TOWARD UNSUPERVISED DISCOVERY OF PRONUNCIATION ERROR PATTERNS USING UNIVERSAL PHONEME POSTERIORGRAM FOR COMPUTER-ASSISTED LANGUAGE LEARNING

Yow-Bang Wang, Lin-Shan Lee

Graduate Institute of Electrical Engineering, National Taiwan University, Taipei, Taiwan(R.O.C.)

piscesfantasy@gmail.com, lslee@gate.sinica.edu.tw

ABSTRACT

In Computer-Aided Pronunciation Training, we hope to specify the type of mispronunciation, or Error Pattern (EP), the language learner has made as a more effective feedback. But derivation of EPs usually requires expert knowledge and pedagogical experiences, which is not easy to obtain for each pair of target and native languages. In this paper we propose a preliminary framework toward unsupervised discovery of EPs from a corpus of learners' recordings. We use Universal Phoneme Posteriorgram, derived from Multi-Layer Perceptron trained with a corpus of mixed languages, as features to bring supervised knowledge into the unsupervised task. We also use Hierarchical Agglomerative Clustering algorithm to explore sub-segmental variation of phoneme segments for distinguishing EPs. We tested K-means (assuming known number of EPs) and Gaussian Mixture Model with minimum description length principle (estimating unknown number of EPs) for EP discovery. Preliminary experimental results illustrated the effectiveness of the proposed framework, although there is still a long way to go compared to human annotators.

Index Terms— Pronunciation Error Pattern Discovery, Universal Phoneme Posteriorgram, HAC, K-means, GMM-MDL, Rand Index

1. INTRODUCTION

Speaking two or more languages is not only advantageous but in fact necessary for people in the era of a globalized world. Many Computer-Aided Pronunciation Training (CAPT) systems have been developed in recent years to meet the strong demand of second language learning [1][2][3][4][5][6]. In such systems, in order to generate useful feedback for language learners to improve their skills, it is preferred to offer not only numerical scores reflecting their language proficiency, but also the specific types of error they have made. Such types of error are usually referred to as Error Patterns (EPs), or the patterns of erroneous pronunciation frequently produced by language learners, usually caused by some articulator mechanism present in the target language but missing in their native languages [6][7]. This implies when N languages are considered, there can be in general N^2 sets of EPs for each pair of native and target languages.

We have been working with Chinese language teachers in our previous works [8][9]. They summarized and defined the most frequent EPs made by Chinese learners based on their expert knowledge and pedagogical experiences. They also made great effort to label a whole set of corpus of Chinese learners' recordings, including manually labeling all the phoneme segments for all utterances as either correct pronunciation or one of the EPs. Such manual labeling process is very time consuming. Some approaches were proposed to automatically derive the rules for EPs. Some began with the pronunciation error rules induced by experts in the literatures of second language learning [6][10], and some compared the orthographic transcriptions with the actual pronunciation annotated by human listeners [11] or free-phone recognition output from an automatic speech recognition (ASR) engine [12]. All these approaches still require either expert knowledge, time-consuming human labeling or reliable ASR results.

On the other hand, substantial effort has been made in recent years on unsupervised speech pattern discovery [13][14][15], with a goal to bypass the need for human annotated data for model training in speech recognition. Because building traditional ASR systems based on the HMM framework for each language and each acoustic condition can be costly, automatically discovering speech patterns based on the acoustic signal characteristics from a corpus becomes an attractive alternative. These situations are similar to the task of EP detection: lack of well annotated corpus. What is more, for EP detection the need for expertise to define and label EPs may be even more difficult and expensive.

In this paper, we learn the experiences of unsupervised speech pattern discovery, and propose a preliminary framework for automatic discovery of EPs from a corpus of learners' recordings without relying on expert knowledge. This is achieved by Universal Phoneme Posteriorgram (UPP) extracted from Multi-Layer Perceptron (MLP) trained with a corpus of mixed languages, plus the Hierarchical Agglomerative Clustering (HAC) for exploring the sub-segmental variation of EPs. Below, section 2 introduces the proposed framework, and section 3 reports the preliminary experimental results. The conclusions are in Section 4.

