Automatic detection and correction of non-native English pronunciations †

W. Menzel §1, D. Herron 2, P. Bonaventura 3, R. Morton 2

[1] University of Hamburg [2] Microsoft Ltd., Cambridge, UK [3] Conversa, Redmond, WA

† 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 <menzel@nats.informatik.uni-hamburg.de>

Abstract

Improving the feedback quality of a computer-based system for pronunciation training requires rather detailed and precise knowledge about the place and the nature of actual mispronunciations in a student's utterance. To be able to provide this kind of information, components for the automatic localization and correction of pronunciation errors have been developed. This work was part of a project aimed at integrating state-of-the-art speech recognition technology into a pronunciation training environment for adult, intermediate level learners. Although the technologies described here are in principle valid for any language pairs, the current system focuses on Italian and German learners of English.

1. Introduction

Computer-based solutions for pronunciation training are becoming increasingly commonplace for foreign language learning purposes (Dalby, Kewley-Port, and Sillings, 1998; Herron, Menzel, Atwell, Bisiani, Daneluzzi, Morton, and Schmidt, 1999; Eskenazi and Hansma, 1998; Witt and Young, 1997, 1998). Nevertheless, currently available solutions offer considerable room for improvement, especially with respect to the feedback generated by the system. In many cases a simple playback facility, the visual presentation of the signal form, a scoring mechanism, or even the identification of the mispronounced word or words might not be sufficient to give students helpful hints on *how to improve* their solutions. Only by finding out the precise nature of the student's mistake is a system in a position to provide detailed error explanations, problem-specific speech stimuli, or individualized suggestions for improvement. Furthermore, such a low-level diagnosis makes it possible to guide the student into specifically tailored opportunities for practice.

In order to demonstrate the potential of contemporary speech technology, components for automatic localization and diagnosis of pronunciation errors have been developed and integrated into a multimedia-based training environment. Although the system is targeted at intermediate Italian and German learners of English, the underlying solutions are mostly generic and could therefore be ported to arbitrary pairs of languages.

The system offers the student a wide range of communicatively relevant exercises (such as question answering) where the answer needs to be constructed from small sets of pre-specified building blocks. The incoming utterance from the student is first recognized by a low perplexity HMM-based speech recognizer (Morton, Whitehouse, and Ollason, 1999) and the result checked for plausibility with respect to the given task. Then, given a successfully recognized utterance, the system tries to find those areas in the signal that are likely to contain pronunciation errors. This so-called *localization* of errors is carried out using the confidence scores provided by the HMM-based recognizer for each word and phone. These confidence scores are computed via Gaussian classifiers that use predictors for broad-phonetic classes to compare the acoustic likelihoods of the recognized phones vs the likelihoods of the most likely state in the model and of the background model. Phones (or words) can be classified as correct or incorrect (localized) by comparing their confidence scores to a threshold.

Diagnosis proper follows as a second stage. It tries to discover how each localized word varied from the canonical pronunciation. A set of mother-tongue specific letter-to-phone and phone-to-phone rules is applied to the target word (in its orthographic and phonemic forms, respectively) in order to produce candidate mispronunciations. Each of these is temporarily added to the recognition dictionary, and the recognizer is then re-run over the same speech signal in forced-alignment mode (at the word level). If the recognizer selects one of the mispronunciations, the student is given specific feedback on what the mistake was (e.g., "you inserted a sound like '***H**ouse*' before that word", to indicate the insertion of an /h/ sound.).

An additional component was developed for the purpose of detecting lexical-stress errors (i.e., putting the stress on the incorrect syllable of a polysyllabic word.) Although this component was included in the demonstration system, there was insufficient data to test its performance, and it will not be discussed in this paper.

This paper gives an overview of the diagnostic components and the quality of results they produce. Section 2 describes the manner in which localization, diagnosis, and feedback were carried out, and Section 3 presents results of offline testing with a corpus of non-native speech.

