AUTOMATIC LOCALIZATION AND DIAGNOSIS OF PRONUNCIATION ERRORS FOR SECOND–LANGUAGE LEARNERS OF ENGLISH*

Daniel Herron §[1], Wolfgang Menzel [1], Eric Atwell [2], Roberto Bisiani [6], Fabio Daneluzzi [4], Rachel Morton [5], Juergen A. Schmidt [3]

[1] University of Hamburg [2] University of Leeds [3] Ernst Klett Verlag [4] Dida*El S.r.l.
[5] Entropic Cambridge Research Labs [6] University of Milan–Bicocca

* This research was supported by the European Commission's R&D programme (DG XIII), under its Language Engineering project ISLE LE4–8353.

§ Universität Hamburg, FB Informatik, AB NatS, Vogt–Kölln–Strasse 30, D–22527 Hamburg, Germany or <herron@informatik.uni-hamburg.de>

ABSTRACT

An automatic system for detection of pronunciation errors by adult learners of English is embedded in a language–learning package. Four main features are: (1) a recognizer robust to non–native speech; (2) localization of phone– and word–level errors; (3) diagnosis of what sorts of phone–level errors took place; and (4) a lexical–stress detector. These tools together allow robust, consistent, and specific feedback on pronunciation errors, unlike many previous systems that provide feedback only at a more general level. The diagnosis technique searches for errors expected based on the student's mother tongue and uses a separate bias for each error in order to maintain a particular desired global false alarm rate. Results are presented here for non–native recognition on tasks of differing complexity and for diagnosis, based on a data set of artificial errors, showing that this method can detect many contrasts with a high hit rate and a low false alarm rate.

INTRODUCTION

The Interactive Spoken Language Education [ISLE] project aims at introducing speech recognition technology into future Computer–Assisted Language Learning [CALL] products for adult learners of English. One of the main goals is to provide an appropriate level of specific feedback in order to point out possible ways to improve pronunciation. Existing courseware products that use speech recognition capabilities are often developed without direct input from the end–user—for example, the feedback to the student is often restricted to a global quality measure without specific advice. Other systems (e.g., [5]) provide more specific feedback, but attempt to detect *what* the error was rather than *where* it was.

ISLE improves on this by localizing errors to specific phones and providing clear feedback to the student (e.g., that an error has occurred, and what the student can do to correct this). ISLE aims to create a natural learning environment in which the student is not responsible for self–diagnosis. Besides providing the student with immediate feedback, long–term performance data (at the exercise, word, and phone levels) is collected to allow the student's performance to be tracked across time. (Figure 1 shows an example from the prototype ISLE interface.) This paper focuses on the technical issues associated with the project and puts aside the many important issues associated with *how* this information is to be used.

STRUCTURE OF THE ISLE SYSTEM

The research efforts of the ISLE project are concentrated on development in four main areas: reliable and robust recognition of non–native speech; localization of pronunciation errors; diagnosis of pronunciation errors; and detection of stress–errors.

Recognition

Recognition of non–native speech is handled by Entropic's IHAPI HMM–based recognition software using native British English acoustic models. Exercise types used in the ISLE system will vary in their complexity but will be chosen such that a certain level of word accuracy is achievable. Results from two representative tasks are shown here. A minimal–pair task was carried out in which speakers read confusable pairs of words in a carrier sentence such as *"I said BAD not BED"*. Each choice word was then recognized from a choice of 30 (simulating a relatively complicated exercise). Table 1 shows results using both monophone and word–internal triphone models.

