Experimental Investigation of the Definition of Reference Accent Distance between Speakers toward Automatic Accent Clustering of Speakers of World Englishes

Tianze Shi (Tsinghua Univ., Univ. of Tokyo), Shun Kasahara, Nobuaki Minematsu,
Daisuke Saito, Keikichi Hirose (Univ. of Tokyo)
stz11@mails.tsinghua.edu.cn,
{shitianze,kasahara,mine,dsk_saito,hirose}@gavo.t.u-tokyo.ac.jp

1 Introduction

English is the only language available for international communication and is used by approximately 1.5 billion speakers. Different speakers exhibit different characteristics of pronunciation, called accents, due to their mother tongues and divergent learning environments of English. It is known that intelligibility of spoken English depends on how different the speaker's accent and the listener's accent are, rather than the native-likeness of the speaker's pronunciation (Flowerdew, 1994). The concept of World Englishes (WE) (Kachru et al., 2009; Jenkins, 2003) regards American English and British English just as two examples of accented English. Native-likeness is not a golden standard to enhance smooth communication in the perspective of WE. Instead, it should be important for a user of English to be aware of the divergence of English accents and where his pronunciation is located in the diversity. Then, creating the global map of WE pronunciations through automatic clustering of them and locating a user's pronunciation in the map will be of practical use for smoother communication. Our previous works (Minematsu et al., 2014; Kasahara et al., 2014) attempted automatic clustering, where the accent distance between two speakers had to be automatically predicted and our predictor was trained by using reference accent distances prepared with IPA transcripts of utterances of training speakers. However, we have to admit that the reference distances defined in the above works were tentative and this paper aims at providing a more appropriate definition for more dialectologically-valid clustering of WE users.

2 Related works

In Minematsu et al. (2014); Kasahara et al. (2014), good efforts were made to cluster individual speakers automatically only in terms of their accents of English pronunciation. Automatic clustering technically requires automatic measurement of the accent distance between an arbitrary pair of speakers. Here, the Speech Accent Archive (SAA) (Weinberger and Kunath, 2011) was used, where all the speakers read a common paragraph and their IPA transcripts are provided and can be compared directly to measure accent distances. In our previous works, an accent distance predictor was trained by using the IPA-based reference accent distances between any pair of the training speakers. The reference distances were obtained by comparing two IPA transcripts by means of Dynamic Time Warping (DTW), where the two transcripts were optimally aligned with some inevitable misalignments. This misalignment was calculated quantitatively as a modified version of Levenshtein distance, which was used as accent distance between the two speakers. The DTW required a phone-to-phone distance matrix, where recordings of phones with diacritical marks from an expert phonetician were used.

However, this definition is not the only way to define the reference distance. For example, in Wieling et al. (2014a) and Wieling et al. (2014b), Pairwise Mutual Information (PMI) based Levenshtein distance and naïve discriminative learning (NDL) were used to derive accent distances. Further, phone distances based on distinctive features were calculated for the purpose of transcript alignment and dialect distance measurement in Nerbonne and Heeringa (1997), Somers (1998), and Kondrak (2000).

In this work, by following our previous works, some variants of the DTW-based method are investigated by visualizing the diversity of WE pronunciations and they are compared to some of the above related works.

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.	pliːˌ kʰɔl stɛla æsk əɹ tu bɹĩŋ ðiːs θĩŋz wɪθ həɹ fɹə̃m n̥ə stɔːɹ sɪks spũːnz əˌ fɪɛʃ snoʊ piːz faːɪv θɪk slæbˌ əv blu ʧiːz ɛ̃nd̥ meɪbi ə snaːk fəɹ hə᷉ bɹʌðəɹ bɔːb wi ɔlso nid ə smɔl plæstɪk sneɪk ɛ̃n ə bɪg tʰɔɪ fɹɔːg fə᷉ ðə kʰɪːdˌ ʃi kə̃n skup ðiˌ θĩŋz ĩntũ θɹi ɹɛd bæːgz ə̃n wi wɪ goʊ miːt həɹ wɛ̃nzdeɪ æt ðə tɹeĩn steɪʃə̃n

Figure 1: The common paragraph read by speakers of the SAA and an example of IPA narrow transcription (a German woman who lived in the U.S. for 8 years labeled as german17 on the website)

p	ə	ɾ	i	z		k	ɔ		l		s	t	ɛ	ɾ	ɑ
p^h		l̥	i	z		k^h	a	ʊ	l		s	t	ɛ	l	ʌ
5.71	4.47	2.31	0.00	0.00		4.37	1.29	0.62	0.00		0.00	0.00	0.00	4.38	0.34

Figure 2: An example of phone-level alignment and cost calculation performed by the DTW algorithm.

