TOEFL.

ISSN 1552-8219

TOEFL iBT Research Report

TOEFLiBT-02
February 2007

Toward an Understanding of the Role of Speech Recognition in Nonnative Speech Assessment

Klaus Zechner

Isaac I. Bejar

Ramin Hemat

Listening.
Learning.
Leading.

Toward an Understanding of the Role of Speech Recognition in Nonnative Speech Assessment

Klaus Zechner, Isaac I. Bejar, and Ramin Hemat

ETS, Princeton, NJ

RR-07-02

Abstract

The increasing availability and performance of computer-based testing has prompted more research on the automatic assessment of language and speaking proficiency. In this investigation, we evaluated the feasibility of using an off-the-shelf speech-recognition system for scoring speaking prompts from the LanguEdge field test of 2002. We first established the level of agreement between two trained scorers. We then adapted a speech engine to the language backgrounds and proficiency ranges of the speakers and developed a classification and regression tree (CART) for each of five prompts based on features computed from the output of the speech recognizer. In a validation on held-out data, we found that while our features are not sufficiently comprehensive to adequately score these prompts, collectively these features appear to capture reliably some aspects of speaking proficiency.

Key words: Assessment, spoken language assessment, automatic spoken language assessment

The Test of English as a Foreign Language™ (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations® (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education. The test is now wholly owned and operated by ETS.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, two-year colleges, and nonprofit educational exchange agencies.

❖ ❖ ❖

Since its inception in 1963, the TOEFL has evolved from a paper-based test to a computer-based test and, in 2005, to an Internet-based test, TOEFL iBT. One constant throughout this evolution has been a continuing program of research related to the TOEFL test. From 1977 to 2005, nearly 100 research and technical reports on the early versions of TOEFL were published. In 1997, a monograph series that laid the groundwork for the development of TOEFL iBT was launched. With the release of TOEFL iBT, a TOEFL iBT report series has been introduced.

Currently this research is carried out in consultation with the TOEFL Committee of Examiners. Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee advises the TOEFL program about research needs and, through the research subcommittee, solicits, reviews, and approves proposals for funding and reports for publication. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Current (2006-2007) members of the TOEFL Committee of Examiners are:

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

Acknowledgments

We want to thank the staff of Multimodal Technologies Inc. in Pittsburgh, in particular Thomas Polzin, for their patience in explaining and helping with their system and the data transcribers, in particular Victoria MacLaren. We are grateful to Sarah Ohls and Dele Kuku for editorial assistance. Paul Deane offered many insights and requests for clarification that have enhanced the report. We are also grateful to reviewers of an earlier version of the report, especially Dominique Estival, Dan Eignor, and Pam Mollaun. We are especially grateful to Xiaoming Xi, who generously contributed many insights and improvements to the report. It goes without saying, however, that any remaining errors are the entire responsibility of the authors.

Table of Contents

Page

Appendixes

List of Tables

List of Figures

Page

The increasing availability and performance of computer-based testing has obvious implications for the assessment of language proficiency (e.g., Chalhoub-Deville, 2001), but doubts remain concerning the feasibility of assessing *speaking proficiency* via computer. By all accounts, the recognition of the speech of language learners is particularly challenging because they are still learning the language, and therefore their speech is likely to contain lexical, syntactical, and other errors. Additionally, their speech can be highly accented. Moreover, for assessment purposes, accurate speech recognition is not sufficient to characterize speaking proficiency based on a communicative perspective. Prosodic characterizations of the speech sample, for example, would be needed as well. This means that research must be planned carefully to determine what investment would be required to reach a sufficiently high level of scoring accuracy by computer without compromising the construct we wish to measure. We conducted a study designed to address that question. Using data from the LanguEdge field test, we first established the level of scoring accuracy between trained scorers. We then adapted a speech engine to the language background and proficiency level of a portion of the sample of speakers available to us for this analysis and developed a classification tree for each prompt to score speaking samples based on variables extracted by the speech recognition engine. The tree was then validated on a portion of the sample that had not been used in estimating the classification trees.

The key finding is that it is possible to capture aspects of speaking proficiency as measured in LanguEdge by means of an automated scoring procedure. The following chart illustrates the results. We scored the five speaking prompts in Form 1 of LanguEdge. Three of these—admire, culture and technological change—are so-called independent tasks; the other two—water and expression—are so-called integrated tasks. The integrated tasks involve an additional modality—listening in the case of water, and reading, in the case of expression—whereas the independent tasks require only speaking and a minimal amount of reading.

In Figure 1, we plot exact agreement for human raters and automated scoring under two conditions: (a) assuming that the speech engine can recognize the spoken responses with 100% accuracy (human/ computer–test aligned) and (b) based on the actual level of recognition (human computer test–unaligned). Because on a 5-point scale there can be a high level of significant agreement—specifically 20%—even if the two scores are unrelated, we include in the plot chance agreement between raters.

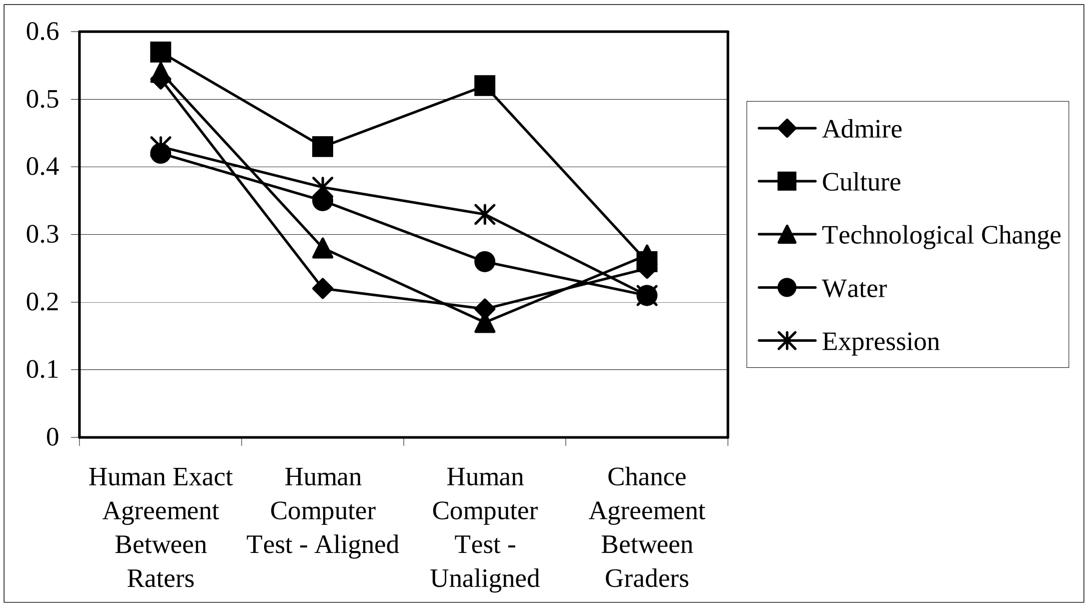

***Figure 1.* Exact agreement between two human scorers, computer and human scorers, and chance agreement.**

The data are plotted to inform the question identified earlier, namely to provide a sense of the investment that would be required in speech technology research to match the accuracy of human raters. The contrast of accuracy based on human scores versus accuracy based on computer scoring, assuming that perfect recognition informs how adequate the set of variables for characterizing speech is. We use a combination of content, lexical, fluency, and rate of speech variables. Clearly, as can be seen in the chart, there is room for improvement toward better agreement with human raters. Interestingly, results are more encouraging for the two integrated tasks in the sense that the level of agreement between rater and computer does not drop as much compared to the level of agreement between raters.

The results for the independent tasks are mixed. The culture prompt results are in line with those of the integrated tasks, with the exception that the results for computer scoring based on actual recognition are unexpectedly high. The other two independent tasks, technological change and admire, follow a similar pattern—with the scoring tree performing at chance or worst, even assuming 100% recognition.

To our knowledge, the results are the first investigation of the feasibility of automated scoring of speaking samples from language learners based on spontaneous responses. Despite the significant challenges that must be overcome in order to score the spontaneous speech of language learners, the results shown above suggest that investing in speech technology is reasonable. Much additional research is needed to understand the prompt attributes than enhance amenability to automated scoring, to improve speech recognition performance, and to identify a more comprehensive set of variables to characterize speaking proficiency. So long as steady progress can be shown, however, the investment is reasonable because of the potential cost savings to the growing number of testing programs offering speaking assessment, and the resulting increased competitiveness afforded by the ability to score speaking tests at a lesser cost.

Toward an Understanding of the Role of Speech Recognition in Nonnative Speech Assessment

The increasing availability and performance of computer-based testing has obvious implications for the assessment of language proficiency in general (e.g., Chalhoub-Deville, 2001), but doubts remain concerning the feasibility of assessing *speaking proficiency*:

> … it is questionable whether the full range of individual and interactive speaking performances that language educators are interested in will be adequately *elicited* in computerized formats; likewise, it is doubtful that the complexities of such performances and the inferences that we make about them will be captured by *automated scoring* and speech recognition technology. (Norris, 2001, p. 99, emphasis added)

This comment touches on two central aspects of evidence-centered design (ECD) (Mislevy, Steinberg, Almond, & Lukas, 2006), a framework that we find useful for discussing assessment design. Specifically, the above quote touches on two of the three main components of an assessment: (a) task design and (b) evidence identification and aggregation. Task design is the component that we normally associate with *test development*, with the exception that in an evidence-centered design context, items are explicitly designed to elicit the evidence called for by the goals of the assessment, and, importantly, the item-development process does not occur until the evidentiary implications of the goals of the assessment are well understood. In other words, we must be very clear about the evidence we seek before we can design the items to elicit that evidence. Norris (2001) argues that a computer-based delivery may preclude eliciting such evidence. In reality, such a conclusion depends in part on the goals of the assessment, and ultimately it is an empirical question whether task designers can rise to the challenge. The second doubt raised by Norris is closer to the purpose of this report. That is, assuming that it has been possible to design computer-delivered tasks that appropriately elicit evidence called for in an assessment of speaking proficiency, is it possible to score the responses, or what is equivalent from an ECD perspective: Is it possible to develop automated procedures for identifying evidence of speaking proficiency? We explore that question in this report.

What makes it possible to even think of the feasibility of automated scoring of speaking proficiency is the significant advances in automated speech recognition (ASR) systems that have occurred over the past few years (e.g., Jurafsky & Martin, 2000). Indeed, aspects of speaking proficiency can already be scored automatically (e.g., Bernstein, 1997; de Jong & Bernstein, 2001) in cases *where responses can be anticipated*. It is perhaps in that context that Norris (2001) suggested that a computer delivered and scored speech task was not feasible because tasks in which responses can be anticipated would not seem to elicit evidence about speaking proficiency when broadly defined. The American Council on the Teaching of Foreign Language (ACTFL) provides proficiency guidelines for speaking (as well as writing, reading, and listening).[1] A reading of these proficiency guidelines suggests that the novice level could perhaps be measured through tasks that elicit a limited range of speech, and therefore it may be feasible to measure that level of performance with tasks that call for anticipated responses. For higher levels of proficiency, however, spontaneity and adaptability to unique situations become increasingly important in the ACTFL guidelines. Therefore, the goals of the assessment appear to determine whether identifying evidence of speaking-proficiency tasks can be automated. In particular, it seems useful to think of potential tasks for eliciting speaking proficiency as requiring more or less spontaneity; that is, greater or lesser predictability of responses.

Pronunciation evaluation and training is a case where anticipated response rather than spontaneous speech could be a more efficient approach to assessing speaking proficiency. For example, Dalby and Kewley-Port (1999) compared two ASR architectures and concluded that, for evaluating pronunciation, different architectures might be better suited depending on the goals of the assessment. Because the architecture of an ASR system used for assessing speaking proficiency is not independent of the goals of the assessment, especially within an ECD perspective, the following section discusses selectively aspects of ASR.

A Very Brief Introduction to Speech Recognition

An ASR system can be conceptualized as a system whose input are acoustic signals in digital form and whose output is the best statistical estimate as to what sequence of words correspond to the input signal. (Jurafsky & Martin, 2000; Rabiner & Jauang, 1993; and Rabiner, 1989, offer introductions to speech recognition. For a brief history of ASR, see Young, 2001.[2]) ASR systems for the consumer market are typically designed to maximize transcription accuracy because a common use is dictation. Therefore, an important metric in comparing ASR systems is their *word error rate* (see Jurafsky & Martin, 2000, p. 271).

The architecture of an ASR system has become fairly standardized. The goal of such a system can be seen as transcribing the acoustic signal into a textual representation, and is mediated by two models. One model, the acoustic model (AM), associates probabilities with *speech units*, called *phones*, that represent a given phoneme. This is more complex than it sounds, since each phoneme may have multiple realizations, or *allophones*, depending on the context, such as whether it occurs in an initial position, between syllables, or at the end. Moreover, in continuous speech, words are not neatly separated as they are in print, which further complicates the recognition problem. Additionally, speaker attributes such as regional or foreign accents affect phone pronunciation.

The first stage of the recognition process is to associate probabilities with each phone so as to cover contiguous time intervals in the speech signal (see Jurafsky & Martin, 2000, p. 267). To this end, the speech signal is sliced in short, 10 millisecond intervals, or frames, and spectral features[3] are extracted for each frame. The features in turn are fed to a statistical model that associates probabilities with each possible phone for that time slice. These probabilities are referred to as *phone likelihoods*. The increasing capabilities of speech recognition systems are due in part to their use of hidden Markov models (HMMs) for representing each phone (see Rabiner, 1989, for a tutorial). Through such modeling it is possible to account for temporal variations of the same phone, for example. Each entry in a lexicon containing the words that are expected to be recognized by the system is expressed as a sequence of HMMs representing the phones assumed by that entry. Some level of pronunciation variation can be handled in this manner, as noted by Strik, Helmer, and Cucchiarini (1999, p. 233), by simply including pronunciation variants, or different sequence of phones represented as HMMs, for any given word. As the number of such variants increases, the recognition process becomes more resource-intensive, since many more possibilities need to be considered during the recognition process. For systems that are expected to perform speech recognition in real time, this could present a problem. However, it may not be an issue in a context where recognition can take place as a post-processing step. As noted by Strik et al. (p. 235), further pronunciation modeling can be incorporated in the recognition process by adaptating the acoustic model; that is, the statistical model linking acoustic features to phones. This requires training data that contains alternative pronunciations and their corresponding transcriptions.

In addition to the acoustic model, the recognition and transcription of speech is mediated by a second model, the language model (LM), and can be thought of as prior information useful during the decoding task. The role of the language model is to encode prior information about the words that are likely to be spoken. The language model takes the form of frequency distributions for single words, pairs of words (bigrams), and triples of words (trigrams)—collectively, n-grams. These probabilities "indirectly encode syntax, semantics and pragmatics ..." (Deshmukh, Ganapathiraju, & Picone, 1999).

In addition to the language model, which contains the prior information about the probability of words being spoken, and the acoustic model, which contains the observation likelihoods of phones, a lexicon is needed for the recognition process as well. The lexicon contains the normative information on the pronunciation of words in the form of phoneme sequences. The AM, LM, and the lexicon are used jointly to decode the signal. The decoding process entails a search through alternative transcriptions of the signal in order to locate the most likely transcription. The search mechanism is computationally complex, since the beginnings and ends of words are not given in advance and, therefore, different word boundaries have to be considered to determine a ranked list of possible transcriptions. Deshmukh, et al. (1999) provide a thorough review of approaches to this task.

Symbolically, the fundamental equation of speech recognition is an application of Bayes theorem:

$$\arg\max_{W} P(W|S) = \frac{P(S\,|\,W)P(W)}{P(S)}, \qquad (1)$$

which simplifies to

$$\arg\max_{W} P(W|S) = P(S\,|\,W)P(W) \qquad (2)$$

because *P(S),* the probability of the signal, is constant within the recognition task.

