NCME 2006
Last updated 03/06/06

Extracting meaningful speech features to support diagnostic feedback: an ECD approach to automated scoring

Xiaoming Xi
Klaus Zechner
and Isaac Bejar

Educational Testing Service (ETS)

Paper presented at the annual meeting of National Council on Measurement in Education (NCME) held in San Francisco, California, on April 6-10, 2006

Abstract

Although the operational scoring of the TOEFL iBT speaking section features the overall judgment of an examinee's speaking ability, the evaluation of specific components of speech such as delivery (pace and clarity of speech) and language use (vocabulary and grammar use) may be a promising approach to providing diagnostic information to learners.

This study used an evidence-centered design approach (ECD) to extract features that provide evidence about the quality of responses to TOEFL iBT speaking tasks through the use of speech and NLP technologies. The computed features were identified through a detailed explication of the rubrics and confirmed by content specialists. Classification trees were developed to model human holistic scores and delivery and language use scores, and validated on independent samples.

We will discuss the feasibility of extracting meaningful speech features amenable to diagnostic feedback.

Introduction

The speaking section of the Test of English as a Foreign Language Internet-Based Test (TOEFL® iBT) is designed to measure the academic English speaking proficiency of non-native speakers who plan to study at universities where English is spoken. This test represents an important advancement in the large-scale assessment of productive skills. However, it poses particular challenges to learners in parts of the world where opportunities to learn speaking skills are limited. First, the learning environments in those countries may not be optimal for acquiring speaking skills. Second, English teachers in those parts of the world may not be adequately trained to teach speaking skills and to provide reliable feedback on their students' speaking performance. This calls for research that would support the design of effective learning products that can serve the diverse needs of learners, promote better learning, and improve teaching practices. An important requirement for such learning products is that they be capable of providing instant feedback to learners. This project explores automated evaluation of TOEFL® iBT speaking performances that supports instant feedback capabilities for potential speaking products.

Previous work

The technologies that support automated evaluation of speaking proficiency are automated speech recognition (ASR) and natural language processing (NLP) tools (for a survey of these technologies see, e.g., Jurafsky & Martin, 2000). The application of these technologies to responses to the TOEFL® iBT Speaking test poses challenges because this test elicits spontaneous speech and the scoring rubrics, based on which the responses are evaluated, draw on models of communicative competence. In addition, TOEFL® iBT speaking has been developed to support the learning and teaching of academic speaking skills. Therefore, a useful automated evaluation system should provide feedback on test takers' performances. To meet those challenges, accurate recognition of spontaneous accented speech and use of features that indicate various aspects of speaking proficiency in the rubrics for score prediction are necessary.

The application of ASR to speaking proficiency is a fairly recent development. One successful area of research has focused on automatically scoring the pronunciation of non-native speakers. Franco et al. (2000) present a system for automatic evaluation of pronunciation of native and non-native speakers of English and other languages at the phone level and sentence level (EduSpeak). Candidates read English texts and a forced alignment between the speech signal and the ideal path through the Hidden Markov Model (HMM) is computed. From this, the log posterior probabilities for pronouncing a certain phone at a certain position in the signal are computed to yield a local pronunciation score. This score is then

combined with other automatically derived measures such as the rate of speech or the duration of phonemes for pronunciation evaluation.

A company recently acquired by Harcourt, Ordinate has developed an approach based on highly-constrained speech, elicited through test tasks such as reading sentences aloud and answering questions that require short responses containing a few words (Bernstein, 1999). Scores in sentence mastery, fluency, pronunciation, and vocabulary based on these tasks are provided by means of ASR. Given the learning and teaching orientation of the revised TOEFL® test, a potential drawback of Ordinate's approach is that the tasks used in their assessments do not call for spontaneous speech production and under-represent the domain of speaking proficiency. For purposes of predicting speaking proficiency such an approach may be adequate. In fact, evidence suggests that correlations between the automated scores on Ordinate's speaking tests and human scores on other performance-based speaking tests are high (e.g., Bernstein, 1999). However, our goal is not simply to provide a score but also to provide feedback on students' performance on tasks that elicit spontaneous speech so that students could advance their speaking proficiency. This requires that their speaking performances be characterized in more detail. Fortunately, there is a growing body of research that informs our work. For example, Cucchiarini et al., 1997a, 1997b) have focused on the fluency features of free speech that can be extracted automatically from the output of a typical ASR engine. However, fluency features alone are far from adequate in representing a full model of speaking proficiency suggested by current models of communicative competence (Bachman, 1990; Bachman & Palmer, 1996).

Our present study explores the automated evaluation of TOEFL® iBT Speaking responses through automated means. It intends to characterize more aspects of speaking proficiency with speech features motivated by well-defined scoring rubrics. The approach we take is evidence-center design (ECD) (Mislevy, Steinberg, Almond, & Lukas, in press; Williamson, Mislevy & Bejar, in press). Specifically, our goal is to identify *evidence* in students' speaking performance that serve to characterize their communicative competence (Bachman, 1990; Bachman & Palmer, 1996). This approach not only leads to defensible automated scores but is also consistent with the goal of developing learning products. In other words, the same set of features used to characterize performance on the TOEFL® iBT Speaking can potentially serve as the basis to provide diagnostic feedback and guide the student towards improved performance. The synergy between automated scoring and learning has been discussed previously by Bejar and Braun (1994).

Zechner, Bejar, and Hemat (in preparation) made the first attempt to use speech technologies to extract speech features that provide some evidence about the overall quality of responses to TOEFL® iBT prototype speaking tasks. Building on their work, this project uses an analytic approach, whereby the

relative importance of speech features that characterize different aspects of speaking proficiency is determined based upon their relationships to the human-assigned analytic scores.

Analytic rubrics of Delivery, Language Use, and Topic Development (see descriptions of the rubrics in the Method section) were developed and tested for TOEFL® iBT Speaking (Xi & Mollaun, in press). In this project, we focused on the automated evaluation of Delivery and Language Use as well as the overall quality of TOEFL® iBT Speaking responses. We started by analyzing the Delivery and Language Use rubrics with the goal of formulating features that could be realized computationally by means of speech and NLP technologies. The identification and computation of the sub-features of these dimensions were informed by second-language learning and assessment literature, and extensive feedback from test developers and expert raters.

