Quantitative, Notional, and Comprehensive Evaluations of Spontaneous Engaged Speech

Garry Molholt, María José Cabrera, V. K. Kumar, and Philip Thompsen
West Chester University of Pennsylvania

ABSTRACT

This study provides specific evidence regarding the extent to which quantitative measures, common sense notional measures, and comprehensive measures adequately characterize spontaneous, although engaged, speech. As such, the study contributes to the growing body of literature describing the current limits of automatic systems for evaluating spoken proficiency, provides examples of the essential nature of various notional and comprehensive variables, supports continued development of hybrid systems, and includes suggestions for the possible utilization of additional variables for automatic analyses. Data for this study were gathered and analyzed as follows. After 4 weeks of activities related to career development, 20 native English speaking college freshmen made recordings in English explaining their career preferences. Three experiments were then conducted. Experiment 1 analyzed the recordings according to current quantitative analyses used in fully automatic evaluations of fluency. Experiment 2 examined the recordings through a perception study according to common everyday notions of fluency. Experiment 3 analyzed the recordings according to an adaptation of the comprehensive rubrics used by the Educational Testing Service (ETS) for evaluating oral proficiency. The comprehensive evaluation (Experiment 3) provided the most insight, and temporal quantitative measures (Experiment 1) provided the least insight concerning the proficiency of the 20 speakers.

KEYWORDS

Spoken Proficiency, Fluency, Quantitative Analysis, Notional Analysis, Comprehensive Analysis

INTRODUCTION

We are currently faced with three characterizations of speech proficiency. According to the strongest assertions of those who advocate fully automatic quantitative computer analysis of speech features: "At the overall score level, Versant English Test machine-generated scores are virtually indistinguishable from scoring that is done by careful human transcriptions and repeated independent human judgments" (Balogh, 2008). According to a more general common or everyday notional definition—within a broad conception—spoken fluency equals proficiency (Hilton, 2008). Finally, from a more comprehensive standpoint, as in the rubrics established by the Educational Testing Service (ETS), spoken proficiency involves several aspects of delivery, language use (including vocabulary and grammar), and topic development, not all of which are currently amenable to effective automatic analysis (Xi, Zechner, & Bejar, 2006).

In order to understand the relationships among variables used in these three different approaches, we devised three experiments for analyzing the engaged spontaneous speech of 20 native English speaking college freshmen. The first included detailed quantitative analyses of the temporal aspects of the speech samples, along with some additional nontemporal quantitative measures. The second, from a more general perspective, was a perception study fo-

cusing on the notions of clarity, confidence, and fluency. The third study involved a classroom adaptation of the categories of the comprehensive ETS rubrics used by human raters. Conclusions regarding the efficacy of the three approaches and the need for further research were drawn from correlations of the rank orders of the speakers as calculated for each of the three methods: quantitative, notional, and comprehensive. The results of our study showed that the comprehensive analysis provided essential insight into proficiency which was not available from either the notional or the quantitative analyses. In addition, the notional study provided better overall insight than the quantitative study, especially with respect to automatically analyzed temporal quantitative variables.

There are several reasons for investigating the speech of native speakers of English before considering the speech of nonnative speakers of English. If we want to establish targets for nonnative speakers, we should have a clear understanding of the variation among native speakers. In fact, when trying to establish targets for nonnative speakers to achieve, we may want to limit the ranges of some of the parameters within the native speaking population rather than simply assuming that any native speaker is a good model, even those among college students. This will become clear from listening to files, including stronger proficiency, average proficiency, and weaker proficiency (see the online version of this article for links to speech samples).

There are also many reasons for investigating spontaneous engaged speech, as opposed to prepared speech or spontaneous speech on topics that are completely new to the speaker. For this study, engaged speech refers to speech on a topic the speaker has been involved with for a significant period of time. This is different from speech describing a cartoon the speaker has never seen before or speech responding to a hypothetical question the speaker has never thought of before. Engaged speech allows the speaker to reflect on depth of content while still being spontaneous. As long as the speech was not prepared in advance, it represents the speaker's competence and not merely a type of performance. When we conduct job interviews, participate in promotion committees, or conduct admission interviews for various programs (all high-stakes situations), we expect the candidates to be engaged in their topics, fluent in their speech, and yet spontaneous in their responses. We would not generally focus on questions that are unrelated to the situation at hand or on simple questions requiring low entropy or expected one-word answers (Molholt, Cabrera, Kumar, & Thompsen*,* 2009).

DATA COLLECTION

In order to create shared experiences for 20 college freshmen so that they could make recordings of spontaneous engaged speech, we included a 4-week unit on the topic of career development in a freshman composition class. After an orientation from the director of our career development center*,* the students participated in a job fair held at the student union. Then all the students took a 90-minute online career aptitude test, called FOCUS, which provided them with detailed information on their interests, relevant careers, required training, and pay scales, along with many other details. After participating in class discussions and presentations, students wrote an essay on the entire process of developing their ideas about careers and then a second essay regarding their current career choices. Finally, the individual students came to a quiet office and discussed their essay during office hours. It was during this time that the recordings were made for subsequent analyses of the speech files. Students were not told in advance that they would be making recordings about their current career preferences.

RECORDINGS

After the students discussed their essays in the office, they were asked to explain their current career preferences. The students spoke into a Shure PG81 microphone connected to an M-Audio preamp. This in turn was connected to a PC running KayPentax MultiSpeech software with the input sampling rate set at 22,050 Hz. 16 bits. The duration of most recordings was about 35-45 seconds and provided sufficient information to determine the relative salience of the various speech variables. More data would be required, however, for full-scale proficiency evaluations. Since the students had not been previously informed that they would be making a recording, they did not have a prepared script to recite or read from. Students signed a form giving the authors permission to use their data.

Three different types of analysis (quantitative, notional, and comprehensive) were applied to the recorded samples of the engaged speech of the 20 students.

QUANTITATIVE ANALYSIS

Background Literature on Quantitative Analysis

There has been extensive interest regarding the extent to which face-to-face oral proficiency interviews could be replaced by automatic computer analyses of speech. As early as 1986, Molholt and Pressler started looking at the feasibility of automatic fluency checking in a project with ETS (Molholt & Pressler, 1986). Cucchiarini and colleagues (Cucchiarini & Strik, 1999; Cucchiarini, Strik, & Boves, 2000, 2002) have investigated the significant differences between quantitative analyses of read and spontaneous speech among second language speakers. Though they found that the correlations between human ratings and computer ratings are much lower (r = .65) for spontaneous speech than for read speech (range of r = .81 to .93), they concluded that certain quantitative measures of timing are important to our perceptions of fluency. These measures include pace (phonemes per second from start to finish), articulation rate (phonemes per second not including the duration of internal pauses), phonation-time ratio (the amount of time devoted to speech sounds within an utterance, as opposed to the total amount of time including silence), and the mean length of runs (duration of stretches of speech with no pause equal to or greater than 200 ms).

From the various publications and presentations representing Versant, however, several different pictures emerge. For example, on the one hand, it is claimed that "At the Overall score level, Versant English Test machine-generated scores are virtually indistinguishable from scoring that is done by careful human transcriptions and repeated independent human judgments" (Balogh, 2008; Versant, 2008c). On the other hand, it is claimed that "Versant also had a correlation coefficient of [only] 0.75 with the [more advanced] interview test of ILR (Interagency Language Roundtable)" (Versant, 2006). To understand the difference between these characterizations, we need to look at what Versant measures. As stated in Versant (2008b),

> For the Versant English Test, responses to four item tasks are currently used for automated scoring. These are: reading aloud, repeating sentences, building [finishing] sentences, and giving short answers to questions. In scoring, there is exactly one correct word sequence expected for each response to the read and repeat items. Expert judgment was used to define correct answers to the short-answer question and sentence-build items.

Although this precludes any chances of eliciting actual spontaneous speech, it is somewhat beyond read speech. Thus, it stands to reason that correlations between machine ratings of Versant evaluations and the more advanced ILR interview test reported by Versant (2006) would fall somewhere between the correlations between machine and human ratings for read speech and for spontaneous speech reported by Cucchiarini et al. (2002).