3 Material: Speach Accent Archive

The Speech Accent Archive (SAA) [1] provides readings of a common elicitation paragraph, which were collected from approximately 2,000 international speakers including many non-native speakers. This archive is very useful because it provides not only speech samples but also their IPA transcripts with diacritical marks. The elicited 69-word paragraph and one example of its IPA narrow transcription is shown in Figure 1. The 69-word paragraph breaks down to 221 American English phonemes according to the CMU Pronunciation Dictionary (Rudnicky, 2007). Approximately 100 base phones are used in the SAA corpus while, considering diacritical marks, the number of different kinds of phones is more than 550.

4 Algorithms and methods used in the experiments

4.1 Dynamic Time Warpping (DTW) algorithm

When comparing two IPA transcripts, we apply the DTW algorithm as we did in our previous studies. Given a phone-to-phone distance matrix and two phone sequences, the DTW algorithm is able to generate an alignment with minimized total distance or misalignment. We define the reference distance between two speakers as the average of 69 word-level DTW alignment distances. The DTW is used widely in taking alignment and calculating similarity between two temporal sequences such as DNA analysis. Figure 2 shows an example of the DTW algorithm performed on two transcripts.

4.2 Calculation of acoustic model based phone-to-phone distances

In order to apply the DTW algorithm, a definition of the distance between two phones is essential, where comparing acoustic features between the phones is a direct approach. In our previous works, the most frequent 153 phones in the SAA were produced by an expert phonetician and, using these data, we constructed a three-state Hidden Markov Model (HMM) for each phone, where each state in the model was characterized as Gaussian distribution of the acoustic features of the phone. Here, Mel-frequency cepstral coefficients (MFCC) were used. The phone-to-phone distance was defined as average of the three state-to-state distances between the two phone HMMs. In our previous works, the Bhattacharyya distance (BD) was used as distance metric between two states (distributions).

In statistics, however, different kinds of distance metrics are often used for different purposes. In this work, four metrics of BD, symmetric Kullback-Leibler divergence (KL), Hellinger distance (HD), and average Mahalanobis distance (MD) are compared in calculating the accent distance between two speakers. The analytical forms of the four metrics between two Gaussian distributions are given in the following

[1] `http://accent.gmu.edu`

equations (Sooful and Botha, 2002).

$$D_{BD}(p,q) = \frac{1}{8}(\mu_p - \mu_q)^T \left(\frac{\Sigma_p + \Sigma_q}{2}\right)^{-1} (\mu_p - \mu_q) + \frac{1}{2} ln \left(\frac{|(\Sigma_p + \Sigma_q)/2|}{\sqrt{|\Sigma_p||\Sigma_q|}}\right) \quad (1)$$

$$D_{KL}(p,q) = \frac{1}{2}(\mu_q - \mu_p)^T \left(\Sigma_p^{-1} + \Sigma_q^{-1}\right)(\mu_q - \mu_p) + \frac{1}{2} tr\left(\Sigma_p^{-1}\Sigma_q + \Sigma_q^{-1}\Sigma_p - 2I\right) \quad (2)$$

$$D_{HD}^2(p,q) = 1 - e^{-D_{BD}(p,q)} \quad (3)$$

$$D_{MD}(p,q) = (\mu_p - \mu_q)^T (\Sigma_p \Sigma_q)^{-1} (\mu_p - \mu_q) \quad (4)$$

where p, q are two Gaussian distributions, μ and Σ represent the feature mean vector and covariance matrix of a distribution respectively.

4.3 Calculation of American English phoneme-to-phoneme distances

While the phone-to-phone distance is a precise and quantitative representation of phonetic difference, we can say that transcript comparison based on the phonetic distances may not be appropriate for automatic distance prediction without IPA transcripts. This is because it is impractical to automatically identify phones in utterances. On the other hand, it is reasonably practical to automatically identify phonemes in utterances and extract phoneme sequences from them using automatic speech recognition technologies. Accent distances based on phone-to-phone distances are expected to give better clustering results than those based on phoneme-to-phoneme distances because phonemes are abstract versions of phones. However, in speech technologies, context-dependent phonemes are often used, where the number of kinds of phonemes is logically N^3 (N=the number of phonemes) and it is larger than the number of kinds of phones found in the SAA. In this study, in addition to the phone HMMs constructed with a phonetician's recordings, context-independent phoneme HMMs (monophones) and context-dependent phoneme HMMs (triphones), which are trained by using other speakers, are used to calculate phoneme-to-phoneme distances.