ECD-Based Automated Scoring of Speaking

Figure 1 is a schema decomposing ECD-based automated scoring. It assumes that the assessment has followed ECD design principles. In particular, the goal of the assessment has been defined with sufficient precision such that we can pinpoint in a task response what constitutes *evidence* relevant to spoken proficiency. Moreover, we assume that the work product is the examinee's response to tasks that have been designed with the goal of eliciting evidence about spoken proficiency.

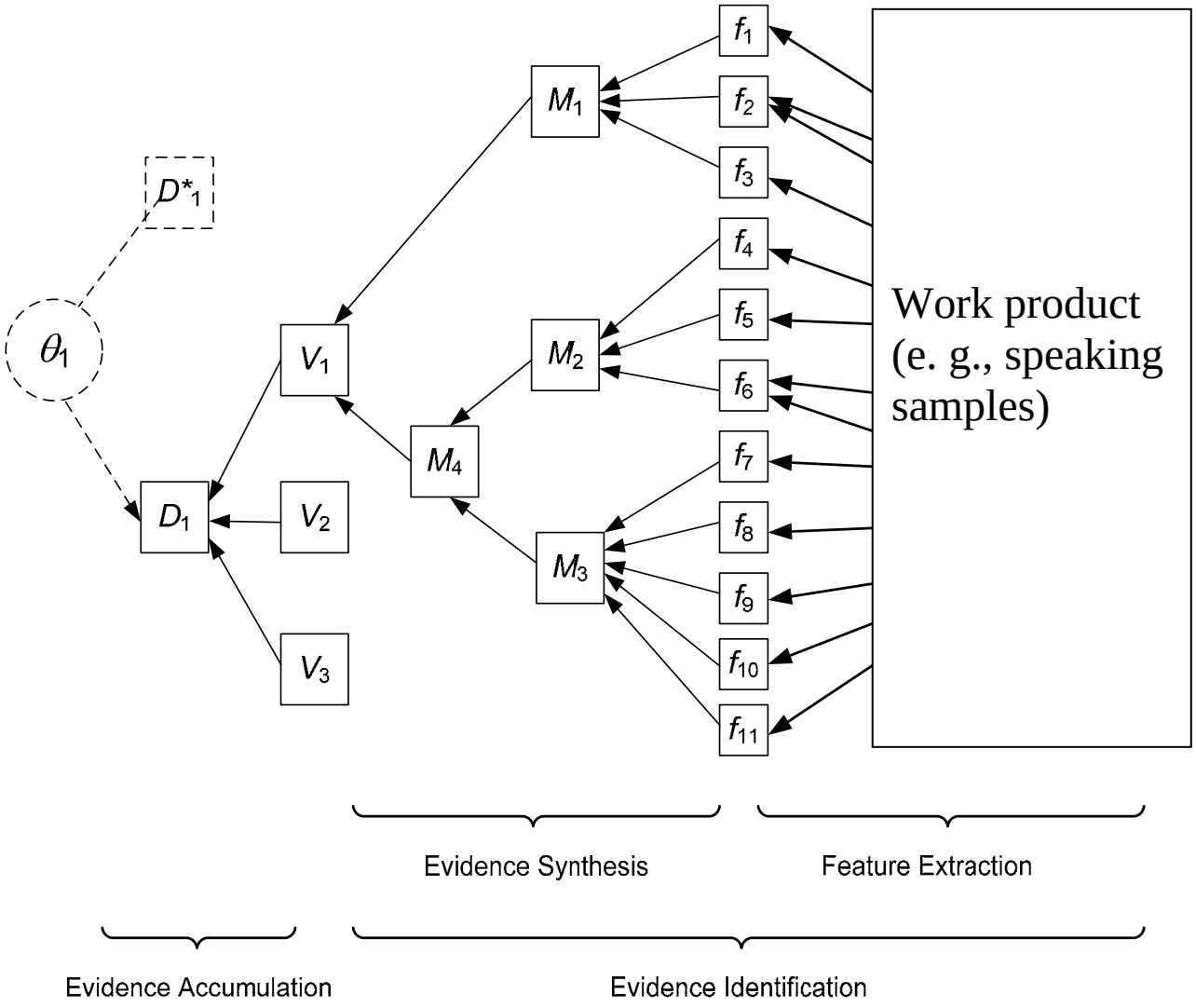

Figure 1. **A schema decomposing ECD-based automated scoring (after Braun, Bejar and Williamson, in press)**

According to Figure 1, scoring consists of two major processes: *evidence identification* and *evidence accumulation*. Evidence identification can be decomposed into two subcomponents: *feature extraction* and *evidence synthesis*. Evidence accumulation refers to the process of aggregating scores on multiple work products into an estimate of "ability." Evidence accumulation is indistinguishable regardless of whether automated scoring is used because the input to the process is comprised of scores from a set of work products, regardless of how they were obtained. Evidence identification, however, is fundamentally different in automated scoring and human scoring. Human scoring, as a means of

evidence identification, leaves unspecified both the variables (or features) used by the rater as well as the process for assigning a score to a work product (*evidence synthesis*) because it depends on the raters' interpretation of the scoring guidelines and the mental processes used to assign a score. Construct representation in the case of human scoring, is maintained through the careful training of the raters, close monitoring of each rater, and the process as a whole. By contrast, evidence identification in the case of automated scoring specifies, in detail, the features and how they are to be aggregated; and construct representation is maintained through the set of features and their aggregation into a score.

As suggested by Figure 1, the first step in automated scoring is to decompose the spoken sample into a set of *features*. Features are variables that represent evidence about speaking proficiency, or enable the computation of evidence, about proficiency. It is through a careful selection of features that we can enhance construct representation. As implied in Figure 1, features can be used to define higher-level features which ultimately define the score by one of several methods for *evidence synthesis.*

An important distinction among methods of evidence synthesis is whether they are *supervised* or *unsupervised*. Supervised methods fit a specific statistical model such as linear regression, classification and regression trees, support vector machines, or neural nets, among others. The data to estimate the model consists of the features and the *criterion score* associated with those features. Typically, the criterion score is a score that has been assigned by human scorers although, in general, they can be any score we believe adequately represents the construct. Because supervised methods are fitted to a sample of data consisting of features and criterion scores, the potential for over-fitting exists and the scoring model typically needs to be tested on data that has not be part of fitting the model.

