Automatic Speech Recognition: Reliability and Pedagogical Implications for Teaching Pronunciation

In-Seok Kim
Department of English Language, Dongduk Women's University, Seoul, Korea
iskim@dongduk.ac.kr

ABSTRACT

This study examines the reliability of automatic speech recognition (ASR) software used to teach English pronunciation, focusing on one particular piece of software, *FluSpeak, as a typical example*. Thirty-six Korean English as a Foreign Language (EFL) college students participated in an experiment in which they listened to 15 sentences that appeared in *FluSpeak* and recorded their voices, repeating sentence by sentence. The ASR software analysis of their production was then compared to pronunciation scores determined by native English speaking (NES) instructors. Although the correlation coefficient for intonation was nearly zero, indicating that ASR technology is still not as accurate as human analysis, the software may be very useful for student practice with aspects of pronunciation. The paper suggests a lesson plan for teaching English pronunciation through ASR software.

Keywords

Automatic speech recognition (ASR), English pronunciation, Intonation, Pedagogical implications, Software, Pronunciation training

Introduction

Computer-based pronunciation training has emerged thanks to developments in automatic speech recognition (ASR) technology (described in technical detail in Witt & Young, 1998; Herron, *et al*., 1999; Menzel, *et al.* 2001; Beatty, 2003). However, even as foreign language teachers become increasingly aware of the advantages of using ASR software, they have become concerned with the reliability of machine-scored pronunciation.. This concern stems from their belief that a high degree of agreement should be obtained between automatic and human scores. Finding a high degree of correlation between the two would increase the use of ASR software for pronunciation training. Coniam (1999, p.49; 1998) has already suggested the development of an assessment tool, such as a reading aloud test via voice recognition technology; that is, students read aloud sentences that are scored by the voice recognition software.

The prime purpose of this pilot study is to determine the correlation coefficient between the pronunciation scores of one automatic speech recognition software, *FluSpeak*, and those of NES instructors, using Korean EFL college students as subjects. To this end, this paper will undertake three tasks. First, it will briefly overview the architecture of ASR in pronunciation software, including *FluSpeak*. Second, it will describe the experimental procedures employed to determine the correlation coefficient between the scorings *of FluSpeak* software and those of NES instructors and analyze the data used to determine correlations among the several variables. Finally, it will suggest the pedagogical implications for effectively teaching pronunciation with ASR software, both in the classroom instruction and for self-training. The paper concludes by discussing future directions in determining more accurately the correlation between ASR and human scores.

Architectures and Features of ASR

ASR is a cutting edge technology that allows a computer or even a hand-held PDA (Myers, 2000) to identify words that are read aloud or spoken into any sound-recording device. The ultimate purpose of ASR technology is to allow 100% accuracy with all words that are intelligibly spoken by any person regardless of vocabulary size, background noise, or speaker variables (CSLU, 2002). However, most ASR engineers admit that the current accuracy level for a large vocabulary unit of speech (e.g., the sentence) remains less than 90%. Dragon's *Naturally Speaking* or IBM's *ViaVoice*, for example, show a baseline recognition accuracy of only 60% to 80%, depending upon accent, background noise, type of utterance, etc. (Ehsani & Knodt, 1998). More expensive systems that are reported to outperform these two are *Subarashii* (Bernstein, *et al*., 1999), *EduSpeak* (Franco, *et al.,* 2001), *Phonepass* (Hinks, 2001), *ISLE Project* (Menzel, *et al*., 2001) and *RAD* (CSLU, 2003). ASR accuracy is expected to improve.

Among several types of speech recognizers used in ASR products, both implemented and proposed, the Hidden Markov Model (HMM) is one of the most dominant algorithms and has proven to be an effective method of dealing with large units of speech (Ehsani & Knodt, 1998). Detailed descriptions of how the HHM model works go beyond the scope of this paper and can be found in any text concerned with language processing; among the best are Jurafsky & Martin (2000) and Hosom, Cole, and Fanty (2003). Put simply, HMM computes the probable match between the input it receives and phonemes contained in a database of hundreds of native speaker recordings (Hinks, 2003, p. 5). That is, a speech recognizer based on HMM computes how close the phonemes of a spoken input are to a corresponding model, based on probability theory. High likelihood represents good pronunciation; low likelihood represents poor pronunciation (Larocca, et al., 1991).

While ASR has been commonly used for such purposes as business dictation and special needs accessibility, its market presence for language learning has increased dramatically in recent years (Aist, 1999; Eskenazi, 1999; Hinks, 2003). Early ASR-based software programs adopted template-based recognition systems which perform pattern matching using dynamic programming or other time normalization techniques (Dalby & Kewley-Port, 1999). These programs include *Talk to Me* (Auralog, 1995), the *Tell Me More Series* (Auralog, 2000), *Triple-Play Plus* (Mackey & Choi, 1998), *New Dynamic English* (DynEd, 1997), *English Discoveries* (Edusoft, 1998), and *See it, Hear It, SAY IT!* (CPI, 1997). Most of these programs do not provide any feedback on pronunciation accuracy beyond simply indicating which written dialogue choice the user has made, based on the closest pattern match. Learners are not told the accuracy of their pronunciation. In particular, Neri, *et al*. (2002) criticizes the graphical wave forms presented in products such as *Talk to Me* and *Tell Me More* because they look flashy to buyers, but do not give meaningful feedback to users. The 2000 version of *Talk to Me* has incorporated more of the features that Hinks (2003), for example, believes are useful to learners:

- A visual signal allows learners to compare their intonation to that of the model speaker.
- The learners' pronunciation accuracy is scored on a scale of seven (the higher the better).
- Words whose pronunciation fails to be recognized are highlighted.

