# Phonetic rules for diagnosis of pronunciation errors

P. Bonaventura (Conversa, Redmond , WA), D. Herron (Microsoft, Cambridge, UK) and W. Menzel (Universität Hamburg, Germany)

## Abstract

The paper describes the use of phonetic transcription rules in a component for diagnosis of pronunciation errors produced by students of a foreign language. The module is part of a computer-aided language learning tool designed to report pronunciation problems by Italian and German non-native speakers of English (Herron et al., 1999). It generates a number of possible pronunciation candidates that are tested against the spoken utterance by means of a HMM-based speech recognizer. The most plausible hypothesis is then taken to derive different kinds of feedback information for the student, including detailed error descriptions, analogous example cases, and specifically tailored follow-up exercises.

## Introduction

Rule-based systems for automatic phonetic transcription have generally been used to generate standard transcriptions for speech synthesis and recognition purposes (e.g., Menzel, 1984; Bonaventura et al., 1998; Albano et al. 1999). In such an application the primary goal consists in finding the most appropriate phonetic string given the orthographic input and perhaps taking into account some properties of the surrounding syntactic context. If, however, a transcription procedure is meant to be used as part of a diagnosis component in a computer-based pronunciation training system, it no longer needs to determine only the single best transcription. Instead, a fairly rich set of pronunciations is required, which can then be checked for plausibility against the available acoustic data. Based on the assumption that students produce their mispronunciations in a generally non-random fashion, pronunciation rules are expected to model the underlying regular patterns of typical non-native spoken utterances. In many cases these regularities refer to either the phonetic or phonological characteristics of the student's mother tongue (L1), while sometimes an influence of the target language (L2) can be identified as well.

Generating mispronunciation hypotheses by means of transcription rules allows one to precisely tailor the space of candidates under consideration. If compared, for example, to the results of an unconstrained phone-loop recognizer, a considerable reduction in general recognition uncertainty can be achieved. Furthermore, rules can be ranked according to error relevance or impact, thus providing additional cues for the selection between competing error hypotheses.

Of course, in order to be able to derive mispronunciation hypotheses at all, the utterance actually spoken by the student needs to be known to the system. Therefore, the diagnostic component has been embedded into a more general context using a low perplexity HMM-based speech recognizer to determine the most likely input string (Morton et al., 1999). Moreover, error localization techniques based on special-purpose scoring schemes have been used to further narrow down the space of hypotheses by focusing the diagnosis on selected parts of the speech signal with particularly low articulatory quality.

## Types of mispronunciations

In order to generate diagnosis candidates, rules are required to represent the most important mistakes made by non-natives, in other words both the most frequent errors and those that would most heavily affect comprehension. The mistakes accounted for by the rules were selected a) on the basis of general phonetic knowledge (Kohler 1977; Hakkarainen 1995; Wieden and Nemser 1991; Moulton, 1963; Brown, 1991); b) from previous non-native speech recognition experiments (Bonaventura et al., 1998); and c) from an English speech corpus produced by non-native German and Italian intermediate-level adult learners of English (Menzel et al., 2000). Mispronunciations due to extra-linguistic factors, such as stuttering, hesitation, or phone inversion were not accounted for by the rules, because they depend on the occasional performance of single speakers and were not likely to be systematically repeated.

In general, three different kinds of problem areas can be distinguished (see Table 1 for the approximate percentage of errors in our corpus accounted for by each rule type):

1. Problems in the pronunciations of non-native sounds,
2. carry-over of pronunciation regularities from the mother tongue (L1), and
3. overgeneralizations of target language (L2) regularities.

Problems in the pronunciation of non-native sounds occur if at the competence level the student is aware of the proper pronunciation in the language to learn (L2), but at the performance level articulatory constraints cause poor performance in the non-native language. Such mistakes would consist in the substitution of phonemes that do not exist in the student's mother tongue (L1) with similar phones in the native language; for example, interdental fricatives [θ] and [ð] are realized as the homorganic stops [t] and [d] or as the fricatives [s] and [z], which are present in German and in Italian. Phenomena typical of Italian mispronunciations are the following:

- short reduced English vowels are approximated using the most similar vowels in L1 (e.g., [ʊ] is realized as [u:], as in *put*)
- diphthongs are split into two single vowels: [ɔɪ] realized as [o i:], as in *boiler*
- consonant clusters not undergoing the native phonotactic constraints are simplified (e.g., [k s θ] becomes [k θ], as in *sixth*)