Unsupervised methods of evidence synthesis require just the features. The outcome of this process is a set of rules for mapping a vector of features to a score. The process can take many forms and rely on analyses, such as clustering or expert judgments, to arrive at the rules that map features to scores. The best known example illustrating this approach is described in Braun, Bejar and Williamson (in press).

In this study, we used a supervised method of evidence synthesis, classification trees (Breiman, Friedman, Olshen, & Stone, 1984). This machine-learning method has the advantage of high interpretability compared to other supervised approaches. It also accommodates the characteristics of the speaking construct as is discussed below and allows the possibility of an "intervened" approach where human experts can modify the scoring rules derived based on data. The next section provides a brief overview of the technique and the rational for using this method for this machine learning task.

A brief review of classification trees

Classification trees

Classification trees are hierarchical, sequential classification structures that recursively partition the observations into different classes. At each decision node, the variable that can best classify the cases into distinct classes at a certain value is selected to perform the partition. For example, if the speech rate is less than three words per second, a response goes to the left node (representing the lower score class) and to the right node (representing the higher score class) if otherwise. Then each of the child nodes may be further partitioned into two more nodes down the tree and the process is repeated until a *terminal node* is reached. In a nutshell, classification trees yield a set of *if-then* logical (split) conditions that permit the accurate prediction or classification of cases.

An advantage of classification trees is that they do not assume that the underlying relationships between the predictor variables and the predicted classes are linear, follow some specific non-linear link function, or that they are monotonic in nature. For example, holistic speaking scores could be positively related to a speech rate if the rate is less than five words per second, but negatively related if the rate is greater than that. That is to say, the tree could have multiple splits based on the same variable, revealing a non-monotonic relationship between the variable and the predicted classes. Moreover, a variable that does not discriminate in the higher-score classes can be used in classifying lower-score classes without impacting the prediction of the higher classes. These characteristics of classification trees contrast with other classification techniques, which use *all* of the important variables for classifying each case. Since the distinguishing speech features for different score classes may be different and the relationship between a speech feature and speaking scores may not be linear, the classification tree becomes a suitable technique for classifying score classes of TOEFL® iBT Speaking. In addition, different patterns of strengths and weaknesses in different aspects of speech may lead to the same score class. This is also compatible with a feature of classification trees that different sets of decision rules may result in the same score class.

Although complex mathematical computations are used in growing the trees, the actual application of the tree is performed using a sequence of simple and easily understood decision rules that are transparent to content specialists. The classification mechanism is intuitively appealing. Thus this classification technique is amenable to incorporating content specialists' input in evaluating and refining the trees that best represent expert raters' decision processes.

Method

The speech recognizer

The speech recognizer can be described as a computer program that reads digitized speech data on one side and outputs the most likely string of words, given the speech signal, on the other side. Both the acoustic properties of speech ("how the phonemes sound") and the statistical properties of spoken language ("which word sequences are likely to occur by themselves?") need to be carefully modeled in order for the program to succeed. The former model is called the *acoustic model* (AM), and the latter the *language model* (LM). The speech recognition program performs a search based on local optimization criteria while tightly coupling AM and LM information as it processes the input data.

The success of speech recognition is typically indicated by a word error rate or by word accuracy. The word accuracy for the corpus of TAST data used in our study, described below, was 72% for the training sample and 50% for the test sample. Note that the training sample was used to train the speech recognizer; therefore the high accuracy was not surprising.

Data

This study used responses to six speaking tasks in a TOEFL® Academic Speaking Test (TAST) from 180 students of various levels and native languages (TAST 2003/180). The first two tasks were *independent tasks* that required the examinees to speak about familiar topics. The four remaining tasks were *integrated tasks* that required the examinees to read and/or listen and then speak.

Of the 180 examinees, 100 had tasks that were completely double-scored and 80 had tasks that were single-scored *holistically*. Following this, a stratified sample of 140 examinees (TAST 2003/140) with various proficiency levels and native language (L1s) were double-scored *analytically* on Delivery and Language Use.

Each response to a speaking task was treated as a unique case in building the scoring models. In other words, generic scoring models were developed for all tasks. This may not be optimal but our small sample did not allow us to explore scoring models specific to a task or a task type. Scores of zero were excluded from the analyses. The number of responses for the training and testing samples for the three scoring models are shown in Table 1. For the Holistic scoring model, approximately 70% of the data were used in training and 30% in testing. No speaker's responses occurred in both the training and the testing samples. For the Delivery and Language Use scoring models, the data sets were not partitioned

into two sets due to the smaller sample sizes. Instead, ten-fold cross-validation was performed on the whole sample.

Table 1: Training and testing sample sizes for the three scoring models

	Training	Testing	Total
Holistic model	742	314	1056
Delivery model	--	--	818
Language use model	--	--	822

Features used in building scoring models

Delivery and Language Use features

Table 2 presents the list of Delivery features, which was expanded from our previous work on automated scoring of prototype TOEFL® iBT speaking responses (Zechner, Bejar & Hemat, in preparation).

Table 2: Delivery Features

Feature Number	Feature label	Class	Feature description
1	Numdff	Fluency	# of disfluencies [eg uh um...]
2	Numsil	Fluency	# of silence events [not including inter-utterance silence events]
3	**Silpwd**	Fluency	Silences per word
4	**silmean**	Fluency	Mean of silences
5	**silstddv**	Fluency	Standard dev. of silence duration
6	**Longpmn**	Fluency	Mean of long pauses
7	**longpstddev**	Fluency	Standard deviation of long pauses
8	**Secpchk**	Fluency	Chunk length in seconds [chunk delimited by 0.5sec or longer pauses]
9	**Wdpchk**	Fluency	Chunk length in words
10	**Wpsec**	Fluency	Speaking rate in words per second
11	Dpsec	Fluency	Disfluencies per second
12	**Silpsec**	Fluency	Pauses/Silences per second

Table 3 contains the Language Use features, which include both vocabulary diversity and sophistication features and grammar accuracy features. The stop words contain the common function words, so the non-stop words can be roughly treated as content words. When computing features 13-16 and 18-37, the stop words were excluded. *Baselist 1* and *Baselist 2* contain the most frequent and the second most frequent 1000 English words and *Baselist3* includes words not in the first 2000 words of English but which are frequent in upper-secondary school and university texts from a wide range of

subjects. Feature numbers 39-44 in Table 3 were computed based on Breland's standardized frequency index (Breland, Jones, & Jenkins, 1994). The higher the number, the less sophisticated vocabulary is used.