More recent ASR programs that have adopted HMM include *Subarashii* (Entropic HTK recognizer used), *VILTS* (SRI recognizer), *FLUENCY* (Carnegie Mellon University SPHINX recognizer), *Naturally Speaking* (Dragon Systems), and *FluSpeak* (IBM ViaVoice recognizer). Those interested in more detailed technological descriptions of each ASR program may refer to Holland (1999) and other articles in the *Calico Journal*, Special Issue, Vol. 16 (1999). *FluSpeak* (MT Comm, 2002a), which was used in this study, will be described in more detail in an attempt to show how HMM based programs are built and how they score learners' pronunciation.

FluSpeak is divided into four types of practice: English Pronunciation Practice with consonants, consonant clusters, vowels, and diphthongs; Intonation Practice; Dialogue Expressions Practice; and a Pronunciation Test that covers the Pronunciation and Dialogue activities. Students can listen to sounds or words with an animated video clip showing a native speaker's mouth and tongue movements. They can then record their voice repeating the sounds or words. When their pronunciation is recognized by the software, their pronunciation accuracy is displayed in a spectrogram on the screen, so that they can compare their pronunciation to native speech visually (see Figure 1).

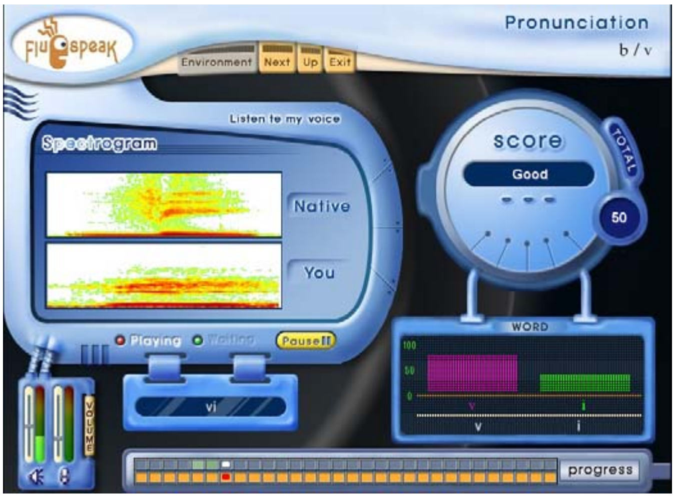
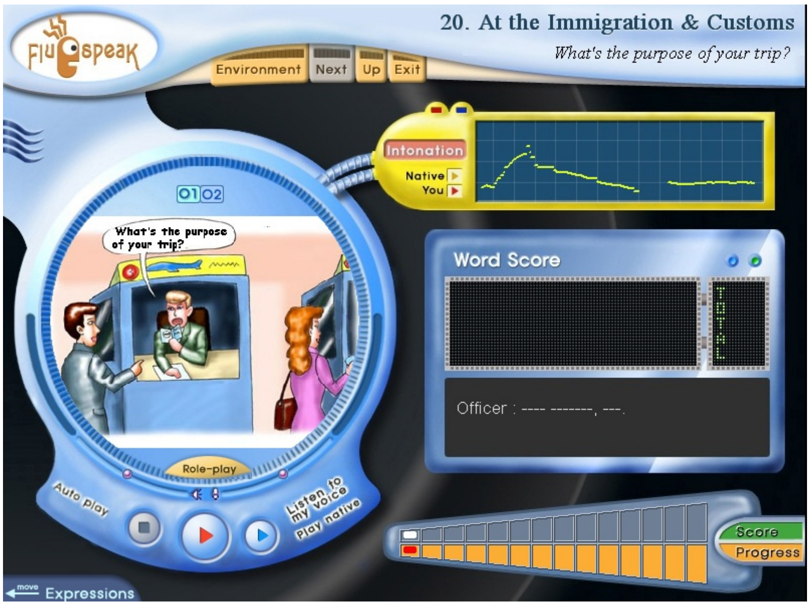


Figure 1. *FluSpeak*: Spectrogram (MT Comm, 2002) [Used with permission]

Besides the video of the native speaker’s face and the spectrogram, visual aids include a graphical display of the vocal tract and speech wave forms.

For Intonation Practice, students listen to the native speaker's pronunciation of a sentence, while seeing the intonation curve shown in yellow on the screen. They then repeat after the native speaker, and the intonation curve of their own pronunciation is visually displayed in red with a score, so that learners can compare their own intonation with the model (see Figure 2, box in upper right).

Dialogue Expressions Practice consists of 40 units of beginning through intermediate dialogues that would be used in ordinary conversations. Students listen to the speaker's model pronunciation as often as they want before recording into the program. When the accuracy of their pronunciation reaches a minimum level of recognition, scores for all individual words in the utterance and their averaged score are displayed on the screen. Until the threshold level of recognition is reached, a voice signal “please try again” is heard. Learners can role-play with the computer by clicking on a button. During practice at this stage, sentences are not displayed in writing to prevent learners from relying on reading the sentences~~,~~ while repeating them. Learners can also check their performance during practice when each unit is completed. Cartoons depicting the context of the dialogue are displayed in the clock-shaped control panel at the left of the screen (see Figure 2).


20. At the Immigration & Customs

What's the purpose of your trip?

Officer : ---- -------, ---.

Figure 2. FluSpeak: pronunciation Practice.. FluSpeak (MT Comm, 2002) [Used with permission]

FluSpeak's automatic pronunciation scoring system -was developed using 30,000 full phonemic models, based on the utterances of 1,500 native speakers of American English and 200 English learners, and on acoustic information about 200,000 words and their pronunciation variants.

