Context-dependent Pronunciation Error Pattern Discovery with Limited Annotations

Ann Lee, James Glass

MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA

{annlee, glass}@mit.edu

10.21437/Interspeech.2014-597

Abstract

A Computer-Assisted Pronunciation Training (CAPT) system can provide greater benefit to language learners if it provides not only scoring but also corrective feedback. However, the process of deriving pronunciation error patterns usually requires linguistic knowledge, or large quantities of expensive, annotated, corpora from nonnative speakers. In this paper we explore the possibility of deriving context-dependent error patterns with limited human annotations. A two-stage labeling mechanism is proposed, which first selects a set of templates for human annotation, and then propagates the labels. To deal with the imbalanced number of correct and incorrect phone-level pronunciations in nonnative speech, pronunciation patterns on an individual learner-level are first summarized, and then corpus-level clustering is done for template selection. The concept of contextual similarity based on a phonemic broad class definition is also proposed for label propagation. For evaluation, we view the task as an information retrieval task, and take advantage of metrics that consider both the importance and the ranking of an error type. Experimental results on a Chinese University of Hong Kong (CUHK) nonnative corpus show that the proposed framework can effectively discover prominent error patterns.

Index Terms: Computer-Assisted Language Learning, unsupervised clustering, graph-based label propagation

1. Introduction

Computer-Assisted Language Learning (CALL) systems have enabled students to learn and practice a second language at their own pace. While learning a new language involves aspects ranging from reading, writing, and speaking, Computer-Assisted Pronunciation Training (CAPT) focuses on the problem of detecting mispronunciation in students' speech.

Over the past two decades, there has been a great amount of research on detecting mispronunciation in nonnative speech [1, 2]. Likelihood-based scoring, e.g. HMM-based log-likelihood scores, or HMM-based log posterior scores [3, 4, 5], have been shown to correlate well with human assessment. However, in light of pedagogical value, a CAPT system can provide greater benefit to students if it can produce scores and also provide feedback on how to adjust pronunciations. Classifiers for common problematic phoneme pairs are built for this purpose [6, 7]. In a more general framework, possible error patterns can be incorporated into a lexicon for recognition [8, 9, 10]. These approaches have been shown to outperform likelihood-based scoring, but possible error types have to be known beforehand.

Identifying possible error patterns can be done in a knowledge-driven manner either by consulting with experienced language teachers, or by carrying out cross-language phonological comparisons between the students' mother tongue (L1) and the target language (L2) [9, 11]. Another way would be to use data-driven approaches that discover error patterns by aligning a human-transcribed L2 dataset with canonical pronunciations from a dictionary [10, 11, 12]. The approaches above however would require either linguistic expertise in L1-L2 pairs, or a fully transcribed L2 corpus, which is expensive, and time-consuming to collect. Recently, Wang and Lee [13] presented results on unsupervised discovery of mispronunciation patterns by clustering universal phoneme posteriorgrams. Note that in this preliminary study, correctly pronounced phones are excluded to avoid a data imbalance problem.

In this paper, we explore the problem of pronunciation error pattern discovery in a more realistic scenario. Given an unlabeled nonnative corpus, assume that there are limited human resources available from whom we can ask for annotations. To deal with the issue that the number of correct and incorrect segments may be highly imbalanced, a two-stage labeling mechanism is proposed. In the first stage, pronunciation patterns on an individual learner-level are first summarized, and then a number of corpus-level pronunciation templates are selected based on the number of human annotations available. In the second stage, the phone labels are propagated on the corpus-level and then on the learner-level in a context-aware manner. In the end, with limited human inputs, the system can not only discover context-indpendent error patterns, e.g. phone substitutions ($\alpha \rightarrow \beta$), but also generate a list of context-dependent error patterns, e.g. phone substitutions under contexts γ and δ ($\alpha \rightarrow \beta$ / γ _ δ). Another issue that we discuss is the importance of each error pattern. Some mistakes are more frequent, and thus should be emphasized more in a CAPT system. As a result, in evaluation, we view the task of error patterns discovery as an information retrieval task, and employ metrics that take the ranking and the relevance of each error type into consideration.

2. Corpus

2.1. Chinese University Chinese Learners of English corpus

The Chinese University Chinese Learners of English (CUCHLOE) corpus [9] is a specially-designed corpus of native Cantonese speakers learning English. There are 100 speakers (50 males, 50 females) in total, who are all university students.

In this paper, we focus on the *minimal pair* set, where all learners read 50 scripts, including 128 pairs of words. The set is fully transcribed by human experts. The human transcription annotates the mispronunciations that the learners have made, and is called the *"surface pronunciation"*. On the other hand, the canonical pronunciation from a lexicon is called the *"underlying pronunciation"*. Fig. 1 shows the distribution of underlying phones in the scripts. Only phones that appear more than 10 times are considered in the experiments.

2.2. Contexts, learners, and pronunciation patterns

By aligning the human transcription with underlying pronunciations, we can detect phone-level pronunciation discrepan-

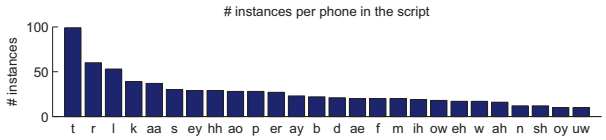

Figure 1: *Number of instances per underlying phone of the 50 scripts in the minimal pair set of the CU-CHLOE corpus*

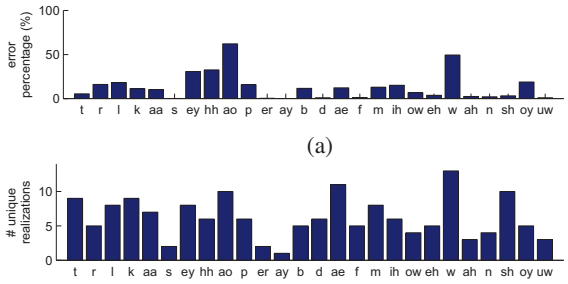

Figure 2: *Error rate and number of unique surface phones for each underlying phone category*

cies (errors) including insertions, substitutions, and deletions. Fig. 2a shows the error rate of each underlying phone group from the entire dataset. Fig. 2b shows the number of unique surface realizations that were produced per underlying phone across the dataset. Note that only surface phones that appear more than 3 times are included in the figure, as oftentimes errors with rare occurrences were caused by misreading. From the two figures, we can see that the distributions of sounds vary a lot from phone to phone. While most of the time the error rate is low ($< 20\%$), suggesting a highly imbalanced ratio between correct and incorrect segments, the few mispronounced segments are distributed into a large number of error types.

Although it seems that a phone may have many error types, if we examine the number of unique pronunciations that an individual learner produces, this number is much lower. As shown in Fig. 3, we can see that on average an individual learner would pronounce a phone in 1 to 3 ways, with a standard deviation less than 1. This indicates that when a learner mispronounces a sound, he/she would tend to repeat the error, rather than producing new types of errors in the future.

Moreover, if the contexts are taken into consideration, the number of unique realizations that an individual learner makes is further reduced. Fig. 4 plots the histogram of number of unique realizations per underlying context-dependent triphone,

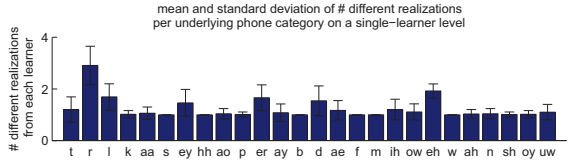

Figure 3: *Average number of unique surface phones for each underlying phone category from individual learners*

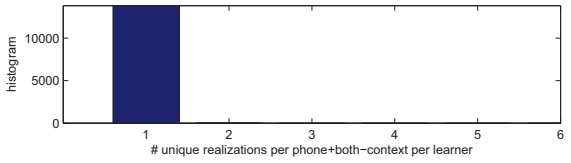

Figure 4: *Histogram of number of unique surface pronunciation per underlying context-dependent triphone from each learner*

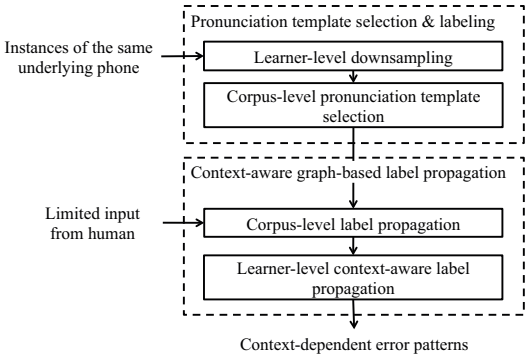

Figure 5: *System flowchart. For all instances in an underlying phone group, downsampling is first done on a single learner basis, and then templates are selected on corpus-level. After obtaining human annotations, the system propagates the labels on corpus-level first, and then on each learner-level.*

accumulated from 100 speakers. Again only phone and context combinations that appear more than 3 times in the scripts are considered here. The figure shows that more than 99% of the time, an individual learner would keep pronouncing an underlying phone in the same way if the context remains the same.

On the basis of the above analysis, in the next section, we design a two-stage framework that takes advantage of learner identity and contextual information to discover pronunciation error patterns with limited human annotations.

3. System Design

Fig. 5 shows the flowchart of our system. After collecting a set of nonnative read speech, and running forced alignment, we categorize all phone segments into groups, based on the underlying pronunciation. For each group, a two-stage labeling is performed. In the first stage, the goal is to select templates that can cover as many of the pronunciation patterns as possible, given the constraint of the number human annotations available. In the second stage, the obtained human annotations are used to annotate the unlabeled phone segments. After the entire dataset is labeled, the system produces a list of context-dependent error patterns. Below we describe each stage in detail.

3.1. Pronunciation template selection

Unsupervised clustering techniques are often applied in dictionary learning for aiding the subsequent supervised learning process. However, algorithms such as k-means tend to produce clusters of relatively uniform size, which is called the "uniform effect" [14]. To deal with the highly skewed distribution in our case, we carry out a two-level template selection process.

As pointed in Section 2.2, the number of pronunciation patterns that an individual learner produces per phone group is relatively limited, compared to the number from the entire corpus. Ideally, a small set of samples would be enough to represent a learner's pronunciations. As a result, clustering-based downsampling is first carried out on the individual learner level. Assume an underlying phone group, p, of the i^{th} learner contains the set of segments $P^i = \{p^i_1, p^i_2, \cdots, p^i_{|P^i|}\}$. After applying an n-class clustering, the original set P^i can then be downsampled to a set of n segments $P^i_d = \{p^i_{d_1}, p^i_{d_2}, \cdots, p^i_{d_n}\}$, where $p^i_{d_j}$'s are selected based on the criterion that segments with the smallest sum-of-distance to all the other segments in the same cluster are chosen as the samples.

The downsampling process can also be viewed as a kind of summarization of each learner's pronunciation patterns. After learner-level downsampling, correctly pronounced segments are

still in the majority for most phone groups, but the skewness of the overall phone pronunciation distribution at the corpus-level will be decreased. Given the constraint that only k human annotations are available for each phone group, another k-class clustering can be done on the downsampled sets, and the final corpus-level templates are selected using the same criterion.

In this two-level template selection process, any unsupervised clustering algorithm can be applied on the two levels. In Section 4.2, we will present experimental results, and discuss the pros and cons of using different clustering algorithms.

3.2. Context-aware graph-based label propagation

After the surface annotations of the templates are obtained, in the second stage, we adopt a graph-based label propagation algorithm to annotate the unlabeled segments. Consider each phone segment as a node v_i in an undirected graph $G = \{V, E\}$. The weights w_{ij} on the edge connecting v_i and v_j can be viewed as the similarity between nodes, and is defined as

$$w_{ij} = exp(-\frac{d^2(v_i, v_j)}{\sigma^2}) = exp(-\frac{d_{ij}^2}{\sigma^2}). \quad (1)$$

If the distance $d(\cdot)$ is defined on a speech representation, e.g. MFCCs or phoneme posteriorgrams, we call it an acoustic distance, denoted as d_{ij}^a, and thus w_{ij} represents the acoustic similarity between two segments, denoted as w_{ij}^a.

In addition to an acoustic similarity, we also propose a concept of *"contextual similarity"* . While each phonetic unit has its own characteristics, a broad class [15], e.g. "Front_Vowel" or "Voiced_Stop", captures some common aspect of a subset of phonetic units, such as manner, or place of articulation. Let B be a set of broad classes. Given a segment v_i and its left and right contexts l_i and r_i, a binary vector b_i of length $2|B|$ can be decoded, where the first $|B|$ elements indicate whether l_i belongs to a broad class or not, and the next $|B|$ elements indicate r_i. The contextual distance d_{ij}^c can thus be computed between b_i and b_j, and the corresponding w_{ij}^c is called the contextual similarity. Table 1 lists all the broad classes that we use. The detailed definition and phoneme examples can be found in [16].

Given a graph, Zhu and Ghahramani [17] proposed an iterative algorithm that propagates a node's labels to all nodes according to their proximity. Let T be a $|V| \times |V|$ probabilistic transition matrix where $T_{ij} = P(j \rightarrow i) = \frac{w_{ij}}{\sum_{k=1}^{|V|} w_{kj}}$, and Y be a $|V| \times C$ matrix, where C is the total number of labels, and Y_{ij} represents the probability that node v_i has the jth label. The algorithm works as follows:

1. All nodes propagate labels for one step: $Y \leftarrow TY$
2. Row-normalize Y so that each row sums to 1.
3. Clamp the labeled data to $Y_{ic} = \delta(v_i, c)$. Repeat from step 1 until Y converges.

On the basis of this algorithm, we first perform corpus-level label propagation on the graph whose nodes are the union of P_d^i from all learners and weights are based on w^a's. Once the elements in P_d^i's are labeled according to the maximum probabilities in Y, learner-level propagation is done on the graphs whose nodes are the set P^i, and weights are an interpolation of w^a and w^c, for each individual learner, respectively. In the end,

Table 1: *Broad classes used in computing contextual similarity*

Manner	Low_Vowel, High_Vowel, Retroflex, Lateral, Nasal, Weak_Fric, Strong_Fric, Closures, Voiced_Stop, Unvoiced_Stop, Silence
Place of articulation	Labial, Alveolar, Palatal, Velar, Front, Back, Mid, Semi_Vowel

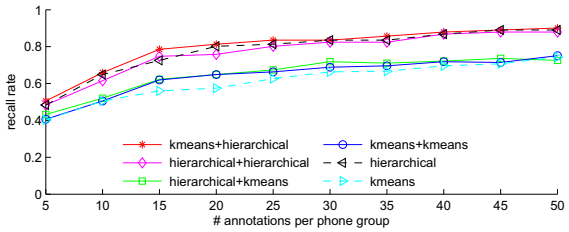

Figure 6: *Recall rate of 4 two-level clustering methods (learner-level+corpus-level) and 2 one-level clustering methods for pronunciation template selection*

the system can output a list of context-dependent pronunciation error patterns ranked by their number of occurrences.

4. Experiments

4.1. Experimental setting

We split the data into 80 speakers (40 males and 40 females) as a test set, including 57,182 phone segments and 26 phone categories, with the remaining 20 speakers used as a development set for parameter tuning. All waveforms are transformed into 39-dimensional MFCCs every 10ms. Forced alignment is done by using a DNN-HMM-based SUMMIT recognizer [18] trained on the TIMIT training set [19]. Within each phone segment, we average the MFCCs at the three regions: 0%-30%(start), 30%-70%(middle), 70%-100%(end), and concatenate the three averaged MFCCs, resulting in a 117-dimensional vector for each phone segment. A Euclidean distance metric is used to compute both the acoustic distance and the contextual distance.

4.2. Pronunciation template selection

Two unsupervised clustering algorithms, k-means and hierarchical clustering, are tested on both the learner-level and corpus-level. The number of groups, n, in the learner-level clustering is tuned using the development set. The number of groups, k, in the corpus-level clustering is a given constraint. The recall rate, which is the number of unique surface annotations obtained, divided by total number of unique surface phones in the dataset, is used for evaluation. Only surface pronunciations that appear more than 10 times are considered, resulting in 91 unique labels.

Fig. 6 illustrates the performance with respect to k. The baselines are k-means and hierarchical clustering done on the corpus-level, i.e. no learner-level downsampling. The results indicate that doing hierarchical clustering on the corpus-level can consistently discover more pronunciation patterns than k-means can. This is because of hierarchical clustering's ability to detect outliers. When the class distribution is less balanced, small clusters can be viewed as a kind of outlier, and hierarchical clustering is better able to separate them from the majority.

Doing k-means on the learner-level, and hierarchical clustering on the corpus-level gives the best performance. With 50 annotations per phone category, which is equivalent to 2.3% of the total number of phone segments, the framework can discover 90% of the context-independent pronunciation error patterns. Although it appears that with more annotations, one-level hierarchical clustering can discover nearly the same amount of pronunciation patterns as a two-level framework, we will show that a two-level design outperforms a one-level framework in discovering context-dependent error patterns.

4.3. Context-dependent error pattern discovery

A CAPT system can provide greater benefit to learners if it can prioritize the learning process. If some errors are more frequent, they should be emphasized more by the system. Fig. 7 sorts the context-dependent error types in the dataset by their

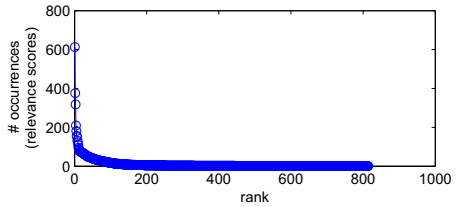

Figure 7: *Context-dependent error patterns ranked by number of occurrences.*

frequency of occurrence according to ground truth human annotations. The long tail implies that only a very small portion of the error types are prominent. Therefore, when evaluating a list of automatically discovered error patterns, it is important to take the prominence of each error type into consideration.

Discounted cumulative gain (DCG) [20], a metric that takes both relevance score and ranking into account, and is commonly used in evaluating information retrieval tasks, can meet our need. Given a ranked list of n retrieved results, DCG can be computed as $DCG_n = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i}$, where rel_i is the relevance score of the i^{th} result. Therefore, if the retrieved results put entities with greater relevance at a higher ranking, its DCG score will be larger. In our case, we treat the number of occurrences as the relevance score for each error type.

After carrying out k-means on the learner-level, and hierarchical clustering on corpus-level for template selection, we then perform label propagation. In addition to graph-based label propagation, we also examine cluster-based label propagation, which assigns the obtained annotations to all other unlabeled segments in the same cluster. Fig. 8 shows the results from 6 combinations of the two-level label propagation methods, together with 3 one-level label propagation based on the templates selected by corpus-level hierarchical clustering only. Two metrics are used: DCG of the output, normalized by DCG of the ground-truth ranking, and precision at n. The former can be viewed as a kind of recall rate weighted by relevance scores and rankings, and the latter is the precision of top n results.

For different two-level label propagation methods, in the higher rank region, carrying out graph-based label propagation on the corpus-level achieves the same level of precision as cluster-based label propagation, but with a much higher DCG score. This is because hierarchical clustering tends to form a few small clusters with one big cluster. Though it helps in discovering unique pronunciation patterns, cluster-based label propagation misses many mispronounced segments, and thus the output is less capable of reflecting the true ranking of error patterns. A two-level framework with graph-based label propagation on the corpus-level also outperforms a one-level framework, due to the learner and contextual information incorporated in it. When the number of human annotations increases from 30 to 50, the performance of the two-level frameworks all improve, while the performance of the one-level frameworks all decrease. This may suggest that a two-level design can also better utilize additional information provided by humans.

For different learner-level label propagation methods, graph-based label propagation based on both acoustic and contextual similarity achieves same level of precisions as other methods, while it achieves the highest DCG scores. With contextual similarity, the labels of phone segments with the same context are more identical, and thus the ranking of the discovered error patterns are closer to the ground truth. Table 2 lists the top 10 results of ground truth error patterns and the results from the best two-level framework with 50 human annotations per phone group. If we view the results from a broad class perspective, they successfully reveal the dominant error types and the contexts that errors often occur, e.g. retroflexion and semivowels in the context of back vowels and word endings.

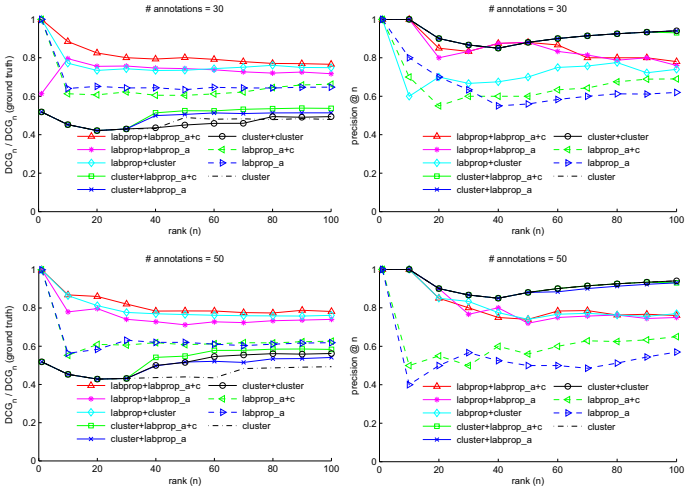

Figure 8: *Performance of context-dependent error pattern discovery of 6 two-level (corpus-level+learner-level) and 3 one-level propagation methods. (cluster: cluster-based label propagation, labprop: graph-based label propagation, a: graph weights w^a, c: graph weights w^c)*

Table 2: *Top 10 context-dependent error patterns from ground truth and the proposed framework (*: deletion)*

Ground truth	Proposed framework ($k = 50$)
$r \rightarrow */aa\_t$	$r \rightarrow */aa\_t$
$r \rightarrow */aa\_\#$	$r \rightarrow */aa\_\#$
$er \rightarrow ax/t\_\#$	$r \rightarrow */ao\_\#$
$r \rightarrow ax/eh\_\#$	$er \rightarrow ax/t\_\#$
$er \rightarrow ax/aw\_\#$	$l \rightarrow */ey\_\#$
$r \rightarrow ax/ay\_\#$	$l \rightarrow */ay\_\#$
$er \rightarrow ax/p\_\#$	$er \rightarrow ax/ay\_\#$
$r \rightarrow ax/ao\_\#$	$er \rightarrow */eh\_\#$
$l \rightarrow */ao\_\#$	$er \rightarrow ax/eh\_\#$
$ey \rightarrow eh/t\_l$	$l \rightarrow */ao\_\#$

5. Conclusion and Future Work

In this paper, we have presented a framework for discovering context-dependent pronunciation error patterns given limited human annotations. We first presented a data analysis indicating learner identity and context provide valuable information for reducing the space of possible pronunciation errors. This also suggests the concept of personalization in CAPT, as different learners have different issues in pronunciation. By applying two-level clustering, and graph-based label propagation, the proposed framework is capable of discovering 90% of context-independent pronunciation patterns, as well as the prominent context-dependent error patterns. Given an unannotated non-native corpus, the prominent error patterns discovered by the framework can serve as a starting point for building a CAPT system, and help learners to start working from their common error types. If additional human annotations become available, the CAPT system can then be incrementally improved.

For future work, as we are focusing on evaluating the quality of the list of discovered patterns now, the next step would be to incorporate the output into a real ASR system to carry out mispronunciation detection. Also, it would be interesting to see if the proposed framework can also discover prominent pronunciation error patterns in different L1-L2 pairs.

6. References

[1] Eskenazi, M., "An overview of spoken language technology for education", Speech Communication, 2009

[2] Witt, S. M., "Automatic error detection in pronunciation training: Where we are and where we need to go", Proc. IS ADEPT, 2012

[3] Kim, Y., Franco, H., and Neumeyer, L. "Automatic pronunciation scoring of specific phone segments for language instruction", Proc. Eurospeech, 1997

[4] Franco, H., Neumeyer, L., Ramos, M., and Bratt, H., "Automatic detection of phone-level mispronunciation for language learning", Proc. Eurospeech, 1999

[5] Witt, S. M. and Young, S. J., "Phone-level pronunciation scoring and assessment for interactive language learning", Speech Communication, 2000

[6] Strik, H., Truong, K., De Wet, F. and Cucchiarini, C., "Comparing different approaches for automatic pronunciation error detection", Speech Communication, 2009

[7] Amdal, I., Johnsen, M. H., and Versvik, E., "Automatic evaluation of quantity contrast in non-native Norwegian speech", Proc. SLaTE, 2009

[8] Kim, J., Wang, C., Peabody, M., and Seneff, S., "An interactive English pronunciation dictionary for Korean learners", Proc. Interspeech, 2004

[9] Meng, H., Lo, Y. Y., Wang, L. and Lau, W. Y., "Deriving salient learners' mispronunciations from cross-language phonological comparisons", Proc. ASRU, 2007

[10] Lo, W. K., Zhang, S., and Meng, H., "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system", Proc. Interspeech, 2010

[11] Cucchiarini, C., Van den Heuvel, H., Sanders, E., and Strik, H., "Error selection for ASR-based English pronunciation training in "My Pronunciation Coach"", Proc. Interspeech, 2011

[12] Hong, H., Kim, S., and Chung, M., "A corpus-based analysis of Korean segments produced by Japanese learners", Proc. SLaTE, 2013

[13] Wang, Y.-B. and Lee, L.-S. "Toward unsupervised discovery of pronunciation error patterns using universal phoneme posteriorgram for computer-assisted language learning", Proc. ICASSP, 2013

[14] Xiong, H., Wu, J., and Chen, J., "K-means clustering versus validation measures: a data-distribution perspective", IEEE Transactions on Systems, Man, and Cybernetics, 2009

[15] Huttenlocher, D. P. and Zue, V., "A model of lexical access from partial phonetic information", Proc. ICASSP, 1984

[16] Chang, H.-A., "Multi-level acoustic modeling for automatic speech recognition", PhD thesis, Massachusetts Institute of Technology, 2012

[17] Zhu, X. and Ghahramani, Z., "Learning from labeled and unlabeled data with label propagation", Technical Report, Carnegie Mellon University, 2002

[18] Glass, J. R., "A probabilistic framework for segment-based speech recognition", Computer Speech & Language, 2003

[19] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V., "TIMIT acoustic-phonetic continuous speech corpus", Linguistic Data Consortium, 1993

[20] Järvelin, K. and Kekäläinen, J. "IR evaluation methods for retrieving highly relevant documents", Proc. ACM SIGIR, 2000