Towards an interactive pronunciation dictionary for learners*

Jong-mi Kim
(Kangwon National University)

Kim, Jong-mi. 2006. Towards an interactive pronunciation dictionary for learners. Studies in Phonetics, Phonology and Morphology 12.3, 513-531. Research towards developing an interactive pronunciation dictionary that features sensitivity to learners’ native phonology is presented for the specific case of Korean learners of English as a Foreign Language (EFL). A speech corpus was collected to determine whether native-language (L1) dependent feedback on pronunciation improved speech quality as indicated by human perceptual tests. The feedback messages address segmental and prosodic problems in the speech of Korean EFL learners. The segmental phonetic variations that were deviated from native English phonological rules were analyzed to trigger the feedback messages on insertion, deletion and substitution problems of phones. The SUMMIT speech recognizer’s capability to model phonetic variations through explicit, context-sensitive, phonological rules was leveraged to detect pronunciation problems common to Korean EFL learners. Prosodic features of the speech were quantified to determine suitability for detection of stress, rhythm, and intonation problems. Taken together, the segmental and prosodic error detection methods were substantiated for generating specific feedback messages to improve pronunciation quality. (Kangwon National University)

Keywords: pronunciation dictionary, non-native speech recognition, feedback effects, SUMMIT speech recognizer, Computer Assisted Language Learning

1. Introduction

Surveys of pronunciation dictionary from language learners have reported that a dictionary with pronunciation exercises is an easily accessible and attractive form of pronunciation reference (Sobkowiak 1999; Kim 2002b). In spite of the findings that non-native accents in foreign languages are mainly derived from native phonology (Lado 1957; Yang 2002; Kim 2002a; Park et al. 2003), widely available audio dictionaries (e.g., Cambridge, Cobuild, Longman, MacMillan, Oxford, and Webster) for language learning are still insensitive to the learners’ native language (L1). Current non-audio pronunciation dictionaries are not so interactive either that learners are unable to receive help

* This work was supported by the Korea Research Foundation Grant (KRF-2003-A00097), and carried out while I was a Visiting Scientist of the Spoken Language Systems Group at the Massachusetts Institute of Technology. I would like to thank all the people who helped me in various ways, although remaining errors are author’s own. Preliminary versions of this work have been presented at the conferences “From Sound to Sense: 50+ Years of Discoveries in Speech Communication,” at MIT, Cambridge, USA on June 11-13, 2003 [abstract only], and the 8th International Conference on Spoken Language Processing "Interspeech 2004: The 8th International Conference on Spoken Language Processing." Several new empirical experiments of segmental, prosodic, and feedback tests have been, however, added since then.
on pronunciation problems, although speech recognition technology has become available to deal with non-native speech sounds (Bonaventura 1998, Goronzy et al. 2004, Jang 2005).

This paper presents research towards developing an interactive pronunciation dictionary that features sensitivity to pronunciation errors derived from careful analysis of second language (L2) phonology for a specific L1. The future dictionary under consideration is one that can record and evaluate learners’ imitation of dictionary pronunciation and instantly provide feedback on non-native accented phone quality, stress, and intonation. The specific case for the purposes of this paper deals with native Korean speakers’ learning English as a Foreign Language (EFL).

Towards this goal, we have designed and collected a speech corpus to address the phonological and prosodic issues for Korean EFL learners. All speech data were force-aligned with the corresponding transcriptions using the MIT SUMMIT speech recognition engine (Glass 2003). Central to our technology is the use of two sets of phonological rules: one for native English phonology, and the other for non-native phonological variations which are expected from learners’ utterances. Differences in the alignments produced with the two sets of phonological rules reveal segmental insertion, deletion, and substitution in non-native speech. These discrepancies in segmental features can trigger feedback messages pointing out specific non-nativeness. Instructions based on measured duration and F0 cues can also be given to learners on the prosodic non-nativeness of pronunciation.

The benefit of our methodology as such yields two folds. On the segmental side, we achieve vocabulary independence in detecting phonetic mispronunciations, as attested in this study for as large data as approximately 13000 words and sentences. Our system does not require pre-recorded L2 words and sentences for acoustic modeling regardless of Korean or Korean-accented English. Instead, we only use L2 phonological rules that are added to the built-in L1 speech recognition system. The built-in L1 system may comprise L1 acoustic training, L1 phonological rules, and L1 phonemic baseforms. Phonemic baseforms of all words in a language can be exhaustively transcribed in a pronunciation dictionary (e.g., CMU Pronouncing Dictionary for American English). Our L2 phonological rules cover untrained L2 speech string, without affecting the recognition performance of L1 speech string. On the prosodic side, we effectively reduce the prosodic variations caused by pragmatic differences by taking advantage of the fact that the non-native speakers in our corpus imitate the native model speaker. We believe our research setting as such is a reasonable simulation of the audible dictionary application. Furthermore, our approach for unlimited vocabulary overcomes the limitation of existing pronunciation-teaching software, such as Dr. Speaking in Korean English, which must limit the vocabulary sets to a small number of pre-recorded words and sentences.
Towards an interactive pronunciation dictionary for learners

