UNSUPERVISED DISCOVERY OF AN EXTENDED PHONEME SET IN L2 ENGLISH SPEECH FOR MISPRONUNCIATION DETECTION AND DIAGNOSIS

Shaoguang Mao[1], *Xu Li*[2], *Kun Li*[2], *Zhiyong Wu*[1,2,*], *Xunying Liu*[2], *Helen Meng*[1,2]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University

[2]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong

msg16@mails.tsinghua.edu.cn, {xuli, zywu, xyliu, hmmeng}@se.cuhk.edu.hk, kli@speechx.cn

ABSTRACT

Second language (L2) speech is often labelled with the native, phoneme categories. Hence, we often observe segments for which it is difficult, if not impossible, to decide on a categorical phoneme label. We refer to these segments as "non-categorical" phoneme units. Existing approaches to mispronunciation detection and diagnosis (MDD) mostly focus on categorical phoneme errors, where one native phoneme is substituted for another. However, non-categorical errors are not considered. To better represent L2 speech for improved MDD, this work aims to discover an Extended Phoneme Set in L2 speech (L2-EPS) which includes not only the categorical phonemes based on the native set, but also non-categorical phoneme units. We apply an optimized *k*-means algorithm to cluster phoneme-based phonemic posterior-grams (PPGs), which are generated through an acoustic-phonemic model (APM). Then we find the L2-EPS based on analysis of the clusters obtained. We verified experimentally that the non-categorical phonemes in L2-EPS can extend the native phoneme categories to better describe L2 speech. Hence L2-EPS can enrich the existing approaches to MDD for better performance.

Index Terms— Mispronunciation detection and diagnosis, mispronunciation patterns, extended phoneme set in L2 speech, unsupervised clustering, phonemic posterior-grams

1. INTRODUCTION

Computer-aided pronunciation training (CAPT) needs mispronunciation detection and diagnosis (MDD). Typically, there are several approaches [1-14]: Methods based on pronunciation scoring are popular and many different types of confidence measures can be used as pronunciation scores [1, 2, 4-9]. This kind of method often works reasonably well on detection tasks, but it does not support diagnosis. Alternative methods, such as extended recognized network (ERN) [15-17], acoustic-phonemic model (APM) [14] perform well. ERN incorporates manually designed or data-derived phonological rules including the canonical phonemic path and common mispronunciation paths to generate possible phoneme paths in a word. APM maps input features with acoustic information and phoneme context information into phone state posterior-grams for better recognition performance in MDD.

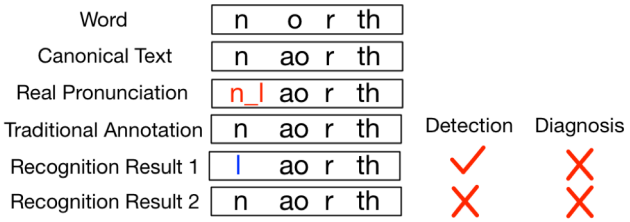

Fig. 1. An example for how non-categorical mispronunciations are wrongly treated in traditional MDD

Most existing approaches to modeling L2 speech can only target categorical phoneme error types based on the native phoneme set, but not the non-categorical errors (i.e., segments for which it is difficult, if not impossible, to label as a single native phoneme category). For example, L2 English speech uttered by native Cantonese speakers often shows that the phoneme /n/ may be mispronounced as a sound that bears resemblance to both /n/ and /l/ (*/n_l/* in Fig.1). In current MDD approaches, they are often coarsely labeled as one of the approximate phonemes. Figure 1 shows an example where the canonical annotation for "north" should be /n ao r th/, but in face of the non-categorical segment that resembles both /n/ and /l/, it may be recognized as either /l ao r th/ (as in Recognition Result 1) which enables mispronunciation detection but inaccurate diagnosis. Alternatively, if the non-categorical segment is recognized as /n ao r th/ (as in Recognition Result 2), it will fail to enable accurate mispronunciation detection or diagnosis.

To solve the above problem, this work investigates the discovery of an Extended Phoneme Set in L2 speech (L2-EPS) which includes both categorical and non-categorical phoneme segments. The objective is to find an improved representation of L2 speech pronunciation patterns. Manual analysis is cumbersome and may not fully and consistently find all instances in L2 speech data. Thus, we propose an automatic approach to discover the L2-EPS in this paper.