2. Implementation

The solutions chosen here do not necessarily represent the theoretically best choices. Rather, they are a compromise between the various factors involved in producing a real system: time (both development time and actual system speed); the need to use certain pre-existing components (e.g., a particular recognizer); and demands placed on the system by the users (the students and teachers.) Thus improvement is possible in many areas, yet the system as a whole functions as planned, and provides a solid demonstration of the possibilities of current technologies and current theories on system-student interaction.

2.1 Localization of errors

Localization is the process of identifying the areas of an utterance that are likely to contain pronunciation errors. This is necessary for two reasons. The first is one of perception; because diagnosis may take a long time, and because the student may often wish to know where errors were made but not how, it is desirable to have the system (relatively) quickly pinpoint errors. The second reason, which is more important for performance, is that diagnosis is a difficult task that is much simplified if the speech signal has been pre-sorted into correct and incorrect regions.

A pronunciation error can be defined as some deviation from a target or model pronunciation. Non-native speech may contain many kinds of pronunciation errors. These may vary in their origin and their degree of deviation from the target, and be more or less serious in the extent to which they hinder communication. For example some errors may merely signal a marked non-native accent while the speech remains intelligible. It would be infeasible and possibly damaging to provide feedback for every possible deviation, as almost every student would quickly become discouraged, even if the system never made a false-alarm type error. Thus a sensible alternative is to select the most severe errors.

Such selection strategies are performed in the system by an error localization component that assigns and sorts confidence scores for each speech segment of interest, which might be a phone, word, or entire utterance. In simple terms, low confidence scores represent increased certainty that the utterance was mispronounced and high confidence scores represent certainty that the utterance was correctly pronounced. Therefore areas of high confidence can be filtered out, leaving segments that have a high probability of containing an error. These segments can be ranked in order of ascending confidence and passed to the diagnosis module, in order that the most serious errors can be attended to first (the confidence can, additionally, be used to determine how much effort the diagnosis makes to decide what, if any, error was made).

Confidence scores are computed based on three likelihood values taken for each frame of speech in the input signal: (1) the acoustic likelihood of the recognized path; (2) the output probability of the most likely ("best") state in the model set; and (3) the acoustic likelihood of the background model. It is assumed that the best state scores provide the reference ceiling and the background model score provides the floor and that the acoustic likelihood should thus lie somewhere in this range. The distance between the acoustic likelihood and the best state defines how close the hypothesized path is to what the person actually said, or how they said it. A Gaussian classifier using those three predictors was trained over a set of predictor values for the correct data and for the incorrect data. For each test observation, the likelihoods that the observation vector came from the error distribution and from the correct distribution were calculated; the ratio of these two likelihoods was the basis for the Gaussian confidence score. Separate classifiers were trained for different classes of phones (e.g., vowels, fricatives, liquids, etc.) in order to increase the modeling accuracy.

2.2 Diagnosis of errors

A novel feature of the system, in comparison to similar existing products, is its ability to detect mispronunciations at a very detailed level. For example, instead of merely providing a score for an entire phrase or word, the system detects and provides feedback on errors on individual phones or groups of phones. A more global score may serve only to reinforce what the student already knows—that he or she had difficulty with a particular word—without proving the specific information necessary to improve pronunciation.

2.2.1 Sources and types of mispronunciations

Based on general and expert knowledge, as well as data collected by the ISLE project from potential users (teachers and students), it was assumed that intermediate level, non-native learners of English will make various types of errors, all of which are manifest as 'pronunciation' errors. These types can be generally described as:

- articulatory difficulties producing particular sounds or clusters of sounds (e.g., the notoriously difficult /th/ sound in English)
- receptive difficulties, because of which the student is unable to perceive and therefore to reliably produce

the distinction between two sounds (e.g., /ih/ and /iy/ for Italian speakers)

- orthographic carry-over from the mother tongue; because so much of language use and learning is written, peculiarities of the student's native orthographic system may interfere with pronunciation of English (e.g., the sequence "IE" is pronounced as /iy/ in German, but can be pronounced in many ways in English.)
- orthographic difficulties of English; because of the high degree of ambiguity in mapping written to spoken English, the student may be expected to mis-apply or mis-generalize 'rules' of English pronunciation.