The ARS operates according to a proprietarial recognition algorithm that MT Comm (2002a)developed over many years. According to a research report released by MT Comm (2002b) these phonemic models realistically represent the system of American pronunciation since they are abstracted from the utterances of a large number of native speakers and have proven to be 99.9% reliable in measuring the pronunciation of native English speakers who did not participate in building the original database.

FluSpeak measures the percentage of similarity of the L2 speaker's intonation pattern to the native speaker's intonation pattern. The software searches for the most probable phoneme to best match an utterance, based on the database of 30,000 phonemic models and 200,000 lexical entries. *FluSpeak* computes the score, based on the percentage of a learner's pronunciation accuracy compared to the model, plus a reliability point. However, since intonation refers to low or high pitches in a sentence, it is extremely difficult to register intonation using only a sound wave. In order to do so, abstracting the pitch of the sound signal is necessary. The value of pitch in a high tone is small (narrow), whereas that in a low tone is large (wide). One can compute the vibration value inversely. An additional problem arises in that the value of pitches varies depending on the individual speaker. Thus, it is difficult to determine the degree of correct intonation for different people based on the absolute value of pitch in a sample. To solve this problem, *FluSpeak* does not determine the absolute value of pitch change, but rather it computes the relative change of pitch determined by the length of a sound.

Methodology

Subjects

Thirty-six students enrolled in the author's intermediate General English Division conversation course for the fall semester of 2003 participated in this experiment. Students were admitted to the course as a sequel to a beginning English conversation course which they took in the spring semester of 2003. These students were enrolled in this General English Division course, required for all freshmen, solely based on completion of prior coursework.

Procedures for Collecting Data

Warm-up session

The experimental session used for the study was preceded by a warm-up session during which students were asked to familiarize themselves with the *FluSpeak* software and try it out on their own for twenty minutes at the multimedia lab of the author's university. They practiced repeating the sentences, recording them into the program, listening to their own voice, and comparing their voice with that of native speakers by seeing the scores and intonation curves displayed on the screen. The purpose of the warm-up session was to eliminate beforehand any technical difficulties that might arise from using the software.

Experimental session

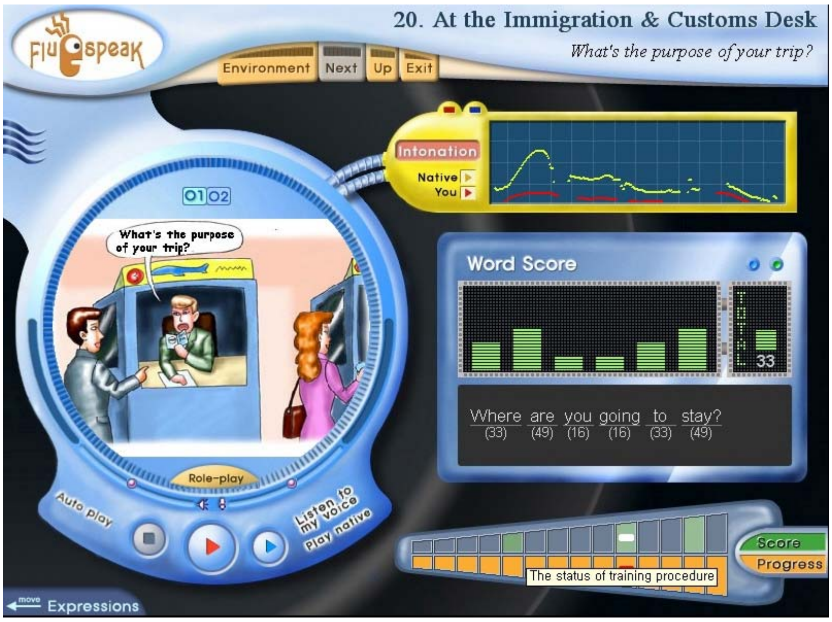
Once students were familiar with how to use the *FluSpeak* software, they were asked to glance through a list of 15 sentences on a sheet. They were told that they should read these sentences, one by one, as naturally as possible once after they heard the native speaker's voice on the program. For having their speech recorded they were told to do the following: 1) They click a yellow square button that appears in a row at the bottom of Figure 3. The first yellow square button plays the first sentence. And they then click the play button on the screen of the program. 2) They look over the sentence on the list given, while listening to it on the program. They repeat the sentence they hear as many times as they want. 3) Once they feel comfortable with repeating the first sentence, they click the second yellow square button and listen to the second sentence by clicking the play button. At this stage students' repeating of the first sentence is automatically recorded.

During the warm-up session the author explained to students that they should read the sentences as connected utterances and as closely as possible to the way that *FluSpeak* read~~s~~ them. Raters were also made aware of the fact that students were supposed to read sentences as connected speech .

A special arrangement was made with the *FluSpeak* producers for students' recordings to be saved into a local file since in the commercial product students' utterances cannot be saved. The experiment used all new sentences unfamiliar to the students but similar to those they had studied during previous coursework. The scores for individual words, bar graphs, and intonation curves were displayed to students as in Figure 3, but only the numeric scores were saved for the purposes of the experiment. Thus, students were able to learn as the experiment proceeded, just as they would have in a normal session with the software. During the experiment students were allowed to practice reading the sentences several times and to save the attempt which they thought was the best. There was no specific time limit set, but the total average time taken for reading the15 sentences turned out to be 30 minutes.

