Automatic Evaluation and Training in English Pronunciation

Jared Bernstein, Michael Cohen, Hy Murveit, Dimitry Rtischev and Mitchel Weintraub
Speech Research Program, SRI International, Menlo Park, California 94025 USA

Abstract

SRI is developing a system that uses real-time speech recognition to diagnose, evaluate and provide training in spoken English. The paper first describes the methods and results of a study of the feasibility of automatically grading the performance of Japanese students when reading English aloud. Utterances recorded from Japanese speakers were independently rated by expert listeners. Speech grading software was developed from a speaker independent hidden-Markov-model speech recognition system. The automatic grading procedure first aligned the speech with a model and then compared the segments of the speech signal with models of those segments that have been developed from a database of speech from native speakers of English. The evaluation study showed that ratings of speech quality by experts are very reliable and automatic grades correlate well ($r > 0.8$) with those expert ratings.

SRI is now extending this technology and integrating it in a spoken-language training system. This effort involves (1) porting SRI's DECIPHER speech recognition system to a microcomputer platform, and (2) extending the speech-evaluation software to more exactly diagnose a learner's pronunciation deficits and lead the learner through an appropriate regimen of exercises.

1.0 Introduction

Computer-assisted foreign language instruction is a natural extension of language laboratory technology that is based on audio tape. Computer-assisted language instruction has been the focus of many research projects and commercial products over the last two decades (Ahmad et al. 1985). Most language instruction systems implemented to date can be grouped into several broad classes. The simplest systems offer highly structured lessons and drills relying solely on text and static pictures. Such simple systems offer only a minor improvement over the traditional printed textbooks. More advanced designs attempt natural language processing, permitting greater flexibility in user's input text and moving away from the highly constrained traditional drill paradigms toward more life-like linguistic interactions. Certain more attractive systems use moving video portraying language use in appropriate cultural contexts.

Still, most computer-based language instruction systems have been designed with a focus on reading and listening comprehension, since it is the practice of these receptive skills that is most easily accomplished using a computer capable of controlling simple text, video, and audio output. Some systems also permit writing practice by accepting constrained textual input from the user. Speaking, however, remains the most difficult aspect of language learning to incorporate into a computer-based instruction system. Thus, the pivotal practice of active conversation skills is still restricted to live classroom instruction and real-life "sink-or-swim" situations.

This paper is concerned with spoken language instruction using a computer capable of both audio input and output. We first describe the methods and results of a study of the feasibility of automatically evaluating the spoken reading performance of Japanese students of English. We then outline our latest efforts for developing a voice-interactive computer-assisted language instruction system capable of speech diagnosis, instruction, and evaluation.

2.0 Automatic evaluation of spoken sentences

The initial objective was to determine the feasibility of automatically evaluating the intelligibility of English sentences read aloud by native Japanese speakers. This technology might be useful, for example, as part of an admission procedure for entering a university. We sought a method for automatically deriving intelligibility scores for spoken sentences that correlated well with those assigned by human expert listeners. In principle, such a method would be useful not only for testing student spoken language skills, but also as a basis for a more comprehensive system providing diagnosis and instruction in speaking the foreign language.

2.1 Method

2.1.1 Materials

Each Japanese speaker read six sentences aloud. Of the six sentences chosen for reading, two were specially designed diagnostic sentences that include one or more examples of extreme vowel and consonant sounds, as well as several sounds and sequences that vary among dialects of English. The other four sentences are sentences for which SRI already had recorded a balanced sample of 40 American English speakers. These sentences, which were designed to provide breadth and depth of coverage of phones and phones-in-context, are:

(1) She had your dark suit in greasy wash water all year. (diagnostic)
(2) Don't ask me to carry an oily rag like that. (diagnostic)
(3) What is England's estimated time of arrival in Townsville?
(4) Show the same display increasing letter size to the maximum value.
(5) How many ships were in Galveston May third?
(6) Is Puffer's remaining fuel sufficient to arrive in port at the present speed?