The paper is organized as follows. We first describe the speech corpus in Section 2, which was designed for a balanced coverage of the major aspects of segmental and prosodic phonology. We then provide a detailed description of our technology in Sections 3 and 4, including automatic methods for detecting segmental and prosodic cues of non-native accentedness. There are two questions addressed in our technology: “Can an automatic speech recognizer be utilized to detect specific segmental errors (Section 3)?” and “Are there easily computable prosodic features that can be used to provide feedback on pronunciation quality (Section 4)?” Afterwards, we evaluate, in Section 5, the effect of L1-dependent feedback on improving learners’ pronunciation. Finally in Section 6, we conclude the findings and discuss the future use of research.

2. Collection of non-native speech corpus

A non-native speech corpus was collected to cover a broad range of L1-interfered productions of English vowels, consonants, syllables, stress, rhythm, and intonation. The phonological distribution of the corpus is summarized in Table 1.

Table 1. Distribution of the data. Parenthesized are the numbers of unique stimuli in the recording list, which were given to multiple learners.

<table>
<thead>
<tr>
<th>Type</th>
<th>Phone</th>
<th>Syllable</th>
<th>Stress</th>
<th>Rhythm</th>
<th>Intonation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>8,176 (1,362)</td>
<td>1,504 (94)</td>
<td>770 (70)</td>
<td>0</td>
<td>0</td>
<td>10,450 (1,526)</td>
</tr>
<tr>
<td>Phrase</td>
<td>0</td>
<td>0</td>
<td>264 (24)</td>
<td>272 (17)</td>
<td>32 (2)</td>
<td>568 (43)</td>
</tr>
<tr>
<td>Sentence</td>
<td>0</td>
<td>260 (20)</td>
<td>371 (33)</td>
<td>569 (32)</td>
<td>1,059 (68)</td>
<td>2,259 (153)</td>
</tr>
<tr>
<td>Total</td>
<td>8,176 (1,362)</td>
<td>1,764 (114)</td>
<td>1,405 (127)</td>
<td>841 (49)</td>
<td>1,091 (70)</td>
<td>13,277 (1,722)</td>
</tr>
</tbody>
</table>

A total of 13,277 speech samples were initially collected for the corpus. Of these, 5,101 data of words, phrases, and sentences were newly acquired for segmental and prosodic investigations in our task domain. The 8,176 words for phone were taken from the SOund RIch DAtabase (Kim et al. 1998, 1999, Kim 2001) to conduct segmental research on the phone level. From this initial corpus of 13,277 speech samples, 201 samples of poor or inconsistent signal quality were discarded, leaving out 13,076 samples available for analysis. The remaining data contained a varied number of native and learner speakers who read each stimulus, which was carefully designed for the specific nature of a given phonological unit. Overall, 5 to 11 speakers were assigned to read the word-level stimuli, and 10 to 18 speakers to the phrasal and sentential level stimuli. The corpus includes approximately 5,000 native English data to compare with the Korean learner English data. Both data share the same recording stimuli list.
The newly acquired data in this corpus was collected according to the following procedure. A native speaker who speaks the main-stream American English was a model speaker. The model speech was recorded and distributed in CD or via the website to learners. A total of 50 learners participated in the experiment. Their English proficiency varied, although most of them were considered to be in the intermediate level based on their academic standing in an English pronunciation class. They had a score of 230 to 350 on the standardized TOEIC listening test. They then received explicit lessons to resolve any potential L1-specific difficulties in pronouncing the stimuli. They were given, on average, a one week practice period prior to the recording. They were instructed to imitate the model speech as closely as possible. The speech of the model speaker and the learners was recorded in a quiet room and digitized at a sampling rate of 16 kHz with the Computerized Speech Lab by Kay Elemetrics.

In the following, we describe the methods which can automatically detect segmental and prosodic cues of non-native accent. The cues are to trigger corrective messages to assist the learners with pronunciation improvement.

3. Detection of non-native segments

Our analysis of the segmental properties of non-native speech used the SUMMIT speech recognition system (Glass 2003), developed by the Spoken Language Systems group at MIT. The SUMMIT system uses context-dependent phonological rules to explicitly encode permissible phonetic variations of a word. Typical rules are unreleasing of stops as in cut down, flapping as in pot of, and cluster simplification as in want a. The rules also include syllabic sonorants in which an unstressed vowel may be deleted to allow syllabicité to the following sonorant consonant as in button. A detailed description of pronunciation variations is provided in Hazen et al. (2002). Various SUMMIT rules are illustrated in Table 2.