Here, *P(S|W)* is the *acoustic model* and computes the likelihood of the signal *S*, given the transcription, *W*. *P(W)* is the *language model* and encodes the prior probability of observing the string of words. *P(W|S)* is the posterior distribution of the transcription.

As there are many potential transcriptions for any given signal, the one with highest posterior probability is interpreted as the transcription of the signal. Put differently, the best guess as to the transcription of the signal is the string of words, W*, among several possible such strings, that maximizes the product of the likelihood that the signal was produced by a string of words and the probability of that sequence of words. Under this statistical conception, speech recognition is "reduced to designing and estimating appropriate acoustic and language models, and finding an acceptable decoding strategy for determining W*" (Young, 2001). An ASR system is a specific implementation of these components.

Before an engine can recognize speech, the language model and acoustic models must be estimated or trained so that they provide the required information for the equations above. In the case of the acoustic model, this requires transcribing a sample of speech and pairing the acoustic and textual representation. Clearly, recognition performance will be affected by the size of the sample and how well the conditions under which the training data were collected match the samples of speech fed to the recognizer. In the case of the language model, this process requires an estimation of the probabilities of observing any given n-gram. A common approach is simply to tabulate the frequencies of n-grams in some relevant corpus, although statistical smoothing of the probabilities thus estimated is necessary.

Given the statistical nature of the recognition process, the performance of an ASR system is affected by a variety of factors related to generalizability; that is, the degree to which the training data are representative of the speech to be recognized. One performance factor is the extent to which the conditions under which the acoustic and language model were obtained match the application of those models to recognition of "new" speech. For example, the environment where the speech sample is captured, the surrounding noise level, the type and quality of the microphone, and the resolution of speech signal, among others, are reflected in the acoustic model. If those conditions are not maintained when recognizing new speech, performance will degrade. Other acoustic factors affect the performance of the acoustic model, including the degree of accentedness and other speech idiosyncrasies. The degree of accentedness, in the case of the application to speaking proficiency, is an aspect of what we wish to measure. The language model also affects performance. The accuracy of the probabilities of observing n-grams clearly depends on the amount of training data. Apart from sample size, the content of the training data should match that of the speech to be recognized.

The search process to locate the most likely transcription is computationally intensive, and therefore its effectiveness depends in part of the computing resources available to the task. For example, some engines allow variations in the weight of the language and acoustic models as well as tradeoffs between speed and accuracy, which allows for more extensive decoding but at the expense of recognition time. These parameters of the recognition engine together with strategies for calibrating the language and acoustic model are important considerations in the application of ASR for assessment purposes, as we shall see below.

The foregoing considerations suggest that the application of speech technology to the assessment of speaking proficiency requires a close coordination between the calibration of the acoustic and language models and the purpose of the assessment. For example, for speaking proficiency assessments that are not adaptive, application of ASR to speech samples could take place offline and therefore, within reasonable limits, recognition speed is not a concern. In this case, a recognition strategy could be adopted that, though more computationally intensive, might yield better results. The same may not be true if the assessment is adaptive or otherwise requires immediate recognition. Yokoyama, Shinozaki, Iwano, and Furui (2003) illustrated an approach where the initial recognition output is processed to enhance the language model. That is, the speech signal is re-recognized with language models enhanced by incorporating the output from the preceding recognition attempt. Clearly, such an approach does not lend itself to assessment applications where immediate recognition is required, but in applications where this is not a requirement it could improve overall performance.

The presence of a multitude of accents among nonnative speakers, not to mention different levels of proficiency, is an unavoidable reality in the assessment of speaking proficiency and could limit the applicability of ASR to the assessment of speaking proficiency in open-ended tasks. However, there is much research addressing these challenges. With respect to accent, methods are emerging that selectively adapt the language and acoustic models. Wang, Schultz, and Waibel (2003) discuss some approaches. Tomokiyo (2001) and Ikeno et al. (2003) describe specific approaches to dealing with heavily Spanish-accented English speech. Wu and Chang (2001) also discuss approaches in which a speaker is placed in an appropriate cluster by means of a short speech sample. If such clusters are based on different accents and proficiency levels, then in principle we can recognize speakers from such clusters more accurately using language and acoustic models specifically based on speakers with similar accents and proficiency profiles. In short, while varying accents and proficiency levels make the recognition task rather more challenging, there is much research in progress that addresses this challenge.

Beyond Recognition

Complex as the recognition process is, there is much more to assessing speaking proficiency than simply recognizing what was said. Even if the recognition accuracy is high, will a textual transcription contain enough evidence to fully assess speaking proficiency? Again, clearly, much depends on the purpose of the assessment. Just which aspects of the multifaceted nature of effective spoken communication are at issue? Within the language learning community speaking proficiency is often defined as a multicomponential construct involving far more than comprehensibility (e.g., Bachman, 1990; Canale & Swain, 1980; Hymes, 1972; and discussion of these and other sources by McNamara, 1996). According to Butler, Eignor, Jones, McNamara, and Suomi (2000, p. 2):

> Communicative competence in oral academic language requires control of a wide range of phonological and syntactic features, vocabulary, and oral genres and the knowledge of how to use them appropriately.

An example of the role of phonology in speaking proficiency is assessing prosodic appropriateness, specifically intonation, in a given context. That is, an utterance such as a question could be pronounced correctly, and therefore recognized, but might lack the conventional intonation associated with questions. From a communicative perspective, correct intonation can be as important or even more important than correct pronunciation. It would seem, however, that the communicative appropriateness of prosody (such as stress and intonation) cannot be judged independently of comprehensibility. For example, to evaluate whether a request is expressed politely as part of spontaneous speech, we first need to determine that a certain text expresses a request. However, certain prosodic features, such as rate and fluidity of speech, are relatively independent of comprehensibility.

To address the oral communicative proficiency of speakers we need to proceed to another stage where a speech sample is transcribed and perhaps phonologically annotated. Interestingly, phonology is still important even at the level of assessing the lexical adequacy of a response. For example, words like *insult* and *torment* take different stress patterns depending on whether they are being used as nouns or verbs. At the highest level of analysis oral genres and the knowledge of how to use them appropriately would need to be addressed. Despite the difficulty of this task, research is beginning to address how prosody can be used to predict speech acts (Searle, 1979), such as asking questions, making requests, and giving commands (Shriberg et al., 1998).

Implications for Assessment Design

The foregoing considerations suggest that achieving automated scoring of spontaneous speech responses will require significant adaptations to current speech technology. The design of the Voice Interactive Training System (Rypa & Price, 1999) illustrates these interactions and tradeoffs between speech technology and assessment design. A communicative perspective in language learning calls for exposure to spontaneous speech by native speakers and elicitation of spontaneous speech from the learner. In the case of the Voice Interactive Training System, the goal was to enable such spontaneous exchanges but also to assess pronunciation in its own right. Achieving both goals requires careful consideration of task design, because spontaneous speech may be challenging from a speech recognition perspective for assessing pronunciation. In particular, a recognizer typically aims for maximum recognition accuracy, and in part that requires broad acceptance of the many ways in which speakers may slightly vary their pronunciation of the same words. In effect, to achieve maximum recognition the acoustic model must learn to ignore slight mispronunciations.

It matters also, of course, how the acoustic model was trained. As noted above, the training of the acoustic model requires a corpus of transcribed speech. If the corpus represents the speech of native speakers, then the acoustic model will have been trained to ignore pronunciation variability among proficient native speakers. When applied to language learners, this model could lead to a very low recognition rate simply because the pronunciation variability among language learners at different proficiency levels is likely to follow quite distinct patterns. An alternative is to train the engine with speech samples obtained from language learners. In this case there would be a strong tension between the dual goals of assessing pronunciation and determining speaking proficiency broadly speaking. For example, learners from different language backgrounds might systematically substitute phonemes from their native language for those of the target language. The acoustic model might be optimized to take advantage of those regularities in the interest of increased recognition. However, if the same acoustic model were to be used to assess pronunciation, it would be clearly at the expense of speeding up the acquisition of pronunciation of the target language because the engine would accept systematic mispronunciations and therefore not alert the learner to those errors. Tomokiyo (2001) discusses at length the adaptation of speech engines to nonnative speakers. She finds, as we do below, that acoustic adaptation is extremely important to improving recognition performance. Similarly, Witt (1999) describes innovative approaches to handling accented speech that are especially suited to a language-learning context. These competing factors led the designers of one system to the following design decision (Rypa & Price, 1999):

> We decided that the desire for score validity should take precedence over the desire for spontaneity when pronunciation was to be scored. Therefore, we developed initial lesson activities in which all pronunciations being scored were based on reading from text. Lesson activities in which pronunciation is not scored can be more flexible … (p. 393)

Other designs come to mind, however. In *interactive* language instruction, it is essential that the recognizer perform quickly, because otherwise the value of interactivity is lost. However, with sufficiently powerful hardware, it would seem feasible to evaluate speech while simultaneously applying different criteria. For example, it if were feasible to recognize speech using different acoustic models, one trained on native proficient speakers and one based on language learners, the difference in recognition *might* be a function of pronunciation proficiency. The value in this approach is that spontaneity and validity of pronunciation scoring could be achieved while making use of communicatively appropriate instructional material.

Another possibility is that, in some situations, reading from text can be combined with a communicative activity. When reading from text, the transcription is provided by the text to be read. For example, a task may call for reading the assignment for the following day to a fellow student. In this case, reading aloud would be interspersed with more spontaneous speech. The reading aloud portion would be scored for pronunciation. A reading-aloud component might be beneficial in general, not just for scoring pronunciation. In fact, it could be the basis of adapting the engine to the speech of each student, which would improve recognition of that student’s subsequent speech samples.

Overview of the Rest of the Report

Although the above examples address the scoring of pronunciation, similar considerations apply to the assessment of speaking proficiency in general. That is, the application of speech technology to assessing speaking proficiency is likely to require adaptation based on an understanding of assessment goals. It remains an empirical question as to how feasible such an adaptation would ultimately be for tasks that call for spontaneous speech. We present results below to begin answering the question. We analyzed spontaneous speaking responses to items from the LanguEdge speaking courseware.

The data analyzed here are a susbset of the data from the field test conducted to evaluate the feasibility of several task types (ETS, 2002a, 2002b) and include speakers from a range of language backgrounds and proficiency levels whose speaking performances were graded by judges. We present results from experiments adapting acoustic and language models and recognition parameters, suggesting recognition accuracy indeed can increase significantly as a result of such adaptations. In addition, we identify several variables to capture aspects of speaking performance meant to be elicited by the LanguEdge tasks. The sole independent external criterion we have to assess how well this collection of variables represents speaking performance is the scores given by graders to these speech samples. Therefore, we obtained classification trees on a development sample and applied them to a sample that had not been included in the development of the tree. We first carried out this analysis based on the transcription of the speech samples, rather than the recognized or hypothesized speech. The level of classification accuracy varied by prompt, but these preliminary results suggest that the variables capture important aspects of the criteria that the graders employed to grade the speech samples. Finally, we apply the classification tree to the actual output of the recognizer on a sample that had not been part of the optimization of the acoustic and language models, or of growing the classification tree. The difference of level of agreement based on the transcription and based on the output of the recognizer provides a realistic indication of the improvements in recognition that will be needed to rely on speech technology as a tool in the assessment of speaking proficiency.

Methods

Data

The speech data used for this project was from 171 students participating in the LanguEdge 2002 Field Test (ETS, 2002a, 2002b). The field test itself was conducted domestically and internationally and involved a total sample of 2,703 students. The language background distribution for the entire sample approximates that of the TOEFL® test-taking population, with

Chinese, Arabic, Korean, Japanese, and French as the most frequent language backgrounds. Data for the field test were collected in 18 domestic test sites as well as 12 international test sites.

The field test included two types of speaking tasks: independent speaking and integrated speaking. Independent tasks “are designed to assess students’ abilities to formulate and communicate ideas on a variety of familiar topics” (ETS, 2002b, p. 5). Three of the prompts are independent. The remaining two prompts are integrated. Integrated speaking tasks address students’ speaking proficiency in an academic context, such as lectures and reading of academic texts. Tables 1 and 2 describe the five prompts and *some* notes on each prompt taken from ETS (2000a). We have highlighted aspects of the topic notes to emphasize the subtle judgments graders are expected to make. The rubrics can be found in ETS (2002a).

The test was administered on personal computers (PCs) that had been equipped with a suitable sound card to enable speech capturing. The speech was recorded with a microphone connected to a PC sound card. The sampling rate was 11.025 kHz and the resolution was 8 bit. This yields about 646 KB of data per minute.

Calibration of the Speech-Recognition Engine

The next few sections describe the data and the procedures used to calibrate the Multimodal speech-recognition engine. First we describe how we apportioned the speaking samples for calibration and testing and then we present the results of several experiments designed to identify the optimal parameters to use for our purposes.

Multimodal Speech-Recognition Engine

For the experiments described below, we use Multimodal Inc.’s Xcalibur speech-recognition engine. It was trained on about 200 hours of native speech of American English, with a training sample rate of 16 kHz. The corpus consists of LDC’s (Linguistic Data Consortium) Broadcast News corpus and a variety of proprietary dictation corpora owned by Multimodal Inc. The acoustic model has 32 cepstral coefficients and 3,150 *codebooks* with 24 Gaussians in 32 dimensions. The language model is a trigram back-off model, which was trained on the same data as the acoustic model: 200 hours of speech, or approximately 120,000 words.

Table 1

Independent LanguEdge Prompts

Prompt description	Topic notes
Independent: admire prompt Describe a person whom you admire. In your response, you should: describe the qualities you most admire in this person explain why you consider these qualities important	Speakers choose to talk about people whom they have personal relationships with, or people they admire from a distance. Most speakers first identify the person they want to talk about, and then list the qualities they admire. Speakers at the 4 and 5 level explain why those qualities are admirable. At lower levels, this part is often not well developed. *Raters should balance the completeness of the response with the linguistic abilities demonstrated in the response.*
Independent: culture prompt Do you agree or disagree with the following statement? "It is important for people to learn about other languages and cultures." Give specific reasons and examples to support your opinion.	The placement of this statement on an ESL test may lead speakers to believe they should agree with the statement, which the majority of speakers did. Some were able to bring their personal experiences into the discussion; others maintained a more objective perspective. In general, the prompt elicited a combination of conceptual and factual information. Many speakers focused on the *culture* aspect of the prompt and had little or nothing to say about *languages*. *Raters should keep these different approaches in mind and reward speakers for what they have done well.*

(Table continues)

Table 1 (continued)

Prompt description	Topic notes
Independent: technological change prompt Technological change has always had a significant impact on people’s lives. What do you think has been a major effect of technology on people’s lives in the past 100 years? In your response, be sure to describe one important technological development explain how it has changed people’s lives	Some frequently mentioned developments include computers, the Internet, the automobile, airplanes, and electricity. Responses that describe something else should be scored on their own merits and not be rewarded or penalized because they differ from the majority. More proficient speakers cover both aspects of the prompt, although they often just name it rather than describe it. Less proficient speakers usually present a more limited explanation of how it has changed people’s lives.

Corpus Creation

We obtained 950 speech samples with a length of 1–1.5 minutes each from the LanguEdge field test. From these files we built a corpus of 170 speakers who had responded to at least one of the prompts and had been intended to be rated by two raters. Due to a variety of technical reasons, a small percentage of these files were not included in the final corpus analysis. The final corpus contained 927 files total in 170 directories.

A professional transcriber transcribed the speech data from the 927 files (see Appendix A for transcription conventions). To facilitate the transcription process, the transcriber was informed about the content of the five items. The transcription yielded about 120,000 words (including speech disfluencies). (For details of the transcription process, see Appendix A.) The transcriptions were converted into an XML-formatted ASR database by means of a mapping script that identified all marked-up segments and then mapped their contents to the new XML database according to mapping specifications. While these specifications are certainly not definable in only one way, they still follow common sense rules about what elements should be in the database to reflect the content of the respective utterances as accurately as possible, but without creating practical problems such as spurious vocabulary growth due to inclusion of word fragments.

