An Automatic Feedback System for English Speaking Integrating Pronunciation and Prosody Assessments

Jeesoo Bang, Sechun Kang, Gary Geunbae Lee

Department of Computer Science and Engineering
Pohang University of Science and Technology, South Korea
{jisus19, freshboy, gblee}@postech.ac.kr

Abstract

We have proposed a computer-assisted language learning (CALL) system, called Postech English Speaking Assessment and Assistant (PESAA) for non-native English learners, especially Koreans, to improve their overall language skills. PESAA is an automatic feedback system for speaking English that integrates both pronunciation and prosody assessments. The system has three error-feedback modules: the pronunciation, rhythm, and phrase break error-feedback modules. PESAA generates scores on each of three modules as well as a combined score. The pronunciation assessment gives feedback that is based on comparing canonical and actual phoneme alignment results. The rhythm and phrase break assessments give feedback that is based on comparing predictions with detected results. English learners can use PESAA to practice pronunciations, rhythm and phrase breaks by themselves. We evaluated PESAA in three different ways: accuracy, correlation of the system’s assessment with human assessments, and user satisfaction from the expected learning effectiveness and user interface. The evaluation showed that PESAA could work as a CALL system well, with good accuracy and positive learning results for the users.

Index Terms: computer-assisted language learning, CALL, language assessment, error feedback

1. Introduction

Foreign language learners must practice a target language frequently and repeatedly to learn the language efficiently and rapidly. Computer-assisted language learning (CALL) systems offer an advantage in foreign language training outside of the classroom. They meet the language learners’ need for practice by providing a private and stress-free environment in which to train comfortably and conveniently, within the bounds of the learners’ schedules and circumstances.

CALL systems can provide instructional materials for pronunciation, grammar, vocabulary, and other areas. Pronunciation is a difficult skill to acquire alone, unlike memorizing grammar or vocabulary. Serious pronunciation problems can even hinder communication and degrade the intelligibility of speech [1]. To learn a foreign language’s pronunciation accurately, language learners should receive appropriate and timely feedback. In addition to pronunciation, prosody is an important skill to acquire in learning languages. Lexical stress, intonation and rhythm help a listener to understand utterances more accurately [2, 3, 4]. Previous research has shown that training in prosodic features is more effective in improving intelligibility than teaching only in the segmental features [5, 6]. However, very few systems address prosodic features, whereas there have been a number of studies on pronunciation training systems. Moreover, there has been little focus on CALL systems that combine segmental or pronunciation features with supra-segmental or prosodic features.

A significant amount of research related to CALL systems has been conducted [7, 8, 9, 10]. The authors in [7] have developed a CALL system that specialized in English pronunciation training for Japanese English learners. The system provides general feedback for a learner through a reading session of role-playing dialogues. The system gives feedback for phonemes, which Japanese learners occasionally provide. The authors in [8] have developed a hand-held pronunciation evaluation device that uses the log-posterior probability of the expected phoneme sequence from automatic speech recognition (ASR) to assess the pronunciation accuracy and that shows the pitch contour to help correct the intonation. The author in [9] proposes a strategy of modeling the pronunciation variation at the syllable level using different subsets of context features. The authors in [10] present a system that assesses spoken English for call center agents with multiple parameters: articulation of sounds, correctness of lexical stress in words and spoken grammar proficiency.

Most of the systems described above focus mainly on pronunciation evaluation, and few of them focus on evaluating syllable stress. We have developed a system called PESAA that integrates pronunciation training with prosody training. PESAA has three assessment aspects: pronunciation, rhythm and phrase breaks. For a single user utterance, each part delivers its assessment result independently of the other parts. The result comprises detailed feedback on phonemes, words or breaks in addition to 0-to-100 scores for each part and for the overall system. PESAA also provides a specific view of each assessment module, which are the pronunciation, rhythm and phrase break, to help learners practice English intensively.

2. System architecture

PESAA is designed as a client-server model, to allow the learners to learn English wherever they want. PESAA has two main components, the user interface and speech processing.