On a more typical exercise such as describing holiday plans, a variety of recognition grammars could be produced. Table 2 shows results for a network of parallel

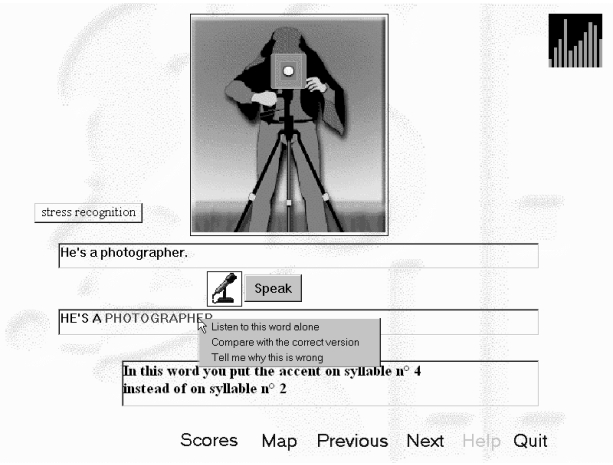

Figure 1: Part of the prototype ISLE user interface

Table 1: Minimal pair recognition accuracy for native UK acoustic models on native and non–native speech

Model set	*Native*	*Non–native*
Monophones	84.2	72.5
Triphones	94.2	75.4

Table 2: Word accuracy on the "holiday" task

Task	*Native*	*Non–native*
Sentence	100.0	98.8
Syntax	99.6	94.8

sentences, and for a more complex grammar with test set perplexity of 2.3. The simpler task approaches 100% non–native accuracy. Other techniques may be investigated to increase accuracy for the more complex tasks.

Localization

In the first–pass "recognition" stage it is critical that the recognizer be tolerant of non–native errors, so that the system can determine the correctness (in terms of truth value) of the student's response. To determine the quality of pronunciation, however, the system will then re–recognize the same utterance in forced–alignment mode using (possibly less tolerant) models. Confidence scores produced by the recognizer are then used to determine possible mispronounced words and phones.

Word–stress error detection

Stress–errors are, regardless of frequency, highly noticeable in foreign–accented speech (see, e.g., [1]). The ISLE system attempts to detect deviations between the expected and the produced stress patterns by comparing the normalized patterns of pitch, energy, and duration over the vocalic regions of each word with known (trained) clusters of stressed and unstressed vowels. Although stress is clearly perceptible to humans, this is a notoriously difficult task for machines, and thus some compensation must be made for the possibility of errors; preliminary tests of the system indicate that it can correctly determine the primary stress of a multi–syllabic word with a word error rate of less than 20%.

Diagnosis of phone–level errors

The major component of the ISLE system is dedicated to detecting and classifying pronunciation errors (as opposed to simply localizing them, which provides a useful but not sufficient degree of information to the student.) The diagnosis method described here relies on the premise that non–native speakers do not, in general, make random mistakes: German learners will make typically "German" mistakes, and so forth ([1]).

First–language–specific diagnosis

To facilitate diagnosis, it is useful to know in advance what errors are expected. Such expected errors can be the product of either phonemic (e.g., difficulty producing a particular phone) or orthographic → phonemic errors (e.g. pronouncing *wilderness* in analogy with *wild* rather than with a short /ɪ/.) The second case requires a system of mapping from orthography to phones, so that expected errors based on the application of incorrect rules can be generated. The first case can be stated more easily as, e.g., "Germans tend to produce a /v/ sound instead of a /w/ sound" and "to devoice word–final stop–consonants" or "Italians have difficulty producing a short /ɪ/ sound" and "tend to insert a schwa at the end of words not ending in a vowel."

The ISLE system detects expected errors by performing an additional (nearly–) forced alignment recognition, and allowing alternative pronunciations of some words. Those alternatives are generated from the errors that might be expected based on the mother tongue. If, in this second recognition pass, an alternative, error–containing pronunciation is recognized by the system (i.e., has a higher acoustic score), the system returns the list of errors.

In designing such a system, several characteristics are desirable. It must (a) very rarely tell the student he has made a mistake when he did not; (b) find enough genuine mistakes to be useful; and (c) not overwhelm him with too much information at once. The third requirement is best dealt with by the user–interface; given a list of errors sorted by severity, it can decide how many the student should be made aware of. The first and second (false alarms [FA] and hit rate, in other words) can be controlled by adjusting the pronunciation probability of the alternative pronunciations. Due to the way in which HMMs are trained, it is likely that performance will differ for different phones, and so it is necessary to tune this bias specifically for each type of error. Some types of errors that the designer might wish to detect could, in fact, turn out to be impossible to detect with a low FA rate and a still–reasonable hit rate.