The correlation of .75 reported by Versant (2006), although significant (p = 0.01), accounts only for 56% of the variance in the relationship between the automatic evaluation and the human ratings, leaving much room (44%) for error. Furthermore, the Versant evaluation has been characterized as only dealing with linguistically simpler tasks (Zechner, Higgins, & Xi, 2007).

A reason for the variation in the claims made by Versant is related to the ability level of the speakers being evaluated. Correlations between machine and human ratings of the fluency of lower level second language speakers are considered to be better than those for higher level speakers. Thus, the upper levels of the ILR ratings would not be adequately represented by the Versant methods.

Versant's *Method and apparatus for voice-interactive language instruction* (US Patent, 1997) is a 29-page document providing explicit explanations of the analysis routines used in their automatic oral proficiency scoring. The Versant procedure relies heavily on quantitative measures of temporal aspects of speech, similar to the work of Cucchiarini and colleagues referred to above, along with quantitative analyses of the proportion of expected versus unexpected units in answers to prompts (see also Versant, 2008a).

Blake, Wilson, Cetto, & Pardo-Ballester, (2008) observed that second language students perform with similar reaction speed to prompts regardless of the type of environment: distance, face-to-face, or blended. They say that the similar reaction speed gives some support for the Versant remote methods, but they do not include references to the actual content of the speech under investigation.

Many recent publications and presentations, such as Audhkhasi, (2009), Deshmukh, Kandhway, and Verma, (2009), and Kondo, Tsutsui, and Nakano, (2008), provide insight into current limitations of fully automatic evaluations of proficiency. Neri, Cucchiarini, and Strik, (2002), noted that the limitations of current automatic speech recognition (ASR) technology imply that error analysis is simply not feasible because the performance levels attained are too poor. Xi et al. (2006) reported that attempts by ETS researchers to automatically analyze proficiency are currently more successful in the areas of fluency, vocabulary diversity and sophistication, and grammatical accuracy. The researchers did not have as much success in the automatic evaluation of intonation, rhythm, pronunciation, vocabulary precision, range, or complexity, or topic development (e.g., coherence, idea progression, or content relevance). Additionally, Xi et al. raised an important question: If the system is not able to identify specific errors, how can it provide a useful proficiency score? As an example of the limitations of ASR, Aylett (2003) noted that even if speech recognizers were able to successfully reach the point of proposing viable choices, they still will have difficulty differentiating between sentences such as "It's hard to recognize speech" and "It's hard to wreck a nice beach." For such sentence choices, the computer is forced to try to decide which would be most likely in the context of the other available utterances, if any are present. Without strictly controlled domains of discourse, cost effective automatic recognition of continuous spontaneous speech (with such problems related to high entropy) still remains a monumental challenge for the future.

Molholt et al. (2009) correlated quantitative measures of speech with perceptions of fluency and found that even though there was a statistically significant correlation, (r = .78, p <

.001), the predictions of the quantitative measures were often strikingly different from the perceptions of fluency. For the temporal measures, they employed the variables judged most salient by Cucchiarini et al. (2002): pace, articulation rate, phonation/time ratio, and mean length of runs. In order to study the possible effects of additional quantitative variables, they added four more variables: overall pitch variation, overall pitch range, phoneme clarity, and the ratio of nonfiller words compared to all word-like sounds, including filled gaps or words used as gap fillers.

Experiment 1: Quantitative Study

Temporal quantitative analyses

The process for the quantitative analyses of the 20 speech files described above and included in Molholt et al. (2009) is as follows. Each file was first transcribed in a spreadsheet, one word per line, including filled gaps. The durations of any gaps, even as small as 3 ms, were entered in the next column after the word or sound that they followed. Then the number of phonemes each word represented was entered next to each word. After that, the number of missing or distorted phonemes was entered. The utterance durations and the gaps were measured manually using the KayPentax MultiSpeech software. The manual measurement procedure was used since it provides much more precise figures than automatic machine measurements when extraneous noise could be a factor.

The temporal quantitative variables included were

1. PACE
   the total number of phonemes per second, including internal gaps, calculated by adding the total number of phonemes and the duration of the gaps
2. PWO
   the articulation rate (pace without gaps)—the number of phonemes per second without including internal gaps
3. PTR
   phonation/time ratio—the relationship between the amount of speech time to the total time of the utterances
4. MLR
   mean length of runs—the average length of stretches of speech without a gap equal to or greater than 200 ms

The results of these analyses of temporal quantitative variables were correlated with the notional analyses and the comprehensive analyses separately, before the additional quantitative analyses were included in further calculations of correlations.

Additional (nontemporal) quantitative analyses

Though variation of intonation (monotone to sing song) has often been included in traditional definitions of fluency, it is frequently left out of automatic analyses of speech. One reason for this exclusion is that noise in the speech signal sometimes distorts the pitch curve. In order to reduce the effects of such noise, each file was checked to identify the actual highest and lowest levels of pitch, and those levels were used as the pitch ranges before obtaining the means and standard deviations of the pitch curves from the MultiSpeech pitch and statistics routines. Since people speak at different pitch levels, the data resulting from these analyses needed to be normalized before the amount of pitch variation could be compared from one person

to another. This was accomplished by dividing the standard deviation of each pitch curve by the mean pitch of that curve. This gives a coefficient of pitch variation (SEMI) which can be compared across individuals. The process also provides values for the pitch range (SEMR) for each speaker (Molholt, Morgan, & Park, 2007).

The phoneme clarity index (PCI) for each speaker was computed by dividing the number of correctly pronounced phonemes by the total number of phonemes represented by all the words in the file. Three types of incorrect phonemes were counted: wrong phonemes (e.g., "sip" vs. "ship"), missing phonemes not the result of blending (e.g., "pres-dent" for "president"), and reduced phonemes creating possible ambiguities or unclear phrases (e.g., "she's disgusted" vs. "she's discussed it").

Finally, the ratio of nonempty filler vocabulary (NEFV) to the total number of word or word-like sounds was computed by dividing the number of words not functioning merely as gap fillers by the total number of words (including fillers). Both PCI and NEFV are related to quantitative analyses of the proportion of expected versus unexpected units in responses to prompts included in the Versant analyses discussed above (see variable definitions in Appendix A).

Raw and normalized data

Table 1 lists the raw data for the eight quantitative measures described above.

Table 1
Raw Quantitative Scores

	Temporal				Nontemporal			
Name	PACE	PWO	PTR	MLR	SEMI	SEMR	PCI	NEFV
Anthony	11.75166	13.45013	0.87721	4.08	3.68	20	0.99492	0.93
Ashley	13.05211	14.63275	0.89198	4.34	2.63	13	0.98361	0.96
Bryan	10.37688	13.31568	0.77930	2.18	1.54	10	0.93386	0.92
Chad	8.20797	11.49011	0.71435	1.67	1.52	9	0.91753	0.97
Jamie H	6.84911	10.56133	0.64851	1.64	2.04	17	0.94052	0.94
Kate	10.93144	13.04590	0.83792	2.66	1.94	13	0.94085	0.93
Lia	9.42088	13.99222	0.67329	1.52	1.24	6	0.88950	0.93
Mark	6.19143	9.49638	0.65198	1.73	1.44	9	0.97541	0.96
Megan B	10.28598	13.74349	0.75096	2.43	2.42	16	0.95320	0.86
Megan M	10.64590	12.32190	0.86398	4.31	1.13	7	0.96000	0.95
Mike B	7.24152	11.60820	0.62383	1.62	1.66	7	0.91935	0.80
Mike H	10.15780	14.06163	0.72238	1.82	1.96	12	0.90425	0.97
Natasha	8.91054	10.42079	0.85507	4.30	1.75	13	0.97188	0.96
Nick	10.32896	12.67773	0.81473	2.85	1.54	9	0.98765	0.93
Rachel	10.72595	12.57125	0.85860	4.14	2.20	21	0.89950	0.89
Steph B	9.00911	11.95765	0.75342	2.57	1.80	8	0.95555	0.87
Steph H	9.14760	12.87379	0.71056	2.12	2.17	14	0.94118	0.94
Steph K	10.82426	14.14312	0.76534	2.53	1.36	7	0.93227	0.96
Steph V	10.36649	13.54886	0.76512	2.96	1.52	8	0.90685	0.88
Tracy	12.02734	15.23148	0.78964	2.10	2.21	15	0.93484	0.92

In order to correlate these data with notional data, and comprehensive data, scores were first converted to *z* scores showing how many standard deviations they were from the population means for each measure. To find the total quantitative scores, the various *z* scores were added together for temporal measures, for the nontemporal measures, and the combination of temporal and nontemporal measures for each speaker. Equal weights were employed (see Table 2).