2.1.2 Speakers

SRI recorded 37 speakers: 6 at SRI, 25 at Stanford University housing, and then 6 more at Stanford. All of the analysis reported below was performed on the 31 speakers comprising the first two groups. None of the 31 speakers had lived in the United States for more than three years at the time of their recording, and only two had been in the United States longer than two years. Most speakers had lived in the United States more than one month and less than 15 months. All speakers are adults - 25 men and 12 women.

2.1.3 Equipment

Recordings were made wherever the speaker was found; e.g., in an office or a living room. The recordings were made using a head-mounted microphone and a high-quality, portable, analog tape recorder. The automatic gain control circuit of the tape recorder was employed. The recorded utterances were digitized at 16,000 16-bit samples per second and were stored on disk.
2.1.4 Ratings

The plan for obtaining human ratings of the quality and intelligibility of the spoken material was to have a small number of experts rate the pronunciation quality of each recorded utterance and then measure the intelligibility of a stratified sample of the speakers.

Two listening tests were administered. The first test presented expert listeners with the same sentence as spoken by each of the 31 readers, so that the listeners could gauge the range of English skill in the sample of speakers. Subsequently, all 186 recorded utterances (6 sentences by 31 speakers) were presented in a random order for rating by the listener. Three expert listeners were instructed to estimate the pronunciation skill (segmental and prosodic) of the speaker. Each expert listener rated the utterances on two occasions, separated by several days.

The second test measured the intelligibility of six speakers selected from the 31 Japanese speakers. These six speakers were selected to cover the range of quality ratings among the 31 speakers. Six balanced forms of the test were administered to three sets of six naive listeners. Each form presented six utterances: one from each speaker and one of each sentence type.

2.1.5 Speech Processing

In the development of automatic grading, the speech signals were processed in a manner similar to that used in discrete density hidden-Markov model speech recognition (Cohen et al. 1990). The sampled speech signal was spectrally transformed on a frame-by-frame basis via a Fast Fourier Transform algorithm. Acoustic features were calculated and quantized for each frame from the discrete Fourier coefficients.

Hidden-Markov models trained on a large number of examples of sentences and words spoken by a diverse sample of American English speakers were used as a stochastic model of the pronunciation of English. Hidden-Markov models that represent phonemes, words, or whole utterances could be formed. Given adequate training, the larger the speech unit, the tighter the model will be (tighter models discriminate more reliably). Two extremes of model size were tested: context-free phoneme models and whole sentence models. Regardless of size, each model consisted of a number of underlying states with each state characterized by (1) discrete probability densities (one for each of the acoustic features) and (2) a set of likelihoods of the transitions into allowed subsequent states. Figure 1 is a diagram of states for the word "she".

2.1.6 Preliminary Separation Studies

Before the human ratings were available for these sentences, a series of studies were done to identify which kinds of states and which features of states were most useful in separating the populations of Japanese and American speakers. These state-feature separators (SFS) were identified for the two diagnostic sentences and were used in the studies reported below.

2.2 Results

2.2.1 Ratings

Rating of the pronunciation quality of the sentences by expert listeners yielded the following results:

- Raters used a seven-point scale adequately to distinguish among the speakers. The average ratings assigned to speakers ranged from 1.8 to 6.5.
- The standard deviation of ratings among all three expert listeners was 0.53 averaged over sentences.

TABLE 1 CORRELATION BETWEEN TWO RATINGS BY EACH LISTENER.

<table>
<thead>
<tr>
<th>Expert</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.98</td>
</tr>
<tr>
<td>2</td>
<td>0.96</td>
</tr>
<tr>
<td>3</td>
<td>0.94</td>
</tr>
</tbody>
</table>

As shown in Tables 1 and 2, the reliability of the ratings was excellent. That is, each listener's ratings of the speech samples was consistent for two sets of ratings that were separated by several days. There was also agreement among the listeners in the ratings given to each speech sample.