Retrieval of students' speech samples and FluSpeak scores

Thirty-three student utterances (the total completing the study) stored in the experimental file were retrieved and saved onto an audio CD. The remaining three student samples were discarded since these subjects did not complete the entire experiment. As explained earlier, the *FluSpeak* program rates English learners' speech by the accuracy of words within a sentence and their intonation, each on a 100 point scale. These two scores for all speech samples were retrieved and averaged into a mean score to see how the *FluSpeak* scores would correlate with native speakers' holistic ratings. Intonation scores were also retrieved separately in order to determine the correlation coefficient between human scores and those judged by *FluSpeak*.



Where are you going to stay?
(33) (49) (16) (16) (33) (49)

Figure 3. The scores for individual words, bar graphs, and intonation curves displayed to students

Development of the rating scale and scoring of students' speech samples by NES instructors

A review of L2 phonology studies reveals that the holistic rating of fluency is the preferred method because of the difficulty of quantifying prosodic deviance in terms of discrete phonetic errors. Witt and Young (1998) define pronunciation quality as based on both phonetic and prosodic features. They believe that for beginners, phonetic characteristics are of greater importance because these cause mispronunciations. But they stressed that as learners' fluency increases, more emphasis should be given to teaching prosody, i.e., intonation, stress and rhythm (p. 26). The author had to develop a new pronunciation rating scale for this experiment since those existing in the literature did not fit the purpose of this study.

Holistic rating of English learners' pronunciation is reported in a number of studies (Adams, 1979, Prator & Robinet, 1985; Wong, 1987; Anderson-Hsieh & Koehler, 1988; Morley, 1991; Anderson-Hsieh, *at al*., 1992; Koren, 1995; Munro & Derwing, 1995; Derwing & Munro, 1997). Most researchers attempt to measure the raters' overall impression of stress, rhythm, phrasing, and intonation. This study adopted a four-point scale without midpoints between the whole numbers, as shown in Appendix B.

The evaluation method (see Appendix B) that the author adopted in this paper consists of a four point scale, ranging from 0 to 3 without midpoints allowed in-between. Points 0 indicates least-native like fluency in that the speaker's pronunciation entails many pronunciation errors and foreign accents and intonation patterns, which makes the listener get lost. In contrast, Point 3 indicates most native-like fluency in that the speaker's pronunciation entails occasional pronunciation errors, but makes her clearly understood. Point 1 indicates the existence of frequent pronunciation errors and foreign accents and intonation patterns that make the speaker's pronunciation somewhat difficult to understand. But Point 2 indicates the existence of some consistent pronunciation errors and foreign accents and intonation patterns, and yet does not make the speaker's pronunciation understandable with some efforts to listen.

Student speech samples were rated by three NES instructors who were teaching at the author's university. These American instructors (two males and one female) possessed higher degrees in TESOL or Linguistics from American universities. They have been teaching English at the General English Division of the university for some five years on the average, so they are familiar with their students' pronunciation. They were given instruction as to how to rate student utterances using a 4-point rubric and did some preliminary rating practice with speech samples before they scored the student productions for this study. Each of the three raters were given an audio CD that contained student speech samples, and were asked to score the speech samples independently, based on the 4-point scale. They were then asked to write points in the column of the scoring sheet. When they rated students pronunciation, they were asked to remain consistent across all items. Especially when they felt hesitant to rate speech samples, they were asked to listen to them several times and come up with a most accurate rating.

The Pearson correlation coefficient was calculated for the set of three scores thus produced for each speech sample. The ratio (*r*=0.49) turned out to be not very high, indicating that the ratings were not consistent across the three scorers. The instructors were then asked to rate the speech samples again, this time discussing the scores when they were not within one point of each other on the rating scale until they arrived at closer agreement. The Pearson correlation coefficient for these adjusted scores went up to 0.69, which means that the scores determined by this adjustment process were relatively reliable. The final scores of all three raters on the experimental data were subsequently averaged into mean scores.

Integration of scores for cross-comparison

FluSpeak scores for word and intonation and their averaged totals were recorded into the table of raw data for ananlysis.

Data Analysis and Discussion

The *WinSPASS* program (SPSS Korea, 2002) was run on the four kinds of pronunciation ratings appearing in Appendix C to calculate correlation coefficients between *FluSpeak* ratings and those of native instructors. Correlation coefficiency between two variables for each speaker was computed in order to assess their consistency.

Correlation between FluSpeak ratings and those of NES instructors at the speaker level

Mean scores for words and intonation, as rendered by the *FluSpeak* program, are juxtaposed with those of native instructors per speaker (Appendix C, last row). The mean score at the word level judged by *FluSpeak* is 46.3 indicating poor pronunciation (less than 50% accuracy), whereas that of the NES teachers is 1.9, or only a little less than 50%. The correlation coefficient between *FluSpeak* scores and NEI's scores at the speaker level is 0.56 (p<0.01), that is, not very high, indicating a mediocre correlation between the two types of scores (see Table 1).

Table 1. The correlation coefficient between *FluSpeak* scores and NEI scores

Variable	**Number**	**Mean**	**SD**	**Correlation Coefficient**
FluSpeak Scores	33	46.3	9.3	***r*=0.56 (p<0.01)**
NEI Scores	33	1.9	.41	

However, as Table 2 demonstrates, the correlation coefficient between *FluSpeak* scores for intonation and NEI scores for general fluency is extremely weak (0.06, p>0.05). This indicates that the accuracy of judging intonation by *FluSpeak* may not be reliable at all.