2.1. User interface

The user interface component records a user’s utterances and shows the feedback and evaluation result of the recorded utterances. This component displays sentences to be recorded for evaluating the pronunciation and prosody; it provides prompts to be shown back to the users with error feedback, and provides assessment results of recorded utterances, by recording a user’s utterances and playing back the recorded utterances. This
component also involves some amount of speech processing, to ensure that the candidate’s speech is recorded at an appropriate volume level and to warn the user in case the recorded speech is too short or too long, after which the times can be modified. Users can select the sentence level and the sentence that they prefer to practice, or they even can type the sentences that they want to practice by themselves.

2.2. Speech processing

The speech processing component, which resides on the server, uses the speech recognition engine to obtain the phoneme alignments and the confidence scores. The speech processing component has three main modules: pronunciation, rhythm and phrase break assessment modules. Each module gives detailed feedback on phonemes, rhythms and phrase breaks with 0-to-100 scores as well as combined scores. The detailed module architecture and score computation are described in Section 3 and Section 4.

3. Module architecture and implementation

3.1. Pronunciation error feedback

![Architecture of pronunciation error feedback](image)

We focused on the differences between Korean and English because PESAA was developed for Korean learners of English. The Korean phoneme inventory differs largely from the English phoneme inventory. For example, Korean does not have the sounds /F/, /V/, /TH/, and /Z/ [11]. Thus, the erroneous phonemes should be pointed out, to train learners in pronunciation effectively.

We built a pronunciation error feedback module to detect the phoneme errors of the input utterances and to provide feedback to the learners. The pronunciation module has four main components: actual phoneme alignment, canonical phoneme alignment, error detection and feedback generation (Fig. 1). Actual phoneme alignment uses an extended recognition network (ERN) [12] that is generated by the pronunciation variants prediction model.

3.1.1. Data

We collected English-reading speech data spoken by Korean English speakers who had high oral proficiency levels. We distributed 6,600 text segments to 170 Korean people learning English; the text segments consist of sentences, words, and phrases. We collected 12,000 speech segments; each person recorded an average of 74 text segments, and two different people recorded each text segment. Two annotators trained in phonetics and phonology curriculum annotated phoneme-level transcriptions of the collected data. The annotators were provided with the automatic phoneme-level transcription that corresponded to the text segment read; they revised the transcription, repeatedly listening to the recording, if necessary. The phoneme-level transcription is represented in ARPAbet symbols. The annotators’ phoneme level agreement had a Fleiss’ kappa value of 0.8685 and had an 86.90% agreement on 9,327 phonemes from 498 sentences. [13]

3.1.2. Mispronunciation detection and feedback generation

The first component, actual phoneme alignment, outputs a phoneme sequence given speech input and text. The actual phoneme decoder uses ERN instead of an unlimited phone loop, which means all possible phoneme sequences. The phoneme alignment on the unlimited phone loop tends to be slow because of the large search space or erroneous because of the pruning. We reduced this problem by replacing the unlimited phoneme loop with the ERN that was generated by our phoneme variant prediction method [14]. The ERN has possible variations of English phonemes that are generally used by Korean speakers of English. Using ERN can reduce the overall phoneme prediction error, and the actual phoneme recognizer can output reliable phoneme decoding results in a shorter amount of time.

The second component, the canonical phoneme decoder, outputs a phoneme sequence given speech input and the text as the actual phoneme alignment. A canonical pronunciation is a reference pronunciation that is in the pronouncing dictionary. We used the CMU pronouncing dictionary to select the correct phoneme transcript for the speech input.

The third component, pronunciation error detection, detects mispronunciations using logistic regression. We used logistic regression because the regression function outputs a real-valued score between zero and one, which can be directly mapped into the decision probability; this probability can be used for CAPT applications. Additionally, the feature functions that were designed are proportional to the mispronunciation probability. Furthermore, it is easy to analyze or control the contribution of each feature using its weight.

The last component, the feedback generation, provides learners’ pronunciation assessment results for input utterances. This component decides whether to provide positive feedback or negative feedback. The feedback generation component compares the confidence of the canonical and actual phoneme alignment results that were generated by canonical and actual phoneme recognizers. The detailed process of determining the feedback decision boundary is discussed in Section 4.

