IMPROVED AND ROBUST PREDICTION OF PRONUNCIATION DISTANCE FOR INDIVIDUAL-BASIS CLUSTERING OF WORLD ENGLISHES PRONUNCIATION

S. Kasahara†, *S. Kitahara*†, *N. Minematsu*†, *H.-P. Shen*‡, *T. Makino*∗, *D. Saito*†, *K. Hirose*†

† The University of Tokyo, Tokyo, Japan
‡ National Cheng Kung University, Tainan, Taiwan
∗ Chuo University, Tokyo, Japan

ABSTRACT

English is the only language available for global communication and is used by approximately 1.5 billions of speakers. It is also known to have a large diversity of pronunciation due to the influence of speakers' mother tongue, called accents. Our project aims at creating a global and individual-basis map of English pronunciations to be used in teaching and learning World Englishes (WE) as well as research studies of WE [1, 2]. Creating the map mathematically requires a distance matrix in terms of pronunciation differences among all the speakers considered, and technically requires a method of predicting the pronunciation distance between any pair of the speakers only by using their speech samples. In our previous study [3], we combined invariant pronunciation structure analysis [4, 5, 6, 7] and Support Vector Regression (SVR) to predict the inter-speaker pronunciation distances. In this paper, several techniques are introduced and examined whether they can increase accuracy and robustness of prediction. Experiments show that the correlation between IPA-based reference distances and the predicted distances is increased from 0.805 to 0.903, which is over the correlation of 0.829 that is obtained by using the phoneme-based ground truth distances.

Index Terms— World Englishes, pronunciation clustering, SAA, IPA transcription, pronunciation structure analysis, support vector regression, f-divergence, phoneme recognition

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 [1, 2] 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.

The ultimate goal of our project is creating a global map of WE 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 easier-to-communicate English conversation partners, who are supposed to have a similar kind of pronunciation. If he is too distant from many of other varieties, however, he may have to correct his pronunciation for the first time to achieve smoother communication with these others.

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. 1. The SAA paragraph and an example of IPA transcription

In this paper, we use the Speech Accent Archive (SAA), which provides speech samples of a common elicitation paragraph read by more than a thousand speakers from all over the world. The SAA also provides IPA-based narrow transcripts of all the samples, which can be used for training a pronunciation distance predictor. To calculate the pronunciation distance between two speakers of the SAA, [9, 10] proposed a method of comparing two IPA transcripts using a modified version of the Levenshtein distance. Although it was shown that the calculated distances had reasonable correlation with the pronunciation distances perceived by humans, it cannot handle unlabeled data, i.e., raw speech. Recently, we proposed a method of predicting the pronunciation distance only using spoken paragraphs of the SAA [3]. The technical challenge is how to make the prediction independent of irrelevant but inevitably involved factors such as differences in age, gender, microphone, channel, background noise, etc. To this end, we used invariant pronunciation structure analysis [4, 5, 6, 7] for feature extraction and SVR for prediction. In training the predictor, reference (correct) distances had to be prepared. In [3], IPA-based phonetic distances calculated through string-based DTW of two IPA transcripts were used. The correlation between the reference distances and the predicted ones was 0.805 and in this paper, several techniques are examined to enhance the performance.

2. THE SPEECH ACCENT ARCHIVE

The corpus is composed of read speech samples of more than 1,800 speakers and their corresponding IPA narrow transcripts. The speakers are from all over the world and they 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 as reference [11]. The IPA transcripts will be used to prepare reference

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

inter-speaker pronunciation distances, 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 involved 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 data collection was done voluntarily by those who had interest in joining the SAA project. 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 only "controlled" diversity. In [12], however, English sentences read by 200 Japanese university students showed a very large pronunciation diversity and in [13], a large listening test by Americans showed that the intelligibility of the individual utterances covered a very wide range. Following these facts, we considered that read speech samples can still show well how diverse WE pronunciations are.

It is well-known that pronunciation diversity is found in both the segmental and prosodic aspects. In this paper, however, we will prepare reference pronunciation distances by using IPA transcripts, meaning that prosodic diversity will be lost. We do not claim that the prosodic diversity is minor but, as was shown in [14], clustering only based on the segmental aspect seems able to show validly how diverse WE are in terms of pronunciation. Reference distances with prosodic features will be treated in a future work.

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. Some of them were found to include a very high level of background noise and/or many pauses, and we manually removed them. Finally, 370 speakers' data were used here and the number of speaker pairs is 68,265 (=370×369 / 2).