Table 2
Quantitative *Z* Scores

	Temporal				Nontemporal				
Name	PACE	PWO	PTR	MLR	SEMI	SEMR	PCI	NEFV	TOTAL
Anthony	1.11	0.46	1.23	1.39	3.09	1.87	1.67	0.25	11.07
Ashley	1.86	1.26	1.48	1.63	1.28	0.29	1.33	1.00	10.13
Tracy	1.26	1.66	0.25	-0.57	0.55	0.74	-0.33	0.00	3.56
Rachel	0.52	-0.13	1.11	1.45	0.53	2.09	-1.33	-0.75	3.49
Natasha	-0.52	-1.57	1.11	1.60	-0.24	0.29	1.00	1.00	2.67
Kate	0.64	0.19	0.86	-0.02	0.09	0.29	0.00	0.25	2.30
Megan M	0.48	-0.30	1.23	1.61	-1.31	-1.06	0.67	0.75	2.07
Nick	0.29	-0.05	0.49	0.17	-0.60	-0.61	1.67	0.25	1.61
Megan B	0.27	0.66	-0.25	-0.25	0.91	0.97	0.00	-1.50	0.81
Steph K	0.57	0.93	0.00	-0.15	-0.91	-1.06	-0.33	1.00	0.05
Steph H	-0.39	0.07	-0.74	-0.55	0.48	0.52	0.00	0.50	-0.11
Mike H	0.20	0.87	-0.62	-0.85	0.12	0.07	-1.33	1.25	-0.29
Bryan	0.32	0.38	0.12	-0.5	-0.60	-0.38	-0.33	0.00	-0.99
Steph V	0.32	0.53	0.00	0.28	-0.64	-0.83	-1.00	-1.00	-2.34
Steph B	-0.47	-0.54	-0.25	-0.11	-0.16	-0.83	0.67	-1.25	-2.94
Chad	-0.93	-0.85	-0.74	-1.00	-0.64	-0.61	-0.67	1.25	-4.19
Jamie H	-1.71	-1.48	-1.48	-1.03	0.26	1.19	-1.33	0.50	-5.08
Lia	-0.23	0.83	-1.23	-1.15	-1.12	-1.28	-1.33	0.25	-5.26
Mark	-2.09	-2.19	-1.48	-0.94	-0.78	-0.61	1.33	1.00	-5.76
Mike B	-1.48	-0.77	-1.85	-1.05	-0.40	-1.06	-0.67	-3.00	-10.28

The z scores in Table 2 are the quantitative scores that were subsequently correlated with the results of the notional and the comprehensive studies.

NOTIONAL ANALYSIS

Background Literature on Notional Analysis

Hilton (2008) provided a concise definition of the commonly understood notion of fluency, stating that, within a broad conception, spoken fluency equals proficiency. Another traditional definition of fluency comes from Richards, Platt, and Weber (1985), who defined fluency as "The features which give speech the qualities of being natural and normal, including native-like use of pausing, rhythm, intonation, stress, rate of speaking, and use of interjections and interruptions" (p. 108). General types of definitions allow for features (e.g., intonation or topic

development) that are not found in the strictly temporal quantitative measures included in the references to Cucchiarini above or in the temporal and nontemporal quantitative measures reported in references to Versant above.

According to Brown (2003, p. 5), "suprasegmentals are crucial for expressing ourselves accurately and for being understood by others. To illustrate, consider the fact that an utterance like 'You are looking very nice tonight.' can be said in a bored manner, sincerely, sarcastically, suggestively… and probably many other ways depending on how stress, intonation, and voice quality are used." So for Brown, too, there is more to fluency and proficiency than merely the temporal aspects.

Experiment 2: Notional Study

In order to study the relationship between quantitative measures and commonly understood notional measures, Molholt et al. (2009) devised a perception study designed to elicit responses from 97 linguistics students to the clarity, confidence, and fluency demonstrated in the speech data of the 20 college freshmen participants in this study (see notional evaluation sheet in Appendix B).

The 97 linguistics students listened to the files in two different sessions. The first session included eight speech files. First, the students listened to all eight files but did not mark their evaluation sheets. Then, they listened to the separate speech files and rated the speech sample for clarity, confidence, and fluency on the scoring sheet. In addition, they could make comments about the ratings they chose.

During the second session in the following week, the remaining 12 files were presented in a similar fashion. Afterwards, the results were tabulated, the various correlations were calculated to compare the quantitative measures to the notional measures. The raw scores were converted to *z* scores, both for the individual scales and total score in which, again, equal weights were employed (see Table 3)

Table 3
Notional *Z* Scores

	Clarity	Confidence	Fluency	Total
Anthony	1.86	2.03	2.23	2.13
Megan M	1.31	1.16	1.16	1.47
Ashley	0.60	1.33	1.17	1.13
Tracy	1.03	0.61	1.06	0.95
Steph K	0.36	0.73	0.70	0.67
Kate	0.47	0.57	0.70	0.64
Nick	0.31	0.85	0.47	0.63
Rachel	-0.14	0.08	0.11	0.07
Megan B	-0.16	0.12	-0.09	0.02
Steph H	-0.12	-0.25	0.14	-0.04
Mark	-0.03	-0.28	-0.23	-0.15
Bryan	-0.34	-0.19	-0.16	-0.18

Natasha	-0.55	-0.25	-0.16	-0.27
Chad	-0.45	-0.36	-0.23	-0.31
Lia	-0.05	-0.75	-0.55	-0.44
Mike H	-1.79	-0.52	-0.83	-0.97
Steph V	-1.10	-1.07	-1.03	-1.04
Mike B	-1.00	-1.25	-1.45	-1.22
Jamie H	-0.93	-1.43	-1.55	-1.30
Steph B	-2.26	-1.60	-1.56	-1.77

The correlations between the total scores and both the confidence and the fluency scores were almost perfect (over .98), and the correlation between the total scores and the clarity scores was .96.

The 97 linguistics students were from four different classes, and the order of presentation of the speech files was changed for each of the four classes in order to control for possible running order effects. The interclass rank order correlations averaged .97. After calculating the proportion (or percentage) of agreement among the 97 raters on each of the anchor points of the Likert scales, we examined the modal proportions. Table 4 reports the averages of agreement on exact (modal) and adjacent (next to the mode) anchor points, along with values two or more anchor points from the mode. The overall average of the exact plus adjacent anchor points was 96.12%, suggesting almost perfect agreement among the raters.

Table 4
Averages of Agreement among the 97 Raters

	Clarity	Confidence	Fluency	Overall average
Exact mode	52.81	54.67	54.67	53.58
Adjacent to mode	42.97	41.70	42.96	42.54
Two or more from mode	4.27	3.54	3.77	3.86