Table 2

Integrated LanguEdge Prompts

Prompt description	Topic notes
Listening: water prompt Discuss the *safe yield* method of managing water resources. In your response, be sure to explain: how the method works why the method is not effective problems associated with the method Use specific details and examples to support your answer.	Some speakers may not give a succinct definition of the safe yield method at the beginning, but arrive at it gradually after some discussion of the purpose of the method and possibly the related problems. It is also common for speakers to intersperse discussion of why the method is not effective with discussion of problems associated with it. These features are not faults, and speakers should be rated on the overall logical flow of the discourse. Some speakers may, at some point, end up confusing the role of humans with the role of nature. No single incidence of this should be weighted too heavily; *the rating should instead focus on a holistic impression of what the speaker understood.* Speakers also frequently use the word *recharge* in their responses, with varied degrees of accuracy. Again, misuse of individual terms should be less important than overall comprehensibility. Some responses may be more difficult to score than others, such as ones that combine inaccurate content with fluent speech. *Raters should attempt to balance content and delivery in determining a holistic score.*

(Table continues)

Table 2 (continued)

Prompt description	Topic notes
Reading: expression prompt In your own words, describe Paul Ekman's research in which he used photographs of people exhibiting various emotions. In your response, be sure to include the following: the purpose of the research the important details of the research the results of the research	Many speakers collapse purpose and findings into one, so raters will *detect a pattern of echoing statements*. Because this is a conceptual passage, successful speakers both summarize and synthesize the main concepts, while others show a pattern of struggling to put abstractions into their own words. Speakers need to be resourceful in selecting what aspects of the research to report; most focus on the earlier study, but especially able speakers may include a summary of the later study as well. The latter inclusion provides breadth for the response, but is not essential. Raters should focus on a well-developed description of the important details of the research. The question asks about the first two paragraphs of the passage, which develop the idea that certain facial expressions reflect basic emotions similarly around the world. The latter part of the passage deals mainly with research suggesting that a person's facial expression can influence their emotional state. Some speakers attempt to include this portion of the text, often inadequately or inaccurately. *Raters should avoid focusing on the inaccuracy of this "additional" information, rather than on content and delivery of relevant information.*

For the purpose of our experiments, we split the corpus into groups by randomly assigning speakers to three corpora: a training corpus, a development set, and a training set. The training corpus (Train) consisted of 669 files; the development set (Devtest) consisted of 124 files; and the evaluation set (Eval) consisted of 134 files. The training set was further split into two sets, one consisting of 90% of the files (Train90), and the other consisting of the remaining 10% (Train10, see below).

Construction of Vocabularies, Pronunciation Dictionaries, and Language Models

The two major adaptation strategies for a speech recognizer are acoustic model adaptation and language model adaptation. This section describes the steps performed to arrive at different language models, which then can be compared using an independent test set.

Vocabularies. We extract all the types; that is, word forms from a randomly chosen subset of the 3,215 utterances of Train90 and from all utterances of the Train10 subcorpus. (The reason for not taking the entire Train90 subcorpus is that, while we construct the language model, we need to simulate the behavior of unknown words). We then sort them alphabetically to form the two vocabularies Train10.Vocab (1,295 words) and Train90.Voc-Reduced (3,295). We use these vocabulary files in subsequently described steps in order to build pronunciation dictionaries and to construct language models.

Pronunciation dictionaries. A pronunciation dictionary enables the speech engine to compose words from strings of basic sounds. In our analysis we decided that an entry in this dictionary would consist of the entry itself followed by the sequence of (possibly) marked phonemes, enclosed in double quotes. The phoneme markings can be stress indicators (used for vowels; 1 for primary stress, 2 for secondary stress, 0 for unstressed syllable) or position indicators (wi = start of word; wf = end of word; ws = one-phoneme word, start = end). An example entry is provided here:

absolutely “ae:2:wi b s ax:0 l uw:1 t l iy:0:wf”

Appendix B explains the phoneme set and provides example words containing each phoneme. We use the multimodal phoneme set since we used that engine in this research (see below for a more complete description of the speech engine). The phoneme set is very closely linked to the phoneme set of the publicly available Carnegie Mellon University (CMU) dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict). A small set of *special phones* is added to that list, which represents various kinds of noises, silence, and some frequent speech disfluencies.

In order to obtain a pronunciation dictionary, we first created a base dictionary with the entire Train90 vocabulary, using available existing encoded pronunciations in the CMU dictionary or in the Multimodal standard dictionary. Entries for which no encoding could be found were manually encoded. In our case, 139 entries had to be manually encoded. Where possible, we used words or word fragments from CMU’s or Multimodal’s dictionary, or worked with close analogues of letter-sound correspondences. The final dictionary contains 3,260 unique entries and 3,921 total entries, including pronunciation variants, that is, additional entries for a given type that in Multimodal's dictionary admit multiple pronunciations. For the 139 entries that we added, a single pronunciation entry was entered. In similar fashion, we also created a pronunciation dictionary based on the Train10 subcorpus that we used later for forced alignment purposes.

Language models. A language model (LM) allows us to estimate the probabilities of words and sequences of n-words. In other words, the language model provides the prior information, or P(W), in equations 1 and 2 discussed earlier. To build such a model, we first need to estimate the probability that a word follows another word, which requires obtaining the frequency of word pairs in a corpus bigram model. A trigram model estimates the probability of a word following a two-word string. However, the observed frequencies need to be adjusted and smoothed, since many specific combinations of words may not be observed in a corpus. This smoothing is a standard component of software for building language models. (See Jurafsky & Martin 2000, chapter 6, for an introduction to the methodology.)

Using the vocabulary of Train90 and the complete training data (all utterances), we built a Kneser-Ney trigram language model (Kneser & Ney, 1995; Ney & Essen, 1993) using software from Multimodal. The software allows for both Kneser-Ney and absolute discounting, two smoothing techniques, but it allows for different cutoff values for bigrams and trigrams (e.g., a cutoff = 1 for trigrams means that only trigrams with a frequency of >1, i.e., at least 2, are kept in the model). Testing both models with different parameters yielded, as measured by the lowest perplexity[4] on an independent test set, the Kneser-Ney language model with bigram cutoff 0 and trigram cutoff 1 as the best model choice. The foregoing LMs were trained using all five prompts in the training set. That is, the language model is not prompt-specific.

Results

Language Model Adaptation Experiments

The experiments described in this and the following sections follow a similar pattern. Usually, when a change in some parameter of the speech recognizer is made, the recognizer needs to be compiled first. Then a master script calls the recognizer decoding script with certain parameters as given by the experimental questions. For each subcorpus and parameter combination, a log-file is generated that contains important statistics about the recognizer's accuracy and errors for every utterance of the corpus. Finally, a script parses these log-files to produce the variables of interest for every speech sample. A sample of transcripts and their corresponding ASR hypotheses of different score levels, as well as a word-error analysis for parts of this sample, are presented in Appendix C.

Experiment 1: Original vs. Corpus-Specific Language Models

In this experiment, we evaluate the recognition performance based on Multimodal's original general-purpose dictation language model and performance, based on the language model described above, which is trained exclusively on the Train90 corpus; that is, excluding the Multimodal language model. (Note that the Multimodal LM is a native-speaker-dictation LM and, as such, not well suited for our corpus of spontaneous utterances by nonnative speakers.) The latter runs were done in two modes: (a) fast and a bit less accurate, emphasizing recognition speed; (b) slow and a bit more accurate that is, emphasizing accuracy. For this tradeoff, Multimodal's script uses the *speed vs. accuracy* (SvA) parameter, which can vary between 0.0 (fast) and 1.0 (slow). In the table below, we provide the *balanced word accuracy,* which is the mean between the reference-based and the hypothesis-based word accuracy. By *reference* we mean the transcription provided by a transcriber, whereas *hypothesis* refers to the transcription by the speech recognizer. The formula is:

$$\text{Wacbal} = 0.5*(C/(C + S + D) + C/(C + S + I)),$$

where C = correct words, S = substitutions, D = deletions, and I = insertions. It is easy to show that the balanced word accuracy is always in [0,1], provided that both reference and hypothesis are nonempty. As can be seen from Table 3 (refer to p. 17 for an explanation of the subcorpora), there is measurable improvement in recognition when the language model is based on data provided by the Train90 corpus. There is additional improvement when the recognition mode is set to maximize accuracy (SvA = 1).

Experiment 2: Language Model Interpolation

This experiment was conducted to determine whether recognition could be improved by mixing (*interpolating*) the language model in the engine with the language model based on the Train90 corpus. An interpolation weight of 0 means that the Multimodal model is ignored; weights greater than 0 give increasing weight to the Multimodal model. A value of 1 means that the Train90-based language model is ignored. We ran the experiment with the following interpolation weights: 0, 0.05, 0.1, 0.25, 0.5, and 1. Table 4 reports the recognition achieved for each of these relative weights of the two language models. As we can see, recognition for weights from 0 to .1 are the same to two decimal places. We chose the weight 0.05. Note that the results for w = 0 and w = 1 do not match the corresponding values in Table 3; this is due to the fact that the recognizer's acoustic models had been adapted before this experiment.

Table 3

Balanced Word Accuracy for Different Conditions and Corpora

Subcorpus	Multimodal LM	Train90-based LM	
	SvA = 0.2	SvA = 0.2	SvA = 1.0
Train10	0.179	0.253	0.275
DevTest	0.127	0.207	0.226
Eval	0.157	0.235	0.260
Corpus average	0.154	0.232	0.254

Parameter Tuning Experiments

In most speech recognizers, there are two parameters that are influential in the recognition process: Lp and Lz. Lp is the *language model penalty*: When it is high, inserting too many words is penalized and the engine will favor shorter possible transcriptions that contain, relatively speaking, longer words. Lz is the language model weight: When it is high, the language model receives a higher weight relative to the acoustic model and vice versa when Lz is low.

Table 4

Recognition Performance as a Function of Interpolation Weights on the Train10 Subcorpus

Interpolation weight							
0.0	0.02	0.05	0.1	0.2	0.5	0.75	1.0
0.2989	0.2974	0.2985	0.2985	0.2946	0.2776	0.2704	0.2064

We varied both of these parameters to determine an *optimal region*, where the recognizer works best for our data. Table 5 shows the results of this experiment and demonstrates that (Lp,Lz)

= (5, 23) or (5, 30) are the best configurations given our training and test material. The default setting of the Multimodal recognizer was (10, 23), which practically performs at par with (5, 23). We chose (5, 23) as the parameters for the work presented in this report.

Table 5

Wacmn as a Function of Tuning Lp and Lz Parameters Evaluated on DevTest Corpus

Lp	Lz				
	5	15	23	30	35
5	.236	.292	.297	.297	.288
10	.221	.283	.296	.290	.274
20	.219	.274	.274	.268	.256
30	.229	.244	.252	.240	.222
40	.209	.223	.222	.214	.199

Acoustic Model Experiments

While the LM determines the *sequence of words* being recognized, the acoustic model (AM) determines the *phoneme sequence* that is recognized based on acoustic feature vectors. It is based on a hidden Markov model (HMM) structure that is trained on a corpus to set the HMM weights. In particular, input vectors are represented as a mixture of Gaussian functions with two parameters each: mean and standard deviation. (For a description of Gaussian mixtures applied to speech recognition see e.g., Jurafsky & Martin, 2000, p. 267). In AM adaptation, predominantly it is the means of these Gaussian mixtures that are being modified based on additional training data. The structure of the HMM does not change during the adaptation process. Intuitively, the means can be seen as representing certain acoustic properties of the original training set. When additional data is used for AM adaptation, these acoustic properties will also change and this will be reflected in the Gaussian means.

The acoustic model adaptation procedure consists of the following steps:

1. Generating labels for all utterances of the training set, using all five prompts for every speaker, through *forced alignment*

In forced alignment mode, the recognizer has access to the transcription of the speech signal and tries to find the optimal path through the HMM in order to generate the known transcription. Labels are simply elementary units, such as words or phones, which are attached to a particular segment of time.

2. Adaptation of the acoustic model by using new input data

A set of previously unseen data is presented, together with its transcription, to an algorithm that tunes the weights of the Gaussian mixtures according to the acoustic properties of the new data in relationship to the data already known.

We iterated this process two times with the two corpora that we have a dictionary for (Train10 and Train90).[5] Table 6 shows the results of recognition runs with zero, one, or two rounds of adaptation for the mixture LM and for zero and one adaptation for the LM using the original Multimodal dictionary. As can be seen, maximum recognition as measured by balanced word accuracy, Wacmn, is reached after one iteration, and thus we use a system with this configuration for the experiments described below.

Table 6

Balanced Word Accuracy Depending on AM Adaptation

	Train10	DevTest	Eval
Mmdict 0 adapt	0.179	0.127	0.157
Mmdict 1 adapt	0.206	0.179	0.188
Mixdict 0 adapt	0.253	0.207	0.235
Mixdict 1 adapt	0.294	0.275	0.291
Mixdict 2 adapt	0.285	0.272	0.284

Note. Mmdict = Multimodal’s original dictionary and LM; Mixdict = mixture of Mmdict-LM and the corpus-specific LM, obtained from the training corpus.

Recognition With the Calibrated System

The foregoing results document the steps taken to adapt the speech engine to our data. After the various experiments described above, we set all system parameters to their optimal value, that is:

1. The language model relied primarily on our corpus, with only a very small weight assigned to the Multimodal language model (Table 4).

2. Only one acoustic model adaptation iteration was used (Table 6).
3. The Lp-Lz parameters were set to (5, 23), (Table 5).

Finally, we set the speed versus. accuracy parameter to 0.2 and 1.0, and report the results in Table 7. As can be seen in Table 7, balanced word accuracy on the Eval corpus, which has not been used in any of the adaptations of the engine, is highest, .34, at SvA = 1.0, and we adopt that parameter value. In other words, the engine as configured and trained can be expected to correctly recognize one in every three words for LanguEdge data.

Table 7

Balanced Word Accuracy of the System With Optimized Parameters in Two Conditions

	Train10	DevTest	Eval
SvA = 0.2	0.299	0.279	0.294
SvA = 1.0	0.345	0.317	0.336

Modeling Proficiency Scores

In this section we report on the relationship between ratings and speech recognition variables. We first report the agreement between raters. Only a portion of the speech samples was rated by two raters. We use the doubly scored samples for assessing the level of agreement between human raters. The level of agreement sets an upper bound on the level of prediction that can be obtained by a combination of speech variables when predicting the human scores. We then present a descriptive analysis of the relationship between the level of recognition accuracy and human scores. The intention of the analysis is purely descriptive because the level of recognition requires that the speech be transcribed so that it can be compared to the transcription produced by the scoring engines. Nevertheless, recognition accuracy is potentially related to speaking proficiency and therefore it would be informative to know whether a global measure of recognition from a speech engine is related to proficiency.

Interrater Agreement

The speech samples were scored following ETS procedures, which are described in the *Handbook for Scoring Speaking and Writing* (ETS, 2002b). In general, the procedure is to train and calibrate raters so that they apply the scoring rubrics in a similar fashion. As noted earlier, the rubrics, no matter how specific, leave ample room for interpretation. The results presented below bear out this prediction. Tables 8 to 13 show the cross-tabulations between two ratings for each of the five prompts, as well as for all the prompts combined. For each such table we can compute the level of agreement as the sum of the diagonal entries, expressed as a percentage of the total number of paired ratings. We refer to that statistic as *exact agreement*. For example, Table 8 is the cross-tabulation for the admire prompt. The exact agreement in this case is 53%, or 37 out of 70 cases. Tables 9 to 13 show the results for the remaining prompts and the combination of all the prompts. The prompt-by-prompt exact agreement ranges from 57% to 42%. For all prompts combined, exact agreement is 49%.