3. REFERENCE INTER-SPEAKER PRONUNCIATION DISTANCES

To train a pronunciation distance predictor, reference inter-speaker distances are needed, which will also be used to evaluate the trained predictor. 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, the fifth author, 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 defined as the average of three state-to-state Bhattacharyya distances. We note here that, since the HMMs were speaker-dependent, all the distances were calculated in a matched condition. 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 69 words. 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 [14] as well as a result of bottom-up clustering of a part of the SAA speakers using IPA-based DTW.

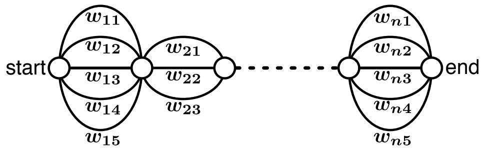

Fig. 2. An example of word-based network grammar

4. THREE BASELINE SYSTEMS

For comparison, we built three baseline systems, two of which corresponds directly to an automated version of the reference distance calculation procedure described in Section 3. The other system was proposed and developed in our previous study [3].

4.1. Naïve automation of reference distance calculation

The calculation procedure is composed of two steps: 1) IPA manual transcription and 2) DTW alignment for distance calculation. In the first two baseline systems, the first process was replaced with automatic recognition of phonemes in input utterances[1]. Here, we used a phoneme recognizer of American English (AE) in this study. Using all the utterances of the 370 speakers as training data, monophone HMMs were trained with the WSJ-based HMMs [15] adopted as initial model. For this training, each IPA transcript was converted into its AE phoneme transcript. 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 AE diphthong.

Since IPA transcription is based on phones and the HMMs were trained based on phonemes, even if we could have a perfect phoneme recognizer, generated transcripts have to be phonemic versions of IPA transcripts: the phoneme-based ground truth transcripts. Conversion from phones to phonemes is an abstraction process and some detailed phonetic information will be lost inevitably. Our first baseline system uses the ground truth transcripts and the pronunciation distance is calculated by comparing two phonemic transcripts based on DTW, conditions of which are the same as those in Section 3 except the local distance matrix. Here, the phoneme-to-phoneme distance matrix is prepared by using the WSJ-based HMMs [15]. Our second system uses transcripts generated from a real phoneme recognizer with word-based network grammar that can cover all the pronunciation variations found in the 370 speakers. Figure 2 shows an example of the network grammar of an n-word sentence, where w_{ij} denotes the i-th word spoken with the j-th pronunciation.

The correlation between the IPA-based inter-speaker reference distances and the phoneme-based distances obtained from the first system was 0.829, meaning that information loss exists to some degrees. The phoneme recognition accuracy of the second system was 73.5% but the correlation was found to be so low as 0.458. This clearly indicates that recognition errors are very fatal.

4.2. Pronunciation distance predictor developed in [3]

As told in Section 1, we need a very robust method to predict the pronunciation distance. The third author proposed a unique method of

[1] As far as we know, there does not exist an automatic recognizerof IPA phones with diacritical marks.

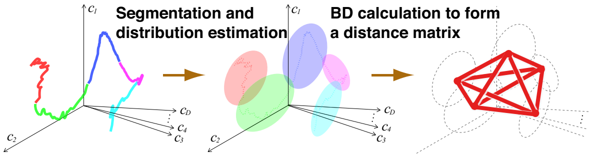
Segmentation and distribution estimation
BD calculation to form a distance matrix
c_1
c_2
c_3
c_4
c_D

Fig. 3. Utterance structure composed only of BDs

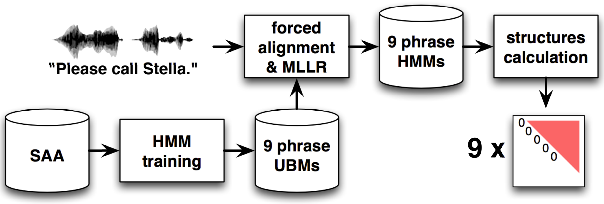

Fig. 4. Procedure to calculate the pronunciation structure

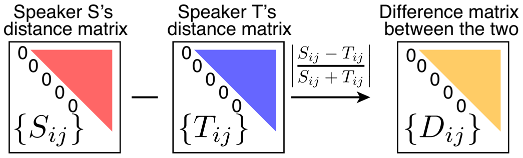
Speaker S's distance matrix
Speaker T's distance matrix
Difference matrix between the two
0
0
0
0
$\{S_{ij}\}$
$\{T_{ij}\}$
$\{D_{ij}\}$
$\left|\frac{S_{ij}-T_{ij}}{S_{ij}+T_{ij}}\right|$