1 The ARPAbet symbol list used in this work can be found in the CMU pronouncing dictionary. Also, http://en.wikipedia.org/wiki/Arpabet provides a parallel representation of ARPAbet to IPA symbols.

2 Available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict, version 0.7a is used in this work.
3.2. Rhythm error feedback

The rhythm error detection and feedback is composed of three main parts: a prediction part, a detection part, and a feedback part (Fig. 2).

From the input sentence, the prediction tool generates a rhythm pattern. The input sentence is analyzed by a part-of-speech tagger (text analysis), and the machine learning features are extracted from the tagged words. From the input speech, the detection part identifies which words are stressed. The input speech is analyzed to extract acoustic parameters (speech analysis). To provide corrective feedback to learners, a comparison is performed between the predicted and detected rhythm patterns. A word gives positive feedback (the sign “O”) or negative feedback (the sign “X”) depending on whether it is correctly or incorrectly uttered with rhythm, respectively. The color of the words represents the rhythmical degree; the closer to the color red, the word is full-stressed and has a full-length vowel; if a word color is black, then the word is not stressed and has reduced vowels, similar to function words.

To build rhythm prediction and detection models, we utilized two types of data: Aix-Machine Readable Spoken English Corpus (Aix-MARSEC) [15] and Korean learners’ English accentuation corpus (KLEAC) [16]. We used Aix-MARSEC corpus to train the rhythm prediction model. Jassem’s narrow rhythm unit (NRIU) notation groups words into rhythmic phrases, and we have considered a stressed syllable that appears in each NRIU for the first time to be a rhythmic word.

KLEAC is used for building the detection model, because non-native English learners have their own prosodic habits when uttering English sentences. To increase the accuracy of the rhythm detection in the learners’ utterances, we adapted acoustic characteristics of non-native learners to the detection model for rhythm. The KLEAC is composed of six hours of speech with 5,500 English sentences produced by 75 native Korean speakers, including orthographic transcription, rhythmic word marks and proficiency labels. The rhythm labels were manually annotated by five phonetic experts. The inter-annotator agreement for rhythms is 87.1% in the KLEAC corpus.

While content words tend to be stressed, function words tend to be unstressed. Several rules for the prediction of rhythmic words were constructed based on the following assumptions: content words are rhythmic words, function words are not rhythmic words, negative auxiliary verbs are rhythmic, and in a sequence of verbs, only the last verb is to be rhythmic. Each rule has its own precedence, to avoid conflicts. However, these rules are not perfect, and erroneous linguistic analysis results can degrade the performance. To compensate for the limitations of the rule-based method, an approach based on machine learning is adopted.

Under the machine learning framework, the existence of rhythm notation for each word can be regarded as a label, and encoded rules with other useful information for classification can be represented into feature vectors. Rhythm labels are affected by surrounding labels, but there is no significant long-distance relationship between them. To reflect these characteristics, we adopted the linear-chain CRF model, which has been widely used in the natural language processing fields [17, 18, 19]. The machine learning features used for the prediction of rhythm are the rules that are mentioned above, part-of-speech tags, and words.

The detection model also adopts the linear-chain CRF model as the prediction model. The detection model utilizes a KLEAC corpus that is different from the prediction model, which utilizes the Aix-MARSEC corpus. This choice was made because the detection model should detect the acoustically unique characteristics of Korean speakers of English. We extracted machine learning features from the acoustic parameters on the middle of the vowel that had the primary word stress. The features are pitch, intensity, duration, and phonetic value.

The corrective feedback for learners is determined by the comparison of predicted and detected rhythm patterns, which is categorized into the following three groups: positive feedback (the sign “O”), negative feedback (the sign “X”), and no feedback (no sign). For the measure that decides the confidence in the feedback, the adjusted score was designed by adopting the output probability of the CRF classifier for each stress label. The adjusted score is calculated by the absolute difference between the probabilities of the predicted and detected rhythm.

