Discriminatively Trained Acoustic Model for Improving Mispronunciation Detection and Diagnosis in Computer Aided Pronunciation Training (CAPT)

Xiaojun Qian[1,2,3], *Frank Soong*[1,2] *and Helen Meng*[1,3]

[1]CUHK MoE-Microsoft Key Laboratory of Human-Centric Computing and Interface Technologies
[2]Speech Group, Microsoft Research Asia, Beijing, China
[3]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong, China

xjqian@se.cuhk.edu.hk, frankkps@microsoft.com, hmmeng@se.cuhk.edu.hk

Abstract

In this study, we propose a discriminative training algorithm to jointly minimize mispronunciation detection errors (i.e., false rejection and false acceptances) and diagnosis errors (i.e., correctly pinpointing mispronunciations but incorrectly stating how they are wrong). An optimization procedure, similar to Minimum Word Error (MWE) discriminative training, is developed to refine the ML-trained HMMs. The errors to be minimized are obtained by comparing transcribed training utterances (including mispronunciations) with Extended Recognition Networks [3] which contain both canonical pronunciations and explicitly modeled mispronunciations. The ERN is compiled by handcrafted rules, or data-driven rules. Several conclusions can be drawn from the experiments: (1) data-driven rules are more effective than hand-crafted ones in capturing mispronunciations; (2) compared with the ML training baseline, discriminative training can reduce false rejections and diagnostic errors, though false acceptances increase slightly due to a small number of false-acceptance samples in the training set.

Index Terms: CAPT, mispronunciation detection and diagnosis, discriminative training, data-driven phonological rule extraction

1. Introduction

The CALL (Computer-Assisted Language Learning) systems for improving second language learners' English have gained wide popularity within the community during the recent two decades. A particular application of CALL, named "Computer-Aided Pronunciation Training" (CAPT), aims at supporting productive training by asking learners to read accordingly to a given prompt, pinpointing pronunciation errors and bringing forth suggestions for improvements. Since this usually involves many rounds of practice, reliable speech recognition technologies can at times serve as a good substitute for human tutors.

Three metrics are commonly used to evaluate the effectiveness of ASR for CAPT: (1) **False Rejections** - the number of words that are recognized as mispronunciations when the actual pronunciations are correct; (2) **False Acceptances** - the number of words that are misclassified as correct yet they are actually mispronounced; (3) **Diagnostic Errors** - the number of erroneous words in the diagnostic feedback for those truly detected mispronunciations, e.g. The learner mispronounces /*n*/ as /*l*/, as in "light" for "night", but the system gives erroneous feedback as "You mispronounced /*n*/ as /*r*/".

There are mainly three problems that prevent ASR technology from being very successful in mispronunciation detection and diagnosis:

1. The classical speech recognition technology based on Bayes' analysis usually lacks a functional form of joint distribution of the observation and the class identity. Even if we know the real data distribution, the ML estimation criterion does not optimize the classification performance due to a different criterion. The case is worsened since non-native speech samples are often limited.

2. Mispronunciation patterns by non-native speakers are often unpredictable. Typical errors include "*insertion*" - addition of one of more sounds to a word, e.g. "poured" may be mispronounced as /*p ao r d ax*/, "*deletion*" - omission of one or more sounds in a word, e.g. "salient" may be mispronounced as /*s ey l y ax n*/, "*substitution*" - replacement of one or more sounds in a word, e.g. "river" may be mispronounced as /*w ih f axr*/, etc. These phenomena are often attribute to the **L1 negative transfer**, and they constitute the sub-problem of mispronunciation modeling.

3. The acoustic-phonetic spaces of native and non-native speakers can be quite different for the same English phoneme under the commonly-adopted ARPABET annotation system. For example, the phone /*b*/ in Mandarin is unvoiced in the syllable "bu", while the phone /*b*/ in the English word "book" is a voiced one, which is non-existent in Mandarin. L2 learners tend to substitute the mother tongue sound for the closest English one. The mismatch may cause confusion in the speech recognition.