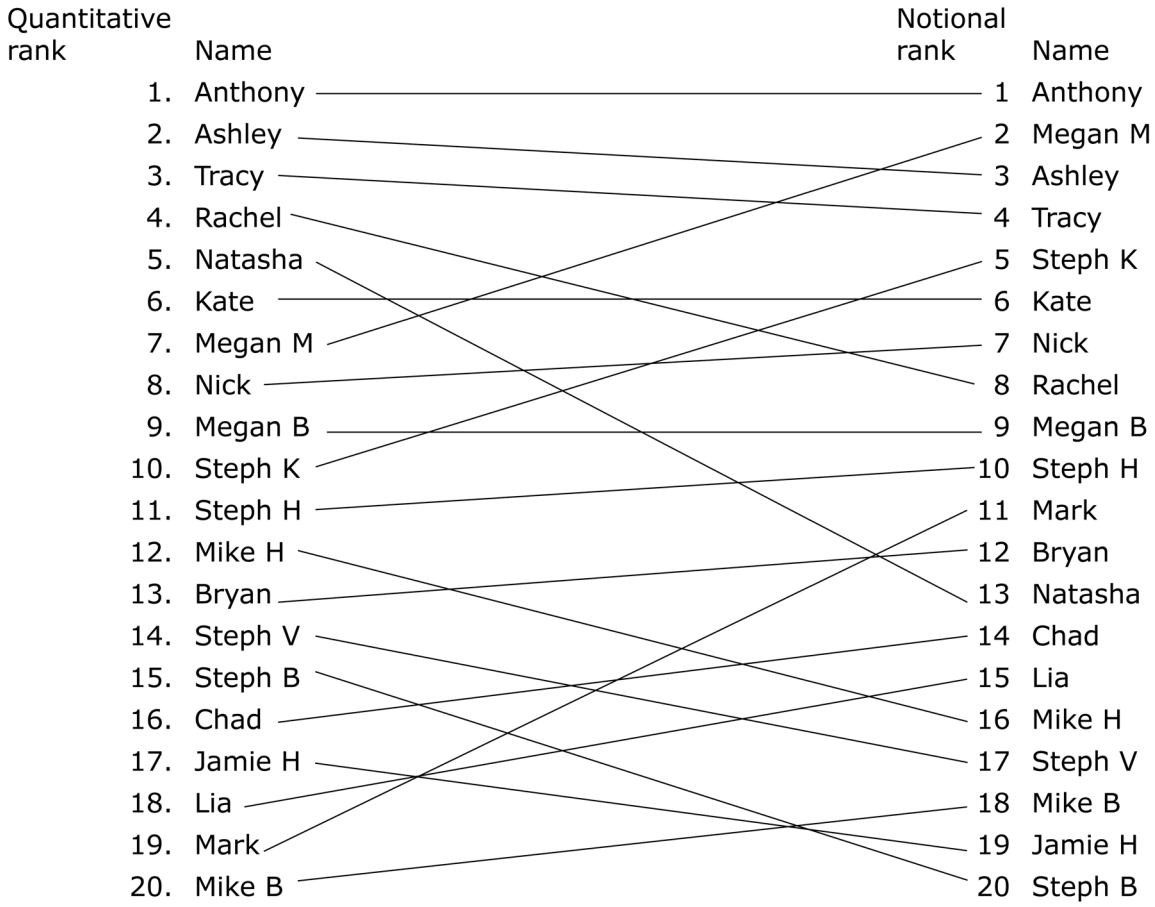
Correlations between the Quantitative and Notional Data

The correlation between the total temporal quantitative measures and the total notional measures was .69, adding the nontemporal measures raised the correlation to .78. Since this is for truly spontaneous speech by native speakers, the combination of temporal and nontemporal quantitative measures performed relatively well, compared to the correlations between quantitative measures and human ratings reported by Cucchiarini and Versant. This is especially true with regard to the speech of advanced speakers, as in the .75 correlation with IRL upper level ratings reported by Versant.

Though the .78 correlation is significant at the .001 level, there is, however, still ample room for improvement. Looking at the rank orders of the speakers as predicted by the quantitative measures, compared to the rank orders of the total scores from the notional study, we can see many crossed lines (see rank orders Figure 1).

Figure 1

Quantitative and Notional Rank Orders



For example, in the quantitative study, #19 Mark, who broke many rules of fluent speech, had a score 14 ranks lower than #5 Natasha. On the notional side, however, #11 Mark was two ranks higher than #13 Natasha. Even though he broke many rules, Mark was judged to be entertaining, and the listeners accordingly gave him a relatively high score. Even though Natasha followed many of the quantitative rules better than Mark, her ideas and vocabulary sounded sometimes childish, and she lost points on the notional side (see the online version of this article for links to speech samples of Mark and Natasha).

The fact that there are many crossed lines in the comparison of the rank orders for the quantitative measures and the notional measures is an indication that the quantitative measures are not adequate for the evaluation of proficiency. Even though the correlations are statistically significant, we would still need better measures if we wanted to make high-stakes decisions regarding hiring, promotion, or program admission.

COMPREHENSIVE ANALYSIS

Background Literature on Comprehensive Analysis

Xi et al. (2006) describe the rubrics used by ETS for human evaluation of oral proficiency.

> *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 content. (p. 12)

Regarding attempts at automatic evaluation of proficiency, Xi et al. (2006) reported that they had some success in extracting durational features related to fluency, some success at identifying diversity and sophistication of vocabulary, and less success at identifying grammatical accuracy. All other categories had lower success as determined by the correlation between machine analyses with human analyses. They noted that it would be premature to put automatic capabilities to high-stakes use before being confident about being able to build an adequate validity model for the features used in the scoring models, and the way those features interact, to provide appropriate evidence about academic speaking proficiency. They also noted that it may not be feasible immediately to implement fully automatic scoring because of the complexity of the problem. Among their comments on the Versant automatic quantitative analyses, they stated that "the tasks used in their assessments do not call for spontaneous speech production and under-represent the domain of speaking proficiency" (p. 4).

Experiment 3: Comprehensive Study

For the comprehensive study, we adapted the categories included in the rubrics developed by ETS (Xi et al., 2006) for a single speaking task in order to analyze the recordings of the 20 students (see comprehensive evaluation sheet in Appendix C). Raters were directed to select a point on the 4-point Likert scale for delivery (fluency, intonation, rhythm, and pronunciation), language use: vocabulary (diversity and sophistication), language use: grammar (range, complexity, and accuracy), and topic development (coherence, idea progression, and content relevance). In addition, they were requested to include written comments explaining the choices they made. Since this was a more demanding task for the evaluators, 30 postgraduate students of linguistics were each given a CD containing the 20 speech files and transcriptions for analysis. These evaluators were not part of the same group that performed the notional evaluations. After full explanation of the instructions were given in class, these students were told to complete and return their signed answer sheets within 2 weeks as a class participation/homework project.

Once the evaluation sheets had been submitted, the results were tabulated and converted to *z* scores (see Table 5). As was the case for the notional study, the total scores for the comprehensive study were calculated using equal weights.