We cluster L2 speech frames [19, 20], and then analyze the clustering results to obtain the L2-EPS. Finally, we design experiments to verify the existence of non-categorical phonemes in the L2-EPS. This work has the following contributions: (1) It proposes a framework to discover and

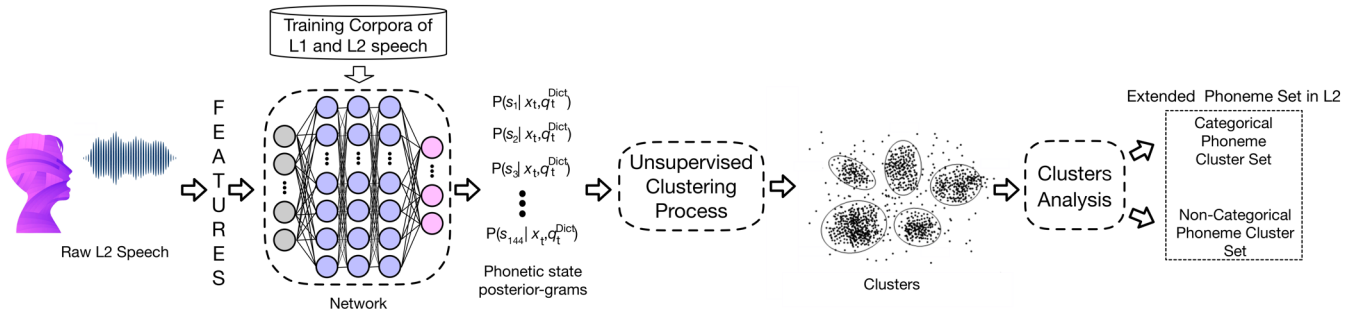

Fig. 2. The framework of the proposed approach

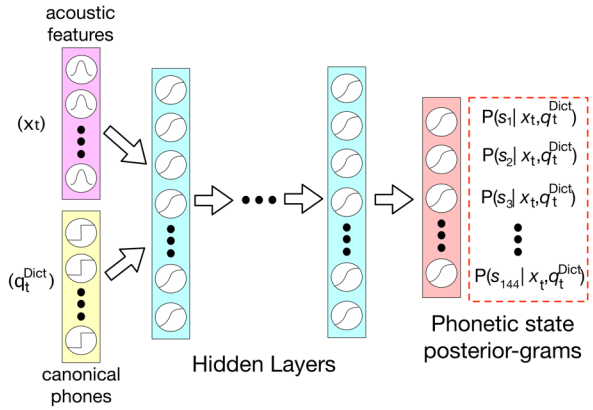

Fig. 3. Diagram illustrating Acoustic-Phonemic Model

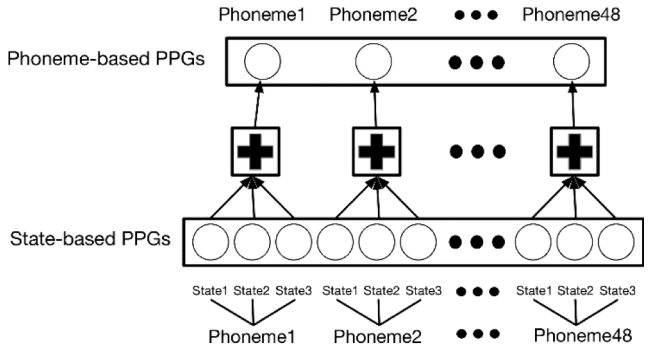

Fig. 4. Schematic diagram of changing state-based PPGs to phoneme-based PPGs

analyze L2-EPS to achieve better coverage of pronunciation patterns in L2 speech. (2) The extended acoustic-phonemic coverage enriches existing methods for MDD and may further improve MDD performance.

2. CLUSTERING FRAMEWORK

The proposed framework is shown in Fig. 2. First, we extract features from the raw L2 speech audio, and then generate Phonemic Posterior-grams (PPGs) with deep neural networks trained with both native (L1) and L2 speech data. PPGs are used to represent articulation of speech sounds in a speaker-normalized space [18, 21-24]. Next, we cluster L2 speech frames based on the PPGs features. Thereafter, we analyze the resulting clusters to label them in terms of categorical phonemes (based on the canonical phoneme set) and non-categorical phonemes (for the L2-EPS).