Table 3: Language Use features

Feature Number	ID	Class	Description
13	Numbase1	Vocab	# of words in Baselist1
14	Numbase2	Vocab	# of words in Baselist2
15	Numbase3	Vocab	Words in Baselist3
16	Nonbase	Vocab	Words not in any baselist
17	Relstop	Vocab	Relative number of stop words in % of Numtok
18	**Relbase1**	Vocab	Relative number of words in Baselist1 in % of Nonstop
19	**Relbase2**	Vocab	Relative number of words in Baselist2 in % of Nonstop
20	**Relbase3**	Vocab	Relative number of words in Baselist3 in % of Nonstop
21	**Relnonbase**	Vocab	Relative number of words not in any baselist in % of Nonstop
22	TypesBase1	Vocab	Number of word types in Baselist1
23	TypesBase2	Vocab	Number of word types in Baselist2
24	TypesBase3	Vocab	Number of word types in Baseline3
25	TypesNonbase	Vocab	Number of word types within words not in any baselist
26	TTRBase1	Vocab	Type-token-ratio within Baselist1
27	TTRBase2	Vocab	Type-token-ratio within Baselist2
28	TTRBase3	Vocab	Type-token-ratio within Baselist3
29	TTRNonbase	Vocab	Type-token-ratio within words not in any baselist
30	FamBase1	Vocab	Number of families covering all words within Baselist1
31	FamBase2	Vocab	Number of families covering all words within Baselist2
32	FamBase3	Vocab	Number of families covering all words within Baselist3
33	**RelFamBase1**	Vocab	Relative number of families in Baselist1 in % of NumBase1
34	**RelFamBase2**	Vocab	Relative number of families in Baselist2 in % of NumBase2
35	**RelFamBase3**	Vocab	Relative number of families in Baselist3 in % of NumBase3
36	TTRGlob	Vocab	Global type-token-ratio in % of all content words
37	RelFamGlob	Vocab	Total number of families in % of all content words within all baselists
38	**Tpsec**	Vocab	Types per second
39	**Wf_media**	Vocab	Median word frequency
40	**Wf_low1**	Vocab	1st lowest word frequency
41	**Wf_low2**	Vocab	2nd lowest word frequency
42	**Wf_low3**	Vocab	3rd lowest word frequency
43	**Wf_low4**	Vocab	4th lowest word frequency
44	**Wf_low5**	Vocab	5th lowest word frequency
45	Wf_10pct	Vocab	Frequency of the 10th percentile
46	Wf_20pct	Vocab	Frequency of the 20th percentile
47	Wf_30pct	Vocab	Frequency of the 30th percentile
48	Wf_40pct	Vocab	Frequency of the 40th percentile
49	Wf_50pct	Vocab	Frequency of the 50th percentile
50	**Total_er**	Gram	Total number of grammar and usage errors
51	**Unique_er**	Gram	Total number of types of grammar and usage errors

Features 1-12 (in Table 2) were used in building the Delivery score models and Features 13-51 in building the Language Use score models. The Holistic score model used both feature sets. Although various length and speech duration variables were extracted, they were not used in building any of the models.

In a separate set of classification tree analyses, only the features in bold face were used in developing the scoring models. This set of variables was selected more rigorously based on construct considerations. Variables that were not controlled for length (e.g. raw counts of word families) were generally excluded.

Linkage of automated features and the scoring rubrics

The holistic rubric of the TOEFL® iBT Speaking contains four band levels (1-4) and three key categories of performance features: Delivery, Language Use, and Topic Development (Figure 2). The descriptors for the four levels (1-4) of Delivery, Language Use and Topic Development were used to create a separate analytic rubric for each dimension. The same 1-4 scale was adopted for each of the analytic rubrics.

Delivery refers to the pace and clarity of the speech. In assessing Delivery, raters considered the speakers' pronunciation, intonation, rhythm, rate of speech, and degree of hesitancy. *Language use* refers to the range, complexity, and precision of vocabulary and grammar use. Raters evaluated candidates' ability to select words and phrases and their ability to produce structures that appropriately and effectively communicated their ideas. *Topic development* refers to the coherence and fullness of the response. When assessing this dimension, raters took into account the progression of ideas, the degree of elaboration, the completeness, and, in the case of integrated tasks, the accuracy of the content.

In Figure 2, different hues of blue are used to indicate our degrees of success in modeling different aspects of speech in the rubrics using speech and NLP technologies. We have been most successful in modeling the fluency aspect of Delivery with automated features such as pacing, speech rate, and frequency and duration of pauses (Features 1-12). Vocabulary diversity and sophistication in Language Use are indicated by Automated Features 13-49 and grammatical accuracy aspect by Features 50 and 51. Although we have computed features that represent aspects of speech such as intonation and rhythm, they need some further modifications to be included in the scoring models.