Table 5

Comprehensive *Z* Scores

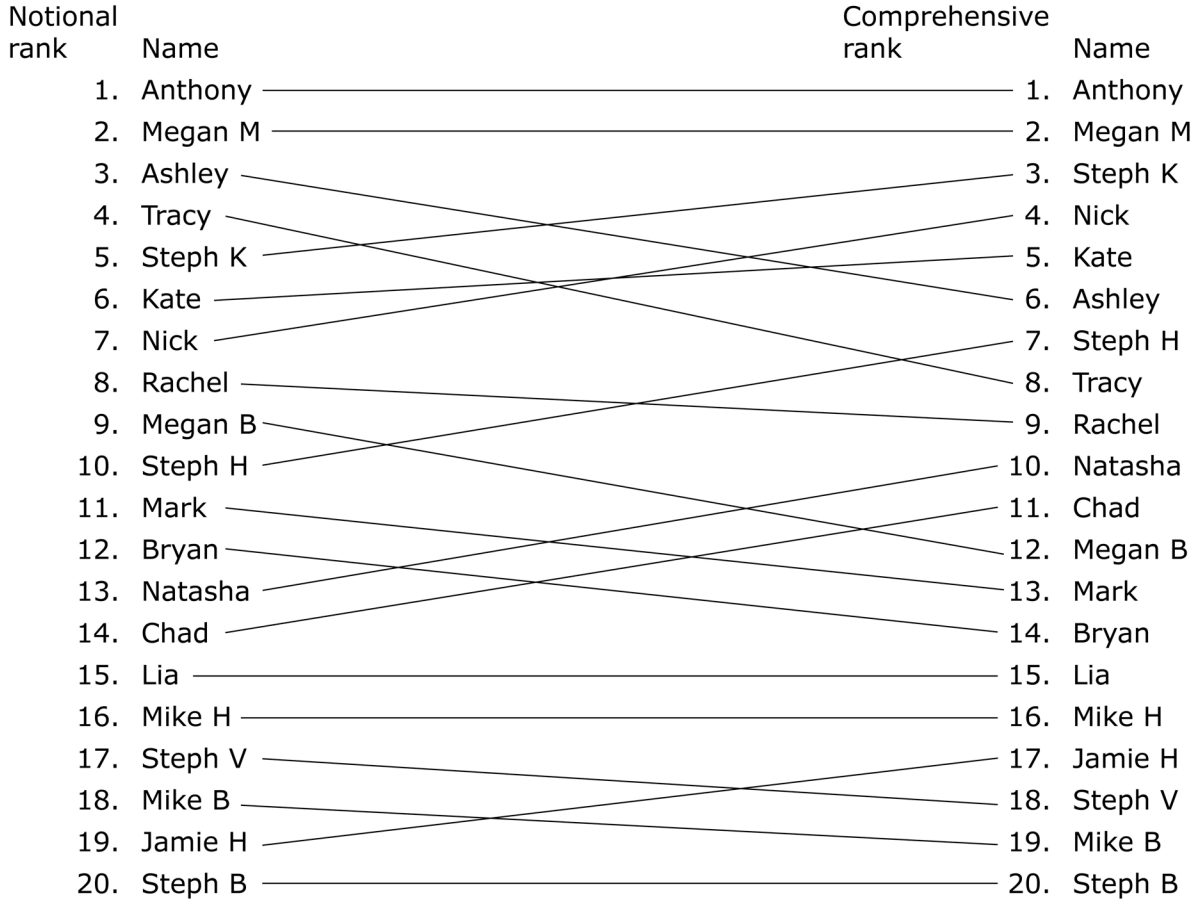
	Delivery				Language use: vocabulary		Language use: grammar			Topic development			
Name	Fluency	Intonation	Rhythm	Pronunciation	Diversity	Sophistication	Range	Complexity	Accuracy	Coherence	Idea Progression	Content Relevance	Total
Anthony	1.72	2.37	1.94	1.81	2.60	2.50	2.57	2.55	2.14	1.94	1.82	1.73	2.34
Ashley	0.78	0.69	0.31	1.13	-0.27	0.56	0.14	-0.27	0.43	0.91	0.38	1.07	0.58
Bryan	-0.28	-0.94	-0.28	-0.25	-0.67	-0.66	-0.64	-0.45	-0.32	-0.19	-0.74	0.07	-0.48
Chad	-0.56	-0.44	-0.33	0.19	-0.20	-0.25	-0.71	-0.86	-0.54	0.63	0.94	0.67	-0.12
Jamie H	-1.39	-1.00	-1.50	-0.50	-0.40	-0.50	-1.00	-0.91	-0.43	-1.00	-0.71	-0.47	-0.91
Kate	0.78	1.37	0.89	0.50	0.33	0.59	0.29	-0.14	-0.07	0.75	0.76	1.07	0.69
Lia	-0.94	-0.19	-0.56	-0.13	-0.93	-0.63	-0.57	-0.91	-0.50	-0.19	-0.82	-0.40	-0.59
Mark	-0.78	-0.44	-0.97	0.19	-0.20	-0.13	0.00	-0.09	0.29	-0.31	-0.88	-1.37	-0.45
Megan B	-0.33	0.16	-0.14	0.19	-1.20	-0.75	-0.18	0.00	-0.43	-0.63	-0.79	-0.60	-0.44
Megan M	1.44	0.88	1.28	1.13	2.13	2.41	1.57	1.64	1.54	1.44	1.41	1.47	1.69
Mike B	-1.56	-1.59	-1.36	-1.25	-0.40	-0.56	-0.93	-0.45	-0.50	-1.50	-1.18	-0.53	-1.11
Mike H	-0.61	-1.81	-0.89	-1.69	-0.53	-0.63	-0.5	-0.45	-1.07	-0.63	0.00	0.27	-0.77
Natasha	0.28	0.38	0.22	0.50	-0.30	-0.91	-0.21	-0.36	0.00	-0.31	0.06	-0.87	-0.09
Nick	0.72	0.44	0.89	-0.31	0.87	0.81	1.00	1.27	0.86	0.69	0.76	0.20	0.72
Rachel	0.44	-0.47	0.11	0.31	0.07	-0.13	0.29	0.18	0.07	0.13	0.12	0.00	0.11
Steph B	-1.69	-0.50	-1.39	-2.34	-1.07	-1.31	-1.50	-1.73	-2.00	-2.06	-1.76	-1.93	-1.73
Steph H	0.78	0.50	0.56	0.13	0.40	0.13	0.79	0.45	0.29	0.56	1.00	0.60	0.58
Steph K	1.06	0.25	1.28	0.63	0.73	0.56	0.93	0.82	1.07	0.88	0.94	0.93	0.92
Steph V	-1.00	0.13	-0.89	-0.75	-1.27	-1.06	-1.14	-1.36	-1.50	-1.44	-1.53	-1.13	-1.16
Tracy	0.67	0.50	1.00	0.50	0.00	0.00	0.07	0.27	0.07	0.19	-0.29	-0.20	0.27

CORRELATIONS AMONG THE THREE STUDIES

The total scores for notional study correlated strongly (r = .95, p < .001) with the total scores of the comprehensive study. The differences in rank order can be attributed to variables beyond the notional study that are included in the comprehensive study (see rank orders in Figure 2).