3. ACOUSTIC-PHONEMIC MODEL GENERATING PHONEMIC POSTERIOR-GRAMS

We use the APM to generate phonemic posterior-grams (PPGs) [21]. Figure 3 illustrates the APM, a deep neural network, that takes in Mel-frequency cepstral coefficients (MFCC) as input acoustic features (x_t), together with binary sequences of canonical phonemes (3 before, 1 current and 3 after) as phonemic features (q_t^{Dict}). The expected canonical phoneme of the moment *t* is obtained by force alignment with canonical transcriptions. The APM outputs phonemic state posterior-grams (PPGs) $P(s|x_t,q_t^{Dict})$. These are vectors that consist of posterior probabilities of every phonemic unit and have previously been used frequently as features to represent the L2 English acoustic-phonemic space [18, 21-24]. The PPGs serve as input features for the subsequent unsupervised clustering process.

4. UNSUPERVISED CLUSTERING PROCESS

We aim to discover non-categorical phonemic segments to extend the canonical phoneme set in L2 speech, based on clustering of the PPGs. We first reduce the dimensions of input PPGs features by transforming the state-based PPGs to phoneme-based PPGs. As mentioned earlier, the PPGs are phonemic vectors that consist of posterior probabilities of every phonemic unit, which can be either phoneme or phoneme state. We sum the state-based probabilities of the same phoneme into a single phoneme-based probability (as shown in Figure 4).

Subsequently, we perform *n*-best filtering on the phoneme-based PPGs by preserving the first n largest values and setting the remaining to zero. It has been shown that filtering can improve clustering performance by decreasing the influence of noise data [19, 26]. The next step is to perform random initialization for k-means clustering and select the best result in ten independent experiments.

5. EXPERIMENTAL SETUP

5.1. Speech Corpus

Our experiments are based on two speech corpora: (1) the CU-CHLOE (Chinese University-Chinese Learners of

English) data set [27] as the L2 speech corpus; and (2) the TIMIT data set as the L1 speech corpus.

We select labeled data of 70 speakers in CHLOE-C as the training set and use another 30 speakers with labeled data as the development set. Note that the TIMIT corpus is also used as part of training data. The training set is used in training networks for PPGs extraction and the development set is used in unsupervised clustering.

5.2. Experimental Setup

Clustering experiments with different configurations are implemented for comparison: (1) The k value in k-means is set from 70 to 120 with step-length being 10. (2) Frame-level features for clustering includes MFCC, state-level PPGs (derived from DNN, LSTM and APM) and phoneme-level PPGs (derived from DNN, LSTM and APM). The n value in n-best filtering (see Section 4) is empirically chosen to be 3. All clustering processes are randomly initialized.

Based on experimentation, we chose the configuration of five hidden layers with 2048 units per layer and tanh as activation function for the APM and DNN. The LSTM is determined to have two hidden layers with 512 cells. 11 frames (5 before, 1 current and 5 after) of MFCC are used as the acoustic features x_t for all networks, with MFCC extracted using 25-ms Hamming window and 10-ms frame shift. 7 canonical phonemes (3 before, 1 current and 3after) are employed as the phonemic features (q_t^{Dict}) for APM.

6. ANALYSIS OF CLUSTERS

6.1 Clustering Results Evaluation

We reference the Davies Bouldin Index (DBI) [28], which is widely used in clustering performance evaluation. It is defined as a function of the ratio of the within cluster scatter, to the between cluster separation. A lower value means that the clustering is better, and we use DBI to evaluate clustering using different setups:

$$DBI \equiv \frac{1}{N}\sum_{i=1}^{N} max_{j\neq i}\frac{S_i + S_j}{d_{i,j}} \quad (1)$$

where N is the number of clusters and S_i, $d_{i,j}$ are defined as following:

$$S_i = \frac{1}{|C_i|}\left(\sum_{X\in C_i}\|X - Z_i\|\right) \quad (2)$$

$$d_{ij} = \|Z_i - Z_j\| \quad (3)$$

where Z_i is the centroid of cluster C_i, $|C_i|$ is the size of cluster C_i, d_{ij} is the distance between Z_i and Z_j (Manhattan distance in our work). According to the results in Table 1, we choose the clustering results with phoneme-based PPGs extracted from APM and k=100 for further experimental analysis.