Table 2. The comparison between *FluSpeak* scores for intonation and NEI scores for general fluency

Variable	Number	**Mean**	**SD**	**Correlation Coefficient**
FluSpeak Scores	33	70.4	13.7	***r*=0.06 (p>0.05)**
NEI Scores	33	1.9	.41	

One major reason for this weak correlation originates in the varying pitch uttered by different people. Since *FluSpeak* measures the percentage of the similarity of L2 speakers' intonation pattern to that of the native English speakers, whose pitch varies naturally, one may assume that its score is likely to be unreliable. The way in which *FluSpeak* measures pitch contributes to its mean score (70.4) for intonation being much higher than that for the word scores (46.3). Another reason which seems to contribute to the unreliability of the *FluSpeak* scoring system is related to construct validity. *FluSpeak* intonation is calculated on an algorithm which measures the intonation accuracy of students' speech samples. However, native speakers' scores are based on evaluating the overall fluency of students' speech, but not just their intonation. Thus, failure to establish any relationship between these two variables is not surprising since a comparison is being made between two scores that measure different objects. One way in which we can solve this kind of problem is to compare *FluSpeak* intonation scores with those that come from humans judging intonation only (if that is possible), a test which was not done in this study.

One may cautiously conclude that the reliability of the *FluSpeak* scoring system is only moderate at the speaker level (0.56). This conclusion appears to be supported by the performance reports for software in several studies which looked at the reliability of their ASR engines. Franco, *et al.* (1997) and Kim, *et al*, (1997) report on correlations at three different levels between human scores and machine scores created by Stanford Research Institute's ASR engine: 0.44 correlation between these two variables at the phoneme level, 0.58 at the sentence level, and 0.72 at the speaker level using 50 sentences per speaker. Ehsani and Knodt (1998) compare machine-human correlations with correlations between human graders: 0.55 at the phoneme level, 0.65 at the sentence level, and 0.80 at the speaker level. Another set of studies with similar results include Rypa and Price (1999), who report a comparable relationship at the sentence level between human-machine score correlations (*r*=0.60 on data drawn from the *VILTS* system they developed) and human-human correlations (0.68) between human scores and those of ASR software reported in Niemeyer, *et al.* (1998). A recent study done by Machovikov, et al., (2002) reflects more or less the same degree of correlation between experts' rating and the ASR system's for speaking 10 Russian digits (approximately 73%). See Table 3 for a comparison of these correlations.

*Table 3***.** Aspects of correlations between human and machine pronunciation scoring

Variable	Franco (1997); Kim et al. (1997): human-machine	Ehasani & Knodt (1998): human-human	Rypa & Price (1999); Niemeyer et al. (1998): human-machine	Rypa & Price (1999): human-human	Machovikov et al. (2002): human-machine
phoneme level	0.44	0.55			73%
sentence level	0.58	0.65	0.60	0.68	
speaker level	0.72	0.80			

The findings of these studies lend support to a belief in the reliability of the *FluSpeak* scoring system despite the apparently low correlations obtained in this small-scale study. The correlation coefficient of *FluSpeak* at the speaker level (*r*=0.56) runs considerably lower than that of SRI's system (*r*=0.80; Ehasani & Knodt,1998), but is comparable to that of *VILTS* (*r*= 0.60; Rypa & Price, 1999). A somewhat low correlation score by *FluSpeak* at the speaker level leads to the speculation that it may be more vulnerable to the idiosyncratic nature of the speaker's pitch. This is a subject for further investigation.

There are other studies that look at other aspects relevant to the reliability of ARS scoring systems. Bernstein (1997) claims that the correlation level between machines and human graders can be as high as 0.85, based on a study conducted at Entropic with 20 to 30 short sentences per speaker. However, one may argue that such a correlation may not be a realistic goal once the system deals with longer sentences, *e.g.*, those consisting of more than five words. A case in point is *PhonePass* (Ordinate, 2000; Berstein & Christian, 1996) which demonstrates the highest correlation coefficient (0.94) between human and machine scoring so far (288 non-native speaker subjects). However, the majority of utterances tested in *Phonepass* are words or phrases. Even when sentences are used, most of them are relatively short and loaded with somewhat easy words. In contrast, utterances tested in *FluSpeak* are relatively long, consisting of up to 12 words in dialogue form, and-furthermore, are loaded with multi-syllabic vocabulary items such as *landing card*, *purpose*, *declaration*, *belongings*, and *agricultural products*. It may be assumed that the greater number of words per sentence and their level of difficulty explains the difference in correlation ratios between *FluSpeak* and other products. This assumption is supported by the result that *FluSpeak* scores for longer sentences (*i.e.*, sentences 1, 3, 7, 10, and 11 in Appendix A) with multi-syllabic vocabulary turned out to be considerably lower than scores for shorter easier sentences. One might assume that an increase in the number of utterances per speaker would increase the correlation coefficient further. However, a decrease in the number of words per sentence may be a more powerful indicator.

The Setting of the Recognition Accuracy Rate

Another aspect of ASR scoring is the setting of the recognition accuracy rate. Ehsani & Knodt (1998) report that certain software, such as *Subarashii* (a Japanese language learning software), is built with a relatively low recognition accuracy rate in order to forgive students' accents. According to their two trial experiments, using 45 students studying Japanese as a foreign language, recognition accuracy rates turned out to be extremely low (46% for the high school version and 36.6% for the Stanford university version). Naturally, the functional accuracy scores reported by the program in each case turned out to be relatively high (66.9% and 71.4% respectively). Ehsani & Knodt argue, however, that near perfect recognition accuracy may not be a necessary requirement for an effective speech or dialogue teaching system (1998, p. 55). However, the claim that recognition accuracy should be lowered at the expense of correcting faulty pronunciation does not appear to have face validity as a pedagogically sound approach. In fact, many ASR products allow users to adjust the difficulty level of sound recognition, depending upon the level of their expectations. Thus, a teacher could adjust the accuracy level very low so as not to discourage beginners, or raise it very high to work with advanced students.