2. PRONUNCIATION ERROR PATTERN DISCOVERY

2.1. Problem definition

Here we assume the task is to discover the EPs for each phoneme given a corpus of learners' voice. We also assume the text transcription of the utterances in this corpus is available, so forced alignment can be performed and the learners' voice are divided into segments corresponding to phonemes. We can thus focus on one phoneme at a time: each time we are given a set of acoustic segments corresponding to a specific phoneme, and the goal is to divide this set into several clusters, each of which corresponds to an EP. Furthermore, because the percentage of correct pronunciation in our corpus is far more than mispronunciation, we excluded the correctly pronounced

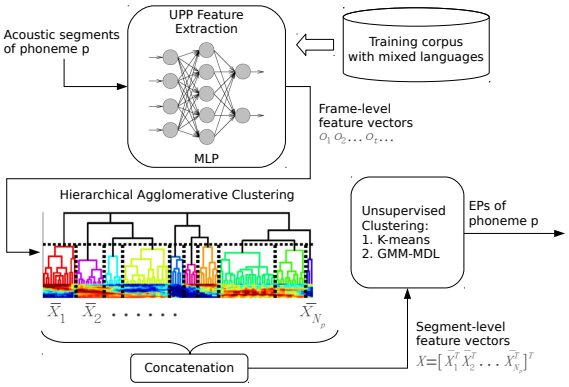

Fig. 1. *Proposed framework for unsupervised EP discovery.*

segments before clustering to avoid data imbalance problem.

2.2. Framework overview

Figure 1 shows the proposed framework for unsupervised EP discovery. First we extract frame-level feature vectors $o_1, o_2, ..., o_t, ...$ for each phoneme segment. In this work we primarily choose to use UPP as the frame-level feature vectors [16]: We train an MLP with some large corpus of mixed languages. The output target is the set of acoustic units for the mixed languages. Then we feed each MFCC frame from learners' recording into this MLP, and the output posterior probability vector is the frame-level feature vector.

Next we apply HAC to merge adjacent frames with similar acoustic features into N_p sub-segments [15]. The number of sub-segments N_p is same for each specific phoneme p, but can be different for different phonemes. The averages of frame-level feature vectors in N_p sub-segments are denoted as $\bar{X}_1, \bar{X}_2, ..., \bar{X}_{N_p}$ respectively. Then we concatenate these averaged feature vectors into one super-vector X as the segment-level feature vector. The segment-level feature vectors corresponding to each phoneme are then clustered into different EPs by an unsupervised algorithm. In this preliminary work we utilize K-means and Gaussian Mixture Model (GMM) with minimum description length (MDL) principle for unsupervised clustering.

The HAC for producing segment-level feature vectors is important here. It not only unifies the dimensionality of feature vectors for each phoneme p, with properly chosen number of sub-segments N_p, the differences among EPs can also be better retained. Because the number of frames in each segment varies, we can not simply concatenate all the frame-level features into the segment-level feature vector, nor average all the frame-level feature vectors into one. The reason for the latter is that the difference between EPs and their corresponding canonical pronunciation can be very subtle, often only by sub-segmental realization, and averaging all frames in a segment may fail to capture such subtle evidence. HAC arranges the frames in a segment into tree-structured hierarchy, in which different threshold of similarity among frames give different numbers of sub-segments out of this one segment. By tuning N_p we can thus optimize the granularity of our feature vectors for each phoneme p.

