specific speaker can find speakers of different gender and different age in the map. As for age, we have a problem. When we collect data, some speakers are reluctant to show their age. Further, what is needed for visualization may not be real age but age perceived by listeners. In this paper, we use a speech corpus of World Englishes, the Speech Accent Archive (SAA) [9]. The speaker attributes of the corpus include real age and it can be used for visualization. By considering real situations of collecting data, however, we develop a method of automatic prediction of perceptual age and apply it tentatively to our task. Objective assessment and subjective assessment are done through comparison between the conventional method and the proposed method.

2. SPEECH ACCENT ARCHIVE AND REFERENCE PRONUNCIATION DISTANCES

The corpus is composed of read speech samples of more than 1,800 speakers and their IPA narrow transcripts. The speakers are from all over the world and they read the common elicitation paragraph, shown in Figure 3, where an example of IPA transcription is also presented. In [1, 2], the IPA transcripts were used to prepare reference inter-speaker pronunciation distances, with which an automatic predictor of the pronunciation distance was trained. In this study, only the data with no word-level insertion or deletion were extracted manually and used. Finally, 370 speakers were available at most for experiments below.

Drawing a map of WE pronunciations is decomposed into two processes of distance prediction between speakers and visualization of the distance matrix. The two processes are independent and, since we want to focus only on the second process in this paper, we use the reference distances [1, 2] for visualization, not the predicted distances, in the following experiments.

Following [10], the reference distance between two speakers is calculated through DTW of their IPA transcripts. Since all the transcripts contain exactly the same number of words, word-level alignment is easy and we only have to treat phone-level insertions, deletions, and substitutions between a word and its counterpart. Since DTW-based alignment of two IPA transcripts needs the distance matrix among all the existing IPA phones in the SAA, we prepared it in the following way. Here 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 calculated 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 the same phone with no diacritical 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 words in the paragraph. 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 [6]. Figure 2 shows a result of bottom-up clustering, dendrogram, of a part of the 370 SAA speakers. In the 370 speakers, 9 German speakers were found and, by adding randomly selected 9 American speakers, the 18 speakers were clustered using the Ward's method [11]. It is clearly shown that accent differences are adequately visualized between the two groups and a larger intra-group diversity is found in the German group.

3. PERCEPTUAL AGE PREDICTION

As explained in Section 1, learners are not good at dealing with a large diversity found in utterances of World Englishes. It should be noted that, in addition to accent differences, those of age and gender can be troublesome to learners. In this study, a method of simultaneous visualization of the diversity in terms of accent, age, and gender is proposed. As for age, however, it is sometimes difficult to obtain from speakers. Then, automatic prediction of perceptual age, not real age, is tentatively examined. The perceptual age of a speaker is defined here as the average age over the ages that are perceived by multiple listeners when hearing the speaker.

In our previous study [12], perceptual age prediction was investigated. A large listening test was done, where 30 subjects guessed the age of about a thousand speakers only by hearing them over head-

Table 1. Experimental conditions for perceptual age prediction

Corpuses	CIAIR-CVC[14], JNAS[15], S-JNAS[16]
UBM	GMM using all the samples in the corpuses
Training	even-numbered 60-sec long samples
Testing	odd-numbered 60-sec long samples
Window	25ms length / 10ms shift
Features	12MFCC + 12ΔMFCC + ΔEnergy log F0
#mixtures	64

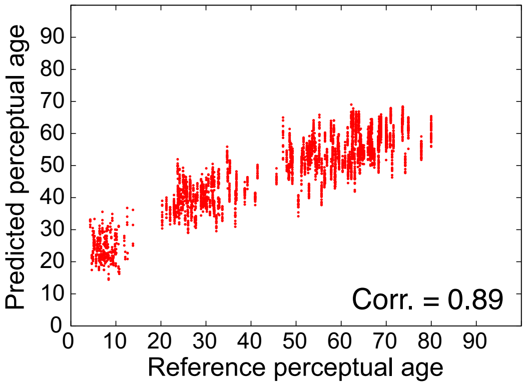

Fig. 4. Correlation of reference and predicted perceptual ages

phones. The reference perceptual age was defined as the average age over the subjects. A Gaussian Mixture Model (GMM) $P(o|\text{spk}_i)$ was trained for speaker i. For new speaker x, his perceptual age was predicted as weighted summation of $\sum_i w_i \text{age}(\text{spk}_i)$, where $w_i = \frac{P(o_x|\text{spk}_i)}{\sum_j P(o_x|\text{spk}_j)}$.