Pedagogical Implications for Teaching English Pronunciation

Teaching pronunciation to EFL students at a low level can be a labor-intensive task for EFL instructors, especially when their classes have 30 to 40 students with a diverse range of proficiency levels. However, ASR pronunciation software such as *FluSpeak* can be used effectively in conjunction with live classroom teaching to develop oral skills. The author has taken the following four steps to blend live teaching with self-training in pronunciation for students enrolled in an intermediate English conversation course.

Step 1: Choral repetition of each sentence after the speaker in FluSpeak software

This step is a tuning-up session where instructors let students know what they are going to learn. If necessary, instructors explain the meanings of key words and useful expressions that need special attention. Students repeat after the model speaker on the software. During this time they are allowed to look at the sentences in the book.

Step 2: Self-training initiated by students

Once students have established some degree of familiarity with the target sentences in class, they can spend more time with the software in the lab, working sentences that are problematic for them individually. When students see the score of their own pronunciation on the screen, they have good reason to try again to reach a higher score. This motivation makes students stick to self-training and use the software for a longer period of time. One teaching tip is that adequate lab time should be allocated for students' self-training with the software. Their practice recordings are, of course, kept in the program file for the instructor's review.

Step 3: Instructor's Q & A session

An instructor takes up a whole class session to practice the dialogue student by student or in chorus. By this time students should feel somewhat confident with speaking the sentences since they have self-trained with them on the ASR software. The instructor asks the question in the dialogue and students respond individually or in group. During this time various other skills such as reading aloud can be practiced.

Step 4: Student pair practice.

Once students are ready to use the sentences in Steps 1 through 3, pairs of students sit near each other and take turns reading the dialogue sentences to their partners.

Step 5: Students' simultaneous repetition with the model pronunciation on the software

In a subsequent lab session students try to repeat the sentences almost simultaneously with the model speaker. At this step students are encouraged not to look at the script of the sentence they are repeating. This is the point where the fluency they worked on during the previous steps becomes evident.

Step 6: Role-play session and other creative skill-using activities

Students are given an opportunity to role-play the dialogues in front of the class without looking at the script. Other creative skill-using activities include making up new dialogues based on the one they learned and pair or group presentation in front of the class.

The lesson plan above exemplifies the case where ASR pronunciation software leads to communicative skills. In the author's experience students feel more confident with speaking in class when they have practiced pronouncing the sentences privately. Also, were instructors to spend much time drilling students with pronunciation of the basic sentences in the dialogue, which is often the case, they would not have a reasonable amount of time to provide the opportunity for communicative practice. Furthermore, instructors tend to agree that this type of pronunciation drill is not always as successful as it should be and rarely can be adequately individualized. Thus, ASR use has two advantages: (1) students feel more confident in their speaking skill with individualized ASR-based training, and (2) human instructors can plan on more motivating communicative activities if they leave the low-level basic pronunciation drills to the ASR software.

At the end of the class that the author taught during this study, he took a survey to determine students' reactions to this ASR-based pronunciation class. The survey showed that an overwhelming number of students (90%) reacted positively to the question, "Do you think that *FluSpeak* helps you improve your English in general ?" Their response to the question "Do you think that *FluSpeak* helps you improve your pronunciation?" was slightly lower (86%), and yet still highly favorable, indicating that students perceived an educational benefit from using the software. However, only 30 % of students answered favorably to the question, "Do you think that the pronunciation and intonation ratings of *FluSpeak* are accurate?" This indicates that students tend to discredit the *FluSpeak* software as a reliable tool for evaluating their pronunciation. One of the reasons for their apparent discontent with this aspect of the software might have something to do with the low pronunciation scores the software gave them. Several students complained about low pronunciation scores from *FluSpeak*, even though their pronunciation seemed to be above average as far as the author could judge them. Their idiosyncratic pitch may have been the culprit.

Conclusions and Future Directions

This study attempts to determine the correlation coefficient between the scores of ASR software and those of human raters, and to speculate on some of the reasons for apparent discrepancies between human and machine raters. An analysis of the experimental data using 36 Korean EFL university students reveals that the correlation coefficient is not high at the word level and near zero at the intonation level. These results hammer home the fact that the present state of technological development falls far below the desired level of accuracy. This, nevertheless, does not indicate that there is no value in adopting ASR software for pronunciation instruction in and out of the EFL classroom. Despite the fact that ASR software has its own limitations, as evinced in this study, it can be used as a valuable tool for teaching pronunciation to EFL students where NES instructors are not readily available. Related to this, the paper has addressed pedagogical and design implications for teaching English pronunciation to EFL students.

To extend this study and validate the pronunciation scoring system of *FluSpeak*, or of any other ASR software for that matter, two research directions are conceivable. First, an experiment with the word recognition error rate of *FluSpeak* can be conducted to determine the accuracy of its scoring system more precisely. Word recognition error rate is widely used as one of the most important ARS measures (Jurafsky & Martin, 2000), and can be computed by a relatively simple formula, *i.e.*, divide the words substituted for correct words, plus words deleted or inserted incorrectly, by the actual number of correct words in the sentence:

$$\text{Word Error Rate (\%)} = \frac{\text{Substitutions + Deletions + Insertions}}{\text{No. of words in the correct sentence}}$$

(Jurafsky & Martin, 2000, p.420)

In this formula, the lower the percentage, the more accurate the word recognition system. As a rule of thumb when the word error rate exceeds 10%, a new algorithm needs to be developed (Jurafsky & Martin, 2000). It would be interesting to compare the correlation coefficient determined in this type of experiment with the word recognition error rate of the ASR software explored above. Second, an experiment is needed to determine whether ASR software accurately judges native speaker pronunciation. If the recognition rate of *FluSpeak* reaches somewhere around 80% for native speakers, one can say that its scoring system is highly reliable.