Fig. 5. Difference matrix derived from two speakers' matrices

representing speech, called speech structure, and showed that acoustic and non-linguistic variations involved in speech can become effectively unseen in the representation [4, 5, 6, 7]. This proposal was done being inspired by Jakobson's structural phonology [16] and infants' sensitivity to distributional properties of sounds [17, 18]. The speech structure is invariant against any kind of continuous and convertible transform and this invariance is due to the transform-invariance of f-divergence [6], which is calculated as

$$f_{\rm div}(p_1,p_2)=\int_{\mathcal{X}} p_2(\boldsymbol{x})\, g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x}, \qquad (1)$$

where $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ are density functions of two distributions on measurable space $\mathcal{X}$. $g(t)$ is a convex function for $t>0$. If we take $\sqrt{t}$ as $g(t)$, $-\log(f_{\rm div})$ becomes the Bhattacharyya distance (BD).

Figure 3 shows schematically the procedure of representing an utterance only by BDs. The utterance is a sequence of vectors and it is converted into a sequence of distributions. Here, any speech event is characterized as distribution. Then, the BD is calculated between any distribution pair and the resulting BD-based distance matrix is invariant speech structure. When this representation is applied to pronunciation analysis, it is called pronunciation structure [5, 7]. In [3], the SAA paragraph was divided into 9 phrases and the pronunciation structure was extracted from each phrase read by each speaker. Figure 4 shows the detailed procedure to calculate the pronunciation structure, where the Universal Background Model (UBM) was trained as HMM for each phrase by using all the 370 speakers. Here, the number of states of the HMM of a phrase is $3N$, where N is the number of phonemes of that phrase. The UBM was adapted to each speaker separately with MLLR adaptation (#classes = 32). By calculating the averaged BD between every pair of the phoneme instance HMMs in each phrase HMM, 9 phrase-based distance matrices were obtained for each speaker. Here, the i-th phoneme instance HMM in a phrase HMM is the three-state HMM spanning from the $(3i-2)$-th state to the $3i$-th state of that phrase HMM. As illustrated in Figure 5, from a distance matrix of speaker S and that of T, we derived a difference matrix D to represent the differences between them. Here, i and j are phoneme instance indexes and $1 \le i,j \le 221$.

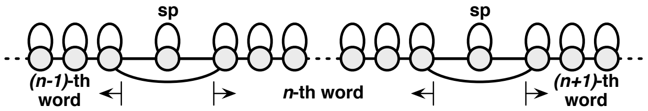

Fig. 6. Explicit modeling of inter-word short pauses

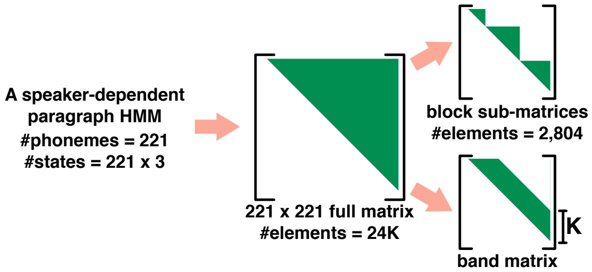

Fig. 7. Paragraph-based full matrix and its band matrix

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

$\{D_{ij}\}$ were obtained for each phrase and all the $\{D_{ij}\}$ were used as input features in SVR to predict the pronunciation distance. The total number of the features was 2,804 and the ϵ-SVR in LIBSVM [19] was used with the radial basis function kernel of $K(x_1,x_2)=\exp(-\gamma|x_1-x_2|^2)$. Experiments showed that the correlation between the IPA-based distances and the predicted ones was 0.805.

5. PROPOSED METHODS AND EXPERIMENTS

To improve the correlation, we did the following examinations. As for detailed conditions including acoustic analysis conditions, we followed those adopted in [3]. Acoustic features used in pronunciation structure analysis were MFCC-based features. 68,265 speaker pairs of the SAA were sorted by the order of IPA-based reference distances and they were divided into two groups of even-numbered pairs and odd-numbered pairs. For training a predictor and testing it, 2-fold cross-validation was done using these two groups.

5.1. Explicit modeling of inter-word short pauses (sp)

Since many speakers of the SAA are non-native speakers, inter-word pauses are often found but they were not explicitly treated in [3]. Then, as shown in Figure 6, by using the WSJ-based monophone HMMs [15], the initial model of a phrase UBM was prepared so that a sp model always existed between consecutive words. When calculating 9 distance matrices, the distances from/to the sp model were omitted. The correlation was slightly raised from 0.805 to 0.811.