<table>
<thead>
<tr>
<th>(left)</th>
<th>core</th>
<th>(right)</th>
<th>realization</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[ ]</td>
<td>dh</td>
<td>dh</td>
<td>whether</td>
</tr>
<tr>
<td>2</td>
<td>[ ]</td>
<td>td</td>
<td>tcl [t]</td>
<td>cut down</td>
</tr>
<tr>
<td>3</td>
<td>{VOWEL SEMIVOWEL}</td>
<td>td {VOWEL SEMIVOWEL hh}</td>
<td>tcl [t]</td>
<td>dx</td>
</tr>
<tr>
<td>4</td>
<td>[n]</td>
<td>td</td>
<td>{tc[l] [ax]}&gt;</td>
<td>want a</td>
</tr>
<tr>
<td>5</td>
<td>{tq td dd}</td>
<td>en</td>
<td>&lt;[tcl dcl] en&gt;</td>
<td>button</td>
</tr>
</tbody>
</table>

Table 2. Sample phonological rules in the SUMMIT speech recognition framework. The curly brackets ‘{ }’ specify the input phonemic contexts on the left and right sides of the given phoneme in the core. The arrows denote the rewrite rules by which the input phonemes in the core are realized in the phonetic surface form on the right side of the arrows. The symbols ‘|’ and ‘[]’ represent alternative forms and optional forms.

\[1\] The model speaker was an English professor at a college, 45, male, from Utah.
The first rule illustrates a rewrite rule of the SUMMIT baseform to an acoustic phonetic form where no allophonic change is involved. By this rule, a dental fricative /θ/ is phonetically realized as [θ]. /ðθ/ is a SUMMIT baseform of [θ] in IPA. The second rule illustrates an optionality in acoustic output by a bracket ‘[ ]’. By this rule, a coda /t/ may have only closure and no burst as in cut down. /td/ is a SUMMIT baseform of a coda /t/. The SUMMIT phonetic form ‘tcl [t]’ denotes an obligatory t-closure followed by an optional t-burst. The third rule illustrates an alternative in acoustic output by a vertical bar ‘|’. A coda /t/ between two vowels may either become a flap [r] as in pot of, or a t-closure followed by an optional t-burst. The SUMMIT phonetic form ‘dx’ is an alveolar flap [r] in IPA. /hh/ is a SUMMIT baseform of [h] in IPA. Although this rule is stress related, the SUMMIT system does not include stress in baseforms. The fourth rule illustrates a deletion. A coda /t/ between /n/ and /ax/ may be deleted if the following vowel is realized as a weak [?] or [?] as in want, or realized as a t-closure if the following vowel is realized as a syllabic [n]. The SUMMIT phonetic forms ‘ax’, ‘ix’, and ‘en’ are [s], [i], and [n] in IPA. The fifth rule illustrates a conditional phonetic distribution. A syllabic nasal /en/ after variants of alveolar stops /t/ in IPA. /td/ and /dd/ may phonetically be realized as either a syllabic [n] preceded by a t-closure, or a d-closure, or as a phonetic sequence of [nn] preceded by a t-burst or a d-burst. An example is button with a SUMMIT baseform of /b ah t eq en/. The SUMMIT base forms ‘tq’, ‘t’, and ‘dd’ refer to a potential glottal stop, a coda /t/, and a coda /d/. We applied, as a few attempts in speech technology and the first dealing with Korean learners’ English to our knowledge, the native phonological modeling framework of a speech recognition system for automatic detection of non-native segmental variations solely by phonological rules, but without any acoustic training from non-native or bilingual speech signal. The advantage of this approach is that it is completely task-independent, since it uses general phonological rules and can thus be used across corpora and especially for new words that are introduced to the system. The SUMMIT phonological rule set was augmented to include typical non-native sound forms of English spoken by Korean learners. We then derived a phonetic transcription of the speech by a ‘forced alignment procedure.’ During this process, the recognizer was configured to find the best-scoring phonetic alignment among the alternatives as determined by the baseform, phonological rules, and the acoustic models. For each non-native utterance, we derived two types of force aligned transcriptions: one from the original L1 rule set and the other from the expanded L2 rule set. Any differences in the two alignments were likely to suggest the existence of non-native phonological variations. When an L1 rule set was used for L2 recognition, a tremendous drop in recognition accuracy was expected.
As much as a 50% drop was reported in German-English (Goronzy et al. 2004). What makes the task more challenging was that our SUMMIT non-native phonological expansions by such generative rule sets in Table 2 were independent of vocabulary contents in the input speech signal.

Figure 1 illustrates the outlined procedure with an example of non-native utterance. The upper phonetic alignment was generated with the expanded L2 recognizer, and the lower one was generated with the original L1 recognizer. As demonstrated in the figure, the L2 rule set better captures the spectral features of non-native sound forms. In fact, the deleted, inserted, and changed phones in the words can’t, decide, and whether were appropriately detected.