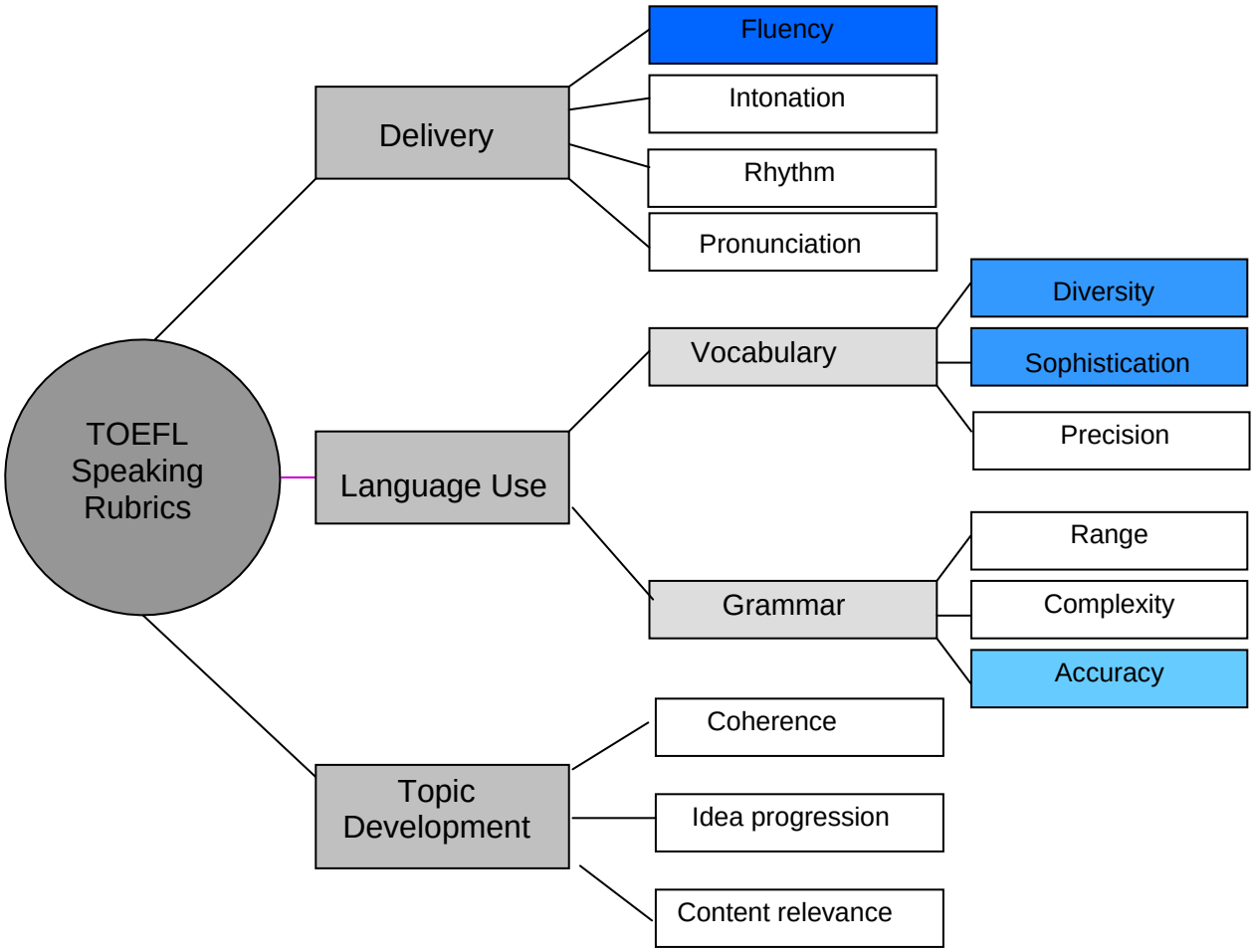

Figure 2: **A graphic representation of the scoring rubrics of TOEFL® iBT Speaking**

Features extracted in Hypo mode and Align mode

For each task response, the above speech features were extracted in two modes: *Hypo* and *Align*. Hypo means that the recognizer is in standard recognition mode where it tries to reconstruct the string of words in the utterance corresponding to the input speech signal. Align means that the recognizer runs in a "forced alignment" mode. Here, the recognizer is told about the word string of an utterance based on human transcriptions of the examinees' responses. The recognizer tries to create a mapping between the transcript and the corresponding input speech signal. The output of this process is a word alignment that has the word strings and the starting and ending time of each word. In other words, the most likely "breakdown" of an utterance into words is obtained given the speech signal and the human transcription. Therefore, the words are obviously all correct using human transcriptions as the criterion; however, the timing information may not be accurate in some cases. The aligned mode may not be practical or efficient in practice because it requires human transcriptions of the speaking responses; however, it

demonstrates that with the existing speech features, if speech recognition were perfect (using human transcriptions as the criterion), what prediction accuracy we would be able to achieve.

Analyses and Results

Classification tree results

CART 5.0 was used to build classification trees to predict Delivery, Language Use, and Holistic scores based on features extracted from the Hypo mode and the Align mode. To build the Holistic scoring models for the Hypo and Align modes, ten-fold cross-validation was performed on the training sample to select the model that yielded the best results on the cross-validation sample. Then the test sample cases were dropped down the best tree to obtain the classification rate. To build scoring models for Delivery and Language Use scores, ten-fold cross-validation was performed on the whole sample and the average classification accuracy on the cross-validation samples was reported. As Brieman et al. (1984) points out, the cross-validation estimates tend to be "conservative in the direction of overestimating classification costs" because the auxiliary cross-validation trees are grown on smaller samples (p.77).

Table 4 presents the agreement rates between automated scores and human score 1 and human-human agreement rate averaged across all six tasks using all features in Tables 2 and 3. Table 5 presents parallel results based on the features in bold face in Tables 2 and 3. As shown in Table 4, the exact plus adjacent human automated agreement rates were fairly close to the corresponding human agreement rates. The human automated score agreement rates still lagged behind those for the human raters. Table 5 demonstrates that with a more selective set of features, the scoring models were worse in predicting Language Use scores and a slight drop in the classification rates for Holistic and Delivery scores was also observed. In addition, with features extracted in the Align mode, the scoring models were more successful in predicting Language Use scores, as demonstrated in both Tables 4 and 5. However, a reverse pattern was seen with regard to the prediction of Delivery scores across the Hypo and Align modes.

Table 4: Agreement rates between automated scores and human score 1 and human agreement rate across all six tasks (using all features in Tables 2 and 3)

	Human 1 & Automated Hypo		Human 1 & Automated Align		Human 1 & Human 2[1]	
	Exact	Exact + adjacent	Exact	Exact + adjacent	Exact	Exact + adjacent
Delivery	51.8%	93.0%	46.7%	93.8%	57.0%	97.1%
Language use	47.0%	91.7%	51.2%	93.1%	53.5%	95.5%
Holistic	50.3%	94.6%	50.3%	93.9%	59.8%	97.3%

[1] The human-human agreement rates for Holistic scores were computed based on the portion of double-rated cases in the data.