Phenomena typical of German mispronunciations are the following:

- [w] is pronounced [v], as in *with*
- palatal affricate [dʒ] realized as the unvoiced correspondent [tʃ], as in *age*
- diphthongs realized as one long single vowel (e.g., [əʊ] is pronounced as [oː], as in *toast*, or [ɪə] as [iː], as in *hearing*)

A carry-over of pronunciation regularities from L1 to L2 takes place mostly at the phonological level, where, for example, Germans devoice English final voiced stop consonants (e.g. *banned*: Brit. pron. ['b æ n d], Germ. pron. ['b æ n t]) thus exhibiting a typical L1 behavior. A similar carry-over can be caused by the spelling-to-sound mapping of the mother tongue. Thus, Italians tend to not pronounce orthographic 'h', or to read digraph 'ng' (when final or before a non-velar sound) as two separate phones [n g] (e.g. *sing*), as suggested by their mother tongue's orthography.

Finally, overgeneralization of L2 regularities occur whenever the speaker chooses a possible pronunciation but in an inappropriate context; e.g. *anger* Brit. pron. ['æ ŋ g ə] non-nat. Ger. pron. ['eɪ n dʒ ə], where the regularity which changes /g/ into /dʒ/ when followed by a front vowel, as in *George*, is overgeneralized. A slightly higher percentage of carry-over mistakes than errors due only to mechanical-articulatory problems is noticeable.

Table 1: Types of English pronunciation errors produced by Italian and German non-native speakers

| **Mother-tongue** | **Type of error** | | |
|---|---|---|---|
| | *Non-native sounds* | *Carry-over* | *Overgeneralization* |
| *German* | 36% | 41% | 22% |
| *Italian* | 39% | 51% | 9% |

Errors due to coarticulation have not been included in these statistics, because coarticulation phenomena were considered as normal effects, not mispronunciations, in continuous speech. Such effects, found across word boundaries, amount to 16.8% and to 4.2%, over 200 German and 200 Italian pronunciations, respectively. The coarticulation effects at word boundary consisted mainly of deletions or assimilations of the initial/final consonant or vowel, due to the influence of the adjacent phones in the neighboring words.

# Transcription rules

Three types of rules were designed in order to cover the mistakes described above:

1) Letter-to-phone rules to account for L2 pronunciations (i.e., to produce the canonical pronunciations.) An example of these rules accounts for the pronunciation of the digraph 'ie' as [iː] (e.g. *briefcase*: Brit. pron. ['b r i: f k eɪ s]):

   ie → [iː]

2) Letter-to-phone rules for L1 that capture the carry-over of the L1 spelling-to-sound mapping into L2, as well as the mistakes due to the overgeneralization of letter-to-phone English regularities. An example of the former rule type is the German rule:

   sch → [ʃ]

   which might describe German-like pronunciation of the English word *scheme* (Brit. pron. ['s k i: m]) as ['ʃ i: m]. An example of the latter type is the rule:

   i → [aɪ] / <CONS>___<CONS>

   which might describe Italian pronunciation of the English *river* (Brit. pron. ['r ɪ v ə]) as ['r aɪ v ə].

3) Phone-to-phone rules that catch problems in the pronunciation of non-native sounds and the overgeneralization of English phonological regularities. An instance of these rules is the following:

    0 → [ə] / <CONS> ____ #

    which might describe insertion of a final schwa in the Italian pronunciation of the English words ending with a consonant, for example in *bad*: Brit. pron. ['b æ d] non-nat. Ital. ['b æ d ə], thus reproducing the most frequent Italian final syllable structure CV.

Rule sets (b) and (c) include also some coarticulation rules to account for phonetic phenomena at word boundaries. An instance of these rules is the following:

{r,n} → 0 / ___ #<STOP>

which deletes a final [r] or [n] before an initial stop, for example, in *imagine conflicts*: Ital. pron. [ɪ 'm eɪ dʒ ɪ # 'k ɒ _n f l ɪ k t s], which simplifies consonant clusters too difficult to pronounce.

Critical words (mainly articles, conjunctions, and pronouns) were accounted for by a small number of ad hoc rules that diagnosed only mistakes on consonants; in fact, vowels were subject to overly complex and idiosyncratic changes, due to the context and to the pronunciation variations by the speakers. Certain words (e.g. *the* or *I*) were likely to be subjected to many small differences that would be too difficult to detect reliably (although the difficult fricative [θ] in *the* was diagnosed.)