In conclusion, ASR pronunciation software is not perfect nor will it be in the immediate future (Nerbonne, 2003). However, it should be born in mind that ASR can be a valuable teaching aid for many foreign language learners. Furthermore, foreign language instructors will come to enjoy how much energy is saved for creative activities in their pronunciation classes. Atwell (1999) tells us that the incorporation of speech recognition technology into CALL software, along with moves to greater learner autonomy and the increase in open learning approaches, may in time offer new ways of constructing the learning experience, while fundamentally changing the balance between classroom and individual learning (p. 29). In the seminal book, *Technology-enhanced Language Learning*, Bush & Terry (1997) envision that from "curricular objectives to lesson planning . . . from teacher training to software applicability, there will be no aspect of foreign language learning that will not be influenced by the technological revolution" (p. xiv). This revolution will make the foreign language instructor's language teaching job more creative, less labor intensive, and even more enjoyable when he or she is willing to embrace the technological changes that have surfaced in the foreign language classroom of the 21st century (Egbert &Hanson-Smith, 1999; Kim, 2003), one case in point being the ASR pronunciation software which this paper has explored.

References

Adams, C. (1979). *English rhythm and the foreign language learner*, The Hague, Holland: Mouton.

Aist, G. (1999). Speech recognition in computer-assisted language learning. In Cameron, K. (Ed.), *CALL: media, design & applications*, Germany: Swets & Zeitlinger, 165-181.

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning, 38* (4), 561-613.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42* (4), 529-555.

Atwell, E. (1999). *The language machine*, retrieved November 10, 2005 from http://www.britishcouncil.org.

Auralog (2000). *AURALANG user manual,* Voisins Ie Bretonneux, France: Author.

Auralog (1995). *Talk to me: User manual*, Voisins le Bretonneux, France: Author.

Beatty, K. (2003). *Teaching and researching computer-assisted language learning*, London: Person Education.

Bernstein, J. (1997). *Automatic spoken language assessment by telephone* (Technical Report No. 5-97), Menlo Park, CA: Entropic.

Bernstein, L., & Christian, B. (1996). For speech perceptions by humans or machines, three senses are better than one. *Paper presented at the International Conference on Spoken Language Processing,* October 3-6, 1996, Philadelphia, PA, USA.

Bernstein. J., Najmi. A., & Ehsani. F. (1999). Subarashii: Encounters in Japanese spoken language education, *CALICO*, *16* (3), 361-384.

Bush, M. D., & Terry, R. M. (1997). *Technology-enhanced language learning*, Lincolnwood, IL, USA. National Textbook Company.

Calico Journal (1999), *Special Issue*, *16* (3), A Journal devoted research and discussion on technology and language learning.

Coniam, D. (1998). Speech recognition accuracy in the speech-to-text process. *TEXT Technology, 8*, 45-55.

Coniam, D. (1999). Voice recognition software accuracy with second language speakers of English*. System, 27*, 49-64.

CPI (Courseware Publishing International) (1997). *See It, Hear It, SAY IT!*, retrieved October 25, 2005 from http://www.usecpi.com.

CSLU (2002). Retrieved October 25, 2005 from http://cslu.cse.ogi.edu/learnss.

CSLU (2003). Retrieved October 25, 2005 from http://cslu.cse.ogi.edu/toolkit.

Dalby. J., & Kewley-Port, D.(1999). Explicit pronunciation training using automatic speech recognition. *CALICO, 16* (3), 425-445.

Derwing, T. M., & Munro. M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in second Language Acquisition, 19*, 1-16.

DynEd (1997). *New Dynamic English CD-ROM Series*, CA, USA.

Edusoft (1998). *English Discoveries CD-ROM Series*, Hebrew, Israel.

Egbert, J., & Hanson-Smith, E. (1999). *CALL environments: Research, practice, and critical issues*, Maryland, USA: TESOL.

Ehsani, F., & Knodt, E. (1998). Speech technology in computer-assisted language learning: Strengths and limitations of a new CALL paradigm. *Language Learning &Technology, 2* (1), 46-60.

Eskenazi, M. (1999). Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language Learning & Technology, 2* (2), 62-76.

Franco, H., Neumeyer, L., Kim, Y., & Ronen, O. (2001). Automatic pronunciation scoring for language instruction. *Proceedings of International Conference on Acoustics, speech, and Signal Processing*, 1471-1474, retrieved October 25, 2005 from http://www.speech.sri.com/people/hef/papers/icassp97_pronunciation.pdf.

Herron, D., Menzel, W., Atwell, E., Bisiani, R., Danelozzi, F., & Morton, R., & Schmidt, J. (1999). Automatic localization and diagnosis of pronunciation errors for second-language learners of English. *Paper presented at the 6th European Conference on Speech Communication and Technology*, September 5-9, 1999, Budapest, Hungary.

Hinks, R. (2001). Using speech recognition to evaluate skills in spoken English. *Working Papers, 49*. Lund University, Department of Linguistics, 58-61.

Hinks, R. (2003). Speech technologies for pronunciation feedback and evaluation. *ReCALL, 15* (1), 3-20.

Holland, V. M., (1999). Tutors that listen: Speech recognition for language learning, *CALICO, 16* (3), 245-250.

Hosom, J.-P., Cole, R., & Fanty, M. (2003). Retrieved October 25, 2005 from http://www.cslu.ogi.edu/tutordemos.