3.3. Phrase break error feedback

The phrase break variation detection and feedback is designed to provide appropriate phrase break feedback when a learner utters a given reference sentence. To achieve this goal, the proposed system comprises prediction, detection and feedback provision parts as the rhythm error detection and feedback. If a reference sentence is given without additional information, then the prediction part determines the appropriate breaks, at which phrase breaks can appear with high probabilities. If a learner utters a reference sentence with a microphone connected to the phrase break system, then the detection part determines the positions at which the phrase breaks of the given utterance are imposed with high probabilities. Based on the determined break positions and the probabilities of the prediction and detection parts, the feedback provision part shows positive, negative or no sign at each juncture.

The overall architecture of the phrase break variation detection and feedback is the same as the rhythm variation detection and feedback. However, the data and features are slightly different from the data and features of the rhythm.

We used the Boston university radio news corpus (BURN) [20] annotated with prosodic markers called Tones and break indices (ToBI) [21]. The ToBI framework is one of the most popular schemes for representing the annotation of prosodic events in an utterance. The ToBI break labeling uses indices between 0 and 4 to represent the disjuncture between successive words. In this paper, we handle the ToBI break indices with the coarse mapping manner, which groups the break indices 3 and 4.
into the presence of breaks and the other indices into the absence of breaks.

We used the BURNC corpus for building both the prediction model and the detection model. To achieve an accurate phrase break prediction, we adopt a linear-chain CRF classifier [22]. We use syntactic and lexical features from the given reference sentence: word identity, POS tag, word class, number of syllables/vowels, and punctuation marks. Similar to in the prediction model, the detection model uses the CRF classifier to detect the phrase breaks from the learners’ utterances. The acoustic features are required to distinguish phrase breaks from utterances. To achieve the high accuracy of phrase break detection, syntactic and lexical features are required in addition to acoustic features [17, 19]. The detection model utilizes the features: duration, pitch mean, and intensity mean of the last syllable of each word, silence after the word and the features that are in the prediction model.

4. Error feedback-based scoring

4.1. Feedback proficiency assessment

Feedback is an important point of the proposed system, and it is important because of its pedagogical effect. The learners should receive appropriate and understandable feedback for each of their utterances, to improve their language skills. The proposed system gives feedback that is based on comparing the predicted and detected results for the pronunciation, rhythm and phrase break. As the feedback systems for all of the three parts are the same, we represent the canonical result as the predicted result and the actual result as the detected result in this section.

The feedback system gives three types of feedback: positive, negative, and ambiguous. If the detected result that is extracted from a learner’s utterance is close to the predicted result, as a reference standard, we give positive feedback. If the result is far away from the reference result, we give negative feedback. Otherwise, we give no feedback. For educational purposes, the positive feedback helps to motivate learning, and the negative feedback helps to correct the mistakes of learners. Incorrect feedback, such as false positives and false negatives, however, adversely affect the reliability of the learning system and the learning motivation, also. Therefore, if the comparison result is not trustworthy, then feedback will not be provided.

The feedback decision equation is

\[
\text{Feedback} = \begin{cases} 
\text{Positive, if } |\pi_{\text{pre}} - \pi_{\text{det}}| < \theta_1 \\
\text{Ambiguous, if } \theta_1 \leq |\pi_{\text{pre}} - \pi_{\text{det}}| \leq \theta_2 \\
\text{Negative, if } |\pi_{\text{pre}} - \pi_{\text{det}}| > \theta_2
\end{cases}
\]

where \(\pi_{\text{pre}}\) and \(\pi_{\text{det}}\) are output probabilities that are hypothesized by the prediction and detection models, respectively, and \(\theta_1\) and \(\theta_2\) are the decision boundaries for the feedback signs.

We conducted experiments in which the feedback results of PESAA were compared to human ratings to compute the correlation coefficients, because the threshold values \(\theta\) are not determined and can vary from 0 to 1. Therefore, regarding the threshold values, we calculated every correlation coefficient value to determine the optimum thresholds and the best correlations in the respective parts.