The judgments of pronunciation quality were robust over sentence content. The average quality rating of the six sentences over all speakers ranged from 3.28 to 3.61, which suggests that the judgments were largely independent of the sentence material.

2.2.2 Intelligibility

A stratified sample of six speakers was selected for intelligibility testing. Table 3 shows the percentage of words correctly spoken (intelligibility score) and the average human-assigned quality rating for these six speakers.

Sentence-intelligibility scores for the speakers were more variable than the quality ratings. Unlike tests of word-intelligibility (which can be designed to achieve considerable precision), developing precise tests of sentence intelligibility is impeded by the listeners' ability to predict words from context.
TABLE 3 REPRESENTATIVE SENTENCE INTELLIGIBILITY SCORES.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Average Quality Rating</th>
<th>Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>INAK</td>
<td>5.33</td>
<td>95%</td>
</tr>
<tr>
<td>JDOT</td>
<td>4.28</td>
<td>84%</td>
</tr>
<tr>
<td>TKIM</td>
<td>3.72</td>
<td>82%</td>
</tr>
<tr>
<td>HOHT</td>
<td>3.27</td>
<td>88%</td>
</tr>
<tr>
<td>KHIR</td>
<td>2.50</td>
<td>67%</td>
</tr>
<tr>
<td>CDOT</td>
<td>1.67</td>
<td>49%</td>
</tr>
</tbody>
</table>

even when individual words are unintelligible. Listeners differ considerably in their ability to predict unintelligible words, and this contributes to the imprecision of the results. Thus, it seems that the quality rating is a more desirable number to correlate with the automatic-grading score for sentences. Furthermore, it agrees with the traditional method of judging pronunciation.

2.2.3 Automatic Grading

Three grading and aligning procedures were evaluated for both wideband and telephone-band speech:

1. Alignment and grading using sentence models.
2. Alignment and grading using phoneme models.
3. Alignment using sentence models and grading using phoneme models.

The most effective automatic grading performance was observed when each spoken sentence was aligned and graded with a model (or a set of models) for that sentence. Sentence model alignments yielded a correlation with the average quality ratings by human experts of 0.81 for wideband speech and 0.76 for telephone-band speech. Table 4 summarizes the results. In the table, "alignment" and "grading" specify the kind of models used in aligning and grading the input speech (either sentence models or phoneme models).

TABLE 4 CORRELATION OF AUTOMATICALLY-COMPUTED SCORES WITH HUMAN RATEDS, (SELECTED FEATURES STATE (SFS) DISTANCES)

<table>
<thead>
<tr>
<th>Alignment:</th>
<th>Sentence</th>
<th>Phoneme</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grading:</td>
<td>Sentence</td>
<td>Phoneme</td>
<td>Phoneme</td>
</tr>
<tr>
<td>320-5600 Hz (wideband)</td>
<td>0.81</td>
<td>0.71</td>
<td>--</td>
</tr>
<tr>
<td>200-3600 Hz (telephone)</td>
<td>0.76</td>
<td>0.11</td>
<td>0.73</td>
</tr>
</tbody>
</table>

At telephone bandwidth, phonetic alignment using context-free phoneme models was degraded as evidenced by the low correlation of 0.11. However, the phoneme models are nevertheless quite robust when properly aligned. When used in conjunction with sentence model alignment, for example, the phoneme models produced an automatic grade that correlates 0.73 with the human quality ratings. Figure 2 shows a scatter plot of automatic grader vs. human judgment scores. These data are for signals band-passed at 200-3600 Hz. The correlation of the data between the two dimensions is 0.76.

A correlation of 0.81 means that the knowledge of sentence-model scores enables us to predict about two-thirds (0.81 * 0.81 = 0.66) of the variation expected in listeners' ratings of quality.

2.2.4 Reliability

In test theory, reliability is an index based on how well two measures of the same variable correlate. Any score is composed of a component of true variance plus a component of error variance. In this case, the true variance is the true difference among the speakers' English pronunciation in terms of the rating criteria used by the listeners. The reliability of a score is an estimate of the true variance component.