Figure 2

Notional and Comprehensive Rank Orders

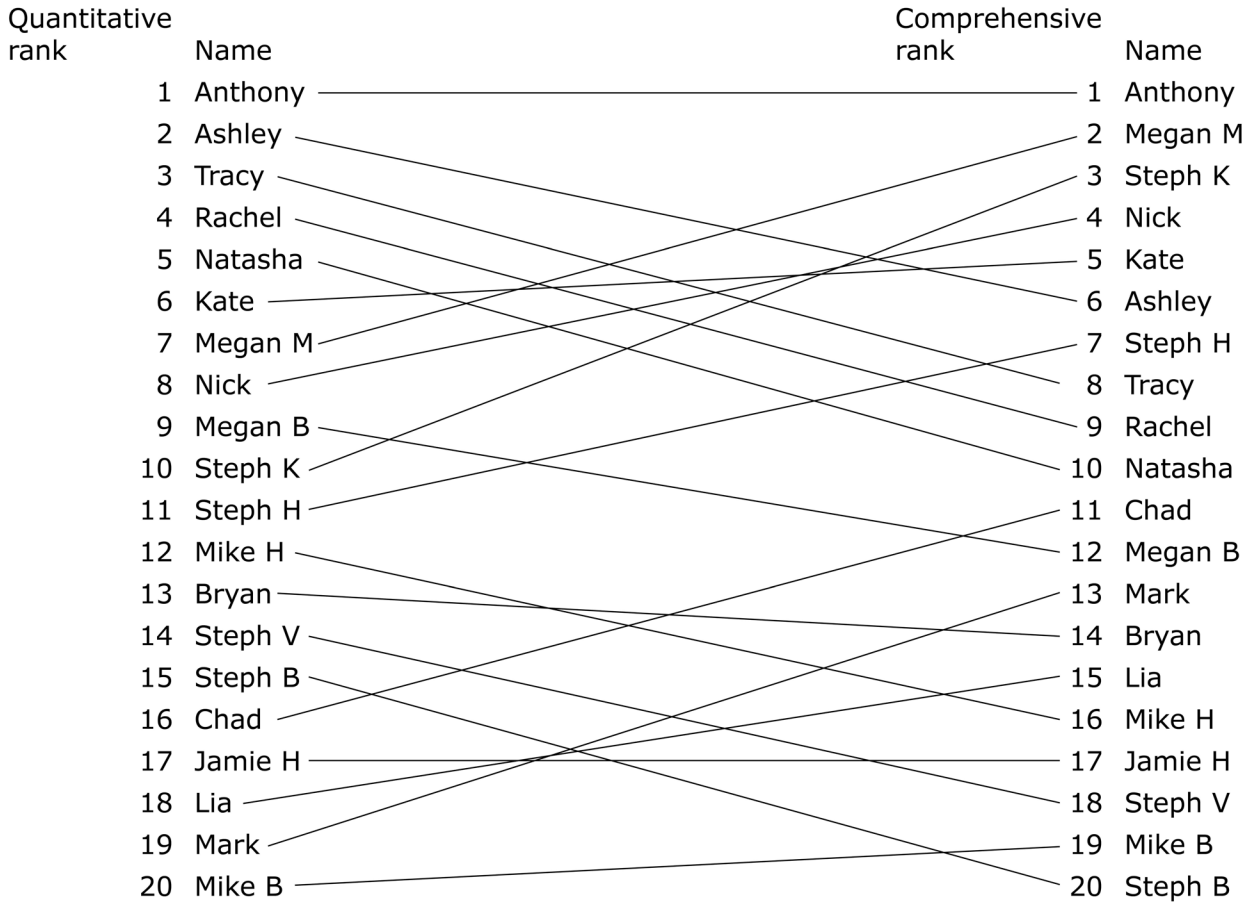


For example, four subjects achieved higher ranks by three levels in the comprehensive study than in the notional study. A review of their *z* scores shows that they all were marked higher in vocabulary, grammar, and/or topic development than in other topics. While Chad's notional *z* scores (clarity, confidence, and fluency) averaged only -0.35, his topic development scores (coherence, idea progression, and content relevance) averaged 0.77. Moving in the opposite direction, while Tracy's notional *z* scores averaged 0.90, her topic development scores averaged only -0.10 (see the online version of this article for links to speech samples of Chad and Tracy). These types of examples provide clear evidence that the comprehensive rubrics are measuring features beyond the scope of the notional analyses.

The quantitative study produced a lower but still significant correlation (r = .72, p < .001) with the total score of the comprehensive study for all the quantitative variables (temporal and nontemporal) but did not correlate significantly (r = .27, p = .246) for the temporal quantitative variables alone (see rank orders in Figure 3).

Figure 3

Quantitative and Comprehensive Rank Orders



By far the most significant ($p < .01$) correlations in this group are between the comprehensive study's delivery section and the quantitative study's temporal phonation/time ratio ($r = .73$), mean length of runs ($r = .64$), and pace ($r = .63$). The total scores for the quantitative nontemporal variables correlated higher with the language use and topic development sections of the comprehensive study variables than did the quantitative temporal total scores. This makes sense because neither vocabulary, grammar, nor topic development are temporal measures. Thus, for all practical purposes, the quantitative temporal variables could be considered to be partially adequate for measuring only some of the aspects measured within the delivery section of the comprehensive study.

Figure 2 shows the relatively clean relationship between the rank orders of the notional study and the comprehensive study; there are not many crossed lines, and the differences in rank have relatively small magnitude. Figure 3 shows the relatively disorderly relationship between the rank orders of the quantitative study and the comprehensive study; there are more crossed lines, and the magnitude of changes in rank is greater. It is interesting to note the different rank orders of Mark and Natasha. While 14 ranks below Natasha (#5) in the quantitative study because Mark (#19) was considered to be entertaining while Natasha was thought to be childish, Mark (#11) was two ranks above Natasha (#13) in the notional study. In the comprehensive study, however, which considered topic development Mark (#13) was three ranks lower than Natasha (#10) because he provided very little content.

Even though all three studies achieved very high degrees of statistical significance, it is clear that the comprehensive study provided better information regarding overall spoken proficiency. This result is because the comprehensive study measures include reference to high entropy tasks which are still beyond the limits of current fully automatic computer systems to adequately analyze. The notional study is next, and the weakest is the quantitative study, especially if restricted to only the temporal measures.

Xi (2008) studied the relationship between the full TOEFL iBT speaking section using the ETS rubrics described above and local ITA (International Teaching Assistant) assessments that were scored on the basis of linguistic qualities and reported a .78 correlation between iBT speaking scores and local ITA screening. This is a relatively high correlation, given the advanced nature of the speaking tasks. When scoring also included teaching skills, the correlation dropped to .70. Regarding the possible uses of TOEFL scores for other purposes, such as screening for employment or granting licenses, Chalboub-Deville and Wigglesworth (2005) noted: "Test users need to undertake local validation research to help make sure that their interpretation and use of test scores are appropriate in their local academic and professional contexts" (p. 385).

DISCUSSION

Correlations within and between the quantitative, notional, and comprehensive studies and comments provided in the literature regarding current limitations of automatic speech recognition, clearly indicate that any recent claims that the technology is already sufficiently developed for making high-stakes decisions are both premature and misleading. While it could be quite convenient to automatically evaluate spoken proficiency (restricting our attention to only those quantitative measures which computers are often able to analyze), in the words of Xi et al. (2006), "the tasks used in their assessments do not call for spontaneous speech production and under-represent the domain of speaking proficiency" (p. 4). Until automatic analysis is more suitable and cost effective for making decisions regarding categories related to vocabulary, grammar, and topic development, a hybrid system relegating some analysis to machines and some analysis to humans is probably still the best approach.

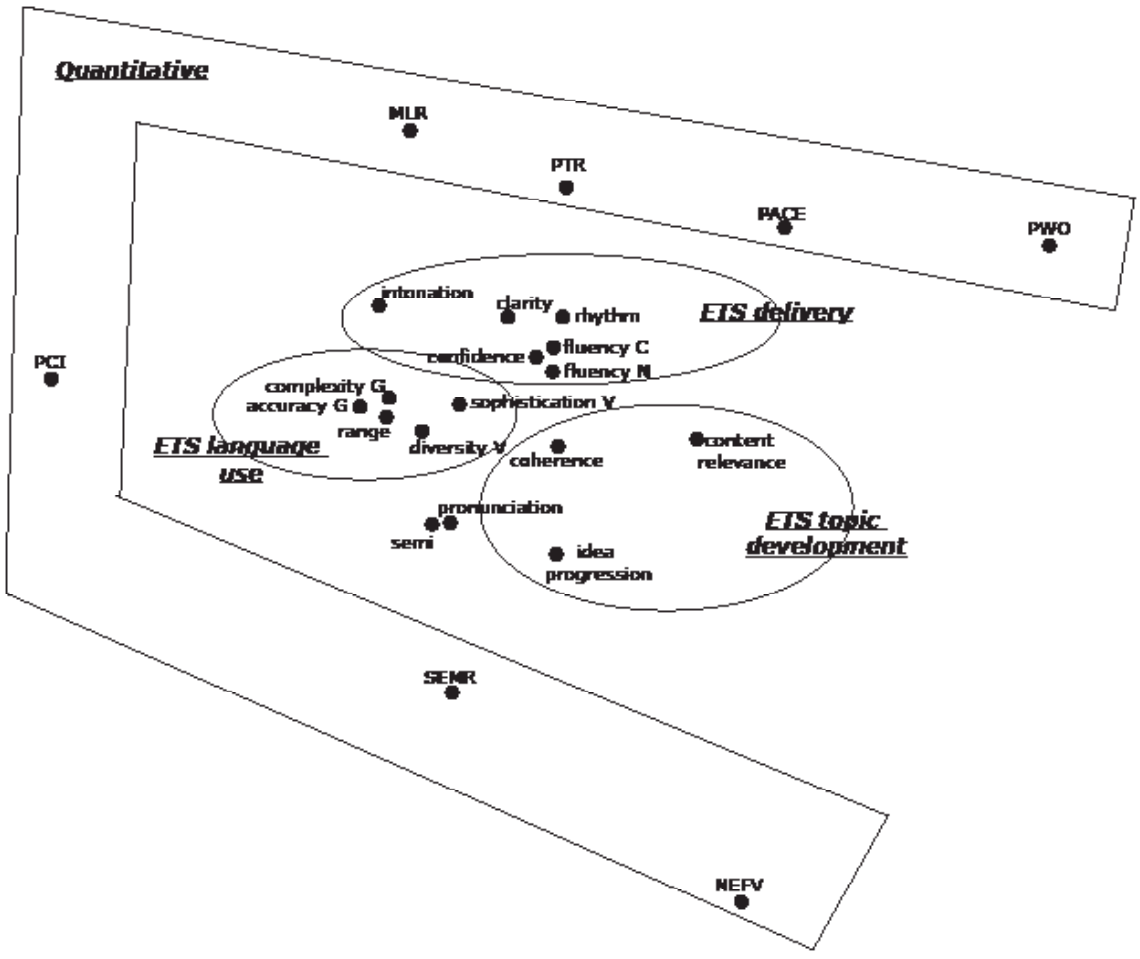
One quantitative area which deserves more attention is the overall variation of intonation (Molholt et al., 2007). By dividing the overall standard deviation of the pitch by the mean pitch of the utterance, we obtain a normalized scale value that allows us to dynamically rate the speech between such points as monotone and sing song, describing the adequacy of the results according to the situation, and to compare the results across different speakers. This type of analysis helps us to bring intonation back into the definitions of fluency and proficiency and certainly would also be useful in studying the speech of nonnative speakers, especially for tone languages.