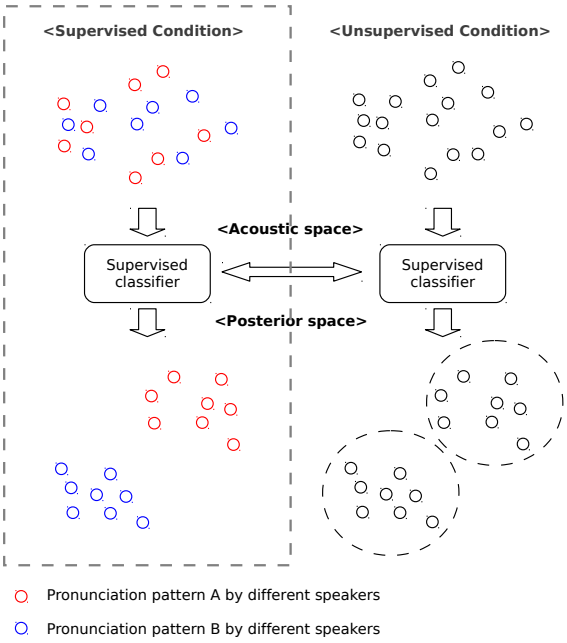

Fig. 2. *The effect of mapping from acoustic space to posterior space in supervised and unsupervised learning.*

2.3. Universal Phoneme Posteriorgram

Posterior probability vector has been widely used in CAPT and unsupervised speech pattern discovery. The well-known Goodness Of Pronunciation (GOP) is calculated based on the posterior probability of target pronunciation [17]. Some works used improved GOP with pre-defined thresholds to find mispronounced segments [18], some further incorporated GOP-based mispronunciation detector with EP network to boost the performance [5][8], and some utilized log-likelihood ratio or posterior probability vectors as input features [9][19][20][21] of discriminative classifiers such as Support Vector Machine (SVM). Also many works of pattern discovery have adopted posteriorgrams as the features for further processing. Some derived posteriorgram with GMM trained with the target corpus [13], and some with MLP trained with another large corpus [14].

The goal of utilizing MLP obtained with supervised training for posteriorgram feature extraction is to bring the information of acoustic space partitioning with known pronunciation units into under-resourced tasks [14][16]. As illustrated in Figure 2, in the upper left the acoustic instances of two different pronunciation patterns A and B scatter over the acoustic space. Because they are produced by many speakers, the speaker variation and pronunciation variation are mixed together. Under supervised condition, the labels of pronunciation patterns are given, and we can train a classifier (e.g. the MLP here) which focus on distinguishing different pronunciation patterns. Therefore in the lower left the two pronunciation patterns A and B become easier to distinguish in the posterior space, despite they are produced by different speakers. However, under unsupervised condition as in the upper right of Figure 2, we are no longer aware of which instance belongs to which pattern. By borrowing the supervised classifier trained with annotated multi-speaker corpus from the left, we can thus similarly map the instances from acoustic space to

posterior space, on which the speaker variation may be removed to a certain degree while preserving the traits of pronunciation variation.

2.4. Unsupervised Clustering Algorithms

In the preliminary work here we use two algorithms for EP clustering: K-means and GMM with minimum description length principle (GMM-MDL)[22]. For K-means we assume the number of clusters k is known, which is the number of EPs of each phoneme summarized by language teachers. Several different distance measures are considered: Euclidean distance $d_{euc}(x, y)$, Cosine distance $d_{cos}(x, y)$ and Symmetric KL Divergence $d_{kld}(x, y)$, where

$$d_{cos}(x, y) = 1 - \frac{x \cdot y}{|x||y|}, \tag{1}$$

$$d_{kld}(x, y) = \frac{1}{2}\sum_{i=1}^{D}(x_i \cdot log\frac{x_i}{y_i} + y_i \cdot log\frac{y_i}{x_i}), \tag{2}$$

x and y are the segment-level feature vectors with dimensionality D, and x_i, y_i are their i-th component respectively.