Jurafsky, D., & Martin, J. (2000). *Speech and language processing*, Upper Saddle River, NJ, USA: Prentice-Hall.

Kim, I.-S. (2003). *Multimedia-assisted language learning: Promises and Challenges*, Seoul: Bookorea.

Kim, Y., Franco, H., & Neumeyer, L. (1997). Automatic pronunciation scoring of specific phone segments for language instruction. *Paper presented at the 5th European Conference on Speech Communication and Technology*, September 22-25, 1997, Rhodes, Greece.

Koren, S. (1995). Foreign language pronunciation testing: a new approach. *System, 23* (3), 387-400.

Larocca, S. T., Moagan J. J., & Bellinger S. M. (1991). On the path to 2X learning: Exploring the possibilities of advanced speech recognition, *CALICO, 16* (3), 295-310.

Machovikov, A., Stolyyarov, K., Chernov, M., Sinclair, I., & Machovikova, I. (2002). Computer-Based Training System for Russian Word Pronunciation. *Computer-assisted language learning, 15* (2), 201-214.

Mackey A., & Choi, J.-Y. (1998). Review of Tripleplay Plus! English. *Language Learning and Technology, 12* (1). 19-21.

Menzel, M., Herron, D., Morton, R., Bomaventura, P., & Howarth, P. (2001). Interactive pronunciation training. *ReCALL, 13* (1), 67-78.

Morley, J. (1991). The pronunciation component of teaching English to speakers of other languages. *TESOL Quarterly, 25*, 481-520.

MT Comm (2002a). *FluSpeak* (*ASR Software)*, Seoul, Korea.

MT Comm (2002b). Computer-Aided English Learning Using the FluSpeak Program. *Paper presented at the 2nd APAMALL conference*, Cha-i University, Taiwan., 82-87.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehension, and interpretation in the speech of second language learners. *Language Learning, 45* (1), 73-97.

Myers, M. (2000). Voice recognition software and a hand-held translation machine for second-language learning. *Computer-Assisted Language Learning, 13* (1), 29-41.

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The Pedagogy-Technology Interface in Computer Assisted Pronunciation Training. *Computer-Assisted Language Learning, 15* (5), 441-467.

Nerbonne, J. (2003). Natural language processing in computer-assisted language learning. In Mitkov, R. (Ed.), *The Oxford Handbook of Computational Linguistics,* Oxford: The Oxford University Press, 670-698.

Neumeyer, L., Franco, F., Weintraub, M., & Price, P. (1998). Automatic text-independent pronunciation scoring of foreign language student speech. *Paper presented at the International Conference on Spoken Language Processing* (ICSLP-96), October 3-6, 1996, Philadelphia, PA, USA.

Ordinate (2000).*Validation summary for PhonePass SET-10*, Menlo, CA, USA: Ordinate Corporation.

Prator, C. H., & Robinet, B. W. (1985). *Manual of American English pronunciation* (4th Ed.), New York: Holt, Rinehart and Winston.

Rypa, M. E., & Price, P. (1999) VILTS: A Tale of two technologies, *CALICO, 16* (3), 385-404.

SPSS Korea (2002). *WinSPSS Software Package*, Seoul: Korea.

Witt, S., & Young, S. (1998). Computer-Assisted pronunciation teaching based on automatic speech recognition. In Jager, S., Nerbonne, J., & van Essen, A. (Eds.), *Language Teaching & Language Technology*, Lisse, The Netherlands: Swets & Zeitlinger, 25-35.

Wong, R. (1987). *Teaching pronunciation: Focus on English rhythm and intonation*, Englewood Cliffs, NJ, USA: Prentice Hall Regents.

APPENDIX A. The List of Sentences Read by Students

Dialogue 1: What's the purpose of your trip?

1. Please give me your landing card and let me see your passport.
2. What's the purpose of your trip?
3. I'm here on business, and I'm going to visit a relative.
4. How long will you be staying?
5. Where are you going to stay?
6. Enjoy your visit.

Dialogue 2: Do you have anything to declare?

7. May I see your passport and declaration card, please?
8. Do you have anything to declare?
9. I don't think so.
10. I have my personal belongings and some presents.
11. Are you carrying agricultural products such as fruits or seeds?
12. Would you mind opening your bag?
13. Is this CD player a gift?
14. No, it's for my own personal use.
15. OK. That'll be all, thank you.

APPENDIX B. Pronunciation Rating Scale

"Overall prosody" measures the raters' holistic impression of stress, rhythm, phrasing, and intonation. A four-point scale, with no midpoints between the whole numbers, is used.

Rating Scale

Least Native-like |----------------|---------------|----------------| Native-like
0 1 2 3

Pronunciation Accuracy Rubric

0 Many pronunciation errors and foreign accents and intonation patterns that cause the speaker's pronunciation of the sentence to be completely unintelligible.
1 Frequent pronunciation errors and foreign [non native-like] accents and intonation patterns that cause the speaker's pronunciation of the sentence to be somewhat unintelligible.
2 Some consistent pronunciation errors and foreign [non native-like] accents and intonation patterns, but the speaker's pronunciation of the sentence is intelligible only with some effort.
3 Occasional nonnative pronunciation errors, but the speaker's pronunciation of the sentence is clearly intelligible with effort from the listener.

Sample Scoring Sheet

Item	2	6	7	8	10	13										
1	1	1	1	2	2	1										

Important points to be aware of when rating students' pronunciation:

Be consistent across all items and all students in rating students' pronunciation.
If you are not sure of a student' pronunciation, listen to the item again.