An Interactive English Pronunciation Dictionary for Korean Learners

Jong-mi Kim[†‡], *Chao Wang*[‡], *Mitchell Peabody*[‡], *Stephanie Seneff*[‡]

[†] Department of English, Kangwon National University, Korea
kimjm@kangwon.ac.kr
[‡] MIT Computer Science and Artificial Intelligence Laboratory
The Stata Center, 32 Vassar Street, Cambridge, MA 02139
{wangc,mizhi,seneff}@csail.mit.edu

Abstract

We present research towards developing a pronunciation dictionary that features sensitivity to learners' native phonology, specifically designed for Korean learners of English-as-a-Foreign-Language (EFL). We envision a future system that can record and process learners' imitation of the dictionary pronunciation and instantly provide segmental and prosodic feedback on accent. Towards this goal, we have designed and collected a speech corpus to address the phonological and prosodic issues of Korean EFL learners. We leverage the SUMMIT speech recognizer's ability to model phonological rules to automatically identify non-native phonological phenomena. These phonological rules were carefully constructed to account for the influence of learners' native language (Korean) on the target language (English). Feedback messages are provided to the learner to point out the non-native phonological variations detected by the speech recognizer in order to help them improve segmental pronunciation. Instructions are also given to the user on the prosodic aspects of the pronunciation, which are based on detected duration and F_0 cues. We evaluated the effectiveness of the feedback mechanism by rating 222 English utterances from six native Korean subjects, before and after receiving native-language dependent feedback messages. Human raters judged 61% of the utterances as improved after feedback.

1. Introduction

Surveys of language learners have reported that a dictionary with pronunciation exercises is an easily accessible and attractive form of pronunciation reference [1, 2]. Such audible dictionaries are widely available, and are published by Cambridge, Cobuild, Longman, MacMillan, Oxford, Webster, etc. However, these dictionaries remain insensitive to a learner's native language, in spite of findings that non-native accents in foreign languages are mainly derived from the native phonology [3, 4, 5, 6, 7].

In this paper, we present our research towards developing a pronunciation dictionary that features sensitivity to learners' native phonology, specifically designed for Korean learners of English-as-a-Foreign-Language (EFL). We envision a future system that can record and process learners' imitation of the dictionary pronunciation and instantly provide segmental and prosodic feedback on accent. Towards this goal, we have designed and collected a speech corpus to address the phonological and prosodic issues of Korean EFL learners. All speech data were force-aligned with their corresponding transcriptions using the MIT SUMMIT speech recognition engine [8]. Central to our technology is the use of two sets of phonological rules, one for native English phonology, and another expanded to cover non-native phonological variations expected from Korean EFL learners. Differences in the alignments produced with the two sets of phonological rules will reveal segmental insertion, deletion, and substitution in the non-native speech, and are used to trigger feedback messages pointing out the specific errors. Instructions are also given to the learner on the prosodic aspects of the pronunciation, which are based on measured duration and F_0 cues.

	Phone	Syll.	Str.	Rhy.	Into.	Total
Word	8,176 (1,362)	1,504 (94)	770 (70)	0 (0)	0 (0)	10,450 (1,526)
Phrase	0 (0)	0 (0)	264 (24)	272 (17)	32 (2)	568 (43)
Sentence	0 (0)	260 (20)	371 (33)	569 (32)	1059 (68)	2259 (153)
Total	8,176 (1326)	1,764 (114)	1,405 (127)	841 (49)	1,091 (70)	13,277 (1722)

Table 1: Distribution of data. The number of unique stimuli are shown in parentheses. (Note: Syll. = Syllable, Str. = Stress, Rhy. = Rhyme, Into. = Intonation.)

Our prosodic scoring methods take advantage of the fact that the non-native speakers in our corpus were trying to imitate examples from a native speaker. We achieve vocabulary independence in detecting phonetic mispronunciations, since our methodology requires only the *phonemic* baseforms of the words. We believe that these research settings are reasonable for the audible dictionary application. Our approach can be contrasted with existing pronunciation teaching software, such as Dr. Speaking [9] in Korean English, which typically have a limited set of words and sentences.

The paper is organized as follows. We first describe our speech database, which was designed to have a balanced coverage of the major aspects of segmental and prosodic phonology. We then provide a detailed description of our technology, including automatic methods for detecting segmental and prosodic cues of non-native accentedness. After that, we evaluate the effect of native-language-sensitive feedback on improving learners' pronunciation. Finally, we summarize our results and point out potential future research.