The alignment result from the L2 rule set in Figure 1 presents various transfer of native phonology. The deletion of /t/ in can’t is expected in Korean phonology, since the language does not allow a consonant cluster in the syllable final position. The vowel inserted after the final /d/ in decide reflects the fact that Korean phonology does not permit a voiced stop in the word-final position. The learner substituted the stop /d/ in place of the dental fricative /θ/ in whether, since Korean does not allow /θ/.

The alignment accuracy was evaluated 1) by two human experts for 542 non-native words, 2) by three human experts for 1,548 native and non-native utterances, and 3) by machine for all 13,076 native and non-native utterances. For the first test set, the author and a speech expert who is a native speaker of American English, evaluated phonetic alignment accuracy of the L1 and L2 rule sets on the randomly chosen data sets of 542 non-native isolated words. Both raters participated in all rating cases and evaluated the accuracy of label choice and boundary placement. The two phonetic alignments based on the

---

3 This rule is optionally present in native English speech, and considered part of universal unmarkedness. The rule application in Korean is obligatory.
Towards an interactive pronunciation dictionary for learners

L1 and L2 rule sets were compared, and a decision was made according to four categories: L1 better, L2 better, similar (i.e., minimally different), and identical. The category ‘L1 better’ means that the phonetic transcriptions produced by the L1 rule set were judged to be better than those by the L2 rule set. Each rating was annotated after both evaluators agreed for label choice and boundary placement. The results are shown in Table 3.

Table 3. Alignment accuracy of phonetic transcription by L1 and L2 rule sets for selected isolated words that were uttered by non-native speakers.

<table>
<thead>
<tr>
<th>Total Utterances</th>
<th>L1 better</th>
<th>L2 better</th>
<th>Similar</th>
<th>Identical</th>
</tr>
</thead>
<tbody>
<tr>
<td>542 (100%)</td>
<td>60 (11.1%)</td>
<td>183 (33.8%)</td>
<td>60 (11.1%)</td>
<td>237 (43.7%)</td>
</tr>
</tbody>
</table>

In Table 3, the alignment accuracy for L1 rule set in Table 3 was 65.9% for L1 rule set ((60+60+237)/542=0.659), and 88.6% for L2 rule set ((183+60+237)/542=0.886). Since the corrective feedback was triggered upon the detected transcription differences from the L1 and L2 rule sets, it is only the 60 cases where the L1 rule set was better (11.1% of the utterances) that could lead to inappropriate feedback.

In order to see if our L2 rule set can successfully be implemented for the non-native accented and unaccented mixed speech, we expanded the evaluation test for a larger number of native and non-native 1,548 words and sentences that were randomly selected from our database. A possibility for our L2 rule set to correctly align both accented and unaccented speech was sought, because some non-native utterances are unaccented and this unaccentedness increases as the L2 proficiency improves. For this evaluation, the two same raters plus one other research assistant compared the transcription results by the L1 and L2 rule sets in the forced alignment mode. The new data was annotated only when at least two of the three evaluators agreed for label choice and boundary placement. The alignment accuracy for this expanded data set of L1 and L2 speech was 81.9% by the L1 rule set, and 86.1% by the L2 rule set. In other words, our L2 rule set was more accurate than the original L1 rule set of SUMMIT by 4.2% for the L1 and L2 mixed utterances. We further noticed that our L2 rule set brought the better alignment accuracy on L2 speech over the original L1 rule set, and yet it did not lower the recognition performance of L1 speech.

We also tested alignment accuracy with the SUMMIT machine for our entire corpus of 13,076 utterances that contain both native and non-native speech. The results show that alignment accuracy by the L1 rule set is 84.7%, and that of the L2 rule set is 89.7%. In other words, our L2 rule set was more accurate than the original L1 rule set of SUMMIT by 5%. We further noticed that the recognizer is very good at picking out the mispronunciations as in [t]/[i] or [æ]/[ae] confusions, when the L2 rules are in place.

To conclude, our L2 rule sets can effectively be implemented without acoustic training, and yet detect more accurately non-native accents than
the L1 rule set as demonstrated by the three evaluation tests. The increased accuracy is not simply achieved by making extra rules of L2 to the existing L1 rule set. Only a carefully designed L2 rule set may prevent an over-generation problem of pronunciation variants. The segmental detection in this section will combine the prosodic detection in the following Section 4, to trigger accurate feedback message in Section 5.

4. Detection of non-native prosody

Prosodic cues, more importantly than segmental cues (Missaglia 1999), play an important role in signaling non-native accent. Prosodic scoring is difficult, however, due to the tremendous variability in the acoustic realization of prosody, and to the lack of an effective model representing the variability. In our speech corpus, we were able to reduce the variability by instructing the non-native learners to imitate the model native speech. The task, therefore, became substantially easier for our scoring engine to detect only the significantly deviated prosodic properties from the model native speech. To that end, we have implemented the automatic transcription method which performs a number of simple calculations for duration and F0 contour. We have also devised a perceptual test to evaluate the accuracy in prosodic transcription, for identifying non-native accent.