## Generating phone error hypotheses

First, the standard transcription for the given word in the target language is obtained by the letter-to-phone rules to account for L2 pronunciations: such rules account for regular English words and for common exceptions. Frequently used loan words and rare foreign words were not accounted for by these rules. The correct letter-to-phone mapping is required because some of the expected errors will result from generating errors by altering one or more of the L2 letter-to-phone rules (e.g., the use of an L1 rule.) Each letter-to-phone rule generates zero or more phones for a sequence of one or more letters (zero phones indicating a 'silent' letter or letters).

Generation of candidate mispronunciations takes place via the application of the L1 letter-to-phone and phone-to-phone rules, which are designed to model effects of orthographic interference and articulatory or perceptual difficulties, respectively. The application of these two types of rules can be summarized as follows:

*letter-to-phone rules*: each letter in the word is accounted for by exactly one L2 letter-to-phone rule in the canonical pronunciation derived above. In a left-to-right process across the letters in the word, each of those canonical rules is allowed to be substituted by a matching L1 letter-to-phone rule. For example, a typical German error can be accounted for by substituting the canonical L2 'w'→/w/ mapping with the L1 'w'→/v/ mapping. A rule matches when (a) the letter (or letters) mapped is identical with the letter (or letters) at the current position in the word and (b) the context matches; contexts can include all of the letters to the left/right of the current position, including the word which precedes or follows the current word. The result of the letter-to-phone error generation is a set of zero or more alternative mappings (e.g., one mapping is the canonical one, and one has the 'w'→/v/ error). Each mapping is itself a set of rules, and each rule is either the canonical L2 rule for that letter (or those letters) or else a matching L1 rule.

*phone-to-phone rules*: after all possible incorrect pronunciations are generated via the language-specific letter-to-phone error rules, the phone-to-phone rules are applied to the entire list of candidates (the standard pronunciation and the alternatives produced by letter-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). Phone-to-phone rules are also contextual, taking into account the phones to the left/right of the current position, including the phones of the previous or following word. Phone-to-phone rules are not allowed to operate either on a phone that was itself generated via the letter-to-phone rules or a phone that formed part of the context of another phone-to-phone rule that was applied in the pronunciation.

Several post-hoc constraints are in effect: (1) no mispronunciations are considered that are valid pronunciations for that word in the pronunciation dictionary of the recognizer; (2) every pronunciation must have at least one vowel; and (3) in the actual system, only one error at a time is diagnosed per word, in order to prevent overwhelming the student with too many different corrections (an error is defined as either an altered letter-to-phone rule or an applied phone-to-phone rule.) The third constraint was only applied in the demonstration system, in order to allow it to run in real-time; the off-line results described below allowed up to three errors per word. This difference did not change the results dramatically, nor did the limitation to one error lessen the utility of the actual system.

Table 2: rules needed for the generation of all errors in the non-native corpus

| **Language** | **Rule type** | |
|---|---|---|
| | *Letter-to-phone* | *Phone-to-phone* |
| *German* | 202 | 74 |
| *Italian* | 360 | 50 |

Table 3: classification of diagnosis results

| **ISLE System's decision** | | **Human annotator's decision** | |
|---|---|---|---|
| | | *Error* | *No error* |
| *Error* | *(same)* | Full hit | False alarm [FA] |
| | *(different)* | Near hit | |
| *No error* | | Miss | Correct acceptance |

# Rule coverage

In order to account for around 2000 mistakes by Italians and Germans, extracted from the non-native speech corpus, 686 rules were needed (see Table 2 for a breakdown by language and rule type). This seemingly large number of rules was necessary because only a small percentage of the mispronunciations by non-native speakers occurred regularly and systematically. The majority of the mistakes were individual realizations, typical of single speakers, often instantiated only once. This diversity in mistakes was due to the fact that, for the purpose of the application, few restrictions were placed on the level of non-native English collected in the corpus, which was used for both training and testing of the recognizer. In fact, such restrictions would have biased the recognition and altered the performance of the tool, which could then only have been used by upper-intermediate level students of English. The disparity in level of competence by the speakers resulted in many different mistakes, mostly not representable by very general rules.