6.2 Cluster Grouping

We compute the PPGs for all the L1 speech frames from the TIMIT data. We also group all the L1 frames into clusters by comparing the Manhattan distance between the frames' PPGs with each generated cluster centroid. If the proportion of speech frames labeled with a canonical phoneme being grouped into the same cluster exceeds a certain threshold (set at 90%), we consider that the canonical phoneme maps to this cluster. For each cluster, if there is only one canonical phoneme mapping to it, we label this it as a Group 1 cluster (i.e. a categorical phoneme cluster). If there are more than one canonical phonemes mapping to a cluster, we label it as a Group 2 cluster (i.e. a mixed categorical phonemes cluster). otherwise, we label the cluster as a candidate non-categorical phoneme cluster, as this implies there is little or no L1 speech frames being grouped into this cluster.

Table 1. Experimental Results in DBI

Features	MFCC	PPGs from DNN		PPGs from LSTM		PPGs from APM	
		State–based	Phoneme–based	State–based	Phoneme–based	State–based	Phoneme–based
k = 70	2.17	1.87	1.62	1.77	1.61	1.53	1.34
k = 80	2.19	1.91	1.57	1.76	1.59	1.57	1.33
k = 90	2.17	1.94	1.60	1.77	1.61	1.49	1.28
k = 100	2.16	1.86	1.58	1.84	1.55	1.35	**1.26**
k = 110	2.19	1.92	1.56	1.74	1.51	1.49	1.29
k = 120	2.18	1.93	1.55	1.67	1.50	1.60	1.43

According to this rule, we can divide all clusters into 3 groups. Group 1 consists of categorical phoneme clusters; Group 2 consists of mixed categorical phonemes clusters; Group 3 consists of candidate non-categorical phoneme clusters. The details of grouping are shown in Table 2.

6.3 Analyzing Clusters by Centroids

Since the centroids of clusters are located in phoneme–based PPGs space, the coordinates of centroids can reflect the clusters' posterior probability distribution on each phoneme. For categorical phoneme clusters, their PPGs have only one peak which is located at the corresponding phoneme bit since their posterior probabilities of this phoneme is extremely high. But non-categorical phoneme clusters' PPGs may have more than one peaks, which indicates their pronunciation may be similar with more than one phonemes. Based on this property, for clusters in Group 3, we give m phonemes, of which corresponding bits' values are the top m, as reference phoneme labels (m=2 in our work).

Table 2. Details of grouping clusters with L1 data

	Description	Requirement
Group 1	Categorical Phoneme Clusters	Only one categorical phoneme maps to this cluster
Group 2	Mixed Categorical Phonemes Clusters	More than one categorical phoneme maps to this cluster
Group 3	Candidate Non-categorical Phoneme Clusters	Clusters not in Group 1 or Group 2

Table 3. Average proportion of four options being selected in some non-categorical phoneme clusters

	/aa_ao/	/aa_ax/	/aw_ax/	/ax_er/	/b_p/	/f_v/	/n_l/	/t_d/	/m_n/
More similar to P_1	50.0%	31.0%	21.0%	46.0%	18.8%	37.2%	21.7%	18.0%	55.6%
More similar to P_2	16.1%	29.4%	35.5%	30.6%	35.2%	45.7%	56.6%	57.8%	27.5%
Equal similarity to P_1 and P_2	9.7%	10.3%	3.2%	3.2%	8.6%	11.0%	8.0%	7.0%	8.8%
Not similar to either P_1 or P_2	24.2%	29.3%	40.3%	20.2%	37.4%	6.2%	13.7%	17.2%	8.1%

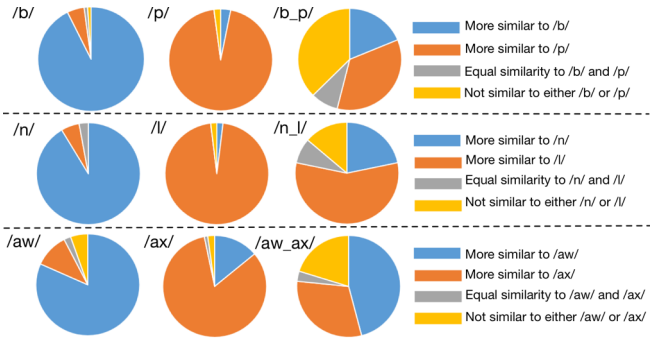

Fig. 5. The statistical results of perceptual tests in different clusters

7. PERCEPTUAL TESTS ON NON-CATEGORICAL PHONEMES

To verify that the phonemes in Group 3 cannot be described with any canonical phoneme in a categorical sense, we designed and ran a set of perceptual tests. Recall from the previous section that each non-categorical phoneme resembles two canonical phoneme P_1 and P_2. Also note that the non-categorical phoneme is marked as /P_1_P_2/.