5.2. MAP-based adaptation applied to UBMs

MLLR-based adaptation is known to work better than MAP-based adaptation when the amount of adaptation data is very small. In [3], since only a single phrase utterance was used to adapt a phrase HMM, which is a sequence of phoneme-instance HMMs, MLLR was adopted. In this section, however, we tentatively examined MAP and it was found that MAP is more effective than MLLR.

D_{ij} was originally proposed in [7] and was used there for linear regression. In the framework of SVR, however, feature normalization is always done as preprocessing. Since we considered that double normalization might lose too many effective features in raw data, we used simpler definition of $D_{ij}=|S_{ij}-T_{ij}|$ in this paper.

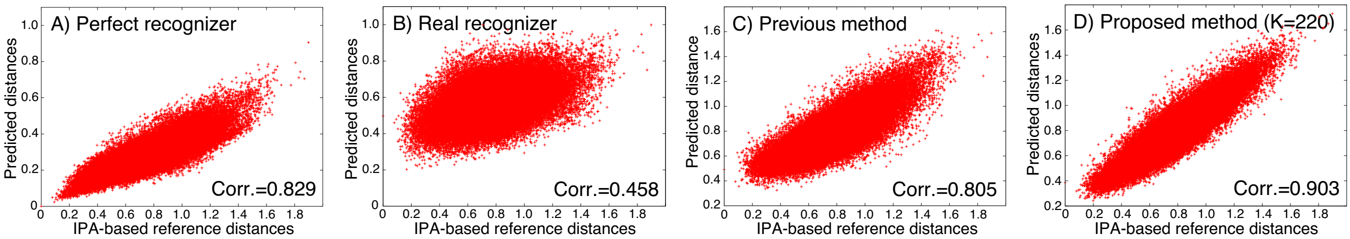

Fig. 9. Correlations between the IPA-based reference pronunciation distances and the predicted distances

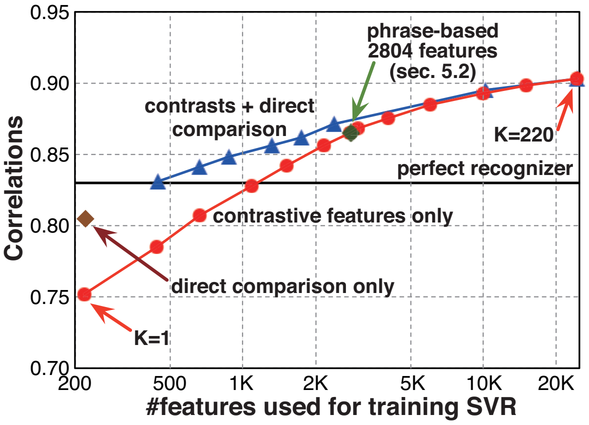

Fig. 8. Correlation improvement by increasing the band width

By using MAP with sp models and unnormalized differences, the correlation was increased up to 0.865.

5.3. Distance matrices calculated from paragraph HMMs

In [3], a pronunciation structure or a distance matrix was built for each of the 9 phrases. In this section, a full and paragraph HMM was examined for each speaker's data in the SAA. By using the paragraph HMM, as shown in Figure 7, a full distance matrix is obtained and a full difference matrix can be used for SVR. In [3], only block sub-matrices in the full matrix were used for SVR. Pronunciation structure is interesting because it captures only local and distant acoustic contrasts and discards absolute acoustic features at all. Then in the experiments of this paper, by using a band matrix, the band width of which is K, we investigated how distant contrasts in the full matrix can contribute to improving the prediction performance (See Figure 7). If K is 10, acoustic contrasts over 0 to 9 phonemes are used for prediction. The larger K becomes, the larger the total number of features used for prediction becomes. Since the SAA paragraph contains 221 phoneme instances, $K \leq 220$.

Figure 8 shows correlation improvement gained by increasing the total number of features by changing K. Here, MAP-adaptation is used with sp models and unnormalized differences. The performance in Section 5.2 was obtained using the phrase-based matrices, namely, 2,804 differential features. If we set K in the band matrix so that the total number of features is similar to 2,804, the correlation becomes very similar to what was achieved in Section 5.2.

Clearly shown in Figure 8, further increase of K can certainly raise the correlation. The maximum correlation, 0.903, was obtained with the maximum number of K, that is 220. Before the experiment, we had been interested in where the correlation peak was found but the peak was not observed at all. This maximum value is much larger than the correlation of 0.829, which was achieved by the phoneme-based ground truth distances. It is very surprising to us that so distant acoustic contrasts, that are 220-phoneme long, are still effective to increase the performance. In [5], Multiple Stream Structure (MSS) was found to be effective in improving the performance of structure-based isolated word recognition. MSS may be able to improve the correlation further in the current task.