A duration difference has been suggested to be a significant indicator of non-native accent in literature, as in Yang (2002), Kim and Lim (2002), and Park et al. (2003). We calculated the duration ratio of the model native speech in comparison with that of the imitated speech by learners for two units: sentences and non-final function words. A longer duration of function words was expected from learners, because Korean is a syllable-timed language where an unstressed syllable does not reduce. We also expected a longer duration for sentences due to L2 disfluency issues, such as hesitations, repairs, and lengthened unstressed syllables. Such a prosodic scoring method of duration detects the non-native speech rhythm and disfluency.

F0 contour is also an effective indicator of non-native accent (Jeong 2003). Flat or opposite F0 slopes were expected to be frequent learner errors for the learners, because Korean does not allow lexical stress. Thus, for selected vowels, we calculated the F0 slope differences between non-native and native utterances. For example, the slopes on the underlined vowels in the following utterance were compared: ‘I can’t see the [blackboard]N’ and ‘It is made of a [black board]NP.’ For the compound noun (blackboard), a greater pitch drop was expected between the first and the second vowels by the English compound stress rule, while a smaller pitch drop or increase was expected for the noun phrase (NP) by the English nuclear stress rule (See Kim (2005a) for acoustic measurements).

For the pitch detection algorithm, we followed Wang and Seneff (2000). Each F0 contour was first normalized with respect to the speaker’s average
Towards an interactive pronunciation dictionary for learners

F0 value. Then, the F0 values centered on each vowel were averaged over a small window of the time domain. The slope was then computed as a ratio of F0 value differences between the center points of two adjacent vowels over the time difference. A significant deviation from the computed F0 slope for corresponding native and non-native speech became the basis for a refinement feedback message. The empirically determined threshold indicated whether the non-native speaker’s pitch slope was too flat, too sharp, or in the opposite direction (drop or rise) relative to the model speaker’s. The prosodic scoring method for F0 slopes as such was useful for detecting the non-nativeness of too flat/sharp/different pitch.

To evaluate the correlation between human perception and the numeric values derived from this prosodic scoring method, we conducted two perception experiments. The first perception experiment was to determine whether perceived anomalies were correlated with the five kinds of prosodic scoring: non-native rhythm, disfluency, flat/sharp/different pitch. The test data consisted of 281 utterances from 24 unique stimuli uttered by 10 to 17 learners. The author with a speech expert who is a native speaker of American English, manually labeled each utterance in terms of the five perceived anomalies: too flat/sharp/different pitch from the model speech, machine-gun rhythm, and L2 disfluency. Each rating was annotated after both evaluators reached an agreement. The prosodic scores of each utterance were not available to the raters at the time of rating. Figures 2 and 3 illustrate a correlation between prosodic scores of duration and F0 and human perception characterized by conditional distributions.

![Sentence duration ratio](image)

**Figure 2.** Fluency perception result, illustrating a correlation between sentence duration and human perception of the fluency level for non-native accentedness as produced by Korean learners of English. The longer the sentence duration is, the worse the fluency is rated.

4 ‘Machine-gun rhythm’ is a colloquial expression referring to the perceptual quality of syllable-timed rhythm.
Figure 3. Pitch perception result, illustrating a correlation between cumulated F0 slope and human perception of the pitch level for non-native accentedness as produced by Korean learners of English. The bigger the cumulative F0 slopes are, the sharper the pitch is rated.

The second perception experiment was to determine whether the feedback messages derived from our prosodic scoring method would be agreeable and motivating for learners. Two native-like bilinguals as the raters participated in this perception test. They were first given 127 utterances with feedback messages and then asked to respond whether the sentences with an error annotation were where learners needed more practices, and whether the sentences without an error annotation were where learners deserved a congratulatory compliment. The two raters were asked to reply with a ‘yes’ or a ‘no’ after each comparative pair of the native model and non-native learner speech samples. The test was repeated twice to standardize the ratings. Both raters answered ‘yes’ for 91.5% of the learner speech samples, and the inter-labeler reliability was 95%.

To conclude, the prosodic scoring method employed in this study can effectively be implemented by a number of simple calculations for duration and F0 contour, and yet accurately detect the non-native prosody that correlates to human perception as shown by the two evaluation test sets. Section 5 discusses the nature and effectiveness of feedback messages.

5. Effectiveness of L1-dependent feedback

The effectiveness of native-language-dependent feedback was tested with simulated feedback messages that would be machine generated from the
Towards an interactive pronunciation dictionary for learners 523

This section will briefly introduce the nature of feedback messages, and discuss in detail how such messages positively evaluate the effect of L1-dependent feedback.