In [12], we applied the GMM-based classical speaker modeling technique to the task of perceptual age prediction. In this paper, the supervector-based technique is applied instead. First, speaker-independent GMM is trained as Universal Background Model (UBM) from a large number of speakers. Then, for speaker i in the database, the UBM is adjusted to him via MAP-based adaptation. After that, mean vectors of speaker i's GMM are concatenated to form a very high-dimensional vector, called supervector. This supervector is often used to represent speaker identity [13]. In this paper, two kinds of UBM-GMM are trained: MFCC-based and F0-based UBM-GMMs. These two models give us two supervectors: MFCC-based and F0-based ones. For perceptual age prediction of new speaker x, we subtract UBM supervectors, bias vectors, from x's supervectors. The resulting differential vectors are used as features and Support Vector Regression (SVR) is used as predictor.

To evaluate the performance of our new predictor, experiments were carried out as two-fold cross validation. Table 1 shows the conditions of acoustic analysis and the features used to train UBM-GMMs. As for speech corpuses and perceptual age labels, we used the same corpuses and labels used in [12]. The prediction performance is shown in Figure 4 and the correlation between the reference perceptual ages and their predicted values is 0.89, only slightly increased from the performance of our old predictor, which was 0.88.

4. PROPOSED METHOD OF VISUALIZATION

Distance matrix $\{d_{ij}\}$ $(1 \leq i, j \leq N)$ in an m-dimensional space can define a unique geometrical shape. The MDS-based chart of $\{d_{ij}\}$ is a result of projecting its geometrical shape onto a two-dimensional space. If the 2-dimensional and projected version of $\{d_{ij}\}$ is denoted as $\{d'_{ij}\}$, the MDS can be viewed as a process of converting $\{d_{ij}\}$ to $\{d'_{ij}\}$ so that difference between them can be minimized. If $N{>}3$ and $m{>}2$, however, this projection process usually causes some distortion, called stress.

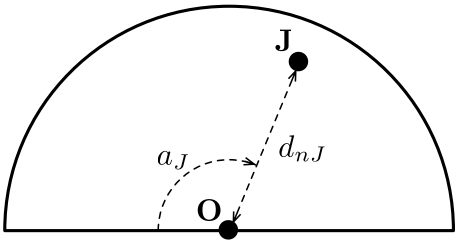

Fig. 5. Visualization of age and pronunciation distance

As told in Section 1, learner n is expected to have main interest in $\{d_{nj}\}(j \neq n)$, which can be visualized in a 1-dimensional space. The age and the gender of a speaker is quantitative and qualitative attributes of that speaker, respectively. In the proposed method, for learner n, two kinds of quantities: $\{d_{nj}\}$ and speaker j's age $\{a_j\}$ $(1 \leq j \leq N, j \neq n)$, are plotted on a two dimensional region and for the two genders, we use two different regions. For simplicity of comparison between the conventional MDS-based scatter chart and the proposed visualization, an upper semicircle is used for the region of the same gender and a lower one is used for that of the other. Figure 5 shows our proposed visualization of $\{a_j\}$ and $\{d_{nj}\}$. Speaker n is plotted at the origin (**O**). Speaker J is plotted at **J**, where a_J (age) is represented as angle and d_{nJ} (pron. distance) is as distance from **O** to **J**. Other speakers are plotted similarly. All the speakers plotted in the upper semicircle are learners who are of the same gender as speaker n's. The other learners are plotted in the lower semicircle.

Figure 6 shows easy and visual comparison between the conventional MDS-based chart and our proposed visualization. Randomly selected 10 male and 10 female speakers out of the 370 SAA speakers are plotted by using the MDS (`cmdscale` function in R [17]) as the leftmost figure. The other two figures are results of our proposed method applied to a male speaker (blue rectangle) and a female speaker (red rectangle). Here real age is used. The proposed method can realize self-centered and simultaneous visualization of the diversity of pronunciation, age, and gender in the 20 speakers[1].

5. ASSESSMENT OF THE PROPOSED METHOD

5.1. Objective comparison

As explained in Section 4, the original MDS-based chart usually includes some distortion or stress. For example in Figure 6, the correlation between distance matrix $\{d_{ij}\}$ in the original space and distance matrix $\{d'_{ij}\}$ in the 2-dimensional and projected space is calculated to be 0.87. If we calculate the correlation separately for speaker n, that is the correlation between $\{d_{nj}\}$ and $\{d'_{nj}\}$, the maximum and the minimum are 0.98 and 0.73, respectively. This means that, if the MDS-based chart is fed back to the 20 learners, some learners will come to pay main attention to distorted results but they cannot be aware of the fact that the presented results include distortion. Pedagogically speaking, this is an extremely critical problem.

[1] Since no face image is included in the SAA, we used face images found in the MORPH face image database [18]. Assignment of a face image to a speaker was done by referring to the real age and the gender of a speaker of the SAA and those of a person of the MORPH.