Many of the sorts of errors expected could be attributed to more than one of these causes. Nevertheless, it is important to note that not all errors stem from the same origin.

2.2.2 How errors are detected

The ideal diagnosis system would, most likely, search for and be able to detect any error. The system is constrained to operate with the phone as the smallest unit, however, so no errors occurring at a level beneath the phone can be independently detected. Furthermore, because in this system only the 41 English phones are used, only insertions, deletions and substitutions of those 41 phones can be detected. In principle, one could find any error that can be described by those 41 phones by implementing a phone-loop recognizer (in order to find the true sequence of phones spoken by the student), yet such an approach is computationally infeasible and unreliable. By constraining the diagnosis to a small subset of the theoretically possible errors, however, a relatively fast and—it is hoped—reliable diagnosis can be performed.

Because in the system the student's mother tongue is known, and because there are presumed to be regularities in the way that those speakers make mistakes, it is sensible to try to predict what errors a student might make based on his or her mother tongue. As described above, such errors can have multiple sources; for all of them, however, it seems possible to describe via rule how a word or pronunciation might be altered. Even a very complete set of rules will generate a candidate set that is small in comparison to what the open phone-loop is effectively using, making this a computationally attractive constraint for a system that must operate in real-time.

Two types of error rules are implemented in the system: letter-to-phone rules, which are designed mainly to capture orthographic errors, and phone-to-phone rules, which attempt to model articulatory or receptive difficulties. The rules have the form:

```
<CONTEXT> <LETTERS> <CONTEXT> → <PHONES>
```

for the letter-to-phone mappings, such that one or more letters are mapped to zero or more phones (zero indicating a silent letter or cluster.) The phone-to-phone mappings have the form:

```
<CONTEXT> <PHONES> <CONTEXT> → <PHONES>
```

in which zero or more phones are mapped onto zero or more phones (zeroes indicating insertion and deletion, respectively). In both cases, the left and right contexts are optional sequences of one or more letters or phones, respectively, and can refer to and extend beyond word boundaries (to model word-initial, word-final, and coarticulatory effects).

2.2.3 Generation of candidate mispronunciations

The first step in producing candidate mispronunciations for a given word is to generate a mapping between the letters in the word and the phones in the correct pronunciation. This is necessary because feedback to the student will always be via the orthographic level, so even if an error is due to a phone-to-phone rule, it must be localized and described at the orthographic level. For example, the word "hotel", pronounced /hh oh t eh l/, is mapped as:

```
H     O     T     E     L
hh    ow    t     eh    l
```

using a set of letter-to-phones rules designed to generate the correct pronunciation for a large number of English words. (These rules do not generate only the unique correct pronunciation, but because both the target letter sequence and phone sequence is known, this is not necessary in this task.) The two types of error rules are then applied to the canonical mapping, in a sequential fashion (letter-to-phone rules followed by phone-to-phone rules.)

The letter-to-phone rules apply in a left-to-right fashion across the letters in the word, as each of the canonical rules is allowed to be substituted by a letter-to-phone rule from the student's mother tongue. The replacement error rule must match both the letter or letters as well as the context, if specified. For example, a student might be expected, based on his mother tongue, to not pronounce orthographic 'H' sounds, resulting in:

```
H     O     T     E     L
      ow    t     eh    l
```

After all possible errors generable with letter-to-phone rules have been produced, the entire set of mappings (i.e., the correct mapping and all of the newly-created error mappings) is subject to the phone-to-phone rules. Phone-to-phone rules map zero or more phones to another set of zero or more phones (zero indicating insertion and deletion, respectively), again respecting a possible phone-based context on the left and/or right.