It should be noted that the notional study, focusing on clarity, confidence, and fluency, has more strength in characterizing spoken proficiency of subjects at an advanced level than the currently proposed automatic quantitative methods. It should also be noted that despite the current difficulties involved in creating a fully automatic system for evaluation of proficiency, there are still many practical methods for utilizing speech analysis in language teaching, especially with respect to using real-time visualizations of speech patterns to communicate with second language learners, (Molholt, 1988; 1998; Molholt & Hwu, 2008).

GUTTMAN'S SMALLEST SPACE ANALYSIS

Guttman's (1982) smallest space analysis (SSA) is a multivariate procedure for understanding the structural aspects of a group of variables. It is useful for discerning how sets of variables are organized to the extent to which they measure similar facets or constructs. The SSA attempts to provide the most parsimonious configuration of relationships among a group of variables represented in Euclidean space. The weighted SSA1 (WSSA1) procedure for symmetrical correctional matrices was used (Amar, 2005) to examine the structure of the eight quantitative, the three notional, and the twelve comprehensive variables in this study. The WSSA1 provides an index of best fit (coefficient of alienation) among the observed relationships between variables and their geometric representation in the Euclidean space. Investigators commonly compute two or three dimensional solutions. The choice of the two or three dimensional solution depends on the interpretability of the solution and the coefficient of alienation which provides a measure of how closely the geometric representation approximates the observed relationships among the variables. The coefficient of alienation can vary between 0 (perfect fit) and 1 (perfect absence of fit). Coefficients of .15 or lower are preferred, although some suggest an upper limit of .20 (Amar, 2005; Donald & Cantor, 1990; Elizur & Sagie, 1999; Schlesinger & Guttman, 1969). Figure 4 shows the three dimensional solution with a coefficient of alienation = .0001, suggesting an excellent fit for the 23 variables included in the correlation matrix.

Figure 4
Smallest Space Analysis



The SSA representation can be easily divided into four major and separate sections, corresponding precisely to (a) the quantitative variables with quantitative temporal variables across the top in the quantitative area and (b) the ETS comprehensive rubrics for topic development, language use, and delivery. As a result, it provides additional evidence that the quantitative temporal variables are perceived to be distinct from the rest and, therefore, cannot be assumed to be inclusive of the rest of the variables under consideration. Within this display, moreover, we can see that the notional variables (clarity, confidence, and fluency) are well inside the area covered by the delivery variables discussed in the comprehensive study.

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

There are several advantages and disadvantages to the three types of analysis under discussion here. Although automatic quantitative analysis is quite convenient, it leaves out too many important variables to be considered adequate for measuring proficiency, especially at advanced levels. While notional analysis is simple and can achieve relatively high correlations with comprehensive analysis, it does not provide the specific details regarding proficiency that an evaluator or a learner would need. Even though the comprehensive approach is time consuming because it requires human evaluations of certain categories, it provides sufficient information that could help committees make decisions, and it does provide specific scores to help examinees understand their level of proficiency. Therefore, for high entropy, high-stakes decisions, the comprehensive analysis appears to be currently the best choice.

This study was a small pilot study that investigated the spontaneous engaged speech of native speakers of English. As such, it only represents one task related to the TOEFL iBT Speaking Test. In order to provide a more secure foundation for defining target ranges for various parameters of English, we need a larger scale baseline study. Once this is accomplished, then comparative studies focusing on other languages and examining nonnative speakers of English and of other languages could be conducted.

There are several measures which could lend themselves to automatic computer analysis that are not yet being fully utilized. For example, while we can see in the smallest space analysis in Figure 4 above that the overall measure of the variation of intonation (SEMI) is in the mainstream of the variables, this kind of variable has not been included in the literature regarding automatic evaluation of fluency. Another overall variation variable that should be considered is the overall variation of amplitude (dB) level. Within a speech file, a high degree of overall dB variation is an indication of choppiness, of long pauses, or a combination of the two. Overall variation of pitch and dB levels are quite useful for differentiating speakers, and both of these could become quite significant for measures of nonnative speech. These variables may also have applications for assessing the speech of the hearing impaired.

A further study could be conducted on the differences in perception between professional and nonprofessional raters of spontaneous engaged speech. In which areas do we find common ground? Are there significant differences? If so, where? What might this mean regarding the composition of committees charged with hiring, promoting, or admitting nonnative speakers? Answers to these questions could provide additional practical insight for those responsible for making high-stakes decisions.

REFERENCES

Amar, R. (2005). *HUDAP manual.* Jerusalem: The Hebrew University of Jerusalem.

Audhkhasi, K. (2009). Automatic evaluation of fluency in spoken language. *IETE Technical Review*, *26*, 108-114.

Aylett, M. (2003). Disfluency and speech recognition profile factors. In R. Elkund (Ed.), *Proceedings of the Disfluency in Spontaneous Speech Workshop* (pp. 49-52). Goteborg, Sweden: Gothenburg Papers in Theoretical Linguistics.

Balogh, J. (2008). *A case for automation in aviation English language assessment*. Upper Saddle River, NJ: Pearson Education, Inc. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/121BCB12-7231-41C6-B051-D3BF98624D7A/0/WhitePaper_CaseForAutomation.pdf

Blake, R., Wilson, N., Cetto, M., & Pardo-Ballester, C. (2008). Measuring oral proficiency in distance, face-to-face, and blended classrooms. *Language Learning & Technology*, *12*(3), 114-127. Retrieved from http://llt.msu.edu/vol12num3/blakeetal.pdf

Branigan, H., Lickley, R., & McKelvie, D. (1999). Non-linguistic influences on rates of disfluency in spontaneous speech. In C. Fossier-Lussier, S. Greenberg, & M. Slaney (Eds.), *Proceedings of the Fourteenth International Congress of Phonetic Sciences San Francisco, California: ICPhS 1999* (pp. 387-390). Retrieved from http://www.ltg.ed.ac.uk/papers/99disfluency2.pdf

Brown, J. D. (2003). Promoting fluency in the EFL classrooms. In S. Yamashita, A. Howard, & C. Rinnert (Eds.), *Proceedings of the 2nd Annual JALT Pan-SIG Conference* (pp. 1-12). Kyoto, Japan: Kyoto institute of Technology.

Chalboub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes*, *24*, 383-391.

Cucchiarini, C., & Strik, H. (1999). Automatic assessment of second language learners' fluency. In C. Fossier-Lussier, S. Greenberg, & M. Slaney (Eds.), *Proceedings of the Fourteenth International Congress of Phonetic Sciences San Francisco, California: ICPhS 1999* (pp. 759-762).

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America, 111*, 2862-2873.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America 117*, 989-999.

Deshmukh, O., Kandhway, K., & Verma, A. (2009). Automatic evaluation of spoken English fluency. In Y. Ma (Ed.), *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (pp. 1-4). Taipei, Taiwan: ICASSP09.

Donald, L., & Cantor, D. (1990). Temporal and trait facets of personnel assessment. *Applied Psychology: An International Review, 39*, 413-429.

Elizur, D., & Sagie, A. (1999). Facets of personal values: A structural analysis of life and work values. *Applied Psychology: An International Review, 48*, 73-87.

Franco, H., Neumeyer, L., Digalakis, V., & Ronen, O. (1998). Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication, 30*, 121-130.

Guttman, L. (1982). Facet theory, smallest space analysis, and factor analysis. *Perceptual Motor Skills*, *54*, 491-493.

Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal, 36*, 153-166.