This paper addresses the problems mentioned above, and is organized as follows: The second section deals with the formulation of an objective function for optimizing False Acceptance, False Rejection and Diagnostic Errors, using discriminative training techniques. In the third section, a data-driven phonological rule extraction algorithm is discusses. Experimental results and discussion will be given in the fourth section. Conclusions are made in the fifth section.

2. Optimization for Mispronunciation Detection and Error Diagnosis

Discriminative training aims to optimize some measures of goodness-of-recognition based on the training data [1]. As an initial step, our work focuses on segmental errors. Suppose we have R training utterances, indexed by $r = 1, 2, \ldots, R$. For the rth utterance, it contains a sequence of K_r words, indexed by $k = 1, 2, \ldots, K_r$. We denote the canonical pronunciation of the kth word by $w_{rk}(0)$, and all the other possible

word mispronunciations predicted for the kth word by $w_{rk}(n)$, $n = 1, 2, \ldots, N_{rk}$.

The expected number of **False Rejections** is:

$$\mathcal{F}_{FR}(\lambda) = \sum_{r=1}^{R}\sum_{k=1}^{K_r}\sum_{n=1}^{N_{rk}} \mathcal{P}_\lambda(w_{rk}(n)|\mathcal{O}_{rk})^\kappa \delta(w_{rk}(0) = t_{rk}) \tag{1}$$

where λ is the set of HMM parameters, t_{rk} is the phonetic transcription for the kth word in the rth utterance. The posterior probability $\mathcal{P}_\lambda(w_{rk}(n)|\mathcal{O}_{rk})$ is defined as:

$$\mathcal{P}_\lambda(w_{rk}(n)|\mathcal{O}_{rk}) = \frac{\mathcal{L}_\lambda(\mathcal{O}_{rk}|w_{rk}(n))\mathcal{P}(w_{rk}(n)|w_{rk}(0))}{\sum_v \mathcal{L}_\lambda(\mathcal{O}_{rk}|v)\mathcal{P}(v|w_{rk}(0))} \tag{2}$$

where $\mathcal{O}_{rk}$ is the observation for the kth word in the rth utterance, v is any possible mispronunciation given $w_{rk}(0)$, $\mathcal{L}_\lambda(\mathcal{O}_{rk}|v)$ is the acoustic likelihood function of $\mathcal{O}_{rk}$ given v, and $\mathcal{P}(v|w_{rk}(0))$ is essentially a conditional unigram language model for the given canonical transcription. $0 < \kappa \leq 1$ is an exponential scaling factor. $\delta(\mathcal{H})$ is an indicator function that equals 1 if $\mathcal{H}$ is true, or 0 otherwise.

Similarly, the expected number of **False Acceptances** is:

$$\mathcal{F}_{FA}(\lambda) = \sum_{r=1}^{R}\sum_{k=1}^{K_r} \mathcal{P}_\lambda(w_{rk}(0)|\mathcal{O}_{rk})^\kappa \delta(w_{rk}(0) \neq t_{rk}) \tag{3}$$

The expected number of **Diagnostic Errors**, complementary to **Correct Diagnosis** under the **True Rejection** category, is defined as:

$$\mathcal{F}_{DE}(\lambda) = \tag{4}$$
$$\sum_{r=1}^{R}\sum_{k=1}^{K_r}\sum_{n=1}^{N_{rk}} \mathcal{P}_\lambda(w_{rk}(n)|\mathcal{O}_{rk})^\kappa \delta(w_{rk}(0) \neq t_{rk}, w_{rk}(n) \neq t_{rk})$$

The objective is to jointly minimize $\mathcal{F}_{FR}(\lambda)$, $\mathcal{F}_{FA}(\lambda)$ and $\mathcal{F}_{DE}(\lambda)$. Consider the following superposition weighted on the posteriors:

$$\mathcal{F}(\lambda) = \mathcal{F}_{FR}(\lambda) + \mathcal{F}_{FA}(\lambda) + \mathcal{F}_{DE}(\lambda) \tag{5}$$
$$= \sum_{r=1}^{R}\sum_{k=1}^{K_r}\sum_{n=1}^{N_{rk}} \mathcal{P}_\lambda(w_{rk}(n)|\mathcal{O}_{rk})^{\kappa_1} \delta(w_{rk}(0) = t_{rk}, w_{rk}(n) \neq t_{rk})$$
$$+ \sum_{r=1}^{R}\sum_{k=1}^{K_r} \mathcal{P}_\lambda(w_{rk}(0)|\mathcal{O}_{rk})^{\kappa_2} \delta(w_{rk}(0) \neq t_{rk})$$
$$+ \sum_{r=1}^{R}\sum_{k=1}^{K_r}\sum_{n=1}^{N_{rk}} \mathcal{P}_\lambda(w_{rk}(n)|\mathcal{O}_{rk})^{\kappa_3} \delta(w_{rk}(0) \neq t_{rk}, w_{rk}(n) \neq t_{rk})$$

If $\kappa_1 = \kappa_2 = \kappa_3 = 1$, we see that:

$$\mathcal{F}(\lambda) = \sum_{r=1}^{R}\sum_{k=1}^{K_r}\sum_{n=0}^{N_{rk}} \mathcal{P}_\lambda(w_{rk}(n)|\mathcal{O}_{rk})\delta(w_{rk}(n) \neq t_{rk})$$
$$= \sum_{r=1}^{R}\sum_{s} \mathcal{P}'_\lambda(s|\mathcal{O}_r)\,\mathrm{RawWordError}(s, t_r) \tag{6}$$

where $\mathcal{O}_r$ is the observation, t_r is the phonetic transcription, s is the competing K_r-word sentence, all for the rth training utterance, and 'RawWordError' is the number of mismatched words between s and t_r. $\mathcal{P}'_\lambda(s|\mathcal{O}_r)$ is defined as:

$$\mathcal{P}'_\lambda(s|\mathcal{O}_r) = \frac{\mathcal{L}_\lambda(\mathcal{O}_r|s)\mathcal{P}(s|t_r)}{\sum_u \mathcal{L}_\lambda(\mathcal{O}_r|u)\mathcal{P}(u|t_r)} \tag{7}$$

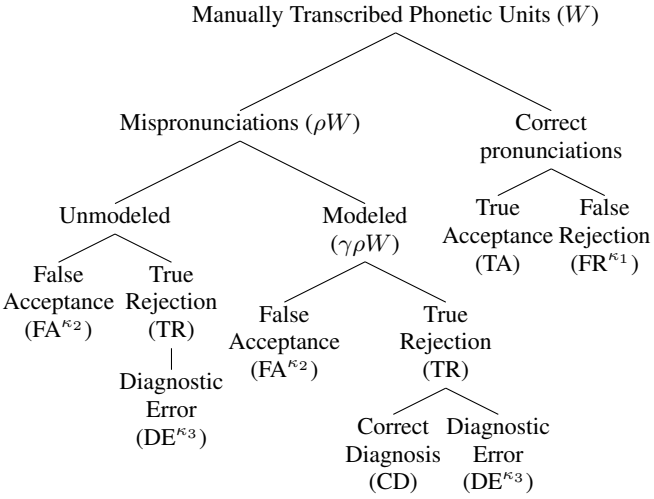

Figure 1: *Hierarchical structure of the mispronunciation detection and diagnosis errors, where W is the total number of units in the training data, ρ is the proportion of mispronounced units in the data, and γ is the proportion of mispronounced units that are modeled.*

where u is any possible utterance mispronunciation given t_r. We see that the minimization of (6) is the same as MWE criterion [1].

The computation of $\mathcal{P}_\lambda(w_{rk}(n)|\mathcal{O}_{rk})$ is guided by a phonetic lattice that contains the canonical phonetic pronunciations of a word, together with its mispronunciations against which we wish to discriminate. The lattice is derived from a set of phonological rules, which are described in section 4.

As it is considered less desirable to reject correct pronunciations, one may exponentially weight the posterior $\mathcal{P}_\lambda(w_{rk}|\mathcal{O}_{rk})$ in $\mathcal{F}_{FR}$ by a factor of $\kappa_1 < 1$ to attach greater importance to FR. By doing so, the cost associate with *False Rejection* can be penalized during $\mathcal{F}(\lambda)$ minimization.