Non–native corpus for training and testing

In order to determine which types of errors are reliably detectable (with a low FA and high hit rate), it is necessary to have a corpus of non–native speech annotated at the word and phone level. The judgments of the human annotators can then be compared with those of the machine system, in order both to train (determine optimal biases and eliminate errors impossible to detect well) and to test the system. The ISLE project has collected a medium–sized (50 speakers) corpus of English speech from non–native (German and Italian) intermediate–level adult learners of English. It will be annotated at the word and phone level by an HMM recognizer operating in near–forced–alignment mode, and then deviations from that annotation will be noted by trained phoneticians.

The time and cost of collecting and annotating even a modest corpus can quickly become overwhelming. Unfortunately, it is clearly difficult both to train and test in any reasonable manner without a far larger amount of data. In addition, before the annotation is completed, it is desirable to have an approximate indication of the performance of the diagnostic component. A partial solution is the use of "artificial" data [3], in which errors are introduced into the pronunciation dictionary in a systematic way, allowing one to test on a relatively large, native data set the ability of the system to detect errors.

DIAGNOSIS OF ARTIFICIAL ERRORS

The experiments described below used the SCRIBE [3] corpus of native English speech; the original recognition dictionary was systematically altered to include errors that are the "opposite" of those that non–native speakers are expected to make. E.g., a German speaker might be expected to say /v/ instead of /w/; thus instances of /w/ were changed in the pronunciation dictionary to /v/ (but the /w/ in *was* remains, and forms the basis for possible FAs). The changes to the dictionary for two example words might be as in Table 3.

Table 3: Original and altered pronunciation

	Pronunciation	
Word	*Original*	*Altered*
was	w ax z	w ax z
very	v eh r iy	**w** eh r iy

When diagnosing these words for the error /w/→/v/, for either word the system can detect or *not* detect the error. Depending on whether the altered pronunciation induced a corresponding error or not, this decision is classified as one of four possible results as in Table 4.

Table 4: Interpretation of results

Altered	*Decision*	
word	*/w/→/v/ error*	*no error*
was	FA	Correct rejection
very	Hit	Miss

The tests reported here used a set of nine rules, all involving the substitution of one phone for another. Three examples of three general classes, detailed in Table 5, were used; they are roughly consistent with some of the errors that German and/or Italian learners of English might make. Tests were performed using the third volume of the SCRIBE corpus, containing 2000 sentences, from 10 speakers with a "South East" British accent, using 2695 unique words. The sentences have a mean length of 9.86 words (and a standard deviation of 3.19).

Artificial errors were then introduced into the recognition dictionary. For every word, each contextually–correct occurrence of the nine "incorrect" phones in Table 5 was changed to the corresponding "correct" phone. Over the 2695 words there was a mean of 0.75

Table 5: Errors introduced into the dictionary

Type	*Phone*		*Frequency per word of*	
	correct	*wrong*	*rule applied*	*rule not applied*
Vowel	/ao/	/ow/	0.08 (0.27)	0.07 (0.25)
	/ay/	/eh/	0.15 (0.37)	0.09 (0.29)
	/ih/	/iy/	0.20 (0.43)	0.35 (0.58)
Pre–vocalic consonant	/dh/	/z/	0.03 (0.18)	0.01 (0.11)
	/th/	/s/	0.10 (0.31)	0.02 (0.12)
	/w/	/v/	0.05 (0.21)	0.07 (0.25)
Word–final devoicing	/d/	/t/	0.10 (0.31)	0.11 (0.31)
	/g/	/k/	0.03 (0.17)	0.01 (0.09)
	/b/	/p/	0.02 (0.13)	0.01 (0.08)

the 2695 words there was a mean of 0.75 changes per word (and standard deviation of 0.82). To facilitate analysis, the 158 words in the corpus that have more than one common pronunciation were not altered. The rules were applied with different frequencies; the last two columns in Table 5 show the mean frequency (and standard deviation) with which each rule *was* applied and with which it was *not* applied—i.e., 10% of words had a word–final /t/→/d/ error applied, and 11% had a word–final /d/ in the original pronunciation (which should not be diagnosed as a /t/→/d/ error).