We calculated correlation coefficients by using the method of the Pearson product-moment correlation coefficient. The human ratings and speech resources used in the experiments are derived from the KLEAC, which provides overall pronunciation and fluency ratings for 75 persons, as assessed by five English experts. We obtained the values 0.49(\(\theta_1\)) and 0.50(\(\theta_2\)) for pronunciation, 0.13(\(\theta_1\)) and 0.41(\(\theta_2\)) for the rhythm, and 0.2(\(\theta_1\)) and 0.7(\(\theta_2\)) for the phrase break at the best correlation.

4.2. Scoring method

We assessed the utterances on a scale of 1 to 100, using the feedback information. The scoring equation is

\[
\text{Score} = \frac{\text{Feedback count of "positive" type}}{\text{Entire feedback count except "ambiguous" type}}
\]

Each module gives its score, and the combined score is the average score of the three module scores.

5. Experiments and results

The proposed system is measured in three different areas: its accuracy, the correlation of the assessment results with human assessments, its user satisfaction on expected learning effectiveness and its user interface (UI). These measures evaluate the proposed system in different ways and infer the usability and the appropriateness of the system as an effective CALL system.

5.1. Prediction and detection accuracy

### 5.1.1. Pronunciation module accuracy

<table>
<thead>
<tr>
<th>Phonoeme recognition</th>
<th>unlimited</th>
<th>simulated</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>75.6</td>
<td>82.4</td>
</tr>
<tr>
<td>Correct pronunciation</td>
<td>P_e</td>
<td>92.2</td>
</tr>
<tr>
<td></td>
<td>R_e</td>
<td>78.8</td>
</tr>
<tr>
<td></td>
<td>F_e</td>
<td>84.9</td>
</tr>
<tr>
<td>Mispronunciation</td>
<td>P_e</td>
<td>26.3</td>
</tr>
<tr>
<td></td>
<td>R_e</td>
<td>53.1</td>
</tr>
<tr>
<td></td>
<td>F_e</td>
<td>35.2</td>
</tr>
</tbody>
</table>

There is no point in calculating the prediction accuracy, because in the pronunciation assessment, the canonical phoneme sequence is generated from the pronunciation dictionary. Thus, we did not calculate the prediction accuracy in the pronunciation module; instead, we calculated only the detection accuracy.

We used the data described in Section 3.1.a, which has 12,000 English speech segments from 170 Korean speakers. The segments consist of words, phrases, and sentences. We generated ERN to limit the phoneme search space with speech segments of 136 speakers, which are 80% of the data. We used a random split n-fold cross validation to calculate the detection accuracy. Additionally, we compared the detection accuracy with and without ERN to verify whether limiting the phoneme search
space is effective or not. We measured the performance with a scoring accuracy (SA) [23], which represents the overall correctness of the decisions. In addition, we computed the precision $P$, the recall $R$, and the F1-score $F$. We computed the precision, recall, and F1-score for each of the mispronunciation and correct pronunciation labels.

The mispronunciation detection method that uses simulated phoneme recognition achieved better results than the method that is based on the unlimited phone loop (Table 1). The accuracy is not very high compared to the F1-score of the correct pronunciation detection. The mispronunciation detection rate appears to be low; however, compared to the recent mispronunciation detection work [24], it is not a low number (precision: 32.7%, recall: 62.7%, F1-score: 42.1). Of course, we cannot directly compare the two values, because the used corpora are different and L1-speakers are also different.

### 5.1.3. Rhythm module accuracy

We trained the rhythm prediction and detection models with the Aix-MARSEC and KLEAC corpora, respectively. We measured the precisions, recalls and F1-scores of our models as well as the accuracies using the five-fold cross validation. The numbers indicate how exactly the proposed models can predict and detect the rhythm in given sentences.

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction</td>
<td>96.6</td>
<td>98.3</td>
<td>96.1</td>
<td>97.2</td>
</tr>
<tr>
<td>Detection</td>
<td>81.2</td>
<td>84.0</td>
<td>85.6</td>
<td>84.8</td>
</tr>
</tbody>
</table>

The proposed work demonstrates that the precision, recall and F1-score values in our models are sufficient to predict and detect rhythms (Table 2). The accuracies of the prediction model and detection model are 96.6% and 81.2%, respectively, with higher F1-scores: 97.2% and 84.8%. The prediction model appears to be the most accurate when it is comparing to the detection model’s accuracy or F1-score. Although the values for the detection model are lower than the values for the prediction model, the numbers are not low values in the classification task when using machine learning methodologies; in the proposed method, the machine learning framework is the CRF.