To put that level of agreement in context it is important to compare it with the agreement that can be expected by chance. That is, with a score scale of 5 points a significant portion of agreements can be expected even if the raters are producing scores at random. The expected number of chance agreements is given by joint probability of the diagonal entries under independence, which is obtained by multiplying the two marginal relative frequencies for each score level and summing them. The resulting sum is the probability of chance agreement. For the admire prompt, the chance agreement is 25%. For each table we also compute kappa (Cohen, 1960), which is a measure of agreement that adjusts for chance agreement.

The most obvious trend in these results is the relatively higher level of agreement that was observed for the independent tasks compared to the integrated tasks. Apparently, the integrated tasks are more difficult to score, which is not surprising since the raters must weigh how well the student understood the prompt and how well he or she spoke.

Although, as noted earlier, speaking proficiency is far more than comprehensibility, it is nevertheless useful to study how the level of recognition of a speech engine relates to scores by raters that should, to some extent, depend on how comprehensible the speaking sample is. In the speech recognition field, the performance of a recognition engine is typically measured by the error rate it achieves on some corpus. Two measures are used: One is recognition word error rate (WER) and the other is mean word accuracy (Wacmn).

$$\text{Wacmn} = [0.5*(C/(C + D + S) + C/(C + I + S))]$$

Table 8

Interrater Agreement for Admire Prompt

	Rater 1	Rater 2					Totals
		1	2	3	4	5	
	1	0	1	1	0	0	2
	2	1	4	4	2	0	11
	3	0	2	13	5	1	21
	4	0	1	6	12	6	25
	5	0	0	0	3	8	11
Totals		1	8	24	22	15	70

Note. Exact agreement is 53%; chance agreement is 25%; and kappa is .37.

Table 9

Interrater Agreement for Culture Prompt

	Rater 1	Rater 2					Totals
		1	2	3	4	5	
	1	0	1	1	0	0	2
	2	1	2	4	0	0	7
	3	0	2	11	1	1	15
	4	0	1	9	21	7	38
	5	0	0	1	6	12	19
Totals		1	6	26	28	20	81

Note. Exact agreement is 57%; chance agreement is 26%; and kappa is .41.

Table 10

Interrater Agreement for Technological Change Prompt

	Rater 1	Rater 2				Totals
		2	3	4	5	
	1	1	0	0	0	1
	2	2	2	2	0	6
	3	2	20	1	3	26
	4	0	9	8	6	23
	5	0	1	6	9	16
Totals		5	32	17	18	72

Note. Exact agreement is 54%; chance agreement is 27%; and kappa is .37.

Table 11

Interrater Agreement for Water Prompt

	Rater 1	Rater 2					Totals
		1	2	3	4	5	
	1	4	0	1	0	0	5
	2	3	6	6	1	0	16
	3	2	8	7	2	1	20
	4	0	2	7	8	3	20
	5	0	0	1	6	6	13
Totals		9	16	22	17	10	74

Note. Exact agreement is 42%; chance agreement is 21%; and kappa is .26.

Table 12

Interrater Agreement for Expression Prompt

	Rater 1	Rater 2					Totals
		1	2	3	4	5	
	1	6	6	3	0	0	15
	2	6	11	12	3	0	32
	3	0	6	21	11	3	41
	4	0	3	13	14	6	36
	5	0	0	5	12	16	33
Totals		12	26	54	40	25	157

Note. Exact agreement is 43%; chance agreement is 21%; and kappa is .28.

Table 13

Interrater Agreement for All Prompts

	Rater 1	Rater 2					Totals
		1	2	3	4	5	
	1	10	9	6	0	0	25
	2	11	25	28	8	0	72
	3	2	20	72	20	9	123
	4	0	7	44	63	28	142
	5	0	0	8	33	51	92
Totals		23	61	158	124	88	454

Note. Exact agreement is 49%; chance agreement is 23%; and kappa is .34.

Recognition Accuracy and Human Scores

We can compute WER or Wacmn for each speaking sample or set of speaking samples for a given student. For the following analysis we computed both for students who had completed all five prompts. We then obtained the correlation matrix with the five corresponding Rater1 scores. We conducted an exploratory factor analysis. A plot of the eigenvalues suggested unambiguously that the number of factors needed to account for the correlations was 2. Two factors were extracted by means of the Principal Factors method. The solution was initially rotated to a Varimax criterion and then by the Promax criterion. The resulting factor pattern matrix appears in Table 14.

Table 14

Factor Pattern Matrix for Rater1 Scores and Wacmn for Each Prompt

Variable	Factor 1	Factor 2
Rater1_admire	0.11	**0.60**
Rater1_culture	0.02	**0.71**
Rater1_technology	0.01	**0.76**
Rater1_water	0.00	**0.77**
Rater1 expression	–0.10	**0.71**
Wacmn_admire	**0.80**	0.02
Wacmn_culture	**0.82**	–0.07
Wacmn_technology	**0.84**	–0.015
Wacmn_water	**0.75**	0.057
Wacmn_expression	**0.75**	0.02

Note. The highest loading in each row is bolded.

Factor 1 is defined by the Wacmn. That is, the variability in Wacmn across all prompts and students is explained by a single underlying variable. The importance of this result is that Wacmn is a reliable indicator of *some* aspect of speaking performance. Similarly, Factor 2 is defined by the Rater1 scores and, because we know that the raters were operating from a scoring rubric, the second factor can be defined as speaking proficiency as measured by Rater1. The correlation between these two factors is 0.22 and therefore 4% of the variance of the Wacmn factor is in common with speaking proficiency as defined by the Rater1. Recall that Wacmn is computed by comparing the human transcription of the speech sample to the transcription obtained by the speech recognition engine. Clearly, the result of this comparison is a reliable albeit weak indicator of speaking proficiency. A second factor analysis was conducted using the same data but a different measure of error rate, WER, computed as:

$$\text{WER} = (S + D + I)/(1 + I + D + S).$$

Two factors emerged again but in this case the proportion of variance in common between the factor defined by Rater1 scores in all five prompts and the WER values in each prompt was 9%. Therefore, variations in the characterization of recognition performance can have more or less in common with proficiency as measured by human raters.

Classification of Speakers Into Categories by Means of a Classification Tree

As noted earlier, speaking proficiency as defined in LanguEdge refers to more than comprehensibility. In this section we present results from an analysis that attempts to classify speakers into proficiency levels using human scores to train a classification tree. We aim to incorporate into this analysis, variables that capture some aspects of speaking proficiency. However, the analysis should be seen as exploratory in that we did not attempt at this stage to incorporate all relevant variables. In general, significant additional research will be required to determine how to realize important aspects of speaking proficiency computationally. The initial results presented here are based on a set of features that are readily computed from the speech engine.

Table 15 shows the variables used in this analysis. We distinguish several classes of variables. One version of these variables is computed on an *aligned* human transcript that has been marked (aligned) with the start and duration of each utterance in 10-millisecond resolution. In aligned mode the engine behaves as if it had perfect recognition as defined by the transcriber we used for these data. Another version of these variables is computed from the actual output of the speech engine. As we shall see below, we use the variables computed both ways. The variables computed in aligned mode simulate perfect recognition, where *perfect* recognition is the transcription obtained by the transcriber. The variables computed in the unaligned case provide a more realistic assessment of how well speakers' proficiency can be classified based on the actual level of recognition we were to achieve, which as has been noted was only .34. The variables are divided into lexical counts: variables capturing the nature of silences between words, speaking rate, and lexical similarity. They are described in Table 15.

Table 15

Definition of the Variables Used to Characterize Speech Samples

Variable	Definition
Human scores	
Rater 1	We used scores assigned by raters as a characterization of speaking proficiency. The scores ranged from 1 to 5, with 5 indicating the highest level of proficiency. Every student we included in our analysis had received a Rater 1 rating by one of several raters. (That is, Rater 1 is not a specific rater.)
Rater 2	A portion was rated by a second rater, Rater 2, to analyze interrater agreement
Recognition error rates	We used two measures of recognition performance to characterize how accurately the speech engine recognized the speech samples. This requires the alignment of the transcribed speech (human transcript) and the output of the recognizer, or hypothesized transcript. An optimization algorithm is then used to resolve the differences between the two transcripts in terms of insertions (I), deletions (D), and substitutions (S). The goal of the algorithm is to simultaneously minimize these three types of errors.
WER	Word error rate is computed as (S + D + I)/(1 + C + D + S).
Wacmn	Mean word accuracy attempts to characterize the recognition performance in a more balanced way by equally weighing the human and machine transcripts. The formula is [0.5*(C/(C + D + S) + C/(C + I + S))].
Utter. counts Numutt	Number of utterances in the response to a prompt. An utterance is defined as an uninterrupted segment of speech, which is to say, uninterrupted speech preceded and followed by silence.

(Table continues)

Table 15 (continued)

Variable	Definition
Lexical counts	
Numwds	The total number of word forms in the speech sample that are found in the pronunciation dictionary
Numdff	A disfluency is an interruption of speech by a class of paralinguistic phenomena (such as *uh* and *um*). Numdff is the number of such phenomena in response to a prompt.
Numtok	Number of tokens = numwds + numdff
Types	Number of unique word forms (e.g., *house* and *houses* are different word forms) in the speech sample
Ttratio	The ratio types/numtok. The inverse of Ttratio can be interpreted as the number of times, on average, a word form is repeated.
Fluency	These variables are meant to characterize the fluency of the speech.
Numsil	A silence is defined as acoustic event that has no discernible phonetic content and can be of variable length. Numsil is the number of such events over the entire speech sample, excluding silences between utterances, which are used by the transcribers to segment a speech sample into a sequence of utterances.
Silpwd	The ratio numsil/numwrd
Silmean	Mean duration in seconds of all silences in a response to a prompt
Silstddv	Standard deviation of silence durations
Length of speech sample	
Segdur	Total duration in seconds of all the utterances

(Table continues)

Table 15 (continued)

Variable	Definition
Rate measures	These words are meant to characterize the rate of speech.
Wpsec	The ratio numwrd/segdur
Dpsec	The ratio numdff/segdur
Tpsec	The ratio types/segdur
Silpsec	The ratio numsil/segdur
Lexical similarity	These measures characterize the lexical similarity of a student's transcript to some corpus. In the present case, the corpus is Train90. Specifically, for each prompt the frequency of word forms in Train90 is obtained. We refer to the resulting word frequencies as *reference content vectors*. To assess the similarity of a given speech sample, a corresponding sample content vector is obtained by tabulating the frequencies of word forms in that speech sample.
Cvfull	The inner product of a speech sample and reference content vectors. The reference content vector consists of the raw frequency of word forms across speech samples of a given prompt in Train90. The speech sample content vector consists of the raw frequency of word forms for a given speech sample.
Cospword	The ratio cvfull/numwords

Classification and regression trees were introduced by Breiman, Friedman, Olshen, and Stone (1984). An overview can be found in Ripley (1996), including discussion of predecessors (Quinlan, 1986). The goal of a classification tree is to classify the data such that the data in the terminal or classification nodes are as pure as possible meaning all the cases have the same true classification; in the present case, this is a score provided by a human rater (the variable Rater1, above). At the top of the tree, all the data are available and are split into two groups based on a split of one of the variables available. Each split is treated in the same manner until no further splits are possible, in which case a terminal node has been reached. Ideally, all the cases in a terminal node have received the same value of Rater1. The procedure aims for that ideal or the closest that it can come to that. In growing the trees there is opportunity to capitalize on chance.

The CART procedure uses an internal cross-validation method based on resampling or growing the tree while holding back a portion of the data. The resulting trees should therefore be robust. Nevertheless, we further validate the trees on a portion of the data that were not used at all in the estimation of the tree. Specifically, we follow these steps for each prompt.

- Estimate a classification tree based on Train90 aligned data. This tree is grown by applying the built-in cross-validation mentioned above. The performance of the tree is measured by a classification table cross-tabulating the score from Rater1 and the classification score provided by the tree. As with the agreement between raters, the diagonal contains the cases of exact agreement. The percentage of exact agreements between Rater1 and the classification scores constitute the classification accuracy of the tree.

- We apply the tree from the previous step to the *aligned* Test corpus (consisting of the Train10, Eval, and DevTest corpora) to assign a score to each record in the Test corpus. It should be noted that the data in the Test corpus were used neither in the optimization and adaptation of the speech engine *nor* in the estimation of the tree. (An exception to this is the sub-corpus Train10, which was used for perplexity-based LM selection and for parameter selection in the tuning experiments.) As noted earlier, the aligned data simulates the condition where the recognizer is performing at the level of a human transcriber. We summarize performance by means of classification tables. Of especial interest is the *exact agreement* between Rater1 and the tree. Exact agreement with the Train90 data is expected to be higher than agreement based on the aligned test corpus.

- Finally, we apply the tree to the *unaligned* Test corpus to assign a score to each record as in the previous step, except that the variables in Table 15 are now computed on the transcript output by the recognizer rather than the transcriber. We summarize performance by means of classification tables. The exact agreement with Rater1 is expected to be lower than agreement based on the aligned test corpus because the data are based on actual recognition, and the data were not part of the training of the speech engine.

- We present for each prompt the estimated tree and a tabular version of the tree that contains the variables and the splits that are used to score the data and provide a sense of the *reasoning* used to assign a specific score.

Independent Tasks

We first present the results for the three independent prompts: admire, culture, and technological change.

Admire prompt. The tree for the admire prompt and the corresponding tabular representation of the tree appear in Figure 2 and Table 16. A score of 1 is assigned if Ttratio is less than or equal to .46. In other words, word forms are repeated approximately twice on average. Higher scores are assigned for Ttratio > .46; that is, when word forms are repeated less often. To obtain a 3 rather than a 2, Numtok higher than 105 is needed. To obtain a 4 rather than a 3, Tpsec needs to be above 1.42 tokens per second. Finally, to obtain a 5 it is necessary to speak for more than 45 seconds or have fewer than 28.5 disfluencies. In general, this tree suggests that higher proficiency is associated with the fewer repetitions of word forms, lengthier responses, higher rate of speech, and fewer disfluencies.

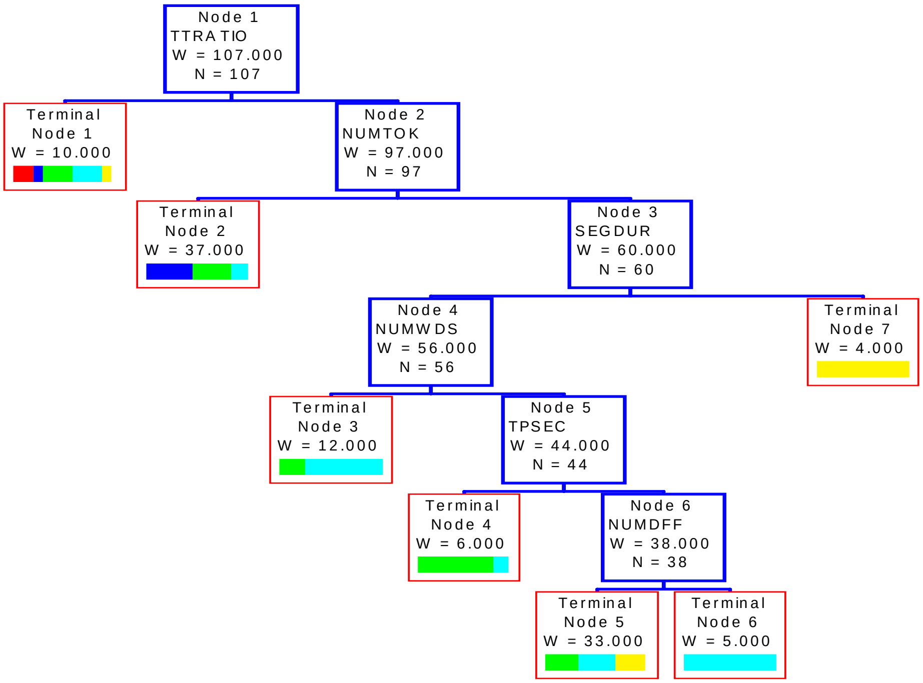

Figure 2. **Classification tree for admire prompt based on aligned Train90 data.**

Table 16

Tabular Representation of the Admire Prompt Tree

Node	Class	*N*	Pathway					
			Ttratio	Numtok	Segdur	Numwrds	Tpsec	Numdff
1	1	10	< = .46					
2	2	37	> .46	< = 105	< = 45.1			
4	3	6	>.46	> 105	< = 45.1	> 92.5	< = 1.42	
6	4	5	> .46	> 105	< = 45.1	> 92.5	> 1.42	> 28.5
3	4	12	> .46	> 105	< = 45.1	< = 92.5		
7	5	4	>.46	> 105	> 45.1			
5	5	33	> .46	> 105	< = 45.1	> 92.5	> 1.42	< = 28.5

Tables 17–19 show the classification accuracy for the admire tree based on the Train90 data and the test data with and without alignment. As noted earlier, the aligned case simulates a condition where the engine recognizes at the same level as a human transcriber, whereas the unaligned data is based on the actual engine recognition. In the case of admire, there are substantial drops in accuracy when the tree is applied to the test corpus with and without alignment.