5 Experiments

5.1 2 sets of phone HMMs for phone-to-phone distance calculation

Two expert male phoneticians, P01 and P02, are asked to pronounce the most frequent 153 phones in the SAA. Each phone is pronounced 20 times and fair attention is paid to diacritical difference within phones sharing the same base phone. Phones that are not pronounced are mapped to the phonetically nearest phone (e.g. the base phone) during DTW calculation. The first 12 MFCCs and 12 ΔMFFCs are used as acoustic features. It should be noted that acoustic features vary depending on speakers. Speaker-dependence in phone-to-phone distances and that in speaker-to-speaker accent distances will be examined.

We have 2 HMM sets of P01 and P02.

5.2 5 sets of American English phoneme HMMs for phoneme-to-phoneme distance calculation

Speaker-dependent monophone HMMs and triphone HMMs are constructed from each of one male and one female native speakers, who are M08 and F12 in ERJ (English Read by Japanese) database (Minematsu et al., 2004). The same acoustic features above are used. Although the number of physically different monophone HMMs is 39, that of physically different triphone HMMs is about 10,000, much larger than that of physically different phone HMMs (153). We also use speaker-independent monophone HMMs trained with the Wall Street Journal (WSJ) corpus (Vertanen, 2006).

We have 5 HMM sets of M08-mono, M08-tri, F12-mono, F12-tri, and WSJ.

Table 1: Correlations in terms of segment-to-segment distances.

(a) Corr. between HMMs for each distance metric

	P01-P02	F12-M08	M08-WSJ
BD	0.66	0.91	0.94
HD	0.86	0.97	0.97
KL	0.56	0.91	0.86
MD	0.29	0.82	0.90

(b) Corr. between distance metrics using P01

	BD	HD	KL	MD
BD	1.00			
HD	0.56	1.00		
KL	0.84	0.54	1.00	
MD	0.71	0.41	0.60	1.00

5.3 Comparison of speaker-to-speaker accent distances

By using variants of segment-to-segment distances to compare IPA transcripts, we can calculate speaker-to-speaker accent distances in multiple ways. It should be noted here that, when we calculate accent distances by using phoneme-to-phoneme distances, the IPA transcripts have to be converted into phoneme sequences in advance using a phone-to-phoneme mapping table (Minematsu et al., 2014; Kasahara et al., 2014). With these multiple definitions of speaker-to-speaker accent distances, we can visualize the diversity of WE pronunciations differently but it may not be easy to assess the adequacy of each visualization.

To compare the variants of segment-to-segment distances quantitatively, we follow previous works (Wieling et al., 2014a,b), where averaged native-likeness of an utterance perceived by more than 1,000 native listeners was compared with that predicted by comparing the IPA transcript of that utterance and those of native speakers' utterances. In those works, PMI-based alignment of IPA transcripts was proposed and in this study, we test use of HMMs and DTW. Similarly to Wieling et al. (2014a,b), we use the same set of 286 utterances and predict the native-likeness of each utterance as average over the distances to all the American English utterances in the SAA.

6 Results and discussion

6.1 Segment-to-segment distances

By using 4 kinds of distance metrics and 7 sets of HMMs, we calculate the segment-to-segment distances in multiple ways. Between some of them, we can calculate the correlation, which is shown in Table 1.

It can be seen that the segment distances correlate well among different speakers (HMMs) and the Hellinger distance seems to give the most speaker-independent definition. In the case of phoneme-based HMMs (M08, F12, WSJ), relatively higher correlations are found if the same distance metric is used. When distances of different metrics are compared, their correlations are not so high (See Table 1b).