Depending on the fraction of mispronunciations that can be explicitly modeled, the posterior scaling can also be applied to $\mathcal{F}_{FA}$ and $\mathcal{F}_{DE}$ to balance their relative importance within the 'mispronunciation' category, i.e. emphasis or de-emphasis. As illustrated in Figure 1, to further take into account the possibility of skewed distribution of correct pronunciations and mispronunciations in the training data, one can use different scaling factors to balance the contribution from each component in the optimization process.

3. Corpus Preparation

Our investigation is based on the CU-CHLOE corpus [2], which contains recordings of 100 Cantonese-speaking learners of English reading minimal pairs, confusable words, phonemic sentences, and the Aesop's Fable "The North Wind and the Sun". We split the whole corpus into training and testing sets, consisting of $5,988$ and $2,587$ utterances respectively. The speech has been annotated by well-trained linguists with the ARPABET phonetic symbols.

4. Data-driven Phonological Rule Extraction

To model the phonological production process of mispronunciations by Chinese (i.e. Cantonese in this work) learners of En-

glish, our previous work [3] used hand-crafted context-sensitive phonological rules:

$$\phi \rightarrow \psi \ / \ \lambda \_ \ \rho \tag{8}$$

which indicates that, the phone ϕ is replace by ψ under the left and right context λ and ρ, respectively. Insertion and deletion errors can be realized by letting either ψ or ψ be ϵ (the null phone). Also, word boundaries are denoted by $\#$.

As we apply these phonological rules in describing the speech of the second-language (L2) learners, we discover the need for ordered applications. For example, our corpus of non-native speech shows that the word "MYTH" can be mispronounced as */m eh f/* rather than */m ih th/*. To generate this mispronunciation, we need ordered rules like "$ih \rightarrow eh \ \ m \_ \ th$", "$th \rightarrow feh\_\#$" or "$th \rightarrow f \ / \ ih \ \_ \ \#$", "$ih \rightarrow \ / \ m \ \_ \ f$".

4.1. Enhanced phonological rules

We extend the rule format in (8) slightly to allow incorporating multiple phones in "$\phi \rightarrow \psi$", while the left and right context are still described by a single phone. Hence the (word-based) phonological rule for "MYTH" will become "$eh \ th \ \rightarrow \ ih \ f \ / \ m \ \_ \ \#$". One possible shortcoming of this extension is the data sparsity problem, due to an increase of the context captured by the rules. However, this enhanced format of phonological rules offers more expressive power to capture interesting phone patterns.

4.2. Data-driven rule extraction

For a given word w: a canonical phonetic transcription and a mispronunciation transcribed by a linguist, we apply "phonetically-sensitive string alignment" [3] to align two phoneme strings. Different costs are assigned by different phoneme pairs in the alignment. In this way, we obtain a set of phone substitutions, deletions and insertions that maps the canonical pronunciation to the mispronunciation. This is illustrated in Table 1.

Table 1: *Phonetically-sensitive string alignment between the canonical transcription and the mispronunciations for the word "NORTH".*

0	n	ao	r	th
1	l	ao		f
2	n	ah		th

The rule extraction procedure scans through each pair. We match as many contiguous phones as possible. For example, from the pair of phone string in rows 0 and 1 in Table 1, we generate the rules "$n \rightarrow l \ / \ \# \ \_ \ ao$" and "$r \ th \rightarrow f \ / \ ao \ \_ \ \#$"; From the pair of phone string in row 0 and 2, we obtain the rule "$ao \ r \rightarrow ah \ / \ n \ \_ \ th$".

4.3. Pruning rules

The procedure above generates a large number of rules. For example, based on a total of $19,474$ mispronounced word tokens in the training set, we obtain $3,200$ rules. There is a need to prune rules that are overly specific, e.g. "$t \ r \ ae \ v \ el \ axr \rightarrow ch \ aa \ f \ l \ ax \ / \ \# \ \_ \ \#$", which only applies to the word "TRAVELER". Rules can also be generated due to transcription errors in the mispronunciations, or misread words. In any case, a large number of phonological rules can generate an excessive number of mispronunciation variants, some of which may never occur, leading to over-generation. As a first attempt, we prune the rules based on their triggering frequency in the training set (i.e. the number of samples supporting this rule) with an occurrence threshold of 2. We also prune rules with high phonetically-sensitive alignment cost per phone with a threshold of 14.5, which is the average cost of all phone pairs. We model the remaining rules $\{r_i\}$ with Finite State Transducers [4], whose "*Union*" operation takes the form:

$$\mathcal{R} = \bigcup_i (r_i) \tag{9}$$

and compose $\mathcal{R}$ with the canonical word transcription to predict possible mispronunciation variants.