Table 17

Cross-Tabulation of Rater1 and Tree Score Based on the Training Corpus (Train90) Aligned Data and Internal v-fold Validation for the Admire Prompt

Actual class	Total cases	% correct	Score				
			1 (*N* = 8)	2 (*N* = 37)	3 (*N* = 8)	4 (*N* = 28)	5 (*N* = 26)
1	2	0.0	0	0	1	1	0
2	18	83.3	1	15	0	2	0
3	36	8.3	2	16	3	7	8
4	36	27.7	3	6	4	10	13
5	15	33.3	2	0	0	8	5

Note. Exact agreement is 33/107 = 31%; chance agreement is 18%; and kappa is .16.

Table 18

Cross-Tabulation of Rater1 and Tree Score Based on Test Corpus (Eval and Devtest) Aligned Data for the Admire Prompt

Actual class	Total cases	% correct	Score				
			1 (*N* = 7)	2 (*N* = 26)	3 (*N* = 7)	4 (*N* = 7)	5 (*N* = 12)
1	3	0.0	0	3	0	0	0
2	8	87.5	1	7	0	0	0
3	20	0.0	3	9	0	5	3
4	20	5.0	3	5	7	1	4
5	8	62.5	0	2	0	1	5

Note. Exact agreement is 13/59 = 22%; chance agreement is 16%; and kappa is .07.

Table 19

Cross-Tabulation of Rater1 and Tree Score Based on Test Corpus (Eval and Devtest) NonAligned Data for the Admire Prompt

Actual class	Total cases	% correct	Score				
			1 (*N* = 7)	2 (*N* = 12)	3 (*N* = 3)	4 (*N* = 26)	5 (*N* = 11)
1	3	0.0	0	3	0	0	0
2	8	25.0	1	2	0	5	0
3	20	10.0	3	2	2	11	2
4	20	30.0	2	3	1	6	8
5	8	12.5	1	2	0	4	1

Note. Exact agreement is 11/59 = 19%; chance agreement is 20%; and kappa is –.02.

Culture prompt. The tree for the culture prompt and the corresponding tabular representation of the tree appear in Figure 3 and Table 20. A score of 1 is assigned if Segdur is less than or equal to 19 seconds. The next three splits (Numtok, Numwds, and Numdff) suggest that higher scores are associated with higher Numtok and Numwds, and lower Numdff; that is, richer vocabulary. The next three variables appear to be used to sort out alternative ways to earn a given score level.

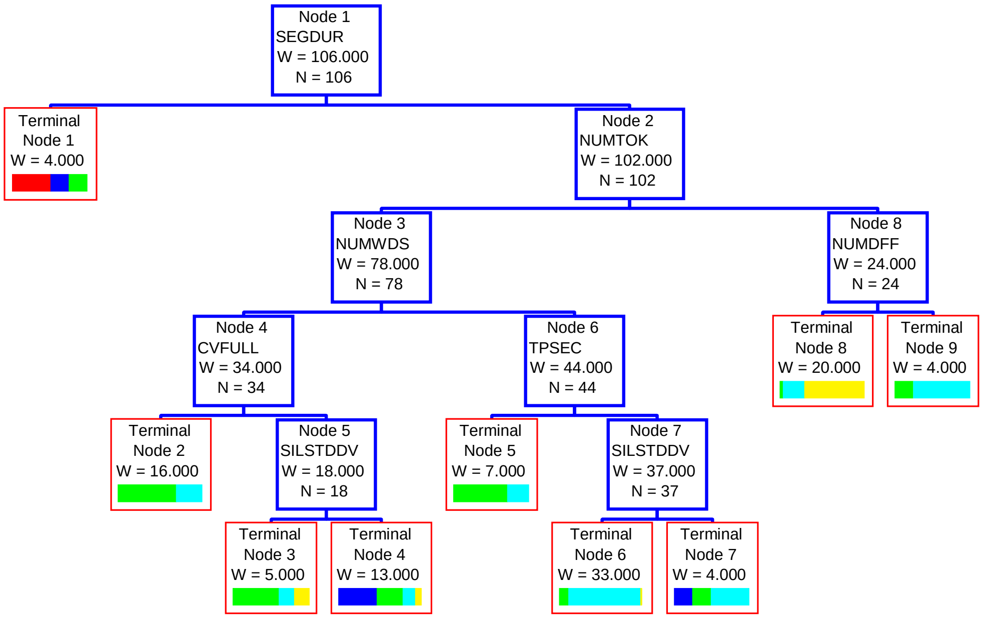

***Figure 3.* Classification tree for culture prompt.**

Table 20

Tabular Representation of Culture Prompt Tree

		Pathway						
Node	Class	Segdur	Numtok	Numwds	Numdff	Cvfull	Tpsec	Silstddv
1	1	< = 19.1						
7	2	> 19.1	< = 141.5	> 91.5			> 1.28	> .1
4	2	> 19.1	< = 141.5	< = 91.5		> 1,678		> .04
2	3	> 19.1	< = 141.5	< = 91.5		< = 1,678		< = .04
3	3	> 19.1	< = 141.5	< = 91.5		> 1,678		< = .04
5	3	> 19.1	< = 141.5	> 91.5			< = 1.28	
6	4	> 19.1	< = 141.5	> 91.5			> 1.28	< = 1
9	4	> 19.1	> 141.5		> 30.5			
8	5	> 19.1	> 141.5		< = 30.5			

Tables 21–23 show the classification accuracy for the culture tree based on the Train90 data and the test data with and without alignment. In the case of culture, unlike the admire results just presented, the classification accuracy remains constant for the aligned test data and actually improves for the unaligned case although this fact should not be over-interpreted, because the kappa does not increase in the same fashion.

Table 21

Cross-Tabulation of Rater1 and Tree Score Based on the Training Corpus (Train90) Aligned Data and Internal v-fold Validation for the Culture Prompt

Actual class	Total cases	% correct	Score				
			1	2	3	4	5
			(*N* = 2)	(*N* = 19)	(*N* = 31)	(*N* = 25)	(*N* = 29)
1	2	0.000	0	1	1	0	0
2	8	25.000	1	2	3	2	0
3	31	38.710	1	8	12	6	4
4	48	33.333	0	6	14	16	12
5	17	76.471	0	2	1	1	13

Note. Exact agreement is 43/106 = 41%; chance agreement is .23, and kappa is .23.

Table 22

Cross-Tabulation of Rater1 and Tree Score Based on Test Corpus (Eval and Devtest) Aligned Data for the Culture Prompt

Actual class	Total cases	% correct	Score				
			1	2	3	4	5
			(*N* = 2)	(*N* = 6)	(*N* = 14)	(*N* = 28)	(*N* = 10)
1	0		0	0	0	0	0
2	6	16.667	2	1	2	1	0
3	10	50.000	0	3	5	2	0
4	32	53.125	0	1	7	17	7
5	12	25.000	0	1	0	8	3

Note. Exact agreement is 26/60 = 43%, chance agreement is 31%, and kappa is .18.

Table 23

Cross-Tabulation of Rater1 and Tree Score Based on Test Corpus (Eval and Devtest) Nonaligned Data for the Culture Prompt

Actual class	Total cases	% correct	Score				
			1 (*N* = 0)	2 (*N* = 6)	3 (*N* = 10)	4 (*N* = 40)	5 (*N* = 4)
1	0		0	0	0	0	0
2	6	0.000	0	0	3	3	0
3	10	40.000	0	1	4	4	1
4	32	78.125	0	4	2	25	1
5	12	16.667	0	1	1	8	2

Note. Exact agreement is 31/60 = 52%, chance agreement is 40%, and kappa is .19.

Technological change. The tree for the technological change prompt and the corresponding tabular representation of the tree appear in Figure 4 and Table 24. A score of 1 is assigned if Wpsec is less than or equal to 1.86. Higher scores are obtained with Wpsec greater than 1.86. Score 2 is assigned if Numutt is < = 2.5; speakers with more utterances get a higher score. When Cospword is greater than 5.22, a score of 3 is assigned. For a score of 4, types need to be > 50.5, and finally, for a score of 5, the number of words has to be greater than 106 (i.e., the speaker has to be quite verbose, compared to speakers with lower scores).

Table 24

Tabular Representation of Technological Change Prompt Tree

Node	Class	Wpsec	Numwds	Numutt	Pathway	
					Cospword	Types
1	1	< = 1.86				
2	2	> 1.86	< = 106	< = 2.5		
3	2	> 1.86	< = 106	< = 2.5	< = 5.22	
4	3	> 1.86	< = 106	> 2.5	> 5.22	< = 50.5
5	4	> 1.86	< = 106	> 2.5	> 5.22	> 50.5
6	5	> 1.86	> 106			

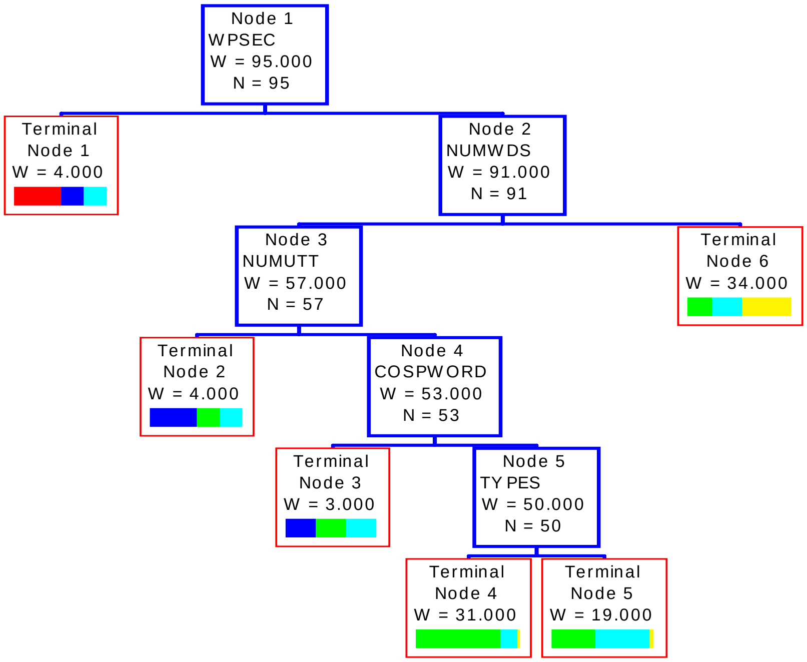

Figure 4. **Classification tree for technological change prompt.**

Tables 25–27 show the classification accuracy for the technological change prompt based on the Train90 data and the test data with and without alignment. In this case, we see results like those for the admire prompt, where classification accuracy and corresponding kappas degrades markedly.

Table 25

Cross-Tabulation of Rater1 and Tree Score Based on the Training Corpus (Train90) Aligned Data and Internal v-fold Validation for the Technological Change Prompt

Actual class	Total cases	% correct	Score				
			1 (*N* = 3)	2 (*N* = 10)	3 (*N* = 38)	4 (*N* = 14)	5 (*N* = 30)
1	2	0.000	0	0	2	0	0
2	4	0.000	1	0	2	0	1
3	43	48.837	0	6	21	8	8
4	28	10.714	2	3	10	3	10
5	18	61.111	0	1	3	3	11

Note. Exact agreement is 33/107 = 30%, chance agreement is 25%, and kappa is .16.

Table 26

Cross-Tabulation of Rater1 and Tree Score Based on Test Corpus (Eval and Devtest) Aligned Data for the Technological Change Prompt

Actual class	Total cases	% correct	Score				
			1 (*N* = 1)	2 (*N* = 4)	3 (*N* = 11)	4 (*N* = 17)	5 (*N* = 20)
1	1	0.000	0	0	1	0	0
2	6	0.000	1	0	3	0	2
3	20	30.000	0	3	6	9	2
4	21	28.571	0	1	1	6	13
5	5	60.000	0	0	0	2	3

Note. Exact agreement is 13/53 = 28%; chance agreement is 22%, and kappa is .08.

Table 27

Cross-Tabulation of Rater1 and Tree Score Based on Test Corpus (Eval and Devtest) Nonaligned Data for the Technological Change Prompt

Actual class	Total cases	% correct	Score				
			1 (*N* = 0)	2 (*N* = 7)	3 (*N* = 7)	4 (*N* = 6)	5 (*N* = 33)
1	1	0.000	0	0	1	0	0
2	6	16.667	0	1	3	0	2
3	20	15.000	0	3	3	5	9
4	21	4.762	0	2	0	1	18
5	5	80.000	0	1	0	0	4

Note. Exact agreement is 11/53 = 17%, chance agreement is 12%, and kappa is .06.

Integrated Tasks

This section presents results from the two integrated prompts: expression and water.

Expression prompt. The tree for the expression prompt and the corresponding tabular representation of the tree appear in Figure 5 and Table 28. A score of 1 is assigned if Types is less than or equal to 39. The scores in the range 2 to 4 are increasingly associated with higher Dpsec and Cvfull. Although it seems counterintuitive that a lower Dpsec would be associated with a score of 2 it is likely that students who spoke less, and less fluently, therefore had fewer disfluencies per second. The deciding variable between a 3 and a 4 is Numsil and, as can be seen, a higher number of silences is associated with a score or 4. Because silences are measured as a count, it is not necessarily counterintuitive that more silences would be associated with a score of 4 rather than a 3, because the longer you speak the more opportunities there are for silences to occur. A score of 5 can be obtained in two ways, by having a smaller Silstddv or a higher Segdur. That is, by speaking more fluidly or longer.