2. Non-native speech data collection

The non-native speech database was designed and constructed to cover a broad range of non-native productions of English vowels, consonants, syllables, stress, rhythm, and intonation. The linguistic distribution of the collected data is summarized in Table 1.

There were a total of 13,277 speech samples in the original corpus. Of these, 8,176 isolated words, taken from SOund RIch DAtabase [10], were used to facilitate segmental research on the phone level, and the remainder, including words, phrases, and sen-

{**left**}	**core**	{**right**}	→ **realizations**	; comments
{vowel}	t	{schwa}	→ tcl t | dx	; flapping
{}	s	{y sh zh}	→ s | sh	; palatalization
{en n}	n	{}	→ [n]	; de-gemination

Figure 1: Representative phonological rules in the SUMMIT recognition framework. The curly brackets "{ }" specify the *input* phonemic contexts on the left and right side of the given phoneme. The arrows denote the rewrite rule by which the input phoneme is realized in the phonetic surface form on the right side of the arrow. The symbols "|" and "[ ]" represent alternative and optional forms respectively.

tences, were newly acquired for prosodic research. 201 samples were discarded from the database because of poor signal quality. The number of speakers reading each stimulus varied, with fewer speakers reading word-level stimuli (5 to 11 speakers), and more speakers reading phrases and sentences (10-18 speakers). The corpus also includes native English and Korean speech to obtain acoustic models of native speech for both languages.

The prosodic database was acquired using the following procedure. A native speaker of General American English played the role of a model speaker. A total of 50 English-language learners participated in the data collection. They vary in fluency, although most of them are considered intermediate in terms of their academic standing in a pronunciation class. The model speech was first recorded and distributed to the learners in CDs or via the Web. The learners then received explicit lessons to resolve any potential difficulties in pronouncing the stimuli, and 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 from both the model speaker and the learners was recorded in a quiet room and digitized at 16 kHz sampling rate using Computerized Speech Lab by Kay Elemetrics.

3. Methodology

In the following, we describe our methods to automatically detect segmental and prosodic cues of non-native accent, which are used to trigger corrective messages to help a learner improve pronunciation.

3.1. Segmental analysis

Our analysis of the segmental properties of the non-native speech used the SUMMIT *landmark-based* speech recognition system [8], developed by the Spoken Language Systems group at MIT. The SUMMIT system uses context-dependent phonological rules to explicitly encode permissible phonetic variations given the phonemic pronunciation of a word. Typical rules are epenthetic silence insertion at locations of voicing change ("sweet"), gemination ("from Maine"), palatalization ("gas shortage"), and rules accounting for unreleased stops ("top down"), or contraction across words as in "wanna" for "want to." The rules also include devoicing in fricatives and stops as well as reduction of vowels with respect to stress alternations. A detailed description of pronunciation variations is provided in [11]. Example rules are shown in Figure 1.

The phonological modeling framework in SUMMIT can be easily adapted for the automatic detection of non-native segmental variations. The phonological rule set was augmented to include the 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 which the recognizer was configured to find the best-scoring phonetic alignment given the alternatives determined by the lexicon, phonological rules, and the acoustic models. For each non-native utterance, we derived the two types of forced aligned transcriptions, one from the original (L1) rule set and the other from the expanded (L2) one. Any differences in the two alignments are likely to suggest the existence of non-native phonological variations. Notice that the non-native phonological expansions are independent of the vocabulary.

Total Utts.	L1 better	L2 better	similar	identical
542	60	183	60	237

Table 2: Results of human judgments of phonetic transcriptions produced by L1 and L2 systems for selected isolated words, spoken by non-native speakers. (Note: Utts.=Utterances)

Figure 2 illustrates the outlined procedure with an example 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 the speech. In fact, the deleted, inserted, and changed phones in the words "can't," "decide," and "whether" were appropriately detected.

The L2 realization in Figure 2 manifests various aspects of phonological influence from the native language. The deletion of /t/ in the word "can't" is expected according to Korean phonology, which does not allow consonant clusters in a syllable final position. The schwa inserted after the final /d/ in "decide" reflects the fact that Korean phonology does not permit voiced stops at word-final position. The speaker substituted the stop /d/ in place of the dental fricative /dh/ in the word "whether," since /dh/ is not present in the native Korean phoneme inventory.