For each non-categorical phoneme cluster, we randomly play 30 audio files (10 audio files for non-categorical phoneme cluster and two related categorical phoneme clusters respectively) to the listening subject and ask him/her to label it as one of 4 options: 1) More similar to P_1; 2) More similar to P_2; 3) Equal similarity to P_1 and P_2; 4) Not similar to either P_1 or P_2. 30 undergraduates majoring in Linguistics or English were invited to participate in these perceptual tests.

After tests, we calculate the average proportion of 4 options being selected among the audio files from each cluster. Results from non-categorical clusters are shown in Table 3. We can see that the average proportions among 4 options mostly lack a majority and listeners cannot predominantly group the audio files in a non-categorical phoneme cluster as any categorical phoneme. It means these non-categorical phonemes indeed exist and cannot be described with any canonical phoneme in a categorical sense.

To further compare the difference between categorical phoneme clusters and non-categorical phoneme clusters, we display the statistical results from some non-categorical phoneme clusters and their related categorical clusters in Figure 5. From the pie charts, we can observe the difference between categorical clusters and non-categorical clusters obviously. According to results, we find that Group 1 clusters are mostly perceived as the corresponding categorical phoneme, but non-categorical clusters lack a majority perpetual vote among the possible reference categorical phonemes.

Table 4. The phonemes in L2-EPS in our work

Categorical Phonemes								Non-categorical Phoneme	
sil	ax	dh	er	ix	n	s	v	aa_ao	eh_ey
aa	ay	dx	ey	iy	ng	sh	vcl	aa_ax	ey_ih
ae	b	eh	f	jh	ow	t	w	ae_ay	f_v
ah	ch	el	g	k	oy	th	y	aw_ax	m_n
ao	cl	en	hh	l	p	uh	z	ax_er	n_l
aw	d	epi	ih	m	r	uw	zh	b_p	t_d

Finally, with the help of linguists, we selected some classical non-categorical phonemes in L2-EPS, which are shown in Table 4. Sample audios are provided in a website (https://sites.google.com/view/l2-eps-cantonese).

8. CONCLUSION

This work aims to discover an Extended Phoneme Set for L2 speech, to achieve better coverage beyond the canonical L1 phoneme set. We apply k-means clustering on phoneme-based phonemic posterior-grams (PPGs) generated through DNN-based acoustic-phonemic model (APM). Then clusters are divided into categorical and non-categorical phoneme group with the help of L1 speech and are further analyzed with cluster centroids. According to experimental results, it is verified that non-categorical phonemes in L2-EPS we find indeed exist and they cannot be described with any canonical phoneme in a categorical sense. L2-EPS includes more complete descriptions on pronunciation patterns in L2 speech, some of which are often ignored by canonical phoneme set, and benefit in improving MDD performance. How to utilize the L2-EPS augment speech recognizer to better solve the MDD problems will be studied in the future.

9. ACKNOWLEDGEMENTS

This project is partially supported by a grant from the HKSAR RGC General Research Fund (project no. 14207315). The research was conducted while the first author was an intern at CUHK.

10. REFERENCES

[1] Neumeyer, L., Franco, H., Weintraub, M., and Price, P., "Automatic text-independent pronunciation scoring of foreign language student speech. In Spoken Language", 1996. ICSLP 96. Proceedings., Fourth International Conference on Vol. 3, pp. 1457-1460, 1996.

[2] Franco, H., Neumeyer, L., Kim, Y., and Ronen, O., "Automatic pronunciation scoring for language instruction.", Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., Vol. 2, pp. 1471-1474, 1997.

[3] Jo, C. H., Kawahara, T., Doshita, S., and Dantsuji, M., "Automatic pronunciation error detection and guidance for foreign language learning.", Fifth International Conference on Spoken Language Processing. 1998.

[4] Franco, H., Neumeyer, L., Ramos, M., and Bratt, H., "Automatic detection of phone-level mispronunciation for language learning.", Sixth European Conference on Speech Communication and Technology. 1999.