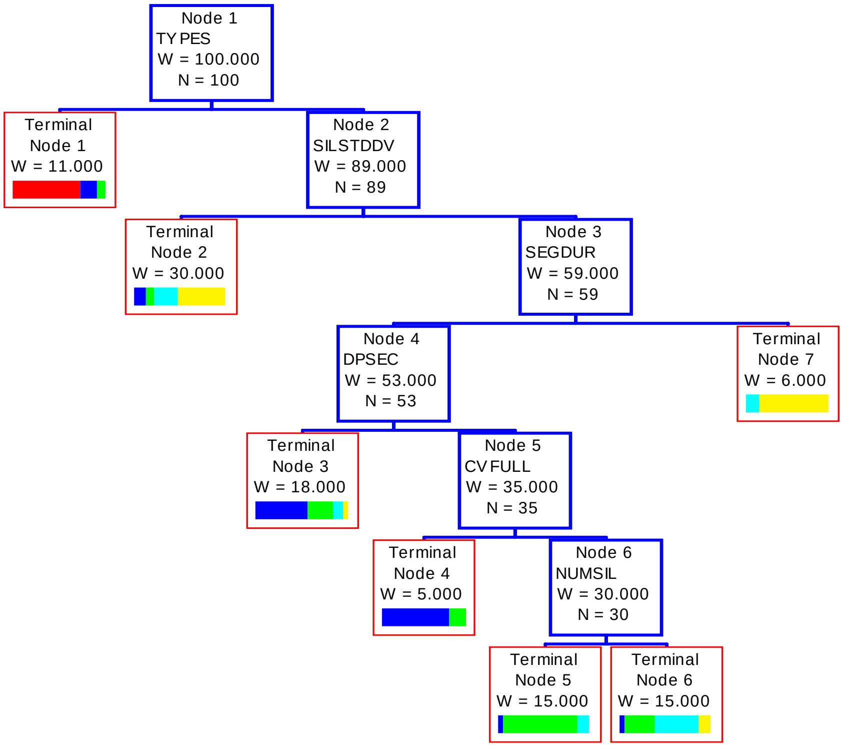

***Figure 5.* Classification tree for expression prompt.**

Tables 29–31 show the classification accuracy for the expression prompt based on the Train90 data and the test data with and without alignment. In this case, classification accuracy does not decline rapidly as with two earlier prompts.

Table 28

Tabular Representation of Expression Prompt Tree

Node	Class	Pathway					
		Types	Silstddv	Segdur	Dpsec	Cvfull	Numsil
1	1	< = 38.5					
3	2	> 38.5	> .04	< = 59.9	< = .35		
4	2	> 38.5	> .04	< = 59.9	> .35	< = 2,342	
5	3	> 38.5	> .04	< = 59.9	> .35	> 2,342	< = 43.5
6	4	> 38.5	> .04	< = 59.9	> .35	> 2,342	> 43.5
7	5	> 38.5	> .04	> 59.9			
2	5	> 38.5	< = .04				

Table 29

Cross-Tabulation of Rater1 and Tree Score Based on the Training Corpus (Train90) Aligned Data and Internal v-fold Validation for the Expression Prompt

Actual class	Total cases	% correct	Score				
			1	2	3	4	5
			(*N* = 12)	(*N* = 18)	(*N* =14)	(*N* = 18)	(*N* = 38)
1	8	87.500	7	0	0	0	1
2	22	18.182	3	4	2	8	5
3	27	22.222	2	8	6	4	7
4	20	10.000	0	4	4	2	10
5	23	65.217	0	2	2	4	15

Note. Exact agreement is 34/100 = 34%, chance agreement is 16%, and kappa is .22.

Table 30

Cross-Tabulation of Rater1 and Tree Score Based on Test Corpus (Eval and Devtest) Aligned Data for the Expression Prompt

Actual class	Total cases	% correct	Score				
			1 (*N* = 8)	2 (*N* = 15)	3 (*N* = 10)	4 (*N* = 7)	5 (*N* = 17)
1	7	85.714	6	1	0	0	0
2	10	30.000	1	3	2	2	2
3	14	28.571	1	3	4	3	3
4	16	6.250	0	6	4	1	5
5	10	70.000	0	2	0	1	7

Note. Exact agreement is 21/57 = 37%, Chance Agreement is 16%, and kappa is .25.

Table 31

Cross-Tabulation of Rater1 and Tree Score Based on Test Corpus (Eval and Devtest) Nonaligned Data for the Expression Prompt

Actual class	Total cases	% correct	Score				
			1 (*N* = 9)	2 (*N* = 16)	3 (*N* = 6)	4 (*N* = 0)	5 (*N* = 26)
1	7	85.714	6	1	0	0	0
2	10	40.000	2	4	1	0	3
3	14	14.286	0	6	2	0	6
4	16	0.000	1	3	2	0	10
5	10	70.000	0	2	1	0	7

Note. Exact agreement is 19/57 = 33%, chance agreement is 12%, and kappa is .25.

Water prompt. Figure 6 shows the CART tree for the water prompt, and Table 32 describes the pathways in this tree. Score 1 is given when the ttratio (type-token ratio) is >.58. Speakers with a high type-token ratio are typically of poor proficiency since they use fewer words and hence also have fewer chances of word repetition. When ttratio < = .51 and also Numwds < = 126, Score 2 is assigned. These speakers talk for a relatively short time (numwds) but still have a relatively higher word repetition rate than those with score of 1. For a score of 3, the ttratio is between .51 and .58 (i.e., slightly lower than for Score 1 and higher than for Score 2), which is a bit counterintuitive. Additionally, the Wpsec (speaking rate) is smaller than 2.29 (about 138 words per minute), which is a rather slow rate. Score 4 also has a ttratio < = .58, and between 133 and 206 words in the response. Score 5 is hard to describe intuitively since it has three different manifestations in the CART tree: (a) as Score 4 but containing between 126 and 133 words; this is rather counterintuitive since one would expect a *higher* number of words here than for Score 4. This is the case for (b), where the response has to contain more than 206 words. Finally, (c) has the same properties as the conditions for Score 3, but here the Wpsec has to be > 2.29, which makes sense. More proficient speakers typically speak faster than less proficient speakers.

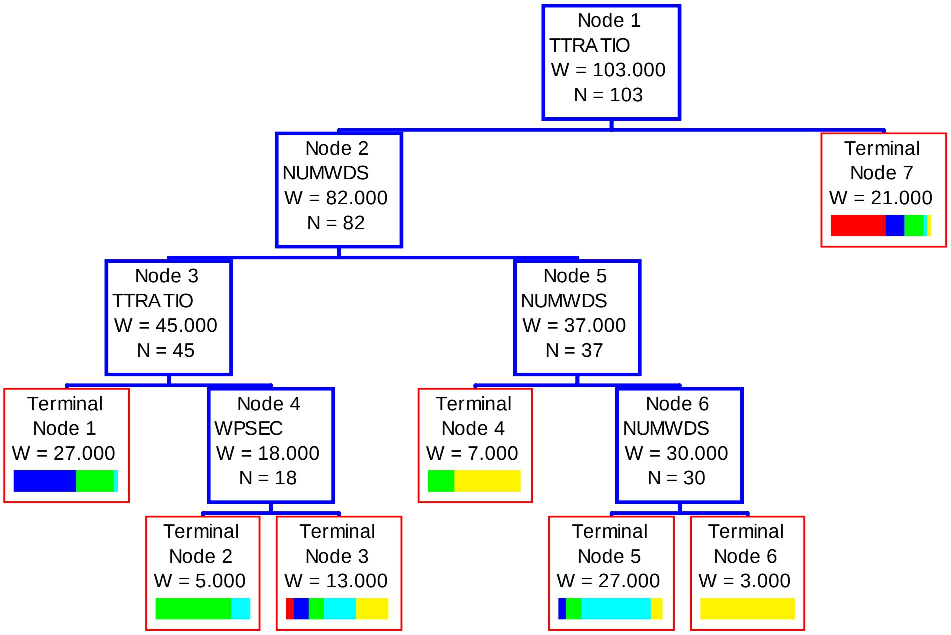

***Figure 6.* Classification tree for water prompt.**

Table 32

Tabular Representation of Water Prompt Tree

Node	Class	Pathway			
		Ttratio	Numwds	Ttratio	Wpsec
1	1	> .58			
1	2	< = .51	< = 126		
2	3	> .51	< = 126	.51 to .58	< = 2.29
5	4	< = .58	133 to 202		
4	5	< = .58	126 to 133		
6	5	< = .58	> 202		
3	5		< = 126	.51 to .58	> 2.29

Tables 33–35 show the classification accuracy for the water tree based on the Train90 data and the test data with and without alignment. For the aligned test data, the classification accuracies vary widely between 6% (score in Class 3) and 80% (score in Class 1). For nonaligned data, the overall accuracy and kappa decrease markedly; in particular, Score Class 3 remains problematic, with a classification accuracy of less than 1%. Part of the reason for this large divergence could be the rather convoluted tree shown in Figure 6.

Table 33

Cross-Tabulation of Rater1 and Tree Score Based on Train90 Corpus Using Nonaligned Data for the Water Prompt

Actual class	Total cases	% correct	Score				
			1 (*N* = 24)	2 (*N* = 21)	3 (*N* = 14)	4 (*N* = 23)	5 (*N* = 21)
1	12	75.000	9	1	1	0	1
2	24	37.500	4	9	6	3	2
3	26	7.692	7	8	2	5	4
4	25	28.000	3	2	2	7	11
5	16	18.750	1	1	3	8	3

Note. Exact agreement is 30/103 = 29%, chance agreement is 17%, and kappa is .15.

Table 34

Cross Tabulation of Rater1 and Tree Score Based on Test XCorpus (Eval and Devtest) Aligned Data for the Water Prompt

Actual class	Total cases	% correct	Score				
			1 (*N* = 13)	2 (*N* = 12)	3 (*N* = 4)	4 (*N* = 15)	5 (*N* = 14)
1	5	80.000	4	0	0	0	1
2	13	23.077	3	3	1	1	5
3	17	5.882	2	8	1	4	2
4	18	50.000	3	1	2	9	3
5	5	60.000	1	0	0	1	3

Note. Exact agreement is 20/58 = 35%; Chance agreement is 17%, and kappa is .21.

Table 35

Cross-Tabulation of Rater1 and Tree Score Based on Test Corpus (Eval and Devtest) Nonaligned Data for the Water Prompt

Actual class	Total cases	% correct	Score				
			1 (*N* = 19)	2 (*N* = 1)	3 (*N* = 3)	4 (*N* = 20)	5 (*N* = 15)
1	5	80.000	4	0	0	0	1
2	13	0.000	6	0	2	3	2
3	17	5.882	5	1	1	6	4
4	18	44.444	4	0	0	8	6
5	5	40.000	0	0	0	3	2

Note. Exact agreement is 15/58 = 26%, chance agreement is 16%, and kappa .is 12).

Integration

To put the results in perspective, we collect the agreement estimates and kappa coefficients for each prompt in corresponding tables, and then we plot the results. Table 36 shows observed exact agreement between raters in the first column. The next three columns are observed exact agreement between the classification tree and Rater1. The order of these three columns is according to the expected level of agreement, where the train-aligned conditions should yield a higher agreement than the test-aligned and test-unaligned conditions. The last column shows the chance agreement between raters. The last row shows the observed exact agreement between raters for all prompts combined and the corresponding chance agreement. Figure 7 displays the information in Table 36. Table 37 shows the corresponding kappa agreement for each prompt, and Figure 8 is the corresponding graphic representation.

Table 36

Summary of Exact Agreement Under Different Conditions for Each Prompt and All Prompts Combined

Prompts	Human perfect agreement between raters	Human computer train-aligned	Human computer test-aligned	Human computer test-unaligned	Chance agreement between graders
Admire	0.53	0.31	0.22	0.19	0.25
Culture	0.57	0.41	0.43	0.52	0.26
Technological change	0.54	0.37	0.28	0.17	0.27
Water	0.42	0.29	0.35	0.26	0.21
Expression	0.43	0.34	0.37	0.33	0.21
All	0.49				0.23

The inclusion of exact agreement between raters as well as chance agreement between raters provides a context for interpreting the agreement between human and computer. That is, the agreement between raters should be higher than the agreement between raters and tree, and the latter should be higher than chance agreement between raters to the extent that the variables capture reliably aspects of speaking proficiency. Moreover, the slope of the trend is potentially informative in characterizing the results. In particular, the drop in agreement between raters and the agreement between tree and computer under the train-aligned condition, the most favorable condition for rater-tree agreement, gives us an indication that the features we extracted are not sufficient to characterize speaking proficiency. At the other end, the fact that for three prompts (culture, expression, and water) the rater-tree agreement remains above chance agreement between raters, even under the test-unaligned condition (the least favorable) suggests that some aspect of speaking proficiency is being captured reliably by the variables.[6] This inference is reinforced by the examination of kappa statistics. Table 37 and Figure 8 suggest a similar pattern of results, with the added bonus that for the culture prompt the unexpectedly high agreement under the test-unaligned condition does not show with kappa.

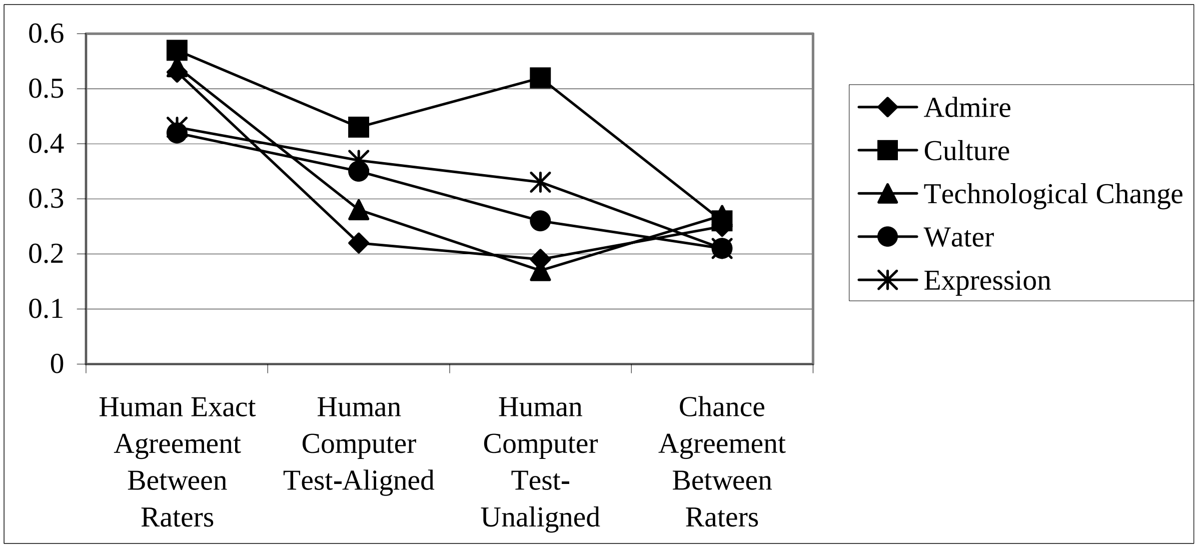

Figure 7. **Agreement measures by condition and prompt.**

Figure 7 shows marked differences between prompts. For two of the prompts, admire and technological change, there is a much larger drop in agreement between human and the test -aligned condition. It is for these prompts that the agreement between raters was highest to begin with. For the other three prompts the drop is not as marked, and it is also the case that raters show the least agreement. This interaction of *scorability* and prompt should be investigated further. Note in particular that it does not appear to be due to the *integrativeness* of the prompts, since both types of prompts are represented in the two patterns of agreement. Nevertheless, scoring of the integrated tasks logically requires an evaluation of the student's comprehension of the stimulus, whether spoken or written, and there were no features among the ones we used designed to capture that variability.