Experiment 1: Word–internal triphone models

Each sentence was then recognized in forced–alignment mode (although with multiple pronunciations) using word–internal triphone models. After recognition, diagnosis was performed using the same model set, to search for any instances of the nine errors. All possible mispronunciations were considered in parallel, so if the word had, e.g., a /w/ and an /ih/, three alternative pronunciations, plus the original one, were considered. Each error–bearing pronunciation was temporarily added to the recognition dictionary as an alternative for that word.

In order to control the hit/FA ratio, biases were assigned to each pronunciation, to make it relatively more or less probable. The bias of the original pronunciation was 1.0, and the bias of each alternative was the product of the biases of each rule that has been applied. In this case, because the biases were not independently altered,

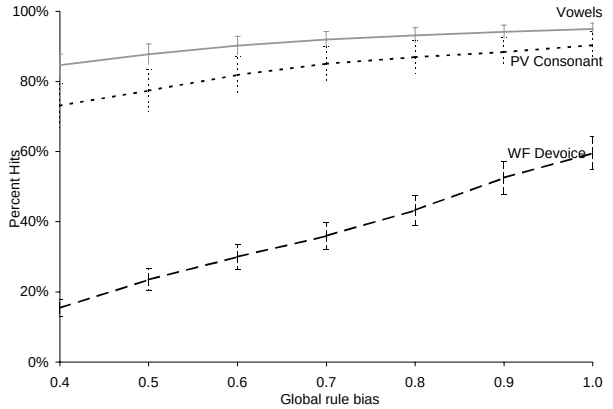

Figure 2: Average hit rate for three classes of phone substitutions, using word internal triphone models.

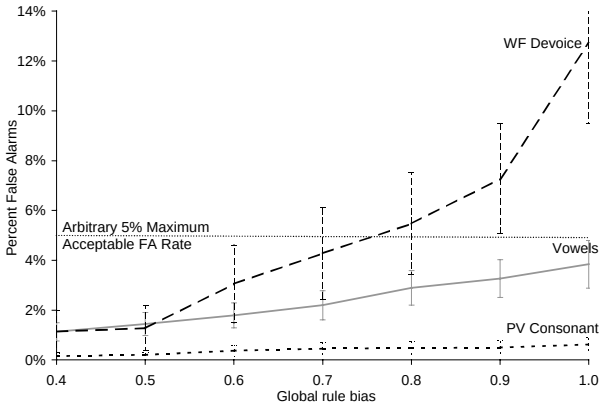

Figure 3: Average FA rate for three classes of phone substitutions, using word internal triphone models.

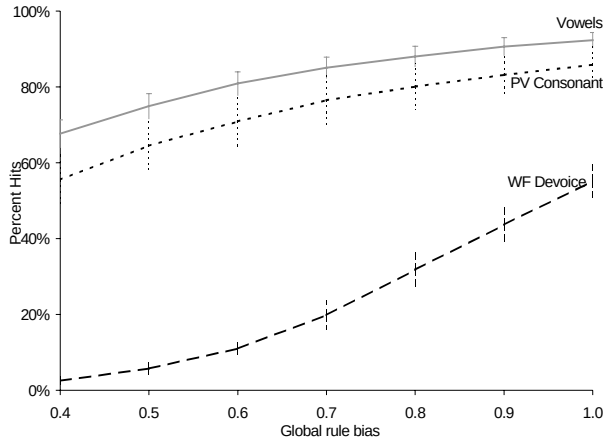

Figure 4: Average hit rate for three classes of phone substitutions, using monophone models.