Table 5: Agreement rates between automated scores and human score 1 and human agreement rate across all six tasks (using features in bold face in Tables 2 and 3)

	Human 1 & Automated Hypo		Human 1 & Automated Align		Human 1 & Human 2	
	Exact	Exact + adjacent	Exact	Exact + adjacent	Exact	Exact + adjacent
Delivery	50.6%	93.5%	45.6%	93.6%	57.0%	97.1%
Language use	44.5%	85.9%	47.6%	92.5%	53.5%	95.5%
Holistic	48.1%	95.2%	49.0%	93.9%	59.8%	97.3%

Decision tree for Holistic scores (Hypo mode; using features in bold face in Tables 2 and 3)

Figure 3 graphically presents the decision rules that led to different Holistic score classes using features in bold face in Tables 2 and 3 extracted in the Hypo mode. The same information is presented in Table 6, in which the decision rules for the same score class are grouped together. As shown in Figure 3 and Table 6, both Delivery and Language Use features were present as splitters in the classification tree for Holistic scores. For example, if the number of words per chunk is less than 9.2, the third lowest standardized word frequency index value is smaller than 52.6, the proportion of less frequent content words is less than 6.1%, and word per second is less than 2.6, a response is classified as a "2" (Terminal Node 1). It is also worth noting that six and eight different sets of decision rules lead to the score classes of 2 and 3 respectively, suggesting that different patterns of strengths and weaknesses may result in the same score. In contrast, there were only four sets of decision rules for the score class of 4 and one for the score class of 1. This is consistent with our understanding that there is more variation in profiles of strengths and weaknesses in the middle-score levels whereas the high- and low-score level students tend to be uniformly strong or weak in different aspects of speech.

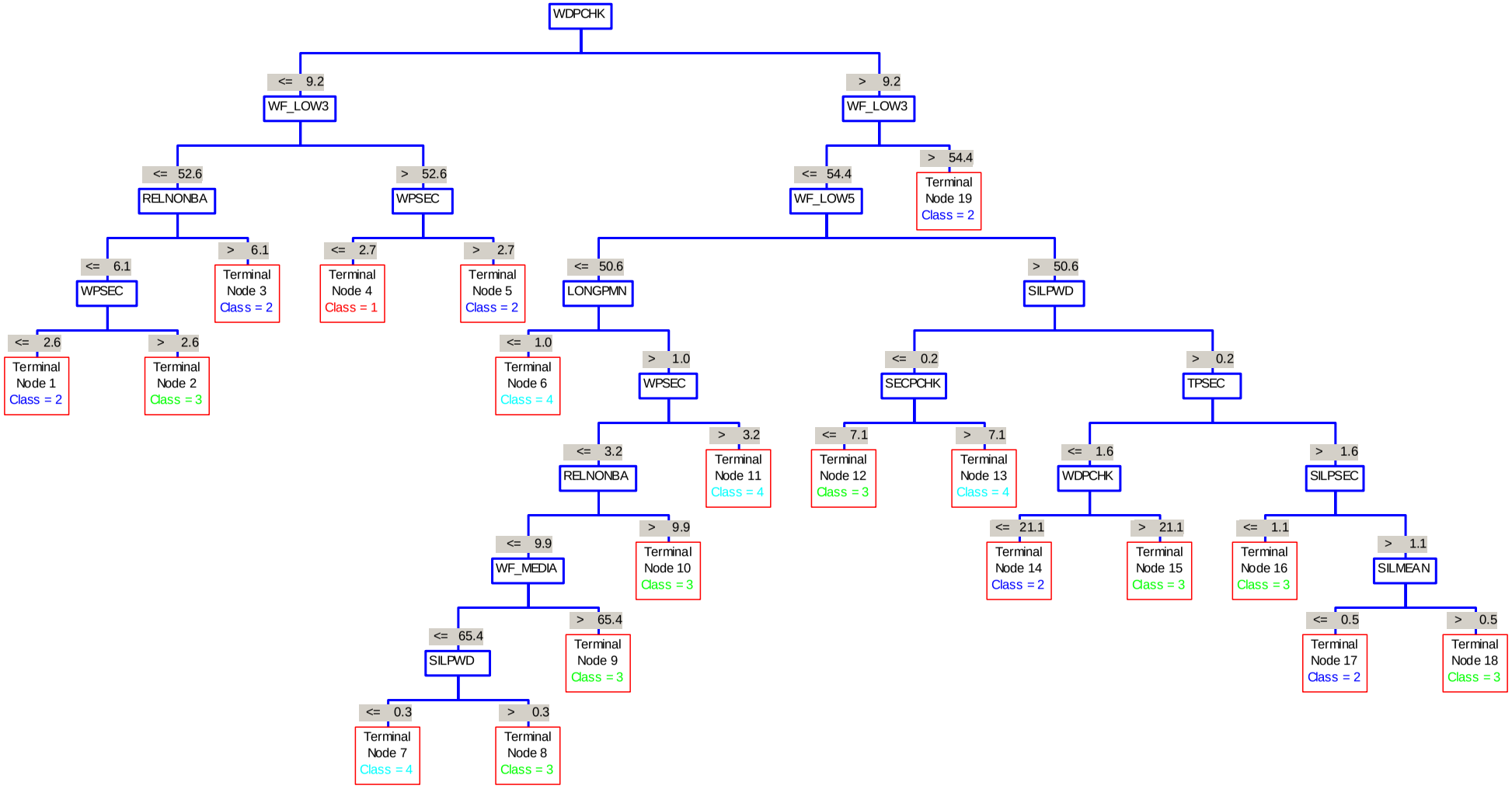

Figure 3: Decision tree for Holistic scores (features extracted based on hypotheses of the speech recognizer and input audio signals)