Table 37

Summary Kappa Statistic for Between Raters and Between Tree and Rater Under Different Conditions

Prompts	Human kappa	Computer-human Human computer train-aligned	 Human computer test-aligned	 Human computer test-unaligned
Admire	0.37	0.16	0.07	-0.02
Culture	0.41	0.23	0.18	0.19
Technological change	0.37	0.16	0.08	0.06
Water	0.26	0.15	0.21	0.12
Expression	0.28	0.22	0.25	0.25

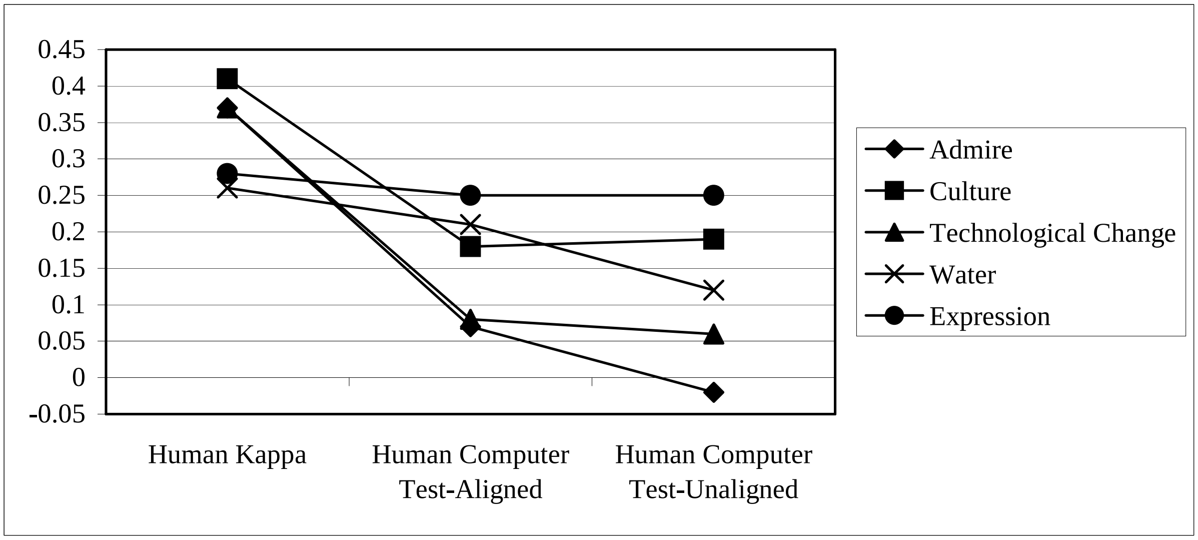

Figure 8. **Kappa by condition and prompt.**

Discussion, Conclusions, and Outlook

Our goal in this report has been to explore the feasibility of automating the scoring of speaking proficiency based on tasks that attempt to elicit communicative competence. As noted elsewhere (Bennett & Bejar, 1998; Williamson, Mislevy, & Bejar, 2006), the incorporation of technology into assessment, and specifically scoring, is a process that is more likely to succeed to the extent that it is guided by assessment principles and the recognition of the multiple and competing tradeoffs that need to be attended to when designing an assessment. Our goal was not to develop an operational scoring system but rather to understand the capabilities of speech technology in an assessment context, and specifically the tradeoffs that the application of that technology for assessment purposes entails. To that end we licensed an off-the-shelf speech-recognition engine and proceeded to adapt it to the tasks of recognizing speech from LanguEdge. The two major adaptation steps that we performed were (a) the introduction of a language model derived directly from our training corpus consisting of spoken responses to five prompts; and (**b**) acoustic adaptation of the Gaussian mixtures using the same training corpus. The latter gave, relatively speaking, a smaller improvement than the former. However, the overall recognition level, at 34%, is low and it is likely improvements are possible through more detailed acoustic modeling than we were able to do for this investigation. What we have learned from the application of the engine can be summarized as follows:

- The application of speech technology to the assessment of speaking proficiency is made especially difficult by the fact that the speaking proficiency potentially affects the recognition accuracy of a speech engine. Indeed the recognition level we were able to reach was approximately one in every three words. By contrast, state-of-the-art engines are able to achieve much higher recognition rates under ideal circumstances. With the present data, in addition to varying speaking proficiencies and accents, there was also variability in the quality of the recording. Some of the sound-quality variability (such as loudness) can be compensated for but there is the potential of other factors, such as speech from neighboring testing stations, to reduce recognition accuracy. Further, acoustic modeling taking into account the mix of language background that TOEFL test takers bring is a reasonable follow-up to this study. Data from such a study would be valuable to further understand the feasibility of automated scoring. In addition, the data from such a study might be valuable in the design of diagnostic or learning products.

- There is some relationship between recognition performance and speaking proficiency, as suggested by the factor analysis of human scores and recognition accuracy. Moreover, recognition accuracy, which is a function of the insertion, substitution, and deletion errors as well as correctly recognized words, reliably picks up individual differences. Nevertheless, recognition accuracy is distinct from proficiency as measured by human raters.
- Granted that recognition accuracy is reliably measured, it needs to be further studied in its own right. In particular, it is important to better understand what it is an indicator of. Of special interest is the possibility that recognition accuracy is, to some extent, determined by the sampling of language backgrounds and proficiency levels of speakers used to adapt the acoustic model. Because we had relatively little data, we used all of it for training or validation. As a result, language groups were not equally represented in the training of the acoustic model, and the potential exists that speakers from the most prevalent groups would be recognized more accurately, other things being equal. That is, if automated scoring is based on speech-recognizer hypotheses (i.e., what a recognizer believes a speaker has said), differential recognition accuracy as a function of language background can lead to differential construct representation and therefore unfairness. In others words, if the recognizer is better able to recognize the speech of students from a given language background, those students are potentially advantaged over students from a different language background. This advantage is potentially unfair if it results from the preponderance of speakers from that language in the adaptation of the language and acoustic models. If, however, the advantage is due to the linguistic proximity of the language in question to English (phonological and syntactic similarities, and differences between English and other languages), then it may not necessarily be an unfair advantage.
- As noted, our aim was not to develop an automated scoring engine at this point but rather to learn from a preliminary application what tradeoffs might be required in developing such a scoring system. For example, we fitted classification trees based on the scores of a single rating. This is clearly suboptimal because of the well known fallibility of human raters (Bejar, Williamson. & Mislevy, 2006). Using additional raters to train a classification tree would certainly improve classification accuracy. Additionally raters could be used in a

different capacity (such as to rate specific components of speaking performance) and then build classification trees for each component skill. Furthermore, classification trees are one of several modeling options. An increasingly attractive option are Bayes nets, which are attractive in their own right as an approach to modeling speech (Bilmes, 2003). In addition Bayes nets can be used for evidence accumulation (Williamson, Mislevy, Almond, & Levy, 2006). That is, the same formalism can be used at all levels of the evidence or scoring process. Another factor to keep in mind is that we grew the trees with aligned data, simulating the case of human-level recognition accuracy, which requires transcription of the speaking samples. We hope to investigate in the future the effect of using unaligned data on accuracy. However, since the current recognition rate is low, our plan is to first attempt to improve recognition rate.

- The classification trees that emerged from the analysis for each prompt were different in content and classification accuracy. In terms of content, the trees typically included several variable types, such as lexical richness, fluency and rate of speech, and content similarity. However, the actual features were chosen empirically as part of the tree-growing process. It should be possible in principle to constrain the trees, or some other evidence identification mechanism, to be held constant across prompts. This has been demonstrated in the assessment of architectural proficiency (Braun, Bejar, & Williamson, 2006) and is a desirable goal. However, to expect prompt-independent evidence models to function equally well across instances, they need to be constructed with that evidence model in mind.
- Human raters' judgments of speech quality may not be accurate. Therefore, other types of validation methods that use criteria other than human scores need to be explored.

In terms of accuracy, the trees for each prompt were not equally accurate. We bracketed the classification accuracy with human agreement on one end and chance agreement on the other end. We also computed the classification accuracy for the scoring trees under three conditions. One condition was classifying the tree based on internal cross-validation using aligned data; the second was applying the resulting tree to the aligned data from an independent corpus; and the third was finally applying the tree to the same unaligned data. We saw an interaction among prompts and scoring conditions. We ruled out *integrativeness* as the explanation for this interaction in that independent and integrated tasks were found among the most and least *scorable* with the classification trees. An issue to consider at this point is the prospect of automating speaking-proficiency scoring in light of the present results. We conclude that additional investment and research effort are called for. That is, we take the present results as encouraging in light of the many optimality conditions that were not present in the study. Moreover, the questions raised by our results, such as the different level of agreement between raters, if pursued, could lead to a better understanding of the speaking construct. Similarly, an accounting of the different degrees of amenability to automated scoring in some prompts motivates a closer examination of the construct.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bejar, I. (2002). Generative testing: From conception to implementation. In S. Irvine & P. C. Kyllonen (Eds.), *Generating items from cognitive tests: Theory and practice* (pp. 199–217). Mahwah, NJ: Lawrence Erlbaum.

Bejar, I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49-82). Mahwah, NJ: Lawrence Erlbaum Associates.

Bennett, R. E., & Bejar, I. (1998). Validity and automated scoring: It’s not only the scoring. *Educational Measurement: Issues and Practice*, *17*(4), 9–16.

Bernstein, J. (1997). Demonstratives and reinforcers in Germanic and Romance languages. *Lingua, 102*(2), 87–113

Bilmes, J. (2003). Graphical models and automatic speech recognition. In M. Johnson, S. P. Khudanpur, M. Ostendorf, & R. Rosenfeld (Eds.), *Mathematical foundations of speech and language processing: Vol. 138. Institute of Mathematical Analysis volumes in mathematics and its applications*. New York: Springer-Verlag.

Braun, H. I., Bejar, I., & Williamson, D. M. (2006). Rule-based methods and mental modeling. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 83–122). Mahwah, NJ: Lawrence Erlbaum Associates.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression tree*s. Belmont, CA: Wadsworth Int. Group.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper* (TOEFL Monograph No. TOEFL-MS-20). Princeton, NJ: ETS**.**

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47.

Chalhoub-Deville, M. (2001). Language testing and technology: past and future. *Language Learning & Technology, 5*(2), 95–98.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

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

de Jong, J., & Bernstein, J. (2001, June). Relating PhonePass overall scores to the Council of Europe framework level descriptors. In N. Noakes (Ed.), *Technology in Language Education: Meeting the Challenges of Research and Practice.* Retrieved August 25, 2006, from http://lc.ust.hk/~centre/conf2001/proceed/dejong.pdf

Deshmukh, N., Ganapathiraju, A., & Picone, J. (1999). Hierarchical search for large vocabulary conversational speech recognition. *IEEE Signal Processing Magazine, 16,* 84–107.

ETS. (2002a). *LanguEdge courseware: Score interpretation guide*. Princeton, NJ: Author.

ETS. (2002b). *LanguEdge courseware: Handbook for scoring speaking and writing*. Princeton, NJ: Author.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: selected readings* (pp. 269–293). Harmondsworth, Middlesex, UK: Penguin.

Ikeno, A., Pellom, B., Cer, D., Thornton, A., Brenier, J. M., Jurafsky, D., et al. (2003, April). *Issues in recognition of Spanish-accented spontaneous English*. Paper presented at the ISCA & IEEE ISCA & IEEE workshop on spontaneous speech processing and recognition, Tokyo, Japan.

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

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language models. *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing, 1,* 181–184.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar, (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15-48).. NJ: Erlbaum.

Ney, H., & Essen, U. (1993). Estimating “small” probabilities by leaving-one-out. In *Proceedings of Eurospeech ’93* (pp. 2239–2242). Grenoble, France: European Speech Communication Association.

Norris, J. M. (2001). Concerns with computerized adaptive oral proficiency assessment. *Language Learning & Technology, 5*(2), 99–105.

Quinlan, J. R. (1986). Induction and decision trees. *Machine Learning, 1*, 81–106.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 37*(2), 257–286.

Rabiner, L., & Juang, B. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice-Hall.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. New York: Cambridge University Press.

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

Searle, J. (1979). *Expression and meaning: Studies in the theory of speech acts*. New York: Cambridge University Press.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., et al. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech, 41*(3–4), 439–487

Strik, H., & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, *29*, 225–246

Tomokiyo, L. M. (2001). *Recognizing nonnative speech: Characterizing and adapting to nonnative usage in LVSCR*. Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

Wang, Z., Schultz, T., & Waibel, A. (2003). *Comparison of acoustic model adaptation techniques on nonnative speech.* Retrieved August 30, 2006, from http://www.cs.cmu.edu/~tanja/Papers/ICASSP03-wang.pdf

Williamson, D. M., Mislevy, R. J., Almond, R. A., & Levy, R. (2006). Bayesian networks. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring for complex tasks in computer-based testing* (pp. 201-258). Mahwah, NJ: Lawrence Erlbaum Associates.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing* (pp. 1-14).[Mahwah, NJ: Lawrence Erlbaum Associates.

Witt, S. M. (1999). *Use of speech recognition in computer-assisted language learning.* Unpublished doctoral thesis, University of Cambridge**,** UK.

Wu, J., & Chang, E. (2001). Cohorts based custom models for rapid speaker and dialect adaptation. In *Proceedings of EuroSpeech '01* (pp. 1261–1264). Grenoble, France: European Speech Communication Association.

Yokoyama, T., Shinozaki, T., Iwano, K., & Furui, S. (2003). *Unsupervised language model adaptation using word classes for spontaneous speech recognition.* Paper presented at the ISCA & IEEE ISCA & IEEE workshop on spontaneous speech processing and recognition, Tokyo, Japan.

Young, S. J. (2001). Statistical modeling in continuous speech recognition. In J. S. Breese & D. Koller (Eds.), *Proceedings of the 17th conference on uncertainty in artificial intelligence* (pp. 562–571), Seattle, WA: Morgan Kaufmann Publishers, Inc.

Notes

[1] See http://www.sil.org/lingualinks/languagelearning/otherresources/actflproficiencyguidelines/ACTFLGuidelinesSpeaking.htm

[2] In this report, we are concerned with so-called speaker independent speech recognition, where the goal is to recognize natural speech based on a large vocabulary without speaker-specific training. By contrast, speaker-dependent speech recognition is oriented to recognizing with a high level of accuracy speech based on a small vocabulary, where the system has been trained to the idiosyncrasies of a specific speaker.

[3] Spectral features are a vector representation of the most essential acoustic characteristics of a speech frame.

[4] Perplexity refers to the level of predictability of the next word, given the previous context. Higher perplexity reflects low predictability.

[5] We use maximum a posteriori adaptation (MAP adaptation) here, as opposed to maximum likelihood linear regression (MLLR) adaptation.

[6] Agreement statistics are known to depend on the distribution of the marginals, which are not necessarily the same under tree and human scoring. Therefore, a comparison of the tree-rater agreement against chance agreement between raters should not be over interpreted.

[7] TRL stands for transliteration.

[8] The speaker identification will either be provided by ETS or designated by MacLaren.

Appendix A

Transcription Conventions (By Victoria MacClaren) Transcription Proposal for ETS

The conventions used in transcription will be a subset of the conventions developed for the Verbmobil project. These conventions make allowances for:

- Human noises
- Hesitations (filled pauses)
- Pausing
- Breathing
- Repetition
- Broken words
- Punctuation – commas, periods, question marks
- Unintelligible words

Words will be spelled in accordance with the *Merriam-Webster Dictionary*.

Transcription Conventions

Punctuation

Three items of punctuation are provided in the transcription: question marks (?), commas (,), and periods (.).

Some general rules apply:

1. A white space is always set *before and after* the punctuation.
2. After a period or question mark, transcription is ***continued in lower case*** unless the word category requires capitalization.
3. The setting of punctuation basically follows the punctuation rules of the specific language.

Capitalization

Proper nouns, the pronoun "I," and abbreviations are capitalized. Letters that are spoken are also capitalized.

Proper names. Proper names are hyphenated to combine them into one entity.

Example:

New-York

Forbes-Avenue

Numbers. Numbers are written out in lexicon form, regardless of usage (i.e., date, adjective, telephone number). Numbers are not hyphenated.

Broken words. Broken words are marked with the equal sign **[=]**. When a speaker begins to say a word, but does not finish enunciating the word, the word ends with **[=]**.

Repetition or Correction

The **[+/. . ./+]** convention is used to symbolize when a speaker either repeats or corrects himself or herself. A repetition may not always be an exact repetition; it may also be a substitution.

There is no white space between the convention markers, only between the words or elements within the mark, and there is no punctuation after the ending bracket.

False Start

A false start is marked as **[-/. . . /-]**. A false start occurs when an individual begins to say something but then changes to another idea, topic, or thought.

There is no white space between the convention markers; only between the words or elements within the mark, and there is no punctuation after the ending bracket.

Neologisms

A neologism is the term used to describe a word that has been made-up or invented by a speaker. It can also be described as a word that does not appear in the dictionary of the primary spoken language, but is also not a foreign word.