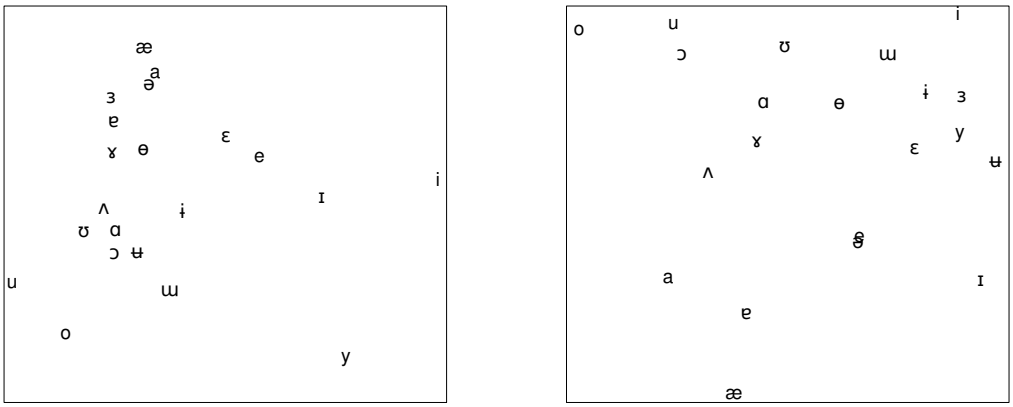
By applying the Multi-Dimensional Scaling (MDS) to a distance matrix of segment-to-segment distances, we can visualize the distribution of the segments. Multiple definitions of the segment-to-segment distance give us multiple visualizations. Two examples of the MDS are shown in Figure 3. For simplicity, we only show the results of vowel-to-vowel distances.

A careful examination suggests that these MDS graphs give certain phonetic clues of relative distance between vowel pairs. Distinction of front vowels and back vowels is shown on the horizontal dimension, while that of open vowels and closed vowels appears on the vertical dimension. Further investigation such as comparative evaluation of these visualizations in terms of their validity is to be done in future work.

6.2 Speaker-to-speaker distances

For each of the segment distance definitions, we perform the DTW algorithm to calculate the speaker-to-speaker accent distance. Correlations are examined in Table 2. Since phoneme-based segment distances correlate very high as shown in the last section, comparison between them is not repeated here. Speaker-to-speaker distances show very high correlations compared to segment-to-segment distances even when different metics are compared (See Table 2b).

We apply the MDS again to visualize the pronunciation diversity (speaker-to-speaker accent distance matrices) and the results of two metrics of M08-mono and P01 are shown in Figure 4 (According to Table 2, results of BD, HD, KL resemble each other, and phoneme-based metrics give similar results).



(a) P01 with BD (b) P02 with KL

Figure 3: Two examples of MDS-based visualization of the vowel distribution

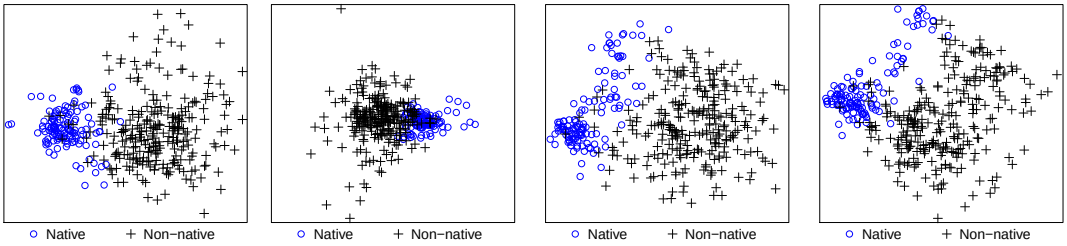
Table 2: Correlations in terms of speaker-to-speaker accent distances.

(a) Corr. between HMMs for each distance metric

	P01-P02	mono-tri (M08)	tri (M08)-P01
BD	0.98	0.98	0.88
HD	1.00	0.99	0.90
KL	0.99	0.99	0.90
MD	0.90	0.93	0.79

(b) Corr. between distance metrics using P01

	BD	HD	KL	MD
BD	1.00			
HD	0.97	1.00		
KL	0.99	0.98	1.00	
MD	0.96	0.89	0.94	1.00



(a) M08-mono, BD (b) M08-mono, MD (c) P01, BD (d) P01, MD

Figure 4: Examples of MDS-based visualization of WE pronunciation diversity.