The feedback generation procedure is the following. Learners first received a recording list of 10 words and 10 sentences (See Appendix). They also received a verbal phonetic lesson on stress placement for relevant morphological and emphatic features. They were then asked to listen to and repeat after the model native speech. The written stimulus list was given to learners as a reference during this ‘listen-and-repeat’ task. An English teacher (the author) listened to the first productions and selected one or two feedback messages among the pre-written list of messages. The pre-written list of the feedback messages was in the learner’s native language Korean. The list contained 20 sets of feedback messages that correspond to the recording stimuli of 10 words and 10 sentences. Each set of feedback messages then consisted of one congratulatory message and two to four correction messages that were derived from our analysis described in Section 3 and 4. For word stimuli, the phone-level feedback was given for insertion, deletion, and substitution. For sentence stimuli, prosodic-level feedback was given on stress placement, rhythm, and intonation.

Table 4 illustrates examples of feedback messages written in Korean provided to the learners. In the bracket are marked typical accents of Korean English as described in the previous section; e.g., phone quality as in ‘add[eu],’ pitch contour as in ‘d[ar]k r[oo]ms [a]re d[ar]k,’ and duration of the reduced function words as in ‘[whether to]’ or in ‘Addition [and] subtraction [are] learned skills.’ The feedback messages also included capitalized letters to mark the core parts of stressed syllables.

Table 4. Translated examples of L1-dependent feedback messages in Korean.

<table>
<thead>
<tr>
<th>Feedback message</th>
<th>Phone</th>
<th>Stress</th>
<th>Rhythm</th>
<th>Intonation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone</td>
<td>add[eu]: You insert the vowel [eu] in this red marked part. Instead, try to say the word add without an insertion at the end. Listen to the native speaker again and repeat as closely as possible.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stress</td>
<td>Not all [dark rooms are dark] rooms: You place a wrong stress in this red marked part. Instead, try to say “not ALL dark ROOMs are DARK rooms.” Listen to the native speaker again and repeat as closely as possible.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhythm</td>
<td>I can’t decide [whether to] wear [my] gray suit [or the] brown [one]: You say these words in the bracket too long and loud. Instead, try to say “I CAN’t deCide whether to WEAR my GRAY suit or the BROWN one.” Listen to the native speaker again and repeat as closely as possible.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intonation</td>
<td>![Addition and subtraction are learned)] [skills]: Overall, you are not using correct English intonation. Instead, try saying “SKILLs” livelier than “ADDition and subTRAction are LEARNed.” Listen to the native speaker again and repeat as closely as possible.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Currently, the feedback messages in the audible dictionary are generated manually depending on the detected cues. An automated process is, however, in need in the future to meet a demand from language learners. We consider a manual generation is acceptable for scientific discussion on the feedback effects in this section, to the extent of evaluating if the simulation is helpful for pronunciation improvement.
After the written L1-dependent feedback was given, learners were asked again to listen to and repeat after the model native speech for the same words and sentences. No verbal explanation was delivered about the written feedback. Individual learners spent from 13 to 69 minutes to practice for the second recording. After the second recording, learners were asked whether they had understood the instructions and whether the instructions had been helpful. All the learners unanimously answered 'yes' to both questions.

Using this procedure, we collected a total of 1,152 utterances that learners produced before and after the feedback. These data sets for testing the feedback effect in this section were different from the data set in Section 2 which were used for segmental and prosodic analyses in Sections 3 and 4. This section discusses in detail how the data positively demonstrates the effect of L1-dependent feedback.

To assess the effectiveness of pronunciation feedback, we conducted three sets of perception experiments: 1) with native speakers, 2) with EFL teachers and learners and 3) by using different methods for rating. These three experiments simulated some representative cases in which an interactive pronunciation dictionary might be used.

The first experiment simulated the case that various Korean learners not enrolled in EFL class would use the pronunciation dictionary to communicate with a native speaker. 6 non-EFL-learners were selected from as diverse backgrounds as possible: sex (3 males and 3 females), age (15 to 45), English proficiency (3 months to 8 years of stay in the United States), education level (high school to Ph.D.), and occupation (various levels and areas of specialization). The speakers spent on average 45 minutes of practice time after the feedback. A total of 222 utterances by the non-EFL-learner were judged by two native speaker non-expert raters. The raters determined whether the first production of speech was non-native accented and whether the second production of speech was better, worse or the same in pronunciation quality. Their judgment was monitored by the author and three other bilinguals for 80% of the data, and was considered that the monitored cases were reliable.6 Figure 4 shows the rating of the utterances in percentage for all data. The rated cases were a total of 222 pairs of utterances; each rater answered 111 pairs of utterances. 18 cases (8%) from the first recording session were judged to be unaccented, and were thus excluded from further analysis.