For example, a word-final consonant may create an intrusive schwa:

H	O	T	E	L
hh	ow	t	eh	l ax

The result is a set of pronunciations, one of which is the correct (canonical) pronunciation, and the rest of which are variant mispronunciations created by the application of one or more error rules (either or both letter- or phone-to-phone). Because the size of this set grows greatly as the number of errors per pronunciation increases, the actual demonstration system was limited to one error per word. The offline results described below allowed up to three errors per word, because speed was not a consideration. It should be pointed out that less than 1% of the errors made by the target speakers contain more than three errors per word (most contain just one.)

2.2.4 *Testing of candidate mispronunciations*

Each localized word generates a set of candidate mispronunciations, and these candidates are added to the recognition dictionary as temporary pronunciations for the respective word. The recognizer is then re-run on the speech signal in forced-alignment (word level) mode, and the recognized pronunciation of each word is examined. If it is one of the generated errors, a record is kept of which rules contributed to the mispronunciation. If the recognized pronunciation was:

H	O	T	E	L
	ow	t	eh	l ax

the student has made two errors: a deletion of the initial /h/ sound, and an insertion of a schwa after the final consonant.

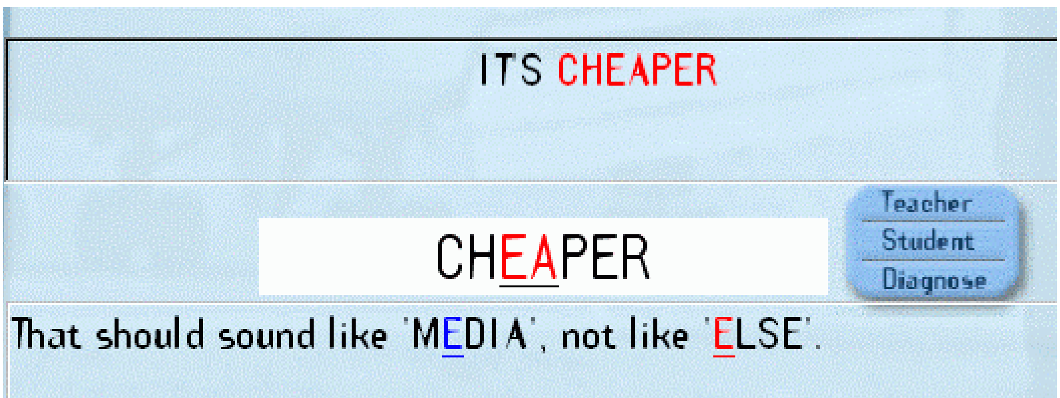
2.3 Feedback for errors

In designing the system, three obvious alternatives presented themselves for how to refer to particular sounds and/or errors (when communicating with the student, that is.) Because the system is designed around the IHAPI HMM-based recognizer (Morton, et al., 1999), which in this case uses a set of 41 phones to represent English, it is easiest to refer to these same symbols. The computational advantages here are more than offset, however, by the disadvantage that the IHAPI phones are neither a well-known standard not particularly easy for the student to understand without training. A better alternative is the IPA system. A many-to-one mapping exists between IPA and IHAPI phones, making it possible to easily convert the IHAPI symbols to IPA symbols before presenting them to the student.

The most difficult, yet perhaps most easily understood system, is to restrict all references to the orthographic level. This has the obvious advantage that the student is without doubt familiar and comfortable with the orthographic level, even in English. It is not trivial, however, to always refer to letters or sequences of letters, when the underlying information is about a sound or a series of sounds.