The first author, together with a native speaker of American English, evaluated the phonetic alignment accuracy on a randomly chosen data set of 542 non-native isolated words. Results are shown in Table 2. The two phonetic alignments (L1 and L2 based) were compared, and a decision was made according to four categories: identical, minimally different ("similar" in the table), L1 better, and L2 better. Each rating was annotated only after both analysts (the researcher and the native speaker) agreed. Since the feedback was triggered upon detected L1-L2 differences, it is only the 60 cases where L1 is better (11% of the utterances) that could lead to inappropriate feedback.

3.2. Prosodic analysis

Prosodic cues play an important role in the perception of non-nativeness in speech, and are perhaps more important than segmental cues. However, prosodic scoring is difficult, due to the tremendous variability in the acoustic realizations of prosody, and the lack of an effective model representing prosody. In our database, we were able to reduce the variability by instructing the non-native speakers to imitate the native example produced by the model speaker. Hence, the problem could be reduced to the substantially easier task of detecting significant deviations of prosodic properties from the native examples. To that end, we have implemented automatic methods for performing a number of simple calculations on duration and F_0 contour, and have devised a perceptual test to evaluate their effectiveness at identifying non-native accent.

A duration difference has been proposed to be a significant indicator of non-native accent, as in [5, 6, 4]. We calculated ratios of the duration of the native speech reference compared with the non-native imitation for three distinct units: non-final function words,

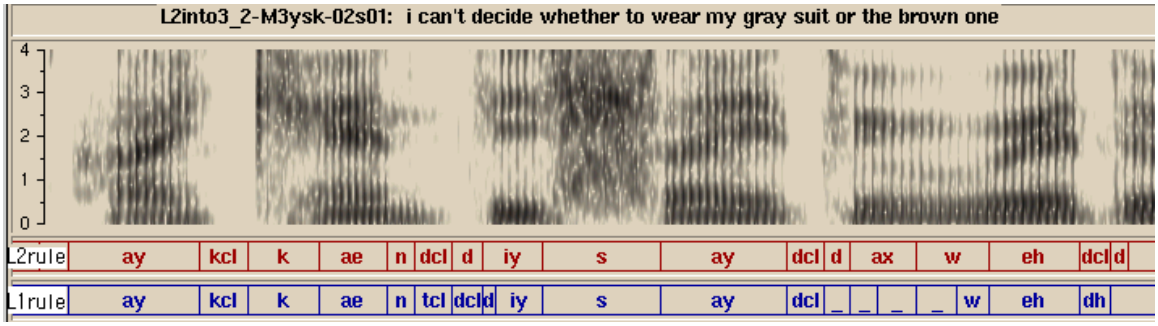

Figure 2: Speech recognition result, illustrating a non-native accent in a Korean learner's English. The alignment with L2 rules is different from that of L1 rules in that the former detects the deleted phone [t] in the word "can'[t]," the inserted phone [ax] at the end of the word "decid[ax]," and the changed phone in the word "whe[d]er."

feet [1] , and sentences. A longer duration of function words is expected for Korean learners, because Korean is a syllable timed language where unstressed syllables would not get reduced. We would also expect a longer duration of sentences, simply due fluency issues, such as hesitations, repairs, and lengthened unstressed syllables. A higher *variance* in foot duration is expected, because the Korean timing beat disregards stress. For function words and sentences, we defined a 20% deviation in length as non-native accent. For foot duration, the standard deviation for selected feet within an utterance was calculated. The threshold for non-native accent depends linearly on the number of feet.

Pitch slope is also an effective indicator of non-native accent [7]. Flat or opposite pitch slopes are expected to be frequent errors for Korean learners, because Korean does not have lexical stress. We used the pitch detection algorithm described in [12]. Each F_0 contour was first normalized with respect to the speaker's average pitch. The slope was then computed as a ratio of F_0 difference between two adjacent vowels over the time difference of their center points. F_0 values were averaged over a small window centered on the vowel. A significant deviation in the computed F_0 slope for corresponding native and non-native speech triggered a refinement feedback message. Empirically determined thresholds indicated whether the non-native speaker's pitch slope is too flat, too sharp, or opposite in direction (drop or rise) compared to that of the model speaker.