Table 6: Decision rules that led to the score classes

Score class	Terminal node	WDPCHK	WF_LOW3	RELNONBA	WPSEC	WF_LOW5	LONGPMN	SILPWD	SECPCHK	TPSEC	SILPSEC	SILMEAN	WF_MEDIA
1	4	<=9.2	>52.6		<=2.7								
2	1	<=9.2	<=52.6	<=6.1	<=2.6								
2	3	<=9.2	<=52.6	>6.1									
2	5	<=9.2	>52.6		>2.7								
2	14	>9.2<=21.1	<=54.4			>50.6		>0.2		<=1.6			
2	17	>9.2	<=54.4			>50.6		>0.2		>1.6	>1.1	<=0.5	
2	19	>9.2	>54.4										
3	2	<=9.2	<=52.6	<=6.1	>2.6								
3	8	>9.2	<=54.4	<=9.9	<=3.2	<=50.6	>1.0	>0.3					<=65.4
3	9	>9.2	<=54.4	<=9.9	<=3.2	<=50.6	>1.0						>65.4
3	10	>9.2	<=54.4	>9.9	<=3.2	<=50.6	>1.0						
3	12	>9.2	<=54.4			>50.6		<=0.2	<=7.1				
3	15	>21.1	<=54.4			>50.6		>0.2		<=1.6			
3	16	>9.2	<=54.4			>50.6		>0.2		>1.6	<=1.1		
3	18	>9.2	<=54.4			>50.6		>0.2		>1.6	>1.1	>0.5	
4	6	>9.2	<=54.4			<=50.6	<=1.0						
4	7	>9.2	<=54.4	<=9.9	<=3.2	<=50.6	>1.0	<=0.3					<=65.4
4	11	>9.2	<=54.4		>3.2	<=50.6	>1.0						
4	13	>9.2	<=54.4			>50.6		<=0.2	>7.1				

Discussion

Classification Accuracy with Classification Trees

The classification trees yielded fairly high classification accuracy with Holistic scores. The exact plus adjacent agreements between human and automated scores were fairly close to those between two human raters. Given that in the operational scoring of TOEFL® iBT Speaking adjacent agreements are acceptable, the prediction was fairly good. However, the human-automated perfect agreements were still lower than those between human raters, suggesting that there may be more error in automated scores if human scores are treated as the "criterion." The classification rates were generally lower on Language Use scores, especially with the more selective set of features, indicating that additional work is needed to better characterize the quality of Language Use.

Although not discussed in this paper, the features used in the scoring models for different task types such as independent versus integrated tasks could well be somewhat different due to theoretical reasons. For example, for integrated tasks, content relevance and accuracy are part of the rubric and should be reflected in the scoring models. For another example, vocabulary features for integrated tasks that involve academic course content may be different than those for tasks that involve campus life. Therefore, building a separate model for each task type may potentially boost the classification accuracy. With a larger sample of data, this is definitely something that should be examined.

In this study, each of the splitters in the classification tree is a single feature. However, classification trees also take linear combinations of features as splitters. It is possible that the classification trees may be more successful in classifying score classes with this model set-up. However, we should also realize that the trees would be less interpretable. Another emerging technique in classification and regression trees, TreeNet (Hastie, Tibshirani, & Friedman, 2001), combines a committee of trees, and has shown superior performance compared to a single-tree model. A drawback, however, is that it is not as easy to derive clear decision rules as with a single-tree model, although TreeNet also provides information about which features are the most important in classifying the cases.

In this study, the scoring models were built with relatively small samples. The prediction accuracy may improve with a larger sample. Another limitation of this study was that the human scores that were modeled were obtained in a pilot study and were more error-prone than those in operational scoring of TOEFL® iBT Speaking, where raters have become more consistent with more familiarity and experience with the test and the rubrics. It is likely that automated scores modeled on less error-prone human scores may be more accurate.

Meaningfulness and Relevance of the Significant Features in the Scoring Models

The speech features used in the scoring models were chosen based on a close examination of the rubrics, test developers' input, and speech analysis literature in second-language learning.

For example, the average length of uninterrupted runs (WRDCHK and SECCHK), a major splitter in both the Holistic and Delivery scoring models, was originally proposed by a TOEFL speaking assessment specialist, and then realized computationally by using the timing convention of half-a-second to delimit the prosodic chunks, i.e., uninterrupted runs of speech. Most of the vocabulary diversity and sophistication features are commonly used in the second-language learning literature to measure the diversity and sophistication of the vocabulary of non-native speakers of English.

As presented in the results section, the decision rules of the classification trees generally make sense from a substantive perspective, although further consultation with a group of test developers is required to see if modifications of the decision rules are necessary. It has to be noted that the decision rules at best represent partial decision processes used by expert raters since some key speech variables such as pronunciation, intonation, rhythm, and content are missing.

One advantage of classification trees, as discussed earlier, is its transparency in the decision rules used to derive the predicted scores. Although it is possible to use it in a completely unsupervised situation (i.e., let the content specialists specify the decision rules for different score classes), given that experts may have difficulty breaking down their thought processes in a way that would be helpful for deriving decision rules for automated scoring, and that they may not be able to articulate all patterns of strengths and weaknesses that lead to specific score classes, it may be more fruitful to either add expert human-rater input when choosing the "right" tree or modify the tree (if necessary) to maximize construct representation while not compromising prediction accuracy significantly. CART 6.0 has incorporated the function of allowing the user to specify splitters. In the future, this added feature will give us more flexibility in modifying the trees to enhance construct representation.

The Promise to Improve Classification Accuracy with More Meaningful Features

The work to extract relevant pronunciation, intonation, pitch, and rhythm features is ongoing. Although we have extracted some content vector variables which indicate the similarity of the vocabulary used in an examinee response with that of different score classes, they were not used in building the Holistic scoring models, because the computation of these features required access to examinee responses to the same speaking tasks. Thus, leaving out the content vector variables, we were able to see how these features would perform in predicting scores without having to train on responses to the specific tasks.

However, in a practice environment where retired TOEFL® iBT Speaking tasks are used, we have access to examinee response data before these tasks are reused for practice purposes. Thus, we can include the content vector variables in the scoring models, which will potentially improve the prediction of the scores.

The grammar and usage features generated by the e-rater® (an automated essay scoring system developed by the Educational Testing Service) algorithms, specifically the total frequency and the number of types of grammar and usage errors did not appear to be important predictors of the scores. There was not much variation among the examinees in this sample on the counts of these errors. Also, most of the errors that could be detected by e-rater® such as subject verb agreement, pronoun errors, missing possessive errors, and wrong and missing article errors do not seriously impair overall spoken communication, although academic writing puts a more stringent requirement on grammatical accuracy. More critical grammar and usage errors in spoken discourse such as awkward phrasing, flawed sentence structures, and the incorrect use of modals (should, can, may, would, etc) are not easily detected but interfere with oral communication.