Holland, M., & Fisher, F. (Eds.). (2008). *The path of speech technologies in computer assisted language learning: From research toward practice.* London: Routledge.

KayPentax, (2009). MultiSpeech Software. Lincoln Park, NJ: Author.

Kondo, Y ., Tsutsui, E., & Nakano, M. (2008). Fundamental research on automatic speech evaluation of L2 English. In T. Koyama (Ed.), *Proceedings of the 13th Conference of the Pan-Pacific Association of Applied Linguistics* (pp. 85-86). Tokyo, Japan: Waseda University.

Molholt, G. (1988). Computer assisted instruction in pronunciation for Chinese speakers of American English. *TESOL Quarterly*, *22*, 91-111.

Molholt, G. (1998). *Accent reduction via acoustic analysis.* Lincoln park, NJ: Kay Elemetrics Corporation.

Molholt, G., Cabrera, M., Kumar, K., & Thompsen, P. (2009, March). *Correlating quantitative measures of speech with perceptions of fluency*. Paper presented at the annual conference of the Computer Assisted Language Instruction Consortium (CALICO), Arizona State University, Tempe, AZ.

Molholt, G., & Hwu, F. (2008). Visualization of speech patterns for language learning. In M. Holland & F. Fisher (Eds.), *The path of speech technologies in computer assisted language learning: From research toward practice* (pp. 91-122). London: Routledge.

Molholt, G., Morgan, J., & Park, J. (2007, May). I*mplementation of automatic fluency checking using freely available toolkits*. Paper presented at the annual conference of the Computer Assisted Language Instruction Consortium (CALICO), Texas State University, San Marcos, TX.

Molholt, G., & Pressler, A. (1986). Correlation between human and machine ratings of test of spoken English reading passages. In C. Stansfield (Ed.), *Technology and language testing* (a collection of papers from the Seventh Annual Language Testing Research Colloquium, held at ETS, Princeton, NJ, April 6-9, 1985) (pp. 111-127). Washington, DC: TESOL.

Neri, A., Cucchiarini, C., & Strik, H. (2002). Feedback in computer assisted pronunciation training: Technology push or demand pull. In E. Voorhees (Ed.), *Proceedings of the International Conference on Spoken Language Processing* (pp. 1209-1212). Denver, Colorado: Interspeech 2002.

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning*, *15*, 441-467.

Richards, J. C., Platt, J., & Weber, H. (1985). *Longman dictionary of applied linguistics.* London: Longman.

Schlesinger, J., & Guttman, L. (1969). Smallest space analysis of intelligence and achievement tests. *Psychological Bulletin*, *71*, 95-100.

Stansfield, C. (Ed.). (1986). *Technology and language testing* (a collection of papers from the Seventh Annual Language Testing Research Colloquium, held at ETS, Princeton, NJ, April 6-9, 1985). Washington, DC: TESOL.

US Patent No. 5634086. (1997). *Method and apparatus for voice-interactive language instruction*. Retrieved from http://www.patentstorm.us/patents/5634086/fulltext.html

VERSANT. (2006). Fully automated spoken English test. Retrieved from http://www.versant.jp/e_seminar.htm

VERSANT. (2008a). Delivery. Retrieved from http://www.ordinate.com/technology/delivery.jsp

VERSANT. (2008b). Scoring. Retrieved from http://www.ordinate.com/technology/scoring.jsp

VERSANT. (2008c). Validation. Retrieved from http://www.ordinate.com/technology/validation.jsp

Xi, X., Zechner, K., & Bejar, I. (2006). Extracting meaningful speech features to support diagnostic feedback: An ECD approach to automated scoring. M. Kolen (Ed.), *Proceedings of the NCME Annual Meeting.* San Francisco, CA: NCME. Retrieved from http://www.ets.org/Media/Research/pdf/speechrater_5.pdf

Xi, X. (2008). *Investigating the criterion-related validity of the TOEFL speaking scores for ITA screening and setting standards for ITAs*. (TOEFL iBT Research Report No. TOEFLiBT-03). Princeton, NJ: Educational Testing Service.

Zechner, K., Higgins, D., & Xi, X. (2007, May). *SpeechRater: A construct-driven approach to scoring spontaneous non-native speech*. Paper presented at the Annual Meeting of the International Speech Communication Association. Retrieved from http://www.ets.org/Media/Research/pdf/speechrater_2.pdf

APPENDIX A

Variable Definitions

A. Quantitative variables (measured by computer analysis)

Temporal
PACE—phonemes per second including gaps
PWO—articulation rate (phonemes per second without gaps)
PTR—phonation/time ratio (duration of actual speech sounds divided by total duration)
MLR—mean length of runs (average duration of stretches of speech with no gaps equal to or greater than 200 ms)

Nontemporal
SEMI—overall variation of pitch from monotone to sing song
SEMR—pitch range from highest to lowest
PCI—phoneme clarity index (number of correctly pronounced phonemes divided by total number of phonemes represented by the words used)
NEFV—non-empty filler vocabulary (number of words which are not fillers divided by the total number of words including fillers)

Total quantitative score = sum of the z scores of the eight quantitative variables using equal weights

APPENDIX B

Notional Study Evaluation Sheet

I give Dr. Molholt permission to use my responses in his research projects.

Name: _____ Date _____

You will hear each file twice. First all files will be played one after the other. Then each file will be played separately, with time for you to mark your answer sheets.

Clarity—Understandable to you
Confidence—Speaker is sure of what is being said
Fluency—Reasonable flow and expression

SPEAKER 1.

General Perception:

CLARITY	Very High	High	Medium	Low	Very Low
CONFIDENCE	Very High	High	Medium	Low	Very Low
FLUENCY	Very High	High	Medium	Low	Very Low

Specific comments:

SPEAKER 2.

General Perception:

CLARITY	Very High	High	Medium	Low	Very Low
CONFIDENCE	Very High	High	Medium	Low	Very Low
FLUENCY	Very High	High	Medium	Low	Very Low

Specific comments:

APPENDIX C

Comprehensive Evaluation Sheet

Name of evaluator: _____ Date: _____

SPEAKING RUBRICS NAME OF SPEAKER _____

DELIVERY

Fluency	Intonation	Rhythm	Pronunciation	
4	4	4	4	
3	3	3	3	COMMENTS?
2	2	2	2	
1	1	1	1	

LANGUAGE USE: VOCABULARY

Diversity	Sophistication	
4	4	
3	3	COMMENTS?
2	2	
1	1	

LANGUAGE USE: GRAMMAR

Range	Complexity	Accuracy	
4	4	4	
3	3	3	COMMENTS?
2	2	2	
1	1	1	

TOPIC DEVELOPMENT

Coherence	Idea progression	Content relevance	
4	4	4	
3	3	3	COMMENTS?
2	2	2	
1	1	1	

ACKNOWLEDGEMENTS

The authors thank all the students of composition and linguistics at West Chester University who participated in these studies. In addition, we sincerely appreciate all of the helpful comments and suggestions of the reviewers.

AUTHORS' BIODATA

Garry Molholt is Professor of linguistics in the Department of English at West Chester University.

María José Cabrera is Assistant Professor of Spanish linguistics in the Department of Languages and Cultures at West Chester University.

Krishna Kumar is Professor of psychology at West Chester University.

Philip Thompsen is Associate Professor in the Communications Studies Department of West Chester University.

AUTHORS' ADDRESSES

Garry Molholt
English Department
517 Main Hall
West Chester University
West Chester, PA 19383
Phone: 610 436 2469
Fax: 610 436 3150
Email: gmolholt@wcupa.edu

María José Cabrera
Languages and Cultures
117 Main Hall
West Chester University
West Chester, PA 19383
Phone: 610 436 2700
Fax: 610 436 3150
Email: mcabrera@wcupa.edu

V. K. Kumar
Psychology Department
062 Peoples Building
West Chester University
West Chester, PA 19383
Phone: 610 436 2348
Fax: 610 436 3150
Email: vkumar@wcupa.edu

Philip Thompsen
Communication Studies Department
528 Main Hall
West Chester University
West Chester, PA 19383
Phone: 610 436 2283
Fax: 610 436 3150
Email: pthompsen@wcupa.edu