The slope difference between non-native and native utterances was calculated on selected vowels for each utterance. For example, slopes on the underlined vowel regions in the following utterance were compared: "Not all [dark rooms]NP are [dark rooms]N." For compound nouns (N), a greater pitch drop is expected between the first and the second vowel in comparison, while a smaller pitch drop or increase is expected for a noun phrase (NP). The refinement feedback message is triggered if the slope is in the opposite direction of the native slope.

Empirical experiments were conducted to evaluate the correlation between the numeric values and human perception, using a total of 127 sentences with 10 stimuli and 10-17 speakers as test examples. An audio-visual perception test was conducted to determine whether all six counts of error annotation correlate with perceived anomalies for the designated portion of each stimulus. For instance, when the underlined vowels in "Not all dark rooms are dark rooms" are measured as "flat," according to our pitch slope computation, the subsequence "dark rooms are dark" really sounds flat. When an annotation indicates slopes in opposite directions, a method involving visual inspection of the pitch contour was used: It would be persuasive if visually presented to a user of the pronunciation dictionary. For this combined test of acoustic and perceptual correlation on the given portion of each stimulus, 97% (all but 3 out of 117 non-native utterances) of the error annotations were rated correct.

The second set of perception tests was done to determine whether the feedback messages derived from this computation on selected portions of each stimulus would be considered reasonable to learners. Two bilingual speakers of Korean and American English participated in the perception test. They were asked to listen to the test samples of all 127 utterances and answer whether the sentences with error annotation need more practice, and whether the sentences without an error annotation deserve congratulatory compliments. They were asked to reply 'yes' or 'no' after each comparison of the native and learner samples. The test was repeated twice to standardize their ratings. Both raters answered "yes" for 91.5% of the learner data, and the inter-labeler agreement rate was 95%.

4. Evaluation of feedback effects

The effectiveness of the native-language-dependent feedback messages was tested on 240 utterances collected from six Korean EFL speakers. The subjects, differing in age group, academic background, and fluency level, each read 20 utterances twice. The stimuli were composed of 10 words and 10 sentences.

The subjects first received the stimulus list as well as a verbal phonetic lesson about the stress placement of the given stimuli depending on different morphological and emphatic compositions. They were then asked to listen to and repeat the model native speech sample of American English. The written stimulus list was also available for their reference during the "listen-and-repeat" task. An English teacher (the first author) listened to their pronunciation of the first production and selected one or two feedback messages written in Korean. Each feedback message consists of one congratulatory message and 2-4 refinement suggestions derived from our analysis described in the previous section. For word stimuli, 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 3 illustrates an English translation of Korean feedback messages provided to the subjects. The bracketed messages resulted from the analysis as described in the previous section, e.g., in terms of the phone quality as in "add[ax]," the pitch slopes as in "d[ar]k r[oo]ms [a]re d[ar]k," and the durations of the reduced function words as in "[whether to]" or of the three foot units of "[Addition and subtraction are learned]."

[1] Each group in brackets in the following sentences represents a foot: "[Deliver] [books] [Friday]" and"[Deliver the] [books by] [Friday]."

Aspects	Feedback message
Phone	add[eu]: You insert the vowel [eu] in this red marked part. Try to say the word "add" without the insertion at the end. Listen to the native speaker again and repeat as closely as possible.
Stress	Not all [dark rooms are dark] rooms: You placed the stress incorrectly in this red marked part. Try to say "not ALL dark ROOms are DARk rooms." Listen to the native speaker ....
Rhythm	I can't decide [whether to] wear [my] gray suit [or the] brown [one]: You say these words too long and strong. Try to say "I CAN't deCIde whether to WEAR my GRAY suit or the BROWN one." Listen to the native speaker ...
Intonation	[[Addition and subtraction are learned] [skills]]: Overall, you are not using correct English intonation. Try saying "aDDItion and subTRAction are LEARNed sKILLs." Listen to the native speaker ...

Table 3: Examples of Korean-dependent feedback message.

Once the Korean-dependent feedback was given in the written sheet, the subjects were once again asked to listen to the native example and repeat it, for the words and sentences that triggered the refinement feedback. No verbal explanation was given on the written feedback. Individual subjects spent from 16 to 69 minutes for the practice and the second set of recordings. After the recording, the subjects were asked if they understood the instructions, and if the instructions were helpful. All the subjects unanimously answered "yes" to both questions.