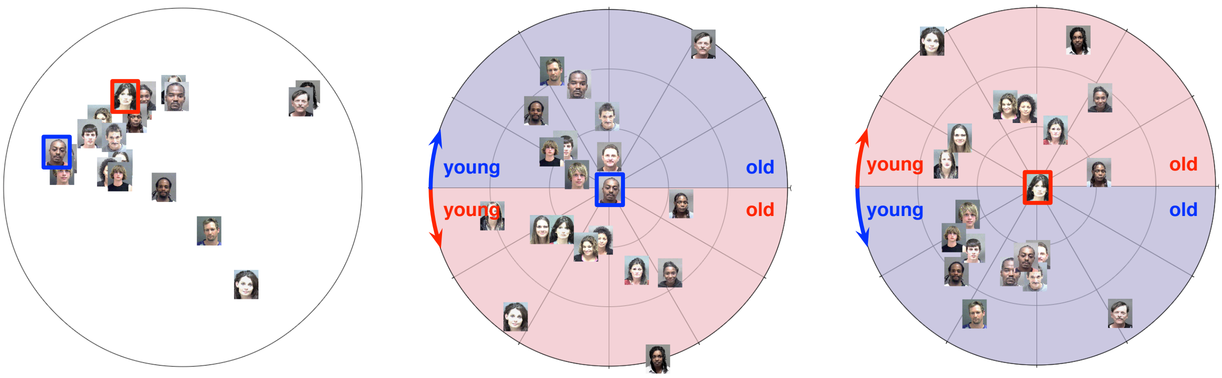

Fig. 6. Conventional and proposed visualization of various kinds of English pronunciation

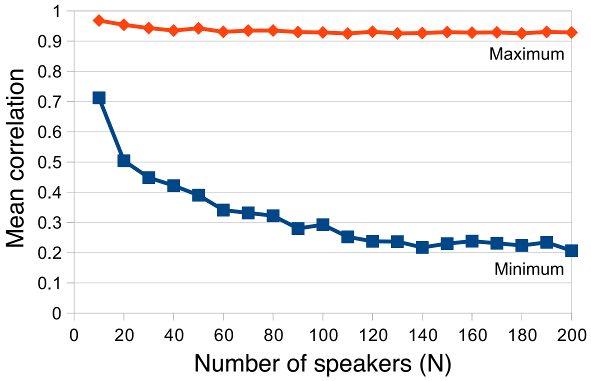

Fig. 7. Correlation reduction by increasing #speakers

It is easily expected that the minimum correlation will be further reduced when we increase the number of speakers (N) used for visualization. Figure 7 shows reduction of the minimum correlation caused by increasing N and it also displays the maximum correlation. Here, N speakers are selected randomly and the minimum correlation and the maximum correlation are calculated as average through repeating random selection of N speakers 30 times. Although the maximum is always high enough, the minimum monotonously decreases. When N is 100, it is approximately 0.3, which means that not a few learners have to face largely distorted results. We can claim that the MDS-based visualization of many learners is very dangerous in classes. In our proposed method, however, since $\{d_{nj}\}$ are used as they are (See Figure 5), our method is guaranteed to cause no distortion at all for any n.

5.2. Subjective comparison

In Section 5.1, it is objectively shown that the MDS-based chart has a severe problem about accuracy of visualizing the pronunciation distance between speakers. Following this finding, two subjective comparison tests were done, in which 30 adults participated as subjects. In the first test, a subject was asked to judge how intuitive or easy-to-understand each visualization method was to know the pronunciation distance between him and others. The 20 speakers selected in Section 4 were used for both methods and it was supposed that the subject was one of them. Rating was done using an 11-degree scale. It should be noted that the MDS's low accuracy of visualizing the pronunciation distance was not explained beforehand. So, accuracy of visualization was ignored and only intuitiveness or easiness-to-understand of visualization was focused on by the subjects. They could listen to the SAA paragraph spoken by the speakers in the visualized results by clicking their faces. The subjects could also change the center-positioned speaker by clicking.

Table 2. Results of subjective comparison

MDS chart	proposed	p-value
6.0	6.3	15.7%

real age	predicted age	p-value
5.9	5.0	4.8%

In the second test, the 20 speakers were plotted by the proposed method with real age and by that with predicted age. The two kinds of charts were presented to a subject and he was asked to judge the degree of validity of visualizing age diversity, independently for each method. It was explained intentionally that the two methods predicted the age of a speaker automatically by using different algorithms. So, we consider that the subjects did unbiased judgment. They could listen to the spoken paragraphs by clicking and rating was done with an 11-degree scale.