Table 3: Correlations between the proposed distance definitions and subject native-likeness scores

	phone-based P01&HD	phone-based P01&KL	phoneme-based F12&KL	Baseline PMI	Baseline NDL
Corr.	**-0.81**	-0.80	-0.78	-0.77	-0.75
Log-transformed Corr.	**-0.83**	-0.82	-0.80	-0.81	-0.82

It can be observed that native speakers and non-native speakers are clearly separated on the first MDS dimension, suggesting that this explains the largest variance found in the distance matrices. Comparative investigation is to be done in future work.

To evaluate the proposed definitions of reference distance, we calculated the correlation between HMM-based distances and subjective native-likeness scores. The results are collected in Table 3. The correlation of our method ($r = -0.81$) outperforms the PMI-based distance metrics ($r = -0.77$) (Wieling et al., 2014a) and NDL-based distance metrics ($r = -0.75$) (Wieling et al., 2014b). As suggested in their works, when we use the logarithmic reference distances, the correlations generally improve. As explained in Wieling et al. (2014a), the agreement of individual raters with the average native-likeness scores is $r = -0.84$. This means that the proposed method is very comparable to human perception of native-likeness.

7 Conclusions

In this paper, we introduced several definitions of HMM-based reference accent distance between WE users. Although some HMMs were trained in a speaker-dependent mode, we found that the segment-to-segment distances and the speaker-to-speaker distances correlate very strongly between different HMMs. This result suggests that the proposed definition is independent of the training speaker. This is reasonable when we consider the transform-invariance of the metrics used in the experiments (Minematsu et al., 2014).

Though we have found some phonetic clues in the MDS charts of the segment-to-segment distances, the phonetic validity needs further examination. As for the MDS charts of the speaker-to-speaker distances, a clear distinction between native and non-native speakers was observed. Further, accent distances calculated based on our proposal showed very strong correlation to human scores.

References

Flowerdew, J. (1994). "Research of relevance to second language lecture comprehension: An overview". *Academic listening: Research perspectives*, (pp. 7–29).

Jenkins, J. (2003). *World Englishes: A resource book for students*: Routledge.

Kachru, B., Kachru, Y., and Nelson, C. (2009). *The handbook of world Englishes*: Wiley-Blackwell.

Kasahara, S., Kitahara, T., Minematsu, N., Shen, H., Makino, T., Saito, D., and Hirose, K. (2014). "Improved and robust prediction of pronunciation distance for individual-basis clustering of world englishes pronunciation". In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 3240–3244).

Kondrak, G. (2000). "A new algorithm for the alignment of phonetic sequences". In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, (pp. 288–295).

Minematsu, N., Kasahara, S., Makino, T., Saito, D., and Hirose, K. (2014). "Speaker-basis accent clustering using invariant structure analysis and the speech accent archive". In *Odyssey 2014: The Speaker and Lanugage Recognition Workshop*, (pp. 158–165).

Minematsu, N., Tomiyama, Y., Yoshimoto, K., Shimizu, K., Nakagawa, S., Dantsuji, M., and Makino, S. (2004). "Development of english speech database read by japanese to support call research". In *the 18th International Congress on Acoustics*, volume 1, (pp. 557–560).

Nerbonne, J. and Heeringa, W. (1997). "Measuring dialect distance phonetically". In *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*.

Rudnicky, A. (2007). "The cmu pronunciation dictionary, release 0.7a". `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

Somers, H. L. (1998). "Similarity metrics for aligning children's articulation data". In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 2, (pp. 1227–1232). Association for Computational Linguistics.

Sooful, J. J. and Botha, E. C. (2002). "Comparison of acoustic distance measures for automatic cross-language phoneme mapping". In *the 7th International Conference on Spoken Language Processing (ICSLP)*.

Vertanen, K. (2006). "Baseline wsj acoustic models for htk and sphinx: Training recipes and recognition experiments". Technical report, Cambridge, United Kingdom: Cavendish Laboratory.

Weinberger, S. H. and Kunath, S. A. (2011). "The speech accent archive: towards a typology of english accents". *Language and Computers*, 73:1, 265–281.

Wieling, M., Bloem, J., Mignella, K., Timmermeister, M., and Nerbonne, J. (2014a). "Measuring foreign accent strength in english. validating levenshtein distance as a measure". *The Mind Research Repository*.

Wieling, M., Nerbonne, J., Bloem, J., Gooskens, C., Heeringa, W., and Baayen, R. H. (2014b). "A cognitively grounded measure of pronunciation distance". *Public Library of Science (PloS) one*, 9:1, e75734.