Nevertheless, it was decided that the familiarity to the student was of the highest importance; thus the system provides feedback to the user by highlighting certain regions, to indicate the locus of the error. In order to describe the error (rather than simply locate it), example words are shown to the student. These example words are high-frequency and of unambiguous pronunciation. Again, one or more letters in the example words are highlighted in order to explain to the student what sounds were substituted, inserted, or deleted. Figure 1 shows an example of feedback from the demonstration system in which the student has mispronounced the word *cheaper* (by substituting the vowel /iy/ with /eh/). The location of the error in the word is indicated to the student by coloring red the letters 'EA', and the student is told that 'that should sound like *media*, not like *else*'. The 'e' is colored blue in the word *media* to demonstrate the correct vowel, and the 'e' in *else* is red, to indicate the sound the student (incorrectly) made. Furthermore, the student can click on either word to hear it

Figure 1: a screenshot from the demonstration system, showing the highlighting and explanation of errors (the error-containing word is located in the white box, and the example correct and incorrect words in the phrase beneath)



pronounced or on the highlighted ‘e’s in order to hear and compare the two vowels.

This type of feedback becomes more complicated with more extreme errors (e.g., substituting one sound with two different sounds), so in some cases the system ‘drops back’ to simply telling the student where the error was, or what the correct sound was, without providing an example of what sound the student produced.

3. Results

Off-line testing of the system was achieved by comparing the diagnostic results of the system to those of the trained linguists who annotated a corpus of non-native German and Italian intermediate-level non-native English (Menzel, Atwell, Bonaventura, Herron, Howarth, Morton, and Souter, 2000). This corpus contains approximately 20 minutes of speech from each of 23 German and 23 Italian speakers. The speech of each subject was annotated by one of six trained annotators at the word, phone, and stress level. Phones, in particular, were scored as either correct, changed (a different phone was substituted), deleted, or inserted. For the purposes of the results described below, the annotators were restricted to using only the 41 native phones of UK English (i.e., those used by the recognizer), because the diagnosis system would in any case be unable to indicate, for example, a substitution of a correct English vowel with an incorrect German vowel.

Because there were multiple annotators operating independently, there was an imperfect agreement them. The degree of disagreement can be measured, in order to calculate both how likely any two given annotators were to find the same error (essentially, an inter-annotator hit-rate), and how likely one annotator was to ‘incorrectly’ find an error that was not found by another annotator (who is considered as the ‘perfect’ annotator). By performing all such pair-wise comparisons between annotators, and averaging the results, a rough expectation can be drawn for upper limits on performance.

3.1 Localization

The classifiers used for the localization module were trained in order to maximize the annotator-relative hit rate while maintaining a constant relative false alarm rate. The false alarm, for example, was fixed at 12.5% of the number of errors; thus, if the corpus contained 5000 phones, of which 1000 were annotated as errors, the system was trained to allow 125 false alarms across those 5000 phones. The annotator-relative hit rate measures the localization hit-rate (the number of phones that the human annotators indicated contained errors) relative to the inter-annotator hit-rate. Thus, if the average inter-annotator hit-rate was 80% (meaning that any two annotators were likely to agree on four out of five errors) and the localization hit-rate was 60% (meaning that 60% of the phones localized as containing errors were marked as errors by the annotators), the annotator-relative hit rate is 75%.

Table 1: results from offline testing of localization with Italian and German annotated data

Language	**Hit rate relative to annotators**	**Accuracy**	
		absolute	**relative to maximum**
Italian	97	44	82
German	67	39	92

A secondary measure used to express localization performance is accuracy, which is defined as:

$$100 * Hits / (Hits + FAs)$$

where *hits* and *FAs* are either the absolute or the relative numbers of hits and false alarms (to produce an absolute or a relative measure of accuracy, respectively.) The relative accuracy is useful because it can be compared across conditions (i.e., even though the Italians and Germans made different numbers of errors, the relative accuracy of localization can be compared). The absolute accuracy, however, is more reflective of the ‘feel’ of the real system (because correct phones are far more likely than incorrect phones, the absolute false alarm rate tends to rise more quickly than the absolute hit rate.) This accuracy can itself be compared to the theoretical maximum possible accuracy, which is computed as:

$$100 * errors / (errors + (dFR * correct))$$

where *errors* and *correct* are the number of incorrect and correct phones in the corpus, respectively, and *dFR* is the desired FA rate.