---

6 The reliability was based on the timely completion of tasks on site and reasonable answering of the commentary questions on the ratings.
Towards an interactive pronunciation dictionary for learners

Figure 4. Ratings on English productions of six Korean speakers after L1-dependent feedback in comparison with those before the feedback as rated by two native English speakers. The learners showed significant improvement in pronunciation quality after feedback ($p<.001$, $n=222$).

The chi-square test showed a significant improvement in both words and sentences. On average, 61% of the utterances were rated to have been improved for the second recording. The speaker *mshm* did not show as much improvement as did the others, perhaps because he spent the shortest time (16 minutes) to practice after receiving the written feedback.

The second experiment simulated the case that Korean students in EFL class would use the pronunciation dictionary to supplement the classroom instruction by the EFL teacher. 17 learners who were enrolled in different intermediate EFL pronunciation classes participated and followed the exact same procedure as in the first experiment. They produced a total of 672 utterances from the 17 EFL learners were evaluated by 5 native EFL teachers at Korean colleges. Of the 672 utterances, 594 utterances from 15 volunteers out of the 17 learners were also evaluated by the volunteer learners themselves. Their judgment was monitored by the author and one EFL graduate student for 80% of the data. It was consented that the judgment was reliable for all monitored cases. Figure 5 shows the ratings of utterances in percentage for all data. The rated cases were a total of 883 pairs by the EFL teachers and a total of 294 pairs by the EFL learners. 39 cases (4%) from the first production of speech were judged to be unaccented by the EFL teachers, and were thus

---

7 The data in Figure 5 have been used as partial evidence to develop classroom pedagogy in Kim (2005b).
8 We excluded 8 utterances which had bad recording quality in at least one of their pairs.
excluded from further analysis. Every EFL learner rater judged that all utterances of their own were accented.

![Feedback effect rated by EFL teachers and learners](image)

Figure 5. Ratings on English productions of 17 Korean speakers after L1-dependent feedback in comparison with those before the feedback as rated by (a) 5 native speaker EFL teachers and (b) 15 EFL learners themselves. The statistics shows a significant improvement for learners in their pronunciation after feedback as rated by both groups of raters ($p<.001$, $n=883$; $p<.001$, $n=294$).

According to the chi-square test, all raters from both groups determined that the utterances after feedback showed significant improvement for both words and sentences. This result is the same as the first experiment in Figure 4. In Figure 5, 55% of the data (that is, a total of 643 pairs out of 1,177 pairs) were improved in the second production speech. It is important to note that the learners positively evaluated their pronunciation improvement, because it may lead to a satisfaction with the use of the dictionary.

The third experiment simulated the case that Korean learners of English would use the pronunciation dictionary to ensure their pronunciation is better understood by various unfamiliar native speakers. To ensure this, the experiment used a ‘blind’ method, in which all speech samples were randomized in the order of display within the pairs and among different pairs. We refer this method as the ‘blind’ method, because the rater has no way of knowing which sample in the utterance pairs is from pre- or post-feedback. 29 speakers participated that consisted of all 23 speakers in Experiments 1 and 2, and six other EFL learners. They spent, on average, 34 minutes of
Towards an interactive pronunciation dictionary for learners

practice time on after the feedback. A total of 978 utterances randomly selected from the entire 1,152 utterances were judged by 24 native speakers.

During this experiment, since the raters did not know which ones were the first utterances, they were asked to respond to different questions from the previous Experiments 1 and 2: 1) Is one more natural than the other? (yes, no) and 2) If yes, which one? (1st utterance, 2nd utterance). 9 The entire process of rating was monitored by the author with an occasional addition of other monitors. The rating was considered to be reliable based on baseline experiments, 10 timely completion of tasks on site, and responses to the commentary questions. To keep a good level of concentration, no raters were allowed to work more than 15 minutes each time and two consecutive times including a break. Table 5 shows all 3,224 ratings of the second question from the whole 978 utterances (489 pairs). Postponed from further analysis were the ‘no’ answers to the first question (n=2416(43%)), ‘Is one more natural than the other?’, that disable the second question, ‘If yes, which one [is more natural than the other]?’. 11

Table 5. Ratings on English productions of 29 Korean speakers after L1-dependent feedback in comparison with those before the feedback as rated with the blind method by 24 native speakers. The raters judged the utterances spoken after feedback to be better than those before feedback. (n=3,224)

<table>
<thead>
<tr>
<th></th>
<th>Word</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>After feedback</td>
<td>591 (56%)</td>
<td>1278 (59%)</td>
</tr>
<tr>
<td>Before feedback</td>
<td>457 (44%)</td>
<td>898 (41%)</td>
</tr>
</tbody>
</table>

Of these 3224 ratings in Table 5, some were statistically significant for the given utterances, but others were insignificant by the chi-square test (p<.05). Figure 6 shows these significant 1,061 ratings out of the 3,224 ratings that were confined to the 173 pairs of distinctive utterances. The observed proportions for the 173 distinctive utterances were 43% (=42) for ‘Before feedback’ and 57% (=56) for ‘After feedback,’ in 98 pairs of words and 33% (=25) for ‘Before feedback’ and 67% (=50) for ‘After feedback’ in 75 sentences. Thus, the learners achieved 8.0% of the improvement rate on average ((56-42+50-25)/489).