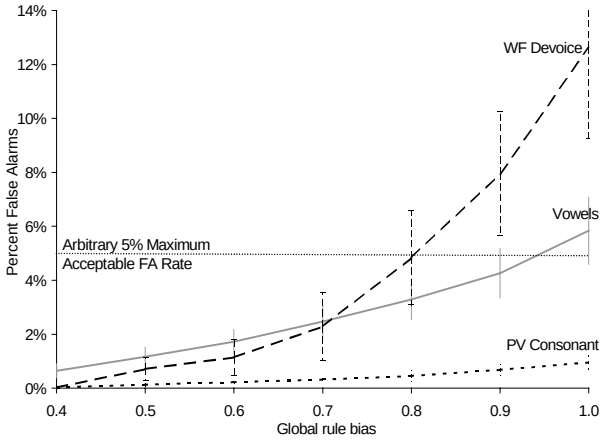

Figure 5: Average FA rate for three classes of phone substitutions, using monophone models.

the bias of each pronunciation was equal to n^r, where n is the rule–bias (ranging from 0.4 to 1.0 at intervals of 0.1) and r is the number of rules applied. Thus pronunciations with multiple errors have a relatively lower probability of recognition than the canonical pronunciation (for $n \neq 1$). It is infeasible to independently adjust the rule biases for a large set of rules, yet this simplified technique should provide an approximation of the proper biases for each rule in order to maintain a particular maximum global FA rate.

The mean hit rate, shown in Figure 2, was computed across the 10 speakers and then averaged across error type (vowel, pre–vocalic consonant, or word–final devoicing). The error bars show the standard error across speakers (which has been averaged within each error–type). Figure 3 shows the FA rate, which has an arbitrary ceiling of 5%—thus the bias that results in the highest hit rate that has a FA rate below 5% is our target.

The diagnosis is in general successful; most importantly, it is not difficult to maintain our target 5% FA rate. Nevertheless, it is clear that certain contrasts (the vowels and the pre–vocalic consonants) are relatively easy to detect, with a high hit rate and a low false alarm rate, while the word–final devoicing contrasts were poorly diagnosed. (It should also be pointed out that the rather large standard errors of the FA rate for the word–final devoicing stem mostly from the /g/→/k/ and /b/→/p/ contrasts, which are both quite infrequent, as noted in Table 5.

Experiment two: Monophone models

Because of their higher contextual specificity, it is assumed that triphone models would be more successful at detecting such errors than monophone models. In order to test this assumption, the same experiment was conducted using monophone models. The results (see Figure 4 and Figure 5) are roughly similar to the triphone models, with only slightly lower performance overall.

CONCLUSIONS

The most important goals for the ISLE system are robust recognition of non–native speech and a low FA rate—no student will use a system that fails to recognize him or that provides discouraging and detrimental feedback. Results show that recognition of non–native speech will still be possible when using moderately complex exercises, which allows the creators of the user interface to design challenging and interesting tasks.

The diagnostic results, although using artificial data, indicate that it should be possible to guarantee a particular global maximum FA rate while still detecting many true errors. It is, of course, impossible to predict actual performance until evaluations can be performed on the human–annotated non–native data, which may have very different characteristics than this artificial data. Nevertheless, certain substitution errors appear to be very easily detectable. Even the word–final devoicing errors, with a usable hit–rate of less than 50%, can *sometimes* be detected. Given that there may be in general too many errors detected in a given sentence, such seemingly poor performance may still be adequate. Various modifications may also increase this rate, e.g., the use of cross–word triphones.

REFERENCES

[1] J. Dalby, D. Kewley–Port, and R. Sillings. Language–specific pronunciation training using the HearSay system. *Proc. Speech Technology in Language Learning 1998*, Marholmen, Sweden, May 1998.

[2] M. Eskenazi and S. Hansma. The Fluency pronunciation trainer. *Proc. Speech Technology in Language Learning 1998*, Marholmen, Sweden, May 1998.

[3] Speech Research Unit, Defence Evaluation and Research Agency. Spoken Corpus Recordings In British English (SCRIBE). Malvern, UK, May 1992.

[4] S. M. Witt and S. J. Young. Language learning based on non–native speech recognition. *In Eurospeech '97*, Rhodes, Greece, Sept. 1997.

[5] S. M. Witt and S. J. Young. Performance measures for phone–level pronunciation teaching in CALL. *Proc. Speech Technology in Language Learning 1998*, Marholmen, Sweden, May 1998.