In comparison to the target human performance, the selected Italian classifiers had a 97% hit rate, and resulted in an absolute accuracy of 44%, or 82% of the maximum possible. The absolute accuracy of the German classifier was 39%, which is 92% of the maximum possible. The German hit rate is lower than the Italian hit rate, reaching 67% of the target hit rate. These results are based on a fixed threshold for false alarms of 12.5% of the training error percentage. If fewer errors are expected in the data, the confidence thresholds can be decreased.

3.2 Diagnosis

For each phrase in the corpus, it is possible to compare the results of human annotation with the system’s diagnosis, on a phone-by-phone level in order to determine how well the system is able to find and

Table 2: classification of diagnosis results

Diagnosis system		**Human annotator’s decision**	
		Error	*No error*
Error	*(same)*	Full hit	False alarm [FA]
	(different)	Near hit	
No error		Miss	Correct acceptance

explain non-native errors. Results were classified into one of five categories, as shown in Table 2; only the FA and hit (full or near) rates are reported below. (Near hits are cases in which the system and annotator agreed that an error had occurred, but not on what the specific error was.)

Figure 2 shows the performance of the system, broken down by phone class. Values shown on the y-axis represent the percentage of phones that were either annotated as having or not having an error, for the hits and false alarms, respectively (in other words, the hit and FA rates are shown in relative terms, normalizing for the prior probability of errors occurring.) The overall mean results are shown on the far right. The inset pie-chart shows the relative size of each phone-class.

Several findings are immediately apparent. Performance is seemingly poor; roughly 25% of the errors are found by the system, and slightly more than 5% of correct phones are incorrectly rejected. It is also clear that performance varies across the phone classes; nevertheless, although the hit rate is, for example, much greater for the vowels than for the liquids, the number of false alarms rises at least as dramatically. Performance for fricatives and semi-vowels is better, yet those classes are very small, and thus account for few of the hits in absolute terms. In fact, when viewed in absolute terms (taking into consideration that the prior probability of an error is far lower—even for the worst of the speakers—than is the prior probability of a correct phone), the results appear far worse, as the FA rate exceeds the hit rate. The absolute FA and hit rate is, of course, what the student will perceive.

Such performance is clearly troubling, as it indicates that students will more frequently be given erroneous, discouraging feedback than they will be given helpful diagnoses. A more fair test of the system, however, can be performed by examining the subset of the corpus that was annotated by all of the annotators, using the effective hit and FA rates for each annotator (relative to the others.) These rates, for each annotator, and for the mean of the annotators, are compared to the rates for the system in Figure 3 (the percentages plotted on the y-axes are relative). The results are certainly better for the human annotators, yet it appears to be a difference in quantity rather than quality. The FA rate of the system is slightly above the mean of the human annotators, yet well within the same range. The hit rate is worse than

Figure 2: relative performance of phone diagnosis for different classes of phones (percentage of phone types shown in inset pie chart)