---

9 Due to the randomized play-list and non-expert raters, a direct translation would be too complicated to answer instantaneously, as in the translation of the first question “Is one of the two utterances in the pair non-native accented?”

10 A base-line experiment was conducted prior to the rating to see if the raters were able to perform the rating task itself. We used dog-barks, dog-growls, and human-barks and asked the two questions: “Is one more natural than the other?” and “If yes, which one?”

11 Notice that Experiment 3 had given more liberty for the raters to stop after the first question of “Is one more natural than the other?” as opposed to the previous experiments that the raters were asked “Is the first utterance before feedback non-native accented or not?” Considering that the time interval for the two comparing productions was as short as 30 minutes on average, we do not expect the same learner speaker could convert non-native accented speech in the first production to a native-like one in the second production.
Figure 6. Ratings on English productions of 29 Korean speakers after L1-dependent feedback in comparison with those before the feedback as rated with the blind method by 24 native speakers. The raters judged the utterances spoken after feedback to be better than those before feedback. \( p < .001, n = 1061 \)

Figure 6 also shows some improvement after feedback. The observed proportions for the two variables were 39\% (=156) for ‘Before feedback’ and 61\% (=243) for ‘After feedback,’ for words and 34\% (=230) for ‘Before feedback’ and 66\% (=504) for ‘After feedback’ for sentences. The Binomial test as a nonparametric statistic technique calculated the general direction of the deviation from randomness. A significance value of \( p < .001 \) was acquired for both words and sentences. This result supports that the deviation from randomness was significant toward one variable ‘After feedback.’ It further seems that, according to the observed proportions of the variables in the two statistic tests, the variable ‘After feedback’ was more proportionately accrued for the independent variable ‘Sentence’ than for ‘Word.’ The result suggests that the test stimulus ‘Feedback’ was a greater factor in the sentence level than in the word level.

To conclude, our L-1 dependent feedback can effectively be implemented in the use of a pronunciation dictionary with written messages and short practice time for ‘listen-and-repeat.’ This type of feedback may bring about improvement on pronunciation quality that correlates to human perception by the EFL teachers, learners and native speakers. We note in particular that the learners evaluate their own production, because it leads to their satisfaction with the dictionary use.
6. Conclusions and future work

An implementation of non-native phonological rules and L1-dependent feedback into the audible lexical dictionary helps learners improve their intelligibility in speech, as evidenced in the perception test results.

The speech corpus in Section 2 that contained a balanced coverage of the major aspects of segmental and prosodic phonology, allowed us to detect in Section 3 and 4 the specific non-native pronunciation problems of Korean learners. In Section 3, we implemented a new set of non-native phonological rules for SUMMIT speech recognizer. As a result, the rules successfully detected specific segmental errors which could subsequently trigger corrective feedback messages for the utterance input. In Section 4, we devised a non-native prosodic scoring method, and the method successfully detected specific prosodic errors to trigger corrective feedback messages. In Section 5, we simulated the L1-dependent feedback triggering, and the results suggested that the learners’ pronunciation was improved with the satisfaction of the learners themselves.

The speech corpus used in this work is available for free within the speech research community by the name, ‘SORIDA 3.’ Permission to make digital or hard copies of part or all of the corpus for personal and classroom use is granted without a fee, provided that copies are not made or distributed for profit or commercial advantages and that copies bear the full citation on the first page.

Appendix: Feedback stimuli

(1) Words:
active, add, addition, bath, confirm, confirmation, dry, field, rich, splendid

(2) Sentences:
a. Add remaining ingredients and bring to a boil.
b. Addition and subtraction are learned skills.
c. Deliver books Friday.
d. Deliver books by Friday.
e. I can’t decide whether to wear my gray suit or the brown one.
f. Not all dark rooms are dark rooms.
g. Please permit me to park there.
h. Teaching languages is harder than learning them.
i. There is no confirmation about the policy.
j. Will you please confirm the policy?
REFERENCES


Dr. Speaking. 2002. [software]. Seoul: Eoneo Inc.


Towards an interactive pronunciation dictionary for learners


Jong-mi Kim
Department of English Language and Literature
Kangwon National University
Chuncheon City 200-701
Republic of Korea (South)
E-mail: kimjm@kangwon.ac.kr

Received: October 10, 2006
Accepted: December 16, 2006