To automatically learn the number of clusters while discovering EPs, we use GMM-MDL algorithm. We trained one GMM for each phoneme p, and then perform maximum-likelihood (ML) classification to assign instances to clusters. The GMM-MDL algorithm is capable of estimating the optimal number of Gaussians in GMM, which is the number of EPs in our application. The objective function to be optimized is:

$$F(S_p, \theta_p) = logPr(S_p|\theta_p) - \frac{1}{2}|\theta_p|log(|S_p|D_p), \tag{3}$$

where θ_p is the parameters of GMM, S_p the set of segment-level feature vectors of dimensionality D_p with size $|S_p|$. $|\theta_p|$ the total number of continuously valued free variables to specify θ_p:

$$|\theta_p| = M_p(1 + D_p + \frac{(D_p + 1)D_p}{2}) - 1, \tag{4}$$

where M_p is the number of Gaussians. Eq. 4 comes from the fact that for each Gaussian there are 1 prior probability, D_p means and $\frac{(D_p+1)D_p}{2}$ variables for covariance matrix. Because the M_p priors sum to one, the overall degree of freedom is reduced by 1. In Eq. 3 the first term on the right hand side is the log-likelihood, and the second term represents the model complexity. So the number of clusters is estimated by the balance between the two considerations.

3. EXPERIMENTS

3.1. Corpus, EP definition and annotation

Our corpus was collected in year 2008 and 2009. 278 learners studying Mandarin Chinese in National Taiwan University from 36 different countries with balanced gender and a wide variety of native languages joined the recording tasks. Each learner was asked to produce a set of 30 phonetically balanced and prosodically rich sentences, each containing 6 to 24 characters. These 30 sentences covered almost all frequently used Mandarin syllables and tone patterns.

The acoustic units for EP definition in this work are Mandarin phonemes represented in Zhuyin. There is a total of 39 canonical Mandarin phoneme units, and 152 EPs were summarized by language teachers based on their expert knowledge and pedagogical experiences, to cover most frequent EPs made by Mandarin Chinese learners. This means in average we have $152/39 \approx 3.9$ EPs per phoneme unit. The definition of EPs includes not only phoneme-level substitution, but also insertion and deletion, and is not limited to any specific corpus including the one mentioned above [8].

Two annotators labeled the surface pronunciation of each acoustic segment in each utterance in the above corpus as correct pronunciation or one of the EPs. We used the labels from one annotator as the reference EPs in our experiments, and the other in finding out the consistency between human annotators.

3.2. Feature extraction and HAC

The training corpus of the MLP for UPP derivation included the ASTMIC Mandarin corpus (read speech produced by 95 males and 95 females with a total length of 24.6 hours), and the training set of TIMIT English corpus (462 speakers from eight dialect regions of the USA, with a total length of 3.9 hours). The MLP training target was the union of the monophone sets of Mandarin and English, consisting of 35 and 38 monophones respectively, without short pause and silence. Logarithm of UPP features (log-UPP) and MFCC (39 parameters, c0 to c12 plus first and second derivatives) were also tested as features.

Three different choices of number of sub-segments N_p divided by HAC was considered: $N_p = 1$, N_{opt} and N_{max}. For $N_p = 1$ we did not divide a segment into smaller sub-segments. For $N_p = N_{opt}$ we tuned N_p for each phoneme p by optimizing the performance. For $N_p = N_{max}$ we set N_p to be the number of frames in the shortest segment, so for the shortest segment we treat each frame as a sub-segment.

3.3. Evaluation Metric

There are many different metrics for evaluating clustering algorithms. Cluster purity is a good example, although it tends to favor larger number of clusters. Here we adopt the Rand Index [23] for its balanced consideration between the similarity within clusters and dissimilarity among different clusters.

We first define the True Acceptance (TA), True Rejection (TR), False Acceptance(FA) and False Rejection (FR) based on all instance pairs as in Table 1 [14]. For example, if an instance pair belongs to the same cluster in both reference and prediction result, it is counted as one TA. We can see TA and TR represent respectively the within-cluster and between-cluster accuracies. The Rand Index is then defined as:

$$RI = \frac{TA + TR}{TA + TR + FA + FR}. \tag{5}$$

Since the mispronounced segments for each phoneme were clustered for EPs individually, we report the Average Rand Index (ARI) over all phonemes p in Mandarin phoneme set P:

$$ARI = \frac{1}{|P|}\sum_{p \in P} RI(p). \tag{6}$$

3.4. Experimental Results

3.4.1. K-means with known number of EPs

Table 2 reports the ARI using K-means with known number of EPs for each phoneme. Different rows represent different features and different distance measures, and different columns represent different numbers of sub-segments N_p derived by HAC. We see the best