Figure 9 shows four correlations of A) perfect phoneme recognizer, B) real phoneme recognizer, C) our previous method [3], and D) our proposed method (K=220). In comparison to C), a large improvement is achieved in the proposed method.

5.4. Use of acoustic distances obtained by direct comparison

We investigated the effectiveness of using acoustic distances obtained by direct and absolute comparison. Since a paragraph HMM is a sequence of 221 phoneme instance HMMs, we can get 221 phoneme-based BDs between a speaker's paragraph HMM and another speaker's. These 221 BDs can be used as additional features to our original feature set based on pronunciation structure analysis. The correlation by using the new 221 features only was 0.805, which is higher than that obtained by using 220 contrastive features (K=1 in Figure 8). In pronunciation structure, by increasing K, the number of effective features is easily increased. In direct comparison, however, 221 is the maximum number. In this sense, contrastive and relational features are very suitable for discriminative models. Figure 8 also shows the effectiveness of the new features when they are combined to the contrastive features. They are effective when K is small but they seem ineffective at all when K is large enough.

6. CONCLUSIONS

This paper proposed an improved method to predict inter-speaker pronunciation distances only by using speech samples. Our proposal is based on combining pronunciation structure analysis for feature extraction and support vector regression for prediction. Experiments showed that our method can show a better performance than the performance obtained by using the phoneme-based ground truth distances and string-based DTW. However, all the examinations were feature-based and, as for regression model, a simple one of SVR with a well-known kernel function was used. We're interested in applying more sophisticated models such as multi-kernel regression [20] and kNN-SVR [21] to our task. We can point out a drawback in the experimental condition in this paper. The experiments were carried out in a speaker-pair-open mode and every speaker was found both in training and testing data. Strictly speaking, the proposed method has to be examined in a speaker-open mode. We can say, however, that we gained a remarkable improvement and if researchers of WE are satisfied with the current performance, it would be better to move forward to collecting data from a larger number of speakers internationally. For that, we have already developed the β-version of an i-OS application for easier collection. We're not sure whether collection from 1.5 billions of speakers is an achievable goal but we're interested in drawing the individual-basis map of WE pronunciations.

7. REFERENCES

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

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

[3] 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.

[4] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," *Proc. IACSSP*, 889–892, 2005.

[5] N. Minematsu, S. Asakawa, M. Suzuki, Y. Qiao, "Speech structure and its application to robust speech processing", *Journal of New Generation Computing*, 28, 3, 299–319, 2010.

[6] Y. Qiao, N. Minematsu, "A study on invariance of f-divergence and its application to speech recognition", *IEEE Trans. on Signal Processing*, 58, 7, 3884–3890, 2010.

[7] M. Suzuki, Y. Qiao, N. Minematsu, K. Hirose, "Integration of multilayer regression with structure-based pronunciation assessment," *Proc. INTERSPEECH*, 586–589, 2010.

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

[9] M. Wieling, J. Nerbonne, J. Bloem, C. Gooskens, W. Heeringa, R. H. Baayen, "A cognitively grounded measure of pronunciation distance," *PLoS ONE*, DOI: 10.1371/journal.pone.0075734, 2014.

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

[12] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, S. Makino, "Development of English speech database read by Japanese to support CALL research," *Proc. ICA*, 557–560, 2004.

[13] N. Minematsu, K. Okabe, K. Ogaki, K. Hirose, "Measurement of objective intelligibility of Japanese accented English using ERJ (English Read by Japa-nese) database," *Proc. INTERSPEECH*, 1481–1484, 2011.

[14] 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.

[15] HTK Wall Street Journal Training Recipe,
http://www.keithv.com/software/htk/

[16] R. Jakobson, L. R. Waugh, *The sound shape of language*, Mouton de Gruyter, 2002.

[17] J. Maye, J. F. Werker, L. Gerken, "Infant sensitivity to distributional information can affect phonetic discrimination," *Cognition*, 82, B101–B111, 2002.

[18] J. F. Werker, F. Pons, C. Dietrich, S. Kajikawa, L. Fais, S. Amano, "Infant-directed speech supports phonetic category learning in English and Japanese," *Cognition*, 103, 147–162, 2007.

[19] C.-C. Chang, C.-J. Lin, LIBSVM, a library for support vector machine, 2001.

[20] A. D. Dileep, "Representation and feature selection using multiple kernel learning," Proc. Int. Joint Conf. Neural Networks, 717–722, 2009.

[21] W.-L. Chao, J.-Z. Liu, J.-J. Ding, "Facial age estimation based on label-sensitive learning and age-specific local regression," *Proc. ICASSP*, 1941–1944, 2012.