5. Experiments

The testing set contains 436 distinct words and $2,822$ different types of word pronunciations, among them, $2,349$ are mispronunciations that have not appeared in the training set; There are a total of $19,366$ word tokens in the testing set, among them, $8,564$ of them are mispronounced.

5.1. Validating the phonological rules

We reference the 51 handcrafted phonological rule set in the our previous word [3]. These rules consider commonly confused phones [2] with contextual constraints and first language phonotactic constraints [5], such as:

$$r \rightarrow \epsilon \ / \ V \ \_ \tag{10}$$

$$r \rightarrow l \ / \ C \ \_ \tag{11}$$

where V stands for vowels and C stands for consonants. In general, they are found in L2 phonology [2].

We compare with the phonological rules extracted and pruned from the training set, in terms of mispronunciations and the number of word mispronunciations modeled in the testing set. Results are shown in Table 2.

Table 2: *The number of mispronunciations by type and the number of mispronunciations by token that are modeled by the two sets of rules in the testing set.*

Rule set	Number by type	Number by token
hand-crafted rules	411	$3,775$
data-driven rules	652	$5,086$

It should be noted that the hand-crafted rules result in $2,239$ possible types of word mispronunciations out of the 436-word vocabulary of the testing set, while the data-driven rules give $4,348$. Among the $2,349$ types of word mispronunciations that are found in the testing set (based on manual phonetic transcription), the hand-crafted and data-driven phonological rules attain precision and recall values as shown in Table 3.

Table 3: *Precision and recall comparison between the hand-crafted rules and the data-driven rules on the testing set.*

Rule set	Precision	Recall
hand-crafted rules	18.36%	17.97%
data-driven rules	15.00%	27.76%

"*Precision*" is defined as the types of modeled mispronunciations over the types of mispronunciation resulted from the rule set, and "*Recall*" is the types of modeled mispronunciations

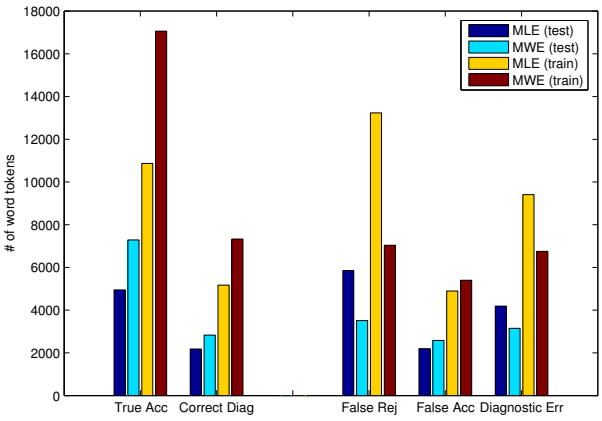

Figure 2: *Recognition performance on both the training and testing set for the data-driven phonological rules.*

over the types of mispronunciations that appear in the testing set. Since our goal is to capture possible word mispronunciations, we tend to emphasize recall over precision. Therefore, we use the data-driven phonological rule set in the acoustic modeling training. It is noted, however, that lower precision implies more complex pronunciations during acoustic Viterbi decoding, possibly leading to more phonetic confusions.

5.2. Recognition results

To address the third problem raised in the introductory section, we train cross-word, tied-state triphone HMMs on TIMIT in Maximum Likelihood, and adapt those with the training set of CU-CHLOE using Constrained Maximum Likelihood Linear Regression [6]. This compensates the mismatch in the acoustic-phonetic feature space between native and non-native speakers. This model is utilized to align the mispronunciation phone lattice. The alignment serves as a baseline in mispronunciation detection and diagnosis. The alignment is also used as a basis for discriminative training.