The rules reflect this typology of mispronunciations: 35% of the rules cover the most frequent and general errors and those that most heavily affect correct perception of the English words, whereas 65% catch less frequent mistakes, in other words, either overgeneralization of L2 regularities (see above) or very specific errors occurring either in rare words or deriving from realizations typical to single speakers (e.g. *graduate* : unique Germ. pron. ['g r æ d ɜː eɪ t]). In an attempt to offset this, the rules were roughly weighted on the basis of the frequency and importance of the mistake they intended to capture, and this weighting was used as a bias that made the system more or less likely to 'find' that particular error (assuring that the most common errors were more commonly and reliably found).

The global rule coverage reflects the distribution of the pronunciation errors: 67% of the rules cover problems in the pronunciation of non-native sounds and carry-over of pronunciation regularities from L1 to L2 (see above), whereas 33% of the rules cover overgeneralization of L2 regularities. Rule percentages do not include coarticulation rules, in order to mirror error percentages reported in Table 1: coarticulation rules amount to 5,6% of the total number of rules.

# Diagnostic results

For each phrase in the corpus, it is possible to compare the results of human annotation (which was carried out by a team of six trained linguists, who marked errors at the word, phone, and stress level) with the system's diagnosis, on a phone-by-phone level in order to determine how well the system is able to find and explain non-native errors. Results were classified into one of five categories, as shown in Table 3; only the FA and hit (full or near) rates are

Figure 1: relative performance of the ISLE phone diagnosis for different classes of phones (percentage of phone types shown in inset pie chart)

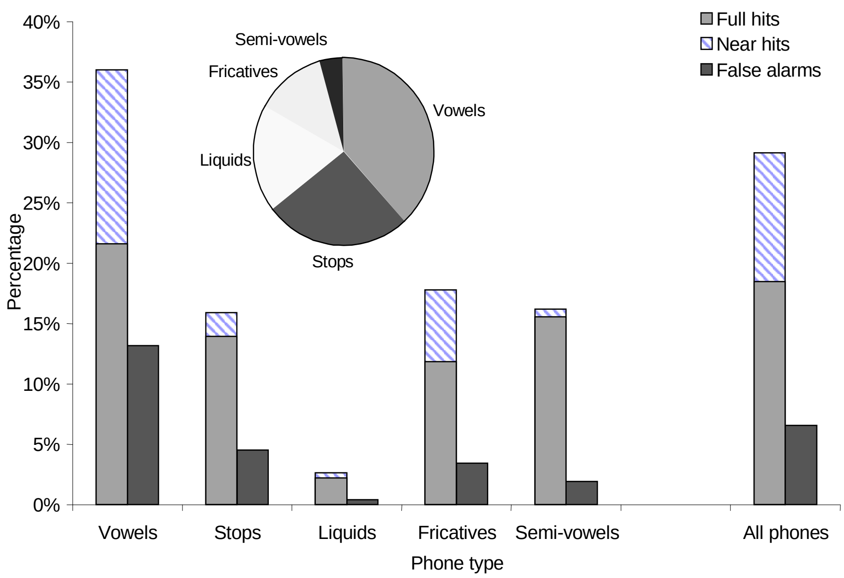
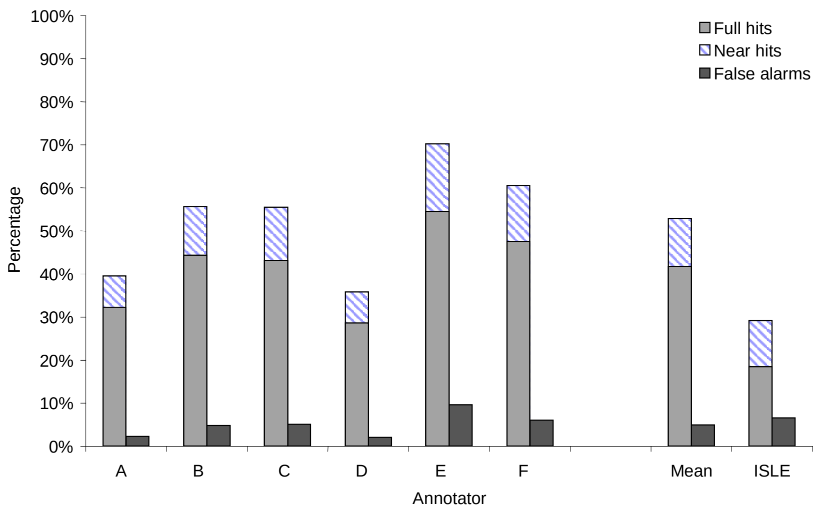


Figure 2: performance of human annotators vs. the ISLE system (relative percentages). A through F are the six linguists who annotated the corpus.