Various strategies are commonly used to estimate reliability. Our strategy was to use repeated ratings by each listener to estimate the reliability of the speech quality ratings. As shown in Table 1, the reliabilities of ratings of speech quality by expert listeners are excellent, ranging from 0.94 to 0.98. Experience with the time alignment of SRI's speech processing routine leads us to conclude that the grading scores would have a retest correlation of near 1.0. The high reliability of the ratings is important for the reason that it provides ample true variance to correlate with the scores from automatic grading.

2.3 Conclusions from the Evaluation Study

The two principal conclusions that can be drawn from the evaluation study are that (1) ratings of speech quality by experts are very reliable and (2) automatic grading using sentence models correlate highly with those expert ratings. Results when using phoneme models with 200-3600 Hz band-limited speech (simulated telephone quality) suggest that high correlations with quality ratings can be achieved, but that further development of the time alignment will be required. We presume that intelligibility would also correlate well with the models if enough listeners and sentences were used to stabilize the intelligibility scores.

The evaluation study has pointed to several important issues, among which are:

- Development of automatic grading has relied predominantly on spectral (segmental) features of the recorded speech sample. A more complete treatment of prosodic features may be needed to improve automatic grading significantly.
- Results with the SFS distances suggest that improved discrimination can be realized with carefully selected states and features. Searching for particularly effective features should be part of further development.
- Cross-gender models of American adults were used for comparison with Japanese adults. Performance could probably be improved by using specific models for male and female speakers. Furthermore, the training and matching may need to be based on a more elaborate sample of speakers (e.g. samples that are stratified by age and size of the speaker).
- There are important trade-offs involved in deciding between phoneme models and sentence models. Sentence-level models are much more expensive to train, but they align the speech signal much better and consistently correlate better with human judgments. On the other hand, phoneme-level models permit an arbitrarily large set of phrases and/or sentences to be constructed with no extra development cost. Achieving adequate performance with phoneme models will require some additional research.

3.0 Instructional Development

3.1 Introduction

Although there are many directions in which SRI's research on pronunciation evaluation might be extended, we are now taking the first steps in the development of instructional systems that can take a more active part in the shaping of spoken language production. In particular, students need to speak the language in a situation in which good models are available and critical feedback is given within some task that has reasonable intrinsic interest.
3.2 Target System

SRI is aiming development toward a language teaching system in which both graphic and spoken forms of language can be taught and tested, and in which both the receptive (listening and reading) and productive (speaking and writing) skills are trained. The target system would ideally have excellent interactive graphics and tools that allow fast development of lessons and the automatic tracking of student progress. However, as a first step, SRI is focusing on a demonstration system that highlights the possibilities of spoken input in language instruction.

3.3 Initial System

The initial demonstration system supports the development of language lessons in which the student has access to some graphical information like a map or a table of data forming the basis for a limited-context "conversation" between the student and the computer. The student is presented with questions and several full-sentence answers in a multiple-choice format. A native pronunciation of the question and all the answers is available for playback at the student's request. The student reads one of the answers into the microphone. The system immediately indicates which answer was picked and how well it was pronounced.

The development of the initial system involves building a small-vocabulary English and Japanese word recognizer in a portable microcomputer environment, improvement of sentence grading and user feedback mechanisms, and collection of speech data for recognizer training. We are also addressing the design of a thematically integrated lesson battery along with a simple and flexible user interface.

4.0 Conclusion

We have demonstrated that automatic evaluation of English pronunciation is possible. Further, it seems that similar technology should be adequate for use in the training of foreign language pronunciation and in the diagnosis of spoken errors. If teaching materials are carefully adapted to use the speech recognition tools that already exist and if speech recognition system components are properly trained and adapted to the requirements of language teaching, many new ways of teaching, analyzing, and evaluating pronunciation will soon emerge.

Bibliography


FIGURE 2 AUTOMATIC GRADE VS. HUMAN JUDGMENT SCORES