Table 1. *The definition of True Acceptance (TA), True Rejection (TR), False Acceptance(FA) and False Rejection (FR) in our experiment.*

For all instance pairs beloning to		Reference clusters	
		same	different
Predicted clusters	same	TA	FA
	different	FR	TR

performance was achieved by log-UPP with cosine distance regardless of N_p being 1, N_{opt} or N_{max}. Although there is still a gap from the high consistency achieved by the human annotator, the results verified that UPP yielded better discriminability among EPs in the posterior space. Also note that the second-best result occurred with UPP and Symmetric KL-Divergence, which conformed the fact that UPP represents posterior probability distribution and KL-Divergence is a good distance measure for it [14].

By comparing different columns we can see that $N_p = N_{opt}$ gave better results than $N_p = 1$ and $N_p = N_{max}$. Note because we should not be able to compare the clustering performance with the reference under unsupervised condition, the results given by $N_p = N_{opt}$ is the upper bound of ARI achievable by properly set the number of sub-segments with the HAC. Yet this still verified that difference among EPs may lie in sub-segmental realizations, which can be better explored with HAC. Simply set $N_p = 1$ gave too coarse features, while partitioning each segment into N_{max} sub-segments may be over-analyzing and introduce noise.

Table 2. *Experimental results in Average Rand Index (ARI) (%) with varying features, distance measures and number of sub-segments using K-means algorithm with known number of EPs for each phoneme.*

Feature	Algorithm	N_p: number of sub-segments		
		1	N_{opt}	N_{max}
MFCC	K-means,d_{euc}	57.78	58.73	57.65
MFCC	K-means,d_{cos}	57.65	58.58	57.39
UPP	K-means,d_{euc}	56.17	57.34	56.25
UPP	K-means,d_{cos}	56.40	57.71	56.70
UPP	K-means,d_{kld}	58.15	58.94	57.47
log-UPP	K-means,d_{euc}	57.03	57.58	56.63
log-UPP	K-means,d_{cos}	**58.46**	**59.43**	**58.36**
Human Annotator		71.05		

3.4.2. GMM-MDL with automatically estimated number of EPs

Table 3 shows the results of ARI using GMM-MDL. The numbers in the brackets are the difference of the automatically estimated number of patterns compared to that summarized by human experts, averaged over all phonemes. We can see similar trend of ARI as in Table 2: log-UPP yielded the best performance, and N_{opt} considering the sub-segmental realization was better. Yet the achieved ARI in Table 3 were lower than those of Table 2, obviously due to the lack of expert knowledge about the number of EPs. Note both UPP and log-UPP yielded 1 to 3 more automatically derived EPs than human-defined EPs in average. In contrast MFCC resulted in less number of clusters. This further showed the better discriminating power of UPP in discovering EPs.

4. CONCLUSION

In this paper we proposed a preliminary framework for unsupervised discovery of pronunciation Error Patterns. We utilized Universal Phoneme Posteriorgram derived from MLP trained with a corpus of mixed languages, to reduce speaker variation while maintaining pronunciation variation. The experimental results showed that Universal Phoneme Posteriorgram successfully boosted the discriminating power among EPs compared to MFCC, in terms of both higher Average Rand Index and more number of clusters.

Table 3. *Experimental results in Average Rand Index (ARI) (%) and average of differences in number of clustered EPs compared to reference EP (in brackets), with varying features and number of sub-segments, using GMM-MDL with unknown number of EPs*

Feature	Algorithm	N: number of sub-segments		
		1	N_{opt}	N_{max}
MFCC	GMM-MDL	53.96 (-0.67)	56.28 (-0.64)	51.72 (-1.18)
UPP	GMM-MDL	**54.64** (2.79)	56.20 (2.59)	54.39 (1.95)
log-UPP	GMM-MDL	54.37 (2.08)	**56.58** (1.49)	**54.58** (0.74)

5. REFERENCES

[1] B. Granström, "Towards a virtual language tutor," in *InSTIL/ICALL Symposium 2004*, 2004.