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 1 shows the performance of the ISLE 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 ISLE diagnostic 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.

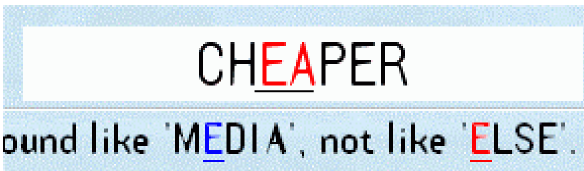
Such performance is clearly troubling, as it indicates that students will more frequently be given erroneous, discouraging feedback than they will be given helpful diagnoses. A more fair test of the system, however, can be performed by examining the subset of the corpus that was annotated by all of the annotators. From that subset, one can calculate 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 ISLE system in Figure 2 (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 ISLE system is slightly above the mean of the human annotators, yet well within the same range. The hit rate is worse than any of the humans, but even so is not exceedingly low.

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.

## Generation of feedback

Because the mapping between letters and phones was at all times preserved, the error-containing orthographic region can be highlighted, in order that the student is easily able to understand the location and nature of the error. In addition, the correct and incorrect sequences of phones (or only one of the two in the cases of deletion or insertion) are described via

Figure 3: a screenshot from the ISLE demonstrator, 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)



example words. For example, Figure 3 shows a screenshot from the ISLE demonstrator, in which the student has mispronounced the word *cheaper* (by substituting the vowel [i:] with [ε]). 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 pronounced, or on the highlighted 'E's in order to hear and compare the two vowels.

# Conclusions

A rule-based transcription component has been devised that generates pronunciation variants to be used as diagnostic hypotheses in a language learning system for pronunciation training; this component has been integrated into a demonstration system[1]. The diagnostic results are nearly competitive with human annotations. In a supervised 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).

Due to the phone error diagnosis 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.

Besides the necessity to optimize and evaluate the system on a considerably extended corpus, the approach can be improved in several directions. Most promising seems the introduction of a rule scoring scheme, 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 show a considerable potential for improving the error detection accuracy of the approach. Such a mechanism would allow one to consider error hypotheses generated by the rules describing the more likely mistakes with priority for the purpose of diagnosis.

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


# Acknowledgements

This research has been supported by the European Commission under the 4th Framework of the Telematics Application Programme (Language Engineering Project LE4 8353).


# References


Albano E. C., Moreira A., Silva A., Aquino P. and Kakinohana R. (1999) Um Conversor Ortográfico-Fônico e uma Notação Prosódica Mínima para Síntese de Fala em Língua Portuguesa. In: E. Scarpa (ed.) Estudos de prosódia, Editora da Unicamp, Campinas.

Bonaventura, P., G. Micca and F. Gallocchio (1998) Speech recognition methods for non-native pronunciation variations, ESCA Workshop, Rolduc, 4-6 May 1998, pp. 17-23.

Bonaventura P., Giuliani F., Garrido J. M. and Ortín I. (1998) Grapheme-to-phoneme transcription rules for Spanish, with application to automatic speech recognition and synthesis, COLING '98 Workshop, Montreal, 16 August 1998, pp. 33-39.

Brown, A. (1991) Teaching English Pronunciation, Routledge, London.

Hakkarainen, H. J. (1995) Phonetik des Deutschen, Wilhelm Fink Verlag, München.

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, Proc. 6th European Conference on Speech Communication and Technology, Eurospeech 99, Budapest, pp. 855-858.

Kohler, K. J. (1977) Einführung in die Phonetik des Deutschen, Erich Schmidt Verlag, Berlin.

Menzel, W. (1984) A Grapheme-to-phoneme transformation for German, Computers and Artificial Intelligence, vol. 3, no. 3, p. 223-234.

Menzel, W., E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton and C. Souter (2000) The ISLE corpus of non-native spoken English. Proc. 2nd International Conference on Language Resources and Evaluation, LREC 2000, Athens, Greece, p. 957-963.

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

Moulton, W. G. (1963) The sounds of English and German, The University of Chicago Press, Chicago.

Wieden, W. and Nemser, W. (1991) The pronunciation of English in Austria, Gunter Narr Verlag, Tübingen.