Lund University, Dept. of Linguistics
Working Papers 49 (2001), 58–61

Using speech recognition to evaluate skills in spoken English

Rebecca Hincks
Dept. of Speech, Music and Hearing, KTH
hincks@speech.kth.se

Abstract

This paper analyzes some of the results of the use of PhonePass, a telephone-based test of spoken English that uses automatic speech recognition. It finds that the test provides sensitive measures of speech rate and phonetic accuracy.

1 Introduction

By the middle of the 1990s, automatic speech recognition had developed to the point where it could be used in language learning applications. Students began to be able to 'talk' with their computers. The American company Syracuse and the French company Auralog both began to market instructional software using speech recognition. Research institutions are now working on finding ways to make software give learners the appropriate corrective feedback.

Another use for speech recognition is in the evaluation of spoken language, as a sort of substitute for human-intensive oral proficiency tests. In California, Ordinate Corporation led by Jared Bernstein uses speech technology to automatically evaluate spoken English by means of its 10-minute PhonePass test administered by computer over the telephone. In the fall of 2000, the PhonePass test was used at KTH to assess student progress after an intensive course in English. This paper examines some of the results of the PhonePass tests and looks at whether this fully automatic test of spoken English achieves its goals.

The Unit for Language and Communication at KTH teaches courses in five foreign languages to undergraduate students of engineering, and courses in Swedish to exchange students. In recent years, the Unit has also been able to offer intensive language training to unemployed immigrant engineers at the commission of the Stockholm County Labor Bureau. In addition to the 200-hour course, each immigrant engineer receives 5 to 10 hours of individual computer- and teacher-assisted pronunciation training. Students of Swedish have been taught with the assistance of IBM's *Speechviewer*, which is particularly helpful in giving visual feedback on prosody (Öster 1999).

2 Method

The Ordinate PhonePass test was used as a measure of student progress in spoken English. Fifteen students took the test at the beginning and end of the ten-week course. A telephone was placed in a soundproof room and both the incoming and outgoing signal were digitally recorded. In addition to the dialogue with the computer, each student was recorded reading a short text containing all English phonemes (Rivers & Temperley 1978). The recordings resulted in a database that can be used to compare each student's production with his or her score given by the Ordinate PhonePass test.

2.1 The PhonePass Test

The PhonePass test uses automatic speech recognition to assess facility in spoken English. It is designed as a simple way for organizations to test the English skills of potential employees or students. To administer the test, an organization purchases test papers from Ordinate Corp. Each test paper is unique, though the items are recombined to make other tests. To take the test, the examinee calls a phone number in California and is connected with a computer. The examinee enters his test paper number on the telephone keypad and then follows instructions. The test results are soon available on the company website.

The test consists of five parts. This paper concerns results derived from Part A, Reading, where the examinee is instructed to read a set of sentences. The recognition engine can assess how the examinee's pronunciation of each word in the sentence compares with acceptable pronunciations, and measure the rate at which the examinee reads.

The speech processing used by Ordinate Corporation was trained with the speech of native speakers and adapted for use by non-native speakers. It uses forced alignment to locate the relevant parts of the speech signal, an HMM-based speech recognizer, a pronunciation dictionary, and an expected-response network constructed from responses collected in over 4000 administrations of the test (Bernstein 1999).

2.1.1 Scoring

The PhonePass test result consists of five sub-scores on a scale from 2 – 8, which are combined to produce an overall score. The weighting of the sub-scores is as follows:

Listening Vocabulary, 30%; Repeat Accuracy, 30%; Pronunciation, 20%; Reading Fluency, 15%; and Repeat Fluency, 5%.

In developing the test, a non-native norming group of 514 speakers of 40 foreign languages was formed. The mean overall score for this group was 5.2, with a standard deviation of 1.15. The standard error of the overall score is 0.2.

2.1.2 Validation

Repeated comparison of the results given by the PhonePass test and those obtained by human-rated measures of oral proficiency show that there is as much correlation between PhonePass scores and averaged human ratings as there is between one human rater and another (Bernstein 1999). The PhonePass test has been validated in relation to the spoken English tests given by a number of language institutes. The average correlation between PhonePass and these tests is 0.71, while the tests themselves had an average inter-rater correlation of .72.

3 Results

Before looking at how individual production is reflected in test results, it is perhaps of interest to know whether the fifteen immigrant engineers improved as a result of the 200-hour course. Unfortunately, they did not, on average, show dramatic improvement in their spoken English as measured by the PhonePass Test. The average Overall score after the first round of tests in October was 4.48; the average Overall score after the second round in December was just .27 points higher at 4.75.

3.1 Reading Fluency

In Part A, Reading, the examinee is asked to read aloud eight of the twelve sentences on the test paper. Data gathered from this part of the test determine the scores for two subscores: Reading Fluency and Pronunciation (Townshend 1998). Students' before and after results on the Reading Fluency subscore are shown in Figure 1.

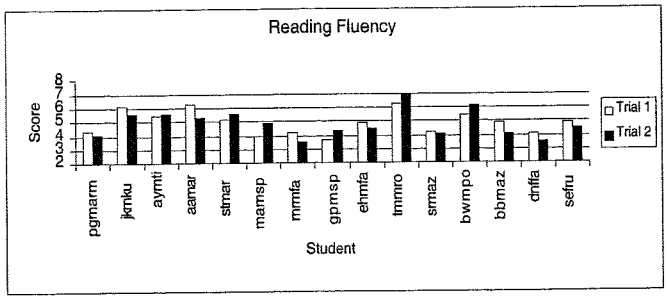

Figure 1. Reading Fluency, before and after, by student.

What do these scores mean on the individual level? Ordinate does not release information as to what specific parameters affect a given subscore. Certain assumptions can however be made; for example, that the speed at which one reads is reflected to some