### 5.1.4. Phrase break module accuracy

We evaluated the prediction model and the detection model of the phrase break module using five-fold cross validation. We used the BURNC corpus for both the prediction and detection model.

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction</td>
<td>90.2</td>
<td>84.0</td>
<td>80.1</td>
<td>82.0</td>
</tr>
<tr>
<td>Detection</td>
<td>91.6</td>
<td>86.1</td>
<td>83.7</td>
<td>84.9</td>
</tr>
</tbody>
</table>

We measured the precision, recall and F1-scores of the models as well as the accuracy (Table 4). To measure the accuracy and F1-score, we used binary classification. The accuracy, precision, recall and F1-score values are considered to be sufficient to accurately predict and detect phrase breaks in our classification tasks.

### 5.2. Automatic assessment

We compared the assessment results of PESAA with human assessments. We used KLEAC data to compute the correlation, because the KLEAC corpus was annotated with proficiency labels. The KLEAC corpus has three types of proficiency labels that five phonetic experts marked: pronunciation, fluency, and overall level. The experts labeled each criterion with a scale of 1 to 5, where 1 is the lowest score and 5 is the highest score. The pronunciation criterion involves accurate pronunciation; the fluency criterion involves the rhythm, break, speed and intonation; the overall criterion is how much the utterances are native-like or not.

The correlation between the pronunciation scores of the system and human assessments was 0.41. The correlation itself is not high; however, the value is not small compared to the results of a previous study [25], which was 0.41. This value represents that the pronunciation feedback by the system is not similar to a human’s assessment. The previous experiments of the pronunciation error detection showed quite good results (82.4% of accuracy); thus, we can infer that the assessment criteria of the pronunciation part are not the same as a human’s assessment criteria.

The correlation values between the prosody score of the system and the fluency score are 0.64 for rhythm and 0.74 for phrase break. We can show that we can judge a person’s fluency in English with only the prosody features. This capability means that the prosodic features are important in language learning, especially in language speaking learning.

### 5.3. User satisfaction

We evaluated the user satisfaction in two ways: feedback satisfaction and UI satisfaction.

#### 5.3.1. Feedback

We designed an experiment for eight Korean university students to utilize the system, to evaluate whether PESAA can be applied in real learning. The students utilized PESAA for an hour, and we collected questionnaires for expected learning effectiveness on a scale of 1 to 5, with positive adjectives anchoring the high end and negative adjectives anchoring the low end.
Table 4. Learners’ questionnaires for expected learning effectiveness on a scale of 1 to 5.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Mean</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do you know about pronunciation, rhythm, and phrase break? (average)</td>
<td>2.79</td>
<td>0.40</td>
</tr>
<tr>
<td>Do you care about pronunciation, rhythm, and phrase break? (average)</td>
<td>3.04</td>
<td>0.68</td>
</tr>
<tr>
<td>Post-test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the system help you to understand about pronunciation, rhythm, and phrase break? (average)</td>
<td>3.25</td>
<td>1.05</td>
</tr>
<tr>
<td>Does the system point correctly to your problems in pronunciation, rhythm, and phrase break? (average)</td>
<td>3.67</td>
<td>0.68</td>
</tr>
<tr>
<td>Does the system help you to improve English proficiency?</td>
<td>3.75</td>
<td>0.19</td>
</tr>
<tr>
<td>Are you interested in using this system to improve your English skills?</td>
<td>3.63</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table 4 shows that the students were not aware of pronunciation, rhythm, or phrase breaks very much (2.79). However, they cared about their English when they were speaking (3.04). These findings mean that the students attempted to speak English intelligibly, even if they did not completely understand the English pronunciation or prosody. After the use of the proposed system, they answered that they understood pronunciation, rhythm, and phrase breaks. The effectiveness of our system was deemed to be good, with many of the students answering that our system helped them to improve their English proficiency (3.63). Additionally, the students answered that they would use the system to improve their English skills. Two of them absolutely agreed (5) with the question, and all of the others were positive at using the system (3 and 4), except for only one user (2). The students appeared to obtain knowledge about pronunciation and prosody, even when the time that they were on the system was as short as an hour. The students were satisfied with our system and reported some improvement in their English skills or English knowledge.