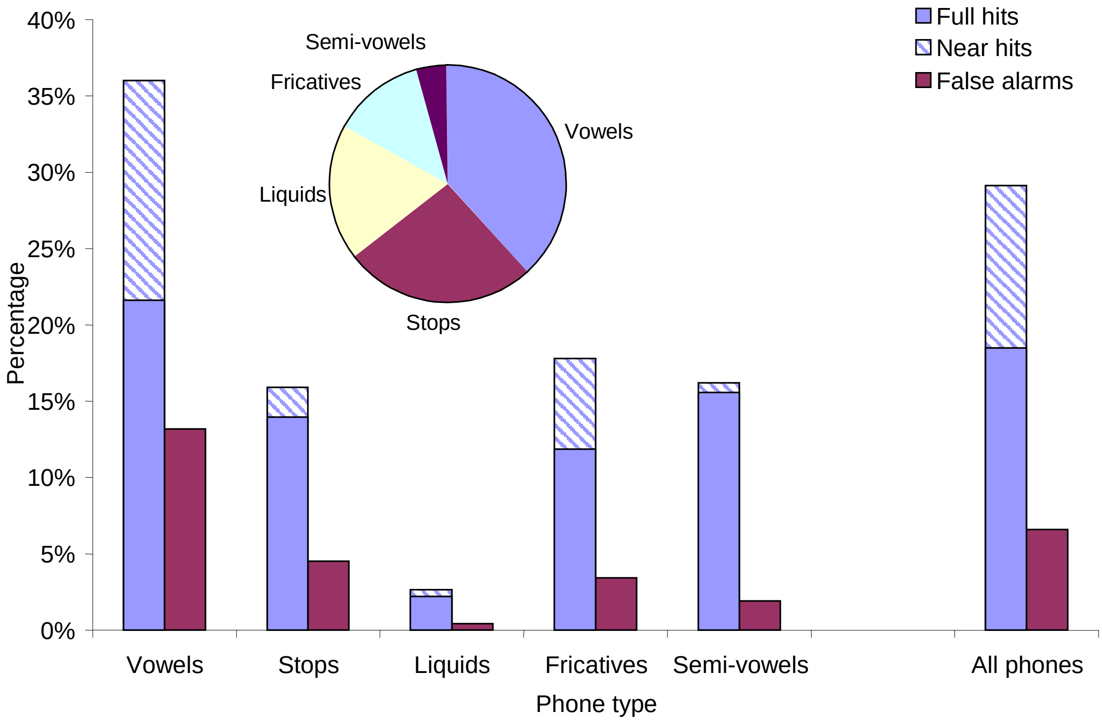
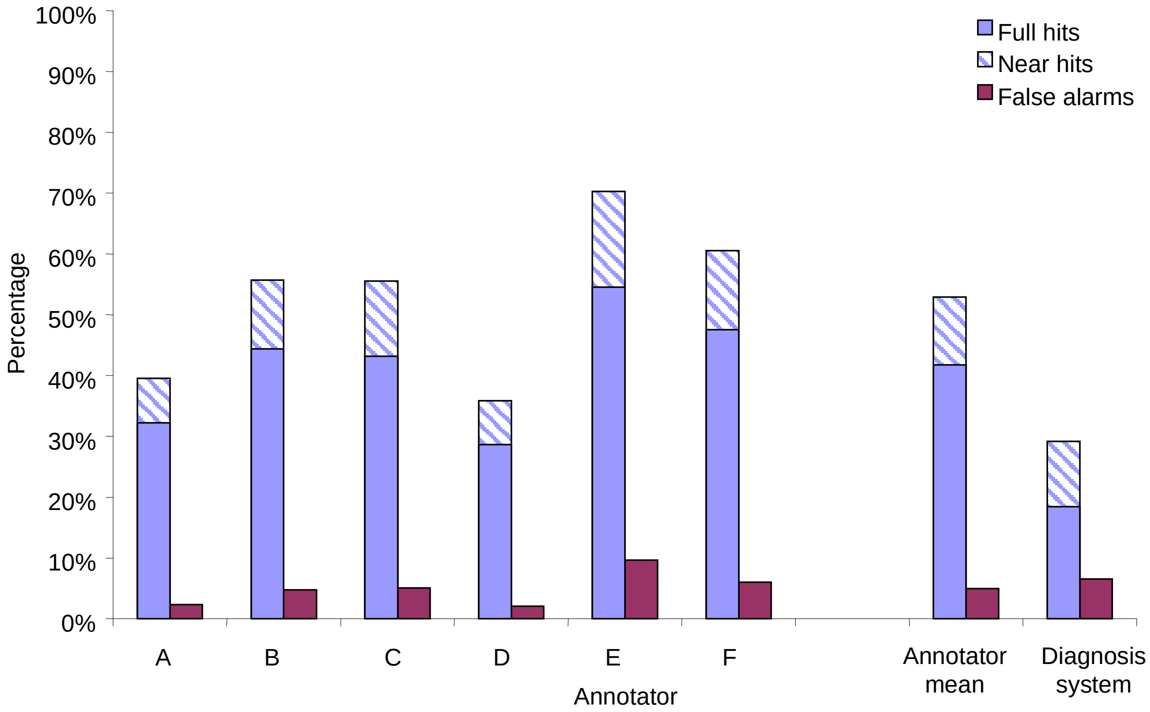