[5] Witt, S. M., and Young, S. J., "Phone-level pronunciation scoring and assessment for interactive language learning.", Speech communication 30.2 (2000), pp.95-108, 2000.

[6] Menzel, W., Herron, D., Bonaventura, P., and Morton, R., "Automatic detection and correction of non-native English pronunciations.", Proceedings of INSTILL(2000), pp.49-56, 2000.

[7] Seneff, S., Wang, C., and Zhang, J., "Spoken conversational interaction for language learning.", InSTIL/ICALL Symposium 2004, 2004.

[8] Zheng, J., Huang, C., Chu, M., Soong, F. K., and Ye, W. P., "Generalized segment posterior probability for automatic mandarin pronunciation evaluation.", Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. Vol. 4, pp. IV-201, 2007.

[9] Zhang, F., Huang, C., Soong, F. K., Chu, M., and Wang, R., "Automatic mispronunciation detection for Mandarin.", Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. IEEE, pp. 5077-5080, 2008.

[10] Truong, K., Neri, A., Cucchiarini, C., and Strik, H. "Automatic pronunciation error detection: an acoustic-phonetic approach.", InSTIL/ICALL Symposium 2004. 2004.

[11] Strik, H., Truong, K., De Wet, F., and Cucchiarini, C., "Comparing different approaches for automatic pronunciation error detection.", Speech communication51.10 (2009), pp.845-852, 2009.

[12] Lee, A., and Glass, J. R., "Context-dependent pronunciation error pattern discovery with limited annotations.", Fifteenth Annual Conference of the International Speech Communication Association. 2014.

[13] Qian, X., Meng, H., and Soong, F., "A two-pass framework of mispronunciation detection and diagnosis for computer-aided pronunciation training.", IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 24.6 (2016), pp.1020-1028, 2016

[14] Li, K., Qian, X., and Meng, H., "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks.", IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.1 (2017), pp.193-207, 2017.

[15] Harrison, A. M., Lo, W. K., Qian, X., and Meng, H., "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training.", SLaTE. 2009.

[16] Lo, W. K., Zhang, S., and Meng, H. "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system.", Eleventh Annual Conference of the International Speech Communication Association, 2010.

[17] Qian, X., Soong, F. K., and Meng, H., "Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT).", Eleventh Annual Conference of the International Speech Communication Association, 2010.

[18] Wang, Y. B., and Lee, L. S., "Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning.", IEEE Transactions on Audio, Speech, and Language Processing 23.3, pp.564-579, 2015

[19] Lee, A., Chen, N. F., and Glass, J., "Personalized mispronunciation detection and diagnosis based on unsupervised error pattern discovery." Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, pp. 6145-6149, 2016.

[20] Yu, D., and Deng, L., "Automatic speech recognition: A deep learning approach.", Springer, 2014.

[21] Sun, L., Li, K., Wang, H., Kang, S., and Meng, H., "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training.", Multimedia and Expo (ICME), 2016 IEEE International Conference on. IEEE, pp. 1-6, 2016.

[22] Lee, A., and Glass, J., "A comparison-based approach to mispronunciation detection.", Spoken Language Technology Workshop (SLT), 2012 IEEE. IEEE, pp. 382-387, 2012.

[23] Wang, Y. B., and Lee, L. S., "Toward unsupervised discovery of pronunciation error patterns using universal phoneme posteriorgram for computer-assisted language learning.", Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, pp. 8232-8236, 2013.

[24] Lee, A., Zhang, Y., and Glass, J., "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams.", Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, pp. 8227-8231, 2013.

[25] Jain, A. K., Murty, M. N., and Flynn, P. J., "Data clustering: a review.", ACM computing surveys (CSUR) 31.3 (1999), pp.264-323, 1999

[26] Boureau, Y. L., Ponce, J., and LeCun, Y., "A theoretical analysis of feature pooling in visual recognition.", Proceedings of the 27th international conference on machine learning (ICML-10), pp. 111-118,2010.

[27] Meng, H., Lo, W. K., Harrison, A. M., Lee, P., Wong, K. H., Leung, W. K., and Meng, F., "Development of automatic speech recognition and synthesis technologies to support Chinese learners of English: The CUHK experience." Proc. APSIPA ASC, pp.811-820., 2010

[28] Davies, D. L., and Bouldin, D. W., "A cluster separation measure.", IEEE transactions on pattern analysis and machine intelligence 2 (1979), pp.224-227, 1979