[2] C. Tsurutani, Y. Yamauchi, N. Minematsu, D. Luo, K. Maruyama, and K. Hirose, "Development of a program for self assessment of japanese pronunciation by english learners," in *Proc. Interspeech 2006*.

[3] K. Zechner, D. Higgins, X. Xi, and D.M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[4] B. Yoshimoto, *Rainbow Rummy: a Web-based game for vocabulary acquisition using computer-directed speech*, Ph.D. thesis, Massachusetts Institute of Technology, 2009.

[5] H. Meng, W.K. Lo, A.M. Harrison, P. Lee, K.H. Wong, W.K. Leung, and F. Meng, "Development of automatic speech recognition and synthesis technologies to support chinese learners of english: The CUHK experience," in *APSIPA Annual Summit and Conference 2011*.

[6] C. Cucchiarini, H. Van Den Heuvel, E. Sanders, and H. Strik, "Error selecion for asr-based english pronunciation training in 'my pronunciation coach'," in *Proc. INTERSPEECH 2011*.

[7] H. Meng, Y.Y. Lo, L. Wang, and W.Y. Lau, "Deriving salient learners mispronunciations from cross-language phonological comparisons," in *Proc. ASRU2007*, pp. 437–442.

[8] Y.-B. Wang and L.-S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *Proc. ICASSP 2012*.

[9] Y.-B. Wang and L.-S. Lee, "Error pattern detection integrating generative and discriminative learning for computer-aided pronunciation training," in *Proc. Interspeech 2012*.

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

[11] D. Luo, X. Yang, and L. Wang, "Improvement of segmental mispronunciation detection with prior knowledge extracted form large l2 speech corpus," in *Proc. INTERSPEECH 2011*.

[12] Q. Shi, K. Li, S.L. Zhang, S.M. Chu, J. Xiao, and Z.J. Ou, "Spoken english assessment system for non-native speakers using acoustic and prosodic features," in *Proc. INTERSPEECH 2010*.

[13] Y. Zhang and J.R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *Proc. ICASSP 2010*. IEEE, pp. 4366–4369.

[14] M.A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Proc. Interspeech 2011*.

[15] C.-A. Chan and L.-S. Lee, "Unsupervised hidden markov modeling of spoken queries for spoken term detection without speech recognition," in *Proc. Interspeech 2011*, pp. 2141–2144.

[16] D. Imseng, H. Bourlard, and P.N. Garner, "Using kl-divergence and multilingual information to improve asr for under-resourced languages," in *Proc. ICASSP 2012*. IEEE, 2012, pp. 4869–4872.

[17] S.M. Witt and S.J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[18] F. Zhang, C. Huang, F.K. Soong, M. Chu, and R. Wang, "Automatic mispronunciation detection for mandarin," in *Proc. ICASSP 2008*, pp. 5077–5080.

[19] J. Jiang and B. Xu, "Exploring the automatic mispronunciation detection of confusable phones for mandarin," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4833–4836.

[20] Y. Chen, C. Huang, and F. Soong, "Improving mispronunciation detection using machine learning," in *Proc. ICASSP 2009*, pp. 4865–4868.

[21] S. Wei, G. Hu, Y. Hu, and R.H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communication*, vol. 51, no. 10, pp. 896–905, 2009.

[22] C.A. Bouman, M. Shapiro, GW Cook, C.B. Atkins, and H. Cheng, "Cluster: An unsupervised algorithm for modeling gaussian mixtures," `https://engineering.purdue.edu/~bouman/software/cluster/`, 1997.

[23] C.D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*, vol. 1, Cambridge University Press Cambridge, 2008.