Figure 3: performance of human annotators vs. the system (relative percentages). A through F are the six linguists who annotated the corpus.



any of the humans, but even so is not far worse than that of the 'worse' human annotator.

Due to the fact that large corpora of non-native speech are not available at the moment, all material present in the corpus has been used to model the non-native phenomena, in order to provide as wide a coverage as possible; for this reason, independent testing of the rules has not been possible at the moment. Furthermore, a more fine grained evaluation might be necessary, which subcategorizes the mispronunciations in the corpus according to how seriously they may disrupt the intelligibility of the spoken utterance.

4. Conclusions

Two components for phone error localization and diagnosis have been presented that are able to identify mispronunciations in spoken utterances produced by learners of a foreign language. They have been successfully integrated into a demonstration system for foreign language learning[1]. Off-line results show that the diagnostic results are nearly competitive with human annotations. Furthermore, in a supervised (on-line) evaluation the diagnostic precision was judged by 13 teachers of English as a second language. When asked to rate the feedback quality on a 5 point scale from very accurate (1) to very inaccurate (5), they assigned a mean of 2.9; six teachers classified it as accurate (2). Apparently, then, actual users are either especially forgiving of the system, perhaps due to an bias towards believing what they are told be a computer, or else they truly felt the quality of instruction was broadly similar to that which they receive from human language teachers. It is also clear that even the false alarms produced by the diagnosis system are potentially useful to intermediate students; because they are rule based, searching for errors that such students tend to make, it is in likely that a relatively high number of false alarms is tolerable in practical (if not theoretical) settings.

Due to its diagnostic capabilities the system is able to offer the student a variety of different feedback options, among them detailed error explanations, model pronunciations in the same or in other contexts and specifically tailored follow-up exercises, all intended to highlight the contrast between the students solution and the target, as well as to reinforce the desired articulatory behavior. This is a clear advance over other currently available solutions, which appear to be much more limited in this respect.

The approach can be improved in several directions. Most promising seems the introduction of a rule-scoring scheme into the rule-based diagnosis, which would allow the system to rank the mispronunciation hypotheses according to their relevance. Initial experiments with a simple biasing mechanism (Herron et al. 1999) have shown a considerable potential for

[1] The demonstration system is downloadable from `http://nats-www.informatik.uni-hamburg.de/~isle/`.

improving the error detection accuracy of the approach. Such a mechanism would allow one to place more emphasis on finding errors that are more likely or that have a greater impact on communication, resulting in a lower false alarm rate and higher hit rate, or in achieving a higher hit-rate on particularly troubling errors.

5. References

Herron, D., Menzel W., Atwell E., Bisiani R., Daneluzzi F., Morton R., Schmidt J. A. (1999) “Automatic localization and diagnosis of pronunciation errors for second-language learners of English”, Eurospeech 99, Budapest, 5-9 September, v. 2, p.855-858.

Menzel, W., E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton and C. Souter “The ISLE corpus of non-native spoken English” (2000). In Proc. LREC2000 (2nd International Conference on Language Resources and Evaluation), Athens, Greece, 31 May-2 June 2000.

Morton, R., Whitehouse, D., Ollason, D. (1999) “The HAPI book”. Entropic Cambridge Research Laboratory, Ltd.

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.

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

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

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.