A total of 222 utterances were judged by two native speakers, to determine whether the first production is non-native accented, and whether the second production is better, worse or the same. Their judgment was monitored by the first author of this paper and three other near-native speakers on 80% of the data, and was considered to be reliable for all monitored cases. Figure 3 shows the ratings of utterances in percentage on all data. Nine utterances (45%) from speaker $foch$'s first recording session were judged as native-like, and were thus excluded from the plot.

As shown in Figure 3, all the subjects showed substantial improvement in clarity. On average, 61% of the utterances were rated improved in the second rendering. Speaker $mshm$ did not show as much improvement as the others, perhaps because he spent the shortest time (only 16 minutes) on the task after feedback.

5. Conclusions and future work

An implementation of non-native phonological rules and native-language-sensitive feedback in an audible lexical dictionary allows speakers to improve their intelligibility in speech, as demonstrated in our preliminary results. Currently, the feedback messages are generated manually depending on the detected cues. We plan to fully automate the process in the future. We have thus far ignored cross-lingual confusions, e.g., confusions between English and Korean vowels. This problem can be addressed by augmenting the English acoustic models with distinctively different Korean acoustic models and expanding the phonological rules to allow cross-lingual confusions. The Korean speech in our database can be used to train Korean acoustic models. We also plan to investigate more sophisticated prosodic scoring mechanisms.

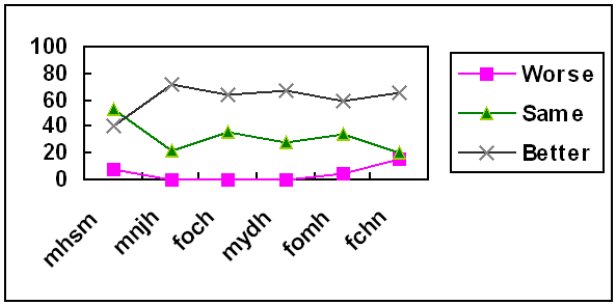

Figure 3: Effectiveness of native-language-sensitive feedback messages. Learners' pronunciations were rated by two native English speakers, comparing utterances spoken before and after feedback. Learners are arranged in increasing order of time spent during the second session (after feedback). (Note: m=male, f=female.)

6. Acknowledgements

The work was supported by the Korea Research Foundation Grant (KRF-2003-A00097). The work has benefited from valuable comments from Kenneth Stevens, Stefanie Shattuck-Hufnagel, Wodzimierz Sobkowiak, and Suzanne Flynn.

7. References

[1] J-M. Kim. English pronunciation dictionaries for Korean learners. *Jungang Journal of English Language and Literature*, 44(2):59–91, 2002.

[2] W. Sobkowiak. *Pronunciation in EFL machine-readable dictionaries*. Motivex, Poznan, 1999.

[3] R. Lado. *Linguistics Across Cultures*. Univ. of Michigan Press, Ann Arbor, 1957.

[4] K-S. Kim and U. Lim. *An acoustic analysis and English pronunciation teaching*. Seoul: Hankook, 2002.

[5] B. Yang. An acoustical study of English word stress produced by Americans and Koreans. *Speech Sciences*, 9(1):77–88, 2002.

[6] K-M. Park, O-H. Lee, and J-M. Kim. English rhythm in foreign language education for Korean learners (abstract only). In *Proc. PAAL*, Okayama, Japan, 2003.

[7] K. Jeong. *Accentuation of English noun compounds and phrases by Korean learners*. Master's thesis, Kangwon National University, Chuncheon, Korea, 2003.

[8] J. Glass. A probabilistic framework for segment-based speech recognition. *Computer Speech and Language*, 17, 2003.

[9] *Dr. Speaking*. [software]. Eoneo Inc., Seoul, 2002.

[10] J-M. Kim, S. A. Dyer, and D. D. Day. Construction of a speech translation database. In *Proc. LREC*, pages 1071–1079, Granada, Spain, 1998.

[11] T. J. Hazen, I. L. Hetherington, H. Shu, and K. Livescu. Pronunciation modeling using a finite-state transducer representation. In *PMLA Workshop*, pages 99–104, 2002.

[12] C. Wang and S. Seneff. Robust pitch tracking for prosodic modeling in telephone speech. In *Proc. ICASSP*, Istanbul, Turkey, 2000.