A neologism is commonly a slang word or it may be the creation of a nonnative speaker who has made a grammatical error. It is marked with the asterisk **[*]**.

Header

A header is used to identify the transcription. The header will list the date the transcription was completed, the transcription ID, the accompanying WAV file ID, as well as any comments made by the transcriber. A semicolon precedes each line of the header.

Example:

;Date: 01/04/2003

;WAV ID: 750175indSpeak2cultures.wav

;TRL[7] ID: e003a

;Comments:

Turn Identification

In order to ease reference, a turn identification is used. The turn identification appears before each turn. It identifies the speaker, the transcription, and the turn's number.

A turn identification appears as such:

- e123a_000_AAP:
    - **e**: The first letter will represent the language, in this case, English.
    - **123**: The number given to each speaker's directory.
    - **a**: A code for the specific transcription within the directory. There are as many as five files per directory, so codes a, b, c, d, and e will be used.
    - **000**: The turn number. Turn numbers begin with 0. The next would be 001, and so forth.
    - **AAP**: The speaker identification.[8]

The identification ends with a colon [:] and is then followed by one white space before the transcription of the turn begins. Each piece of the identification is separated by an underscore.

Turn Marking

Turns will be marked within the transcription. Longer sentence fragments are marked as INCOMPL.

Example:

e123a_000_AAP: 32058 262122

the person I admire most is my mother .

Pause

A pause is marked when a speaker stops speaking and no other vocalizations are made. Pausing is marked only when the halting in speech is substantial.

Pauses are marked with the symbol **[**<P>**]**.

Breathing

Breathing is marked with the symbol **[**<B>**]**.

Human Noises

Four human noises are marked: lip smacking **[**<Smack>**]**, clearing of the throat **[**<Throat>**]**, coughing **[**<Cough>**]**, and laughing **[**<Laugh>**]**. An additional category, noice **[**<Noise>**]**, is used when a human makes a noise that does not fit into the other four categories.

Filled Pauses

Three filled pauses, or hesitations, are marked:

1. <uh> : purely vowel hesitation
2. <hm> : purely nasal hesitation
3. <uhm> : vowel and nasal hesitation

Additionally, **[**<hes>**]** is used to mark a hesitation that does not fit into these other three categories.

Unidentifiable

When an articulation is difficult to understand, the unidentifiable mark **[**<%>**]** is used. This mark symbolizes any length of speech that is difficult to understand for the transcriber. This mark, minus the carats, may be attached to a word to symbolize that the word is difficult to understand.

Example:

> I am going to go downtown to go% shopping%.

In this example, the transcriber found *go shopping* difficult to understand. Reasons for this difficulty may include scraping of the microphone by a speaker or coughing.

Technical Disruptions

When an utterance is broken due to technical reasons, particularly a late recording start, the mark <T_> is used. The underscore represents the part of the utterance that is missing.

Example:

> <T_>ternet is one of the most influential inventions of the twentieth century.

Incomplete Turns

A turn that the speaker leaves incomplete is marked at the end of the turn with INCOMPL. A turn that the transcriber is unsure is complete or not is marked at the end with <*T>t. The transcriber may be unsure of the completion of a turn due to an unidentifiable utterance at the end of the sentence.

Additionally, a turn that is left incomplete because the recording finished before the speaker completed the turn is marked with <*T>t.

Faults With Recording

In the case where the audio is disturbed, the instance will be marked with the dollar sign **[$]**. Additionally, the begin and end time stamps of the occurrence will be noted in a comment following the turn. These time stamps will be marked with <TECH>, for technical disturbance.

Example:

The experiment was +/$to/+ to see about the facial expression.

;<TECH> 123 456

Comments Regarding the Speaker

The gender and proficiency level of the speaker will be marked in the Comments section of the header as such:

<GEND>f : female speaker

<GEND>m : nale speaker

<PROF>X : proficiency of the speaker on a scale of 0**]**5.

- A mark of 0 proficiency represents the inability of the transcriber to assess the speaker's proficiency due to severe technical disruptions or the absence of articulation in the audio file.
- A mark of 1 represents very poor proficiency. A mark of 5 equals high proficiency.
- All scores are subjective to the transcriber.

Comments Regarding the Audio File

The audio quality of each file will be assessed on a scale of 0**]**5. This comment will be made in the Comments section of the header as such:

<QUAL>X, where X represents one of the scores described below.

- A mark of 0 quality represents the inability of the transcriber to assess the recording quality because no articulation is heard on the file.
- A mark of 1 quality indicates that long stretches of the audio are unintelligible due to a poor signal quality.
- A mark of 2 represents much noise and tape repeats, but the contents of the file are for the most part manageable.
- A score of 3 means the quality is fair, but there is some noise and occasional or brief tape repeats.
- A score of 4 refers to a good audio quality, but with some minor noise.
- A score of 5 is of very good quality, with no audible noise.

Table A1

Transcription Convention Summary

Convention	Meaning
.	Period
,	Comma
?	Question
=	Broken word
+/…/+	Repetition/correction
-/…/-	False start
*	Neologism
<Laugh>	Laughing
<Cough>	Coughing
<Throat>	Clearing one’s throat
<Smack>	Lip smack
<Noise>	Other human noises
<P>	Pausing
<B>	Breathing
<uh>	Vowel hesitation

(Table continues)

Table A1 (continued)

Convention	Meaning
<hm>	Nasal hesitation
<uhm>	Vowel + nasal hesitation
<hes>	Other hesitation
<%>	Unidentifiable
%	Difficult to understand
<*T>t	Incomplete turn (transcriber unsure)
$	Affected by technical disturbance
INCOMPL	Incomplete turn
<_T>, <T_>	Incomplete utterance
<TECH>	Technical disturbance
<GEND>f	Speaker is female
<GEND>m	Speaker is male
<QUAL>X	Audio quality
<PROF>X	Proficiency of speaker

Appendix B

Set of Phonemes (Used by the Carnegie Mellon University Dictionary *and* Almost Identically by the Multimodal Recognizer)

Table B1

List of Phonemes in cmudict.0.1

Phoneme	Example	Translation
AA	Odd	AA D
AE	At	AE T
AH	Hut	HH AH T
AO	Ought	AO T
AW	Cow	K AW
AY	Hide	HH AY D
B	Be	B IY
CH	Cheese	CH IY Z
D	Dee	D IY
DH	Thee	DH IY
EH	Ed	EH D
ER	Hurt	HH ER T
EY	Ate	EY T
F	Fee	F IY
G	Green	G R IY N
HH	He	HH IY
IH	It	IH T
IY	Eat	IY T
JH	Gee	JH IY
K	Key	K IY
L	Lee	L IY

(Table continues)

Table B1 (continued)

Phoneme	Example	Translation
M	Me	M IY
N	Knee	N IY
NG	Ping	P IH NG
OW	Oat	OW T
OY	Toy	T OY
P	Pee	P IY
R	Read	R IY D
S	Sea	S IY
SH	She	SH IY
T	Tea	T IY
TH	Theta	TH EY T AH
UH	Hood	HH UH D
UW	Two	T UW
V	Vee	V IY
W	We	W IY
Y	Yield	Y IY L D
Z	Zee	Z IY
ZH	Seizure	S IY ZH ER

Table B2

Additional Phonemes Used by Multimodal

Phoneme	Meaning
/hu/[a]	human noise (eg sneeze, cough, lipsmack)
/nh/ [a]	nonhuman noise (eg microphone click, door)
ax	Weak A, as in: butter: B AH T AX
en	Weak n-syllable, as in: mitten: M IH T EN

[a] Noise phoneme.

Appendix C

Sample Transcripts, Hypotheses, and Word Errors

Example Utterances and Words From Different Score Levels

We randomly selected five subjects whose ratings are from 1 to 5, respectively, and where the two ETS raters agreed on the score. The selection is as follows:

Table C1

Subjects' Test Items Selected by Scores, Their Native Language (If Known), and Their Gender

Score	Subject ID	Native language	Gender
1	170011	German	Female
2	930021	Unknown	Male
3	240079	Chinese	Female
4	930087	Unknown	Male
5	920087	Unknown	Female

There are only two items that occur for each speaker in our speech database: Item 2 (culture) and Item 3 (technical change). We choose to pick the latter, so in the ongoing discussion it can be assumed that it is a recording of this item of the respective speaker. We first present some utterances of each speaker, including the entire original disfluency markup.

Table C2

Sample Utterances for Different Speakers From the Human Transcriptions

Speaker	Turn ID	Start time (frame number)	End time (frame number)	Text
170011:	E069_1_0000	0	80176	+/take/+ <hm> <Smack> <B> take place <hes> everyday now life and I think they are very very important for us . <B>

(Table continues)

Table C2 (continued)

Speaker	Turn ID	Start time (frame number)	End time (frame number)	Text
	e069_1_0001	80600	286335	<uhm> they change our life everyday . and <uhm> <P> <Smack> <B> I like to speak about the computerizing . <B> <uh> computerizing started for <B> <P> thirty years ago and it's still going on . and so I think it's very important process for all of us . <B>
930021:	E095_1_0000	0	142761	<uh> +/in/+ <P> in the last years . for example, <uh> +/there aren't/+ there are not telephone , mobile , washing machine , microwave , Internet.
	e095_1_0001a	150557	219954	+/the/+ <P> these things are not used in <P> one hundred years ago . <B>
	e095_1_0002	225015	253730	and today , we use these thing .
240079	E138_1_0000	0	90600	and <uhm> a major <uh> technology changes in last <uh> one hundred years . <uh><B>
	e138_1_0001	94832	202001	<uhm> people watch TV to get the <uh> news <uh> and <hes> for the entertainment <uh> <%> <B> <*T>t
	e138_1_0002	203060	410842	+/they/+ they get relaxed <uhm> <hm> <Smack> by watching +/al=/+ all +/k=/+ kinds of <uh> entertainment <uh> <B> programs <B> and they get *informations <hes> by <B> listening% to the news channel .

(Table continues)

Table C2 (continued)

Speaker	Turn ID	Start time (frame number)	End time (frame number)	Text
930087	E080_1_0000	0	169477	I think one very important <B> technological development was <uh> in the area of <uhm> communications . it generally% has <uhm> a very big impact <B> on people's lives in the last one hundred years <%> <*T>t
	e080_1_0001	174367	291900	in the way the people communicate <uh> <%> way with the mobile phone , for example . <%> <B> <%> on the television . <B> <uh>
920087	E035_1_0000	0	300224	<T_>gy changed everybody's lives <P> to the negative , unfortunately . it was the atom bomb that was invented <B> <hes> during World-War-Two . and that totally changed everybody's outlook <P> on life and on society and living together because it just made it possible <P> for so many people to be killed at once without any warning or anything .
	e035_1_0001	309455	495428	and this technological change had a great impact +/on art/+ on the arts , in general , <B> on literature , on visual-arts , on music . and it's just that the% invention had a huge impact <B> on people's lives .

(Table continues)

Table C2 (continued)

Speaker	Turn ID	Start time (frame number)	End time (frame number)	Text
	e080_1_0001	174367	291900	in the way the people communicate <uh> <%> way with the mobile phone , for example . <%> <B> <%> on the television . <B> <uh>
920087	E035_1_0000	0	300224	<T_>gy changed everybody's lives <P> to the negative , unfortunately . it was the atom bomb that was invented <B> <hes> during World-War-Two . and that totally changed everybody's outlook <P> on life and on society and living together because it just made it possible <P> for so many people to be killed at once without any warning or anything .
	e035_1_0001	309455	495428	and this technological change had a great impact +/on art/+ on the arts , in general , <B> on literature , on visual-arts , on music . and it's just that the% invention had a huge impact <B> on people's lives .

We can make the following observations:

While Speaker 170011 talks mostly in short *sentences* that are often ungrammatical, Speaker 920087’s syntactic constructions are much better and so can convey the intended content to the listener quite well.

Also, in Speaker 930087’s transcription, several *nonintelligible* markers (<%>) are found; this is mostly due to the poor audio quality of this recording.

We now look at the performance of the speech recognizer for each recording. (The balanced word accuracy is the mean of the accuracy with respect to the reference and with respect to the hypothesis; as discussed in the main section of the report.)

Table C3

Balanced Word Accuracy for Five Speakers Ranging in Scores From 1 to 5

Speaker	Balanced word accuracy	Score
1 – 170011	0.341	1
2 – 930021	0.316	2
3 – 240079	0.374	3
4 – 930087	0.138	4
5 – 920087	0.517	5

These results indicate that, in some cases, the speech recognizer's accuracy is increasing with increasing (better) scores. Here, the poor audio quality of the 930087 recording causes the very low recognition rate and therefore does not fit into this picture.

In order to obtain word error information, we performed an alignment between the reference (human transcript) and the ASR hypothesis for these five recordings, and will present two short examples here. They should be read as follows: Line REF is the reference; line HYP is the hypothesis; line Eval contains the error type (S = substitution, I = insertion, D = deletion). Capitalized words spot recognition errors. Noncapitalized words were correctly recognized. We use the NIST's (National Institute for Standards and Technology) *sclite* scoring program, which was written for and used by most if not all speech-recognition evaluations in the past.

For determining the string differences, we use sclite's standard algorithm, the Levenshtein distance, which assigns weights or cost functions to every possible case (S, D, I, and C = correct).

Example 1 (170011, Score = 1):

File: 170011indspeak3techchange_0

Channel: 1

Scores: (#C #S #D #I) 5 14 2 2

REF: TAKE %HM %HUMAN %breath take ***** PLACE   %MUMBLE EVERYDAY NOW

```
HYP:  **** HATE HAND   %breath take PLAYS %BREATH AND     THEY     KNOW
Eval: D    S    S                  I     S       S       S        S
>> REF:  LIFE and I   think **** they ARE VERY VERY IMPORTANT FOR US
%BREATH
>> HYP:  NICE and AND think THAT they *** READ CAN %BREATH   OF IS
%MUMBLE
>> Eval: S        S         I         D   S    S    S         S   S  S
```

For this speaker, 170011, not too much is recognized correctly in the first utterance. There are also a couple of interesting misrecognitions between words that sound alike, such as *take-hate*, *place-plays*, or *they-day*, Also note that in many cases, as we just demonstrated here, the misrecognized word belongs to a different word category (part-of-speech) than the original word in the reference.

Example 2 (920087, score = 5)

```
Speaker sentences 2: 920087indspeak3techchange   #utts: 3
id: (920087indspeak3techchange-000)
File: 920087indspeak3techchange_0
Channel: 1
Scores: (#C #S #D #I) 38 16 0 24
REF:  %MUMBLE changed ******* everybody's lives ******* ******* to the
HYP:  AND     changed %BREATH everybody's lives %MUMBLE %MUMBLE to the
Eval: S               I                         I       I
>> REF:  negative unfortunately IT      was THE     ATOM bomb that ** **
>> HYP:  negative unfortunately %BREATH was BECAUSE OF   bomb that IF AT
>> Eval:                        S           S       S              I  I

>> REF:  **** ** WAS INVENTED %BREATH %mumble ** *** during world-war-two
>> HYP:  ONCE IT AND THEN     THE     %mumble IT THE during world-war-two
>> Eval: I    I  S   S        S               I  I
```

For Speaker 920087, much more material is correctly recognized and among the correctly recognized words there are more content words, compared to 170011. Furthermore, as part 1 of this utterance shows, some errors are related to omissions or insertions of disfluencies, such as pauses or breathing. Of course, one could also measure the word accuracy when these special noncontent words are ignored. In an experiment, however, we found no significant difference in the scores between our current way of evaluation and an evaluation without special words. (Gains in some examples are offset by losses in other examples).

Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL
programs and services, use one of the following:

Phone: 1-877-863-3546
(US, US Territories*, and Canada)

1-609-771-7100
(all other locations)

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

*America Samoa, Guam, Puerto Rico, and US Virgin Islands