In discriminative training, the FA, FR and DE are treated with equal weights. The acoustic models are refined under the "Minimum Word Error" criterion, and compared with the acoustic model trained under the ML criterion in Figure 2. In the testing set, the performance gain in terms of the sum of FA, FR and DE has decreased from $12,238$ to $9,244$ out of all the $19,366$ word tokens. This is very promising and confirms the benefits of discriminative training.

In particular, discriminative training gives a significant boost in True Acceptance and Correct Diagnosis, as well as a significant reduction in False Rejection and Diagnostic Error. However, there is a slight increase in False Acceptance.

To be more specific, FR and TA offset each other among the correct pronunciations. Given the fixed number of mispronounced words in the testing set, the increase in CD lowers the sum of FA and DE. However, the minimization of the sum of FA and DE is achieved at the cost of sacrificing FA. This is because FA has relatively small contribution to the objective function due to a smaller number of samples (see Figure 2). On the other hand, the growth of FA errors is caused by a sharper increase of FA in the unmodeled mispronunciations than in the modeled ones (see Figure 3). In other words, for the MWE-trained HMM, the increased FA errors are caused by being unable to model those mispronunciations in the lattice. The problem may be addressed by better modeling to mispronunciations in the lattice.

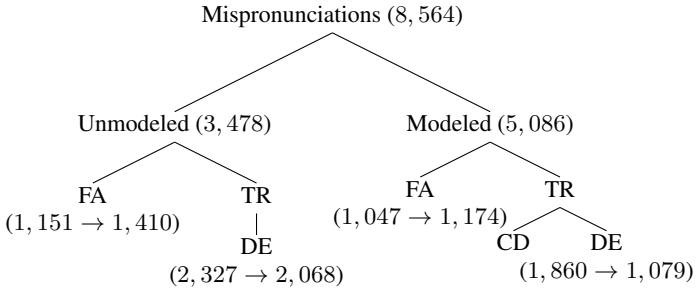

Figure 3: *This is a subtree extracted from Figure 1. It shows a comparison between the number of recognition errors (FA and DE) for mispronounced words on the testing set, before and after MWE training.*

6. Conclusions

Detecting and diagnosing mispronunciations and avoiding false alarms is an important problem in CAPT. We show that the joint optimization (minimization) of the False Rejections, False Acceptances and Diagnostic Errors can help improve the ability of ASR in CAPT. The sum of the three objectives functions are actually equivalent to the MWE criterion in discriminative training. The mispronunciations are better captured by a set of phonological rules derived automatically from data than the knowledge-driven hand-crafted ones. Compared with the ML-trained HMM baseline, the MWE-trained HMM yields promising results in the overall performance in terms of FA, FR and DE. Better mispronunciation modeling is expected to reap greater benefits from discriminative training.

7. Acknowledgments

The authors would like to acknowledge Ms. Alissa M. Harrison and Dr. Wai-Kit Lo at CUHK for providing the necessary resources. We would also like to thank Dr. Zhi-Jie Yan of MSRA for useful discussions and suggestions, as well as the anonymous reviewers for their comments. This project is conducted under the MSRA-CUHK joint laboratory scheme, and is partially supported by the NSFC/RGC Joint Research Scheme (project no. N_CUHK 414/09).

8. References

[1] Povey, D., "Discriminative Training for Large Vocabulary Speech Recognition", Ph.D. thesis, Cambridge University Engineering Dept., 2003.

[2] Meng, H., Lo, Y.Y., Wang, L. and Lau, W.Y., "Deriving salient learners' mispronunciations from cross-language phonological comparisons", in ASRU, 437 - 442, 2007.

[3] Harrison, A.M., Lo, W. K., Qian, X.J. and Meng, H., "Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training", in SIG-SLATE, 2009.

[4] Kaplan, R.M. and Kay, M., "Regular models of phonological rule systems", in Computational Linguistics, vol. 20, no. 3, 1994.

[5] Harrison, A.M., Lau, W.Y., Meng, H. and Wang, L., "Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer", in INTERSPEECH, 2008.

[6] Gales, M.J.F., "Maximum likelihood linear transformations for HMM-based speech recognition", in Computer Speech and Language, 12(2):75-98, 1998.