Table 2 shows the results of the two subjective comparison tests. Intuitiveness of the proposed method was judged a little bit higher than that of the original method but, according to ANOVA, difference was only observed at the significance level of 15.7%. This means that a significant difference of intuitiveness does not exist between the two and that our proposal is at least as intuitive as the original MDS. In the first test, we allowed subjects to write down comments on both methods. While several subjects admitted that the proposed method can show pronunciation distribution in a more organized way, some others pointed out that if a teacher wants to know pronunciation distribution of his students, the original method will fit the aim better. Both methods seem to have their own pros and cons. As we explained, the first test was done without explaining the inevitable and critical problem of the original method. We can say that the two facts that our proposed method cannot have distortion at all and it is as intuitive as the original method show high effectiveness of our proposed method. As for visualizing pronunciation diversity among students in a class, it will be good to put a teacher in the center in the proposed method. It is expected that his students want to share the result of visualization from the teacher's viewpoint.

Results of the second test showed that the prediction performance of perceptual age was not good enough for this task. The predictor was trained by Japanese corpuses [14, 15, 16] but was tested by the SAA spoken paragraphs [9]. This language gap might have influenced the performance. If automatic prediction of perceptual age is really needed in real application, we have to tune and optimize the prediction method for this task. Otherwise, we have to ask speakers how old they are and their real age will be used.

We're planning to collect the SAA spoken paragraphs from speakers of TED [19], where a large number of native and non-native speakers provide about 15-min English talks. If their talks and pronunciations are plotted from a learner's self-centered viewpoint, it will become a browser of TED talks in terms of English pronunciation, especially designed for that specific learner. By using TED talkers with similar pronunciation to that learner's, he will be able to listen without pronunciation trouble. Further, by using talkers with different pronunciation, he can improve his ability of listening to or generalizing differently accented Englishes. We believe that this browser can help that learner to learn WE in a very efficient way.

6. CONCLUSIONS

This paper proposed a novel method of simultaneous visualization of diversity in pronunciation, age, and gender observed in spoken English, where a specific speaker's self-centered viewpoint is introduced. Through objective comparison, much better accuracy of our proposal was clearly indicated. Through subjective comparison, however, the pros and cons of the proposed method was shown. In comments from the subjects, we found some good future directions to use our proposed method to realize an pedagogically effective tool for learning and teaching WE.

7. REFERENCES

[1] H.-P. Shen, N. Minematsu, T. Makino, S. H. Weinberger, T. Pongkittiphan, C.-H. Wu, "Automatic pronunciation clustering using a world English archive and pronunciation structure analysis", *Proc. ASRU*, 222–227, 2013.

[2] S. Kasahara, S. Kitahara, N. Minematsu, H.-P. Shen, T. Makino, D. Saito, K. Hirose, "Improved and robust prediction of pronunciation distance for individual-basis clustering of world Englishes pronunciaiton," *Proc. ICASSP*, 2014 (to appear)

[3] B. Kachru, Y. Kachru, C. Nelson, *The handbook of World Englishes*, Wiley-Blackwell, 2009.

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

[5] N. Minematsu, "Training of pronunciation as learning of the sound system embedded in the target language," *Proc. The Phonetic Conference of China and Int. Symposium on Phonetic Frontiers*, CD-ROM, 2008.

[6] H.-P. Shen, N. Minematsu, T. Makino, S. H. Weinberger, T. Pongkittiphan, C.-H. Wu, "Speaker-based accented English clustering using a world English archive", *Proc. SLaTE*, CD-ROM, 2013.

[7] M. Pinet, P. Iverson, M. Huckvale, "Second-Language experience and speech-in-noise recognition: the role of L2 experience in the talker-listener accent interaction," *Proc. SLaTE*, CD-ROM, 2010.

[8] D. B. Pisoni, "Some thoughts on normalization in speech perception," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix, Academic Press, New York, 9–32, 1997.

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

[10] M. Wieling, J. Bloem, K. Mignella, M. Timmermeister, J. Nerbonne, "Automatically measuring the strength of foreign accents in English," `http://urd.let.rug.nl/nerbonne/papers/WielingEtAl-Accents-Validating-2013-final1.pdf`

[11] J. H. Ward, "Hierarchical grouping to optimize an objective function, " *Journal of the American Statistical Association*, 58, 236–244, 1963.

[12] N. Minematsu, K. Yamauchi, and K. Hirose, "Automatic estimation of perceptual age using speaker modeling techniques," *Proc. EUROSPEECH*, 3005–3008, 2003.

[13] T. Kinnunen, H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, 52, 1, 12–40, 2010.

[14] `http://research.nii.ac.jp/src/CIAIR-VCV.html`

[15] `http://research.nii.ac.jp/src/JNAS.html`

[16] `http://research.nii.ac.jp/src/S-JNAS.html`

[17] `http://www.r-project.org`

[18] `https://ebill.uncw.edu/C20231_ustores/web/product_detail.jsp?PRODUCTID=8`

[19] `https://www.ted.com`