5.3.2. Questionnaire for user interaction satisfaction

As a qualitative evaluation, the subjective feelings of the testers were surveyed with a questionnaire for user interaction satisfaction (QUIS) style usability evaluation. The QUIS of [26] was created to gauge the satisfaction aspect of the software usability in a standard, reliable, and valid way. QUIS focuses on the user’s perception of the usability of the interface as it is expressed in specific aspects of the interface (i.e., overall reaction to the system, screen factors, terminology and system feedback, learning factors, and system capabilities). Each of the specific interface factors and optional sections has a main component question followed by related sub-component questions. Each item is rated on a scale from 1 to 5, with positive adjectives anchoring the high end and negative adjectives anchoring the low end.

We used the short form of QUIS 5.0 to evaluate the usability of PESAA (Fig. 3). The sections of original QUIS were retained, but some items that were not appropriate were dropped. To evaluate PESAA, we designed an experiment for 18 Korean university students attending an English class who utilized the implemented CALL system as a tool for the class, to aid in learning.

Table 5. Learners’ questionnaires for user interface satisfaction on a scale of 1 to 5 for PESAA. Each major item comprises 4-6 detailed questions.

<table>
<thead>
<tr>
<th>Major items for satisfaction</th>
<th>Mean</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall system</td>
<td>3.72</td>
<td>0.51</td>
</tr>
<tr>
<td>Display content</td>
<td>3.54</td>
<td>0.51</td>
</tr>
<tr>
<td>Terminology</td>
<td>3.48</td>
<td>0.50</td>
</tr>
<tr>
<td>Easy use</td>
<td>3.72</td>
<td>0.36</td>
</tr>
<tr>
<td>Processing speed</td>
<td>3.55</td>
<td>0.56</td>
</tr>
<tr>
<td>Total</td>
<td>3.60</td>
<td>0.49</td>
</tr>
</tbody>
</table>

We asked the experiment participants to utilize the system for 30 minutes per a day over three weeks. After the three weeks, we collected QUIS answering sheets from the participants. According to the learners’ QUIS, the overall satisfaction score on 1-5 was 3.60, with a standard deviation (s.d.) of 0.49 (Table 5). This score can be considered to mean that PESAA is meaningfully useful in real English learning and that it has an appropriate user interface that helps the learners to understand their English assessment results easily.

6. Discussion and conclusions

In this paper, we described a computer-assisted language learning system for non-native English learners, especially Koreans, to improve their overall language skill. We designed a system that can assess a learner’s pronunciation and prosody. The system has three error feedback modules: pronunciation, rhythm, and phrase break error feedback modules. Each module has a prediction and detection model, and when comparing their results, the module generates appropriate assessments with regard to feedback of errors.

Through three types of evaluation, we found PESAA to be a useful CALL system. PESAA effectively predicts and detects pronunciation and prosody features. The accuracy of each module is more than 80%. Additionally, PESAA assesses a learner’s utterances more similar to the way that a person would assess a learner’s utterances. The correlation between human annotators of fluency and the assessment results from PESAA reached 0.58 on average. The learners that used our system...
reported that the system helped them to improve their English skills and that the UI was satisfactory.

7. Acknowledgements

This work was supported by the Industrial Strategic technology development program 10035252, Development of dialog-based spontaneous speech interface technology on mobile platform funded By the Ministry of Trade, industry & Energy(MI, Korea). This work was supported by the MKE (The Ministry of Knowledge Economy), Korea and Microsoft Research, under IT/SW Creative research program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-H0503-1201-1002).

8. References