The use of idiomatic expressions is characteristic of spoken English but it is not an easy task to automatically identify them and evaluate the accuracy of these expressions. As part of our immediate work plan, we are extracting robust collocations from the Michigan Corpus of Academic Spoken English, which can then be used as a norm to evaluate the use of collocations in non-native speech.

Promise in Improving the Recognition Accuracy and the Classification Accuracy

The recognition accuracy of the speech recognizer is certainly crucial to the success of any automated speech scoring system. As previous research has shown (Tomokiyo, 2001; Waibel et al., 2000), a speech recognizer's acoustic model can be favorably adapted to improve the recognition of non-native speech by adding non-native speech to the training set. Our own preliminary findings (Zechner et al., in progress) show significant improvements in the recognition rates for some non-native languages with an adaptation set size of about 10-20 hours of speech. Taking advantage of the large number of speech samples generated by TOEFL® iBT, we can expect greater improvements in recognizing accented speech of particular L1 groups if we adapt the acoustic model extensively to the accented speech of a particular L1 group. With more accurate speech recognition, the extracted speech features, especially Language Use and Topic Development, are likely to be more reliable, leading to more precise prediction. This effort can potentially improve the quality of learning tools designed for these L1 populations.

Using automated speech analysis for providing instant feedback

An automated speech feature which is meaningful from a learning and instruction perspective may not be a significant predictor in the scoring model because there is not much variation on it or it is highly correlated with a feature already in the model. However, it will still offer promise in providing diagnostic information and can be included in a diagnostic evaluation/feedback model.

Automated speech analysis has the potential to support diagnostic information but some additional capabilities need to be developed to enable the diagnostic function. For example, the number of different types of grammatical errors may be an automated feature, but in order to maximize the value of diagnostic information, a display and synthesis of different types of errors may be necessary. In addition, speech features that can be extracted using automated means are very fine-grained and concrete (e.g. mean duration of silences), and a substantial amount of additional work is needed to translate them into feedback information at a level that language learners and teachers find meaningful.

Conclusion and Future Work

This study reports an investigation of ECD-based automated scoring of TOEFL® iBT Speaking. Although conceptually we have mapped out the various sub-features that can potentially be extracted automatically, work in extracting additional features continues to represent a full model of speaking proficiency. In addition, our ongoing work in improving the recognition of non-native speech (Zechner et al., in progress) offers promise in extracting more reliable sub-features. Before we are confident about building an adequate validity model for the features used in the scoring models and the way they interact to provide appropriate evidence about academic speaking proficiency, it would be premature to put these capabilities to use for high-stakes decisions.

We recognize that fully automated scoring for operational use may not be feasible immediately given the complexity of the problem. Nevertheless, the project has provided evidence about the feasibility of developing fully-automated scoring in the future and about the possibility of providing some diagnostic feedback to examinees.

The speech analysis capabilities developed through this project will expand our infrastructure to enable ETS to produce computer-assisted learning tools. This study can provide a scientific basis for conceptualizing and developing English Language Learning (ELL) speaking products and offer the potential for ETS to expand and diversify its ELL products to include computer-assisted teaching and learning tools.

References

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing.* New York: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bejar, I. I., & Braun, H. I. (1994). On the synergy between assessment and instruction: early lessons from computer-based simulations. *Machine-Mediated Learning*, 4, 5-25.

Bernstein, J. (1999). Validation Summary for PhonePass SET-10. Technology reports. www.ordinate.com.

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

Braun, H.I., Bejar, I.I. & Williamson, D. M. (in press). 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*.

Cucchiarini, C., Strik, H., & Boves, L. (1997a, September). *Using speech recognition technology to assess foreign speakers' pronunciation of Dutch.* Paper presented at the Third international symposium on the acquisition of second language speech: NEW SOUNDS 97, Klagenfurt, Austria.

Cucchiarini, C., Strik, S., & Boves, L. (1997b). *Automatic evaluation of Dutch pronunciation by using speech recognition technology.* Paper presented at the IEEE Automatic Speech Recognition and Understanding Workshop, Santa Barbara, CA.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., & Butzberger, J. (2000). The SRI EduSpeak system: Recognition and pronunciation scoring for language learning. Paper presented at the InSTiLL-2000 (Intelligent Speech Technology in Language Learning), Dundee, Scotland.

Hastie, T, Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* Springer Series in Statistics. Springer-Verlag, New York.

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

Mislevy, R. J., Steinberg, L., Almond, R., Lukas, J., (in press). Concepts, Terminology, and Basic Models of Evidence-Centered Design. In Williamson D. M., Mislevy, R. J.& Bejar I.I. (Eds.) *Automated Scoring of Complex Tasks in Computer Based Testing.* NJ: Earlbaum.

Tomokiyo, L. (2001). *Recognizing non-native speech: Characterizing and adapting to non-native usage in speech recognition*. Ph.D. thesis, Carnegie Mellon University.

Waibel, A., Soltau, H., Schultz, T, Schaaf, T., & Metze, F. (2000). *Multilingual Speech Recognition.* In Verbmobil: Foundations of Speech-to-Speech Translation, Wahlster W. (ed.), Springer Verlag, 2000.

Williamson D. M., Mislevy, R. J., & Bejar I.I. (Eds.) *Automated Scoring of Complex Tasks in Computer Based Testing.* NJ: Earlbaum.

Xi, X, & Mollaun, P. (in press). Investigating the utility of analytic scoring for TOEFL Academic Speaking Test (TAST). TOEFL iBT Research Report No 1. Educational Testing Service: Princeton, NJ.

Zechner K., Bejar I., Higgins D., Lawless R., & Futagi Y. (In progress). *Automatic Acoustic Model Adaptation for Non-native Speech Using Elicited Read Sentence With L1-Specific Phonetic Complexity*.

Zechner, K., Bejar, I, & Hemat, R. (In preparation). Towards an Understanding of the Role of Speech Recognition in Non-native Speech Assessment. TOEFL iBT Research Report. Educational Testing Service: Princeton, NJ.