Improving L1-Specific Phonological Error Diagnosis in Computer Assisted Pronunciation Training

Theban Stanley, Kadri Hacioglu

Rosetta Stone Labs, Boulder, Colorado, USA
tstanley@rosettastone.com, khacioglu@rosettastone.com

Abstract
With the increasing use of technology in classrooms, computer assisted pronunciation training (CAPT) is becoming a vital tool in language learning. In this paper, we present a system that takes advantage of data from learners of a specific L1 to better model phonological errors at various levels in the system. At the lexical level, a statistical machine translation approach is used to model common phonological errors produced by a specific L1 population. At the acoustic level, L1-dependent maximum likelihood (ML) nonnative models and discriminative training are explored. In our experiments, use of a Korean language dependent nonnative lexicon gives us diagnostic abilities that did not exist in our baseline configuration. Replacing the native ML acoustic model with the L1-dependent nonnative model produces relative improvements of 27–37% in precision for phone detection/identification tasks. We also propose a constrained variant of minimum phone error (MPE) training which is better adapted to phone detection/diagnosis. This technique produces 5–6% relative improvement in precision in comparison to ML nonnative acoustic models.

Index Terms: language learning, phonological error modeling, machine translation, minimum phone error training

1. Introduction
As the comfort level of students using technology steadily increases, computer assisted pronunciation training (CAPT) is fast becoming a vital part of a successful language-learning curriculum. With the increasing demand for language tutors in the booming economies of Asia, a well-designed CAPT system can be an effective personal tutor for language learners who want to go through the curriculum at their own pace and schedule. Widely used CAPT systems boast infinite patience and present repetitive drills of predesigned content with limited feedback in the form of scores [1] or playback of the prompted sentence. Learners using these systems are forced to deduce their own feedback strategies by closely observing subtle changes in scores or playing back recordings of voicers who are native speakers. This kind of implicit feedback is especially ineffective when learners have perception problems with certain native sounds.

High demand for language-learning products in Asia has created a justifiable and commercially viable trend of modeling L1-dependent phonological errors in CAPT systems. The most common cause for mispronunciations arises from the fact that the target language has phonetic units that are nonexistent in the learner’s mother tongue. This usually leads to the learner replacing the target with the closest phonetic unit that exists in his/her L1 inventory. Deletion and insertion errors commonly occur due to phonotactic constraints that are projected from their L1. This phenomenon as a whole is referred to as L1 negative transfer effect and its modeling is essential in providing accurate and actionable feedback to the learners. Linguistic experiences with a specific L1 population can be used to model these effects but they are often subjective and biased. Most widely used data-driven techniques in automatic extraction of phonological error rules are based on variants of edit distance algorithms [2]. Most of these algorithms are constrained to one-to-one mapping and are ineffective in capturing phenomena that occur across syllable and phonotactic levels. Also, they lack a straightforward approach in generating nonnative pronunciations as they don’t explicitly model priors on rules or interdependency between rules.

Most CAPT systems rely on generic native acoustic models to provide goodness of pronunciation scores based on the acoustic closeness of the learner’s pronunciation to the canonical [1]. This is partially due to the very limited availability of nonnative data and even scarcer availability of reliable annotations. Even if these systems were presented with pronunciation alternatives, they are not tuned to discern the acoustic differences between them. Also, they don’t account for the mismatch in the acoustic-phonetic feature space between native and nonnative speakers.

Classical speech recognition is based on generative approaches that maximize the probability of an observation given its class label. These techniques don’t learn any explicit discrimination between confusable sounds in the nonnative acoustic space.

This paper addresses these issues that are prevalent in most CAPT systems. At the lexical level, we present a data-driven probabilistic framework that takes a holistic approach to extracting and generating nonnative pronunciations. At the acoustic level, we explore different ways of injecting nonnativeness into the acoustic models and also present a variant of minimum phone error training [3] that optimizes on maximizing the discriminability between confusable phonetic units in the nonnative acoustic space.

2. Modeling phonological errors at the lexical level
We adopt a statistical machine translation (MT) approach in generating likely nonnative phone sequences given a canonical pronunciation. Traditional MT is formulated as the problem of generating the best sequence of words in the target language given a sequence of words in the source language. In this work, we reuse the MT framework and reformulate the problem as finding the best sequence of nonnative pronunciation that represents a good translation of the canonical phone sequence [4]. The Bayesian formulation is as follows:
\[ T^* = \arg \max_T P(S|T) \cdot P(T) \]

where \( S \) is the native phone sequence and \( T \) is the nonnative phone sequence. The \( P(S|T) \) term corresponds to the translation model and \( P(T) \) to the language model of the target language in traditional MT. We use Moses [5], a phrase-based MT framework, for phonological error modeling and the following subsections give a brief overview of the process illustrated in Fig. 1.

### 2.1. Phonological error model

The \( P(S|T) \) term in (1) models the phonological transformations between the canonical and nonnative phone sequence. A parallel phone corpus of native and nonnative phone sequences are aligned using Giza++ toolkit [4] which implements the original IBM MT models. The phone corpus is aligned both ways and the one-to-one mappings are loaded into Moses to grow the high-precision alignment points (obtained by intersection of two-way alignment) using various criteria and expansion heuristics [4]. The resulting phonological error model has phone-chunk pairs of varying length with a translation probability associated with each of them.

### 2.2. Nonnative phone language model

The phone language model represents the \( P(T) \) term in (1) and guides the search during the decoding phase by providing prior knowledge on the likelihood of a certain competing phone sequence for an L1 population. In this paper, we used the IRSTLM toolkit [4] to generate 3-gram nonnative phone models with Witten-Bell smoothing.

### 2.3. Decoder

Given a phonological error model and a phone language model, the Moses decoder can generate N-best nonnative phone sequences for any native pronunciation. At each state, the current cost of the hypothesis is computed by combining the cost at the previous state with the translation cost and the language-model cost for the current state. The Moses decoder implements a beam search which prunes away competing paths that fall below a threshold based on histogram pruning. As the competing hypotheses can be of different lengths a future cost needs to be computed for the pruning to work as desired.

At the conceptual level, the MT framework presents a more principled paradigm to model rule probabilities and interdependencies between rules in comparison to edit distance based techniques that need to rely on heuristic based rule selection and application strategies [2]. At the algorithmic level, the MT approach employs a more complex algorithm in comparison to its edit distance based variants. It has the ability to model one-to-one, one-to-many, and many-to-many mappings, which gives it a wider array of tools to efficiently learn the underlying phenomena that occur in nonnative speech. For example, Korean speakers often substitute “i” for “r” in refrigerator, which is discovered and modeled as one rule using the MT approach, while edit distance based methods usually model them as separate independent rules.

### 3. L1-dependent nonnative acoustic modeling

Most CAPT systems use generic native models and provide confidence scores and goodness of pronunciation scores [1] as these models are more practical when catering to a wider nonnative population. These systems typically have lower performance that arises from mismatch in the acoustic-phonetic feature space between native and nonnative speakers of different L1 populations. As these systems don’t explicitly model errors they don’t have the ability to pinpoint and provide actionable feedback. Even if phonological errors are modeled at the lexical level, the generic acoustic model lacks the ability to discriminate between nonnative pronunciation alternatives. L1-specific systems are less common due to the limited availability of well-balanced and annotated L1-specific corpora.

#### 3.1. L1-dependent maximum likelihood nonnative acoustic model

A straightforward approach to creating L1-dependent acoustic models is to incrementally adapt the native model to the learner’s voice. But this can be a risky venture as it can blur the discriminability of the models once we start adapting to false accepts. In this paper, we initially train maximum likelihood native models using the training recipe illustrated in Fig. 2.

We use the HTK toolkit [6] for training these models. We start by training one-mixture monophone models initialized by a flat-start procedure where all the Gaussians have the same global mean/variance followed by realignment and iterations of embedded reestimation using the forward-backward algorithm discussed in [6]. The resulting one-mixture monophone models are cloned over to create triphone models followed by embedded reestimation. The context dependent triphone models are state tied and reestimated. Finally, the number of mixtures in the model is incrementally increased from 1 to 8 with reestimation at each level of mix-up. We train nonnative models by bootstrapping from the native model as shown in Fig. 2. In this paper, two variants for training nonnative models are explored. The first variant uses unannotated data along with a nonnative lexicon for training. This technique is useful when large amounts of unannotated data are available. We also trained nonnative models using phone annotated data. The performance of these two variants is discussed in section 5.
3.2. L1-dependent minimum phone error acoustic model

Many discriminative training (DT) techniques have been proposed that try to optimize the acoustic model parameters on some goodness of measure like word error rate (WER) [3]. Minimum word error (MWE) training has shown to outperform other DT techniques in large vocabulary continuous speech recognition (LVCSR) [3]. The objective function, which is maximized in MPE [3], is given by

\[ f_{\text{MPE}}(\theta) = \sum_{r=1}^{R} \log \frac{P_\theta(s|O_r)}{P(s)} A(s_r,s_r) \]  

(2)

where \( R \) is the number of training sentences and \( O_r \) is the observation sequence. \( A(s_r,s_r) \) is the raw phone transcription accuracy of the sentence \( s \) measured against the reference \( s_r \), \( P_\theta(s|O_r) \) is the scaled posterior probability of the sentence \( s \) given by

\[ P_\theta(s|O_r) = \frac{p_\theta(O_r|s)^k \cdot P(s)^k}{\sum_{u} p_\theta(O_r|u)^k \cdot P(u)^k} \]  

(3)

The parameter \( k \) is a scaling factor on the acoustic and language model log likelihoods and controls the smoothness of the objective function. HTK implements the Extended Baum-Welch algorithm [6] which makes uses of lattices to compactly represent the correct transcription and all other competing transcriptions for the utterance. While training MPE models for LVCSR, the numerator lattice represents the correct word transcriptions phone marked by the canonical pronunciation of the words. The denominator lattice represents the competing hypotheses approximated by a large vocabulary recognizer along with a weak language model. As this technique uses only canonical pronunciations for the correct word transcriptions, it aims at increasing the discrimination between confusable words with similar canonical pronunciations. CAPT systems usually employ grammar recognition as spoken prompts are already known and our task is more concerned with increasing the discrimination between pronunciation alternatives within words.

We propose a constrained variant of MPE training that is more adapted to grammar recognition, which is widely used in CAPT systems. Fig. 3 illustrates the steps involved in the discriminative training of this model, which makes use of phone annotated data to generate lattices. The numerator word lattice is generated from the word level annotations and is phone marked by the corresponding phone level annotations. The denominator word lattice is generated from the predesigned prompt and phone marked using the pronunciation alternatives in the nonnative lexicon. The generated lattices are fed into the MPE trainer to train MPE acoustic models. In this variant of MPE training, it should be noted that the denominator term in (3) is constrained to competing pronunciation alternatives of the words, and \( s_r \) is the annotated phone sequence and not the canonical pronunciation of the word (as in standard MPE training). As \( s \) approaches to \( s_r \), the phone transcription accuracy, \( A(s,s_r) \), increases, which in turn maximizes the objective function in (2). This optimization is expected to provide better precision and recall for error diagnosis tasks.

4. Corpus

For experimentation, we used the Korean learners of English corpus that was collected at Rosetta Stone (known as RS-KLE). The corpus consisted of prompted speech data from an assortment of content that includes minimal pairs (e.g., light/right), stress minimal pairs (content/content), sentences, short paragraphs, loan words, and words with larger constant clusters. The corpus consists of 111 speakers, of which 15 speakers were inter-annotated and another 15 speakers were intra-annotated. The RS-KLE corpus was phonetically annotated by three human annotators using the International Phonetic Alphabet (IPA). The data were roughly split 80-20 at the speaker level with no speaker overlap between the test and train splits. The train split was used to generate the nonnative lexicon and L1-dependent acoustic models. The test split was used for evaluation of phone detection and phone identification tasks.

5. Experiments

Experiments were conducted to explore different configurations along the lexical and acoustic modeling dimensions.

5.1. Lexicon

For the baseline experiments we used a native lexicon which was transcribed in IPA. The lexicon was adopted from the Carnegie Mellon University (CMU) Pronouncing Dictionary [7] and included vocabulary used in Rosetta Stone products. The lexicon consisted of 116,000 unique words. The training split of the RS-KLE corpus was used to train the nonnative lexicon using Moses. We removed diacritics and mapped certain novel phones to the nearest English phone to be compatible with the phone set used for acoustic modeling. The Moses decoder has the ability to generate N-best lists, and based on empirical observation, we decided on an N-best length of 4, which seemed to strike a good balance between undergeneration and overgeneration of nonnative alternatives [4].

5.2. Acoustic models

For baseline experiments, we used an American English native model (context dependent 8-mixture triphone model) coupled with a native lexicon. The American English native model was trained using TIMIT and voicer data from Rosetta Stone’s products. A version of nonnative model was trained from unannotated nonnative data using the bootstrapping technique illustrated in Fig. 2. In order to measure the effect of reliable annotated data on the system performance, another version of nonnative model was trained using phone annotated data. The best performing ML acoustic model was used to generate phone-marked numerator and denominator lattices, which were then used for MPE training as illustrated in Fig. 3. Except for the
baseline experiments, all other acoustic model configurations used a nonnative lexicon to detect and identify phonological errors in the test data.

The system evaluation was performed on phone detection and phone identification tasks. The system was evaluated on overall accuracy, precision, recall, and F-1 score (harmonic mean of precision and recall).

5.3. Phone error detection

Phone error detection is defined as the task of tagging a phoneme as correctly or incorrectly pronounced. Accuracy measures the overall classification performance of the system while precision and recall measure the diagnostic performance of the system. Precision measures the number of correctly tagged mispronunciations over all the tagged mispronunciations of the system. Recall measures the number of correctly tagged mispronunciations over all the tagged mispronunciations by the annotator.

As shown in Table 1, the addition of the MT-based nonnative lexicon introduces diagnostic capabilities that did not exist in the baseline configuration. MT-based lexicons have been shown to outperform lexicons generated using edit distance-based techniques [4]. Training ML nonnative models using phone annotated data seems to make the system more sensitive to phonological errors, which is evidenced as a 27% relative improvement (RI) in precision and 21% RI in F-1 in comparison to ML native models. Constrained MPE training produced a 6% RI in precision and 3% RI in F-1 in comparison to ML native models. A 7% RI in precision and 4% RI in F-1 are discovered by the system. Similar to phone detection, Table 2 shows that the addition of nonnative pronunciations to the lexicon introduces diagnostic capabilities to the system. A 37% RI in precision and a 31% RI in F-1 using annotation-trained ML nonnative models is a testament to the usefulness of reliable phone labeled data. A 7% RI in precision and a 4% RI in F-1 are achieved with constrained MPE training in comparison to the best performing ML nonnative model. Due to the difficulty of this task in comparison to detection, the numbers are on the lower side. The inter-grader numbers show there is still room for improvement with these systems.

Table 2. Phone identification performance of various system configurations

<table>
<thead>
<tr>
<th>RS-KLE</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML native model + native lexicon</td>
<td>80.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ML native model + nonnative lexicon</td>
<td>76.6</td>
<td>21.3</td>
<td>21.8</td>
<td>21.5</td>
</tr>
<tr>
<td>MPE native model + nonnative lexicon</td>
<td>77.4</td>
<td>22.8</td>
<td>22.3</td>
<td>22.5</td>
</tr>
<tr>
<td>ML nonnative model (from unannotated data) + nonnative lexicon</td>
<td>76.6</td>
<td>22.1</td>
<td>23.5</td>
<td>22.8</td>
</tr>
<tr>
<td>ML nonnative model (from annotated data) + nonnative lexicon</td>
<td>79.6</td>
<td>29.2</td>
<td>27.1</td>
<td>28.1</td>
</tr>
<tr>
<td>MPE nonnative model (from annotated data) + nonnative lexicon</td>
<td>80.4</td>
<td>31.1</td>
<td>27.3</td>
<td>29.1</td>
</tr>
</tbody>
</table>

Inter-grader | 82.1 | 36.1 | 40.4 | 38.2 |

5.4. Phone identification

Phone identification is defined as the task of phone labeling the learner’s spoken utterance. While accuracy measures the overall performance, precision measures the error-diagnostic accuracy of the system. Recall measures the coverage of error rules discovered by the system. Similar to phone detection, Table 2 shows that the addition of nonnative pronunciations to the lexicon introduces diagnostic capabilities to the system. A 37% RI in precision and a 31% RI in F-1 using annotation-trained ML nonnative models is a testament to the usefulness of reliable phone labeled data. A 7% RI in precision and a 4% RI in F-1 are achieved with constrained MPE training in comparison to the best performing ML nonnative model. Due to the difficulty of

6. Conclusions

In this paper, we present a system that injects L1-dependent knowledge from data at various levels of modeling for better detection and diagnosis of phonological errors made by language learners. The system produces significant improvements in precision and F-1 with the addition of nonnative lexicons and discriminatively trained nonnative acoustic models. Tighter coupling between the MT and the MPE trainer by replacing the weak language model in (3) with probabilities from the MT system might be a good candidate for further exploration. The recent reemergence of discriminatively trained systems like neural networks is also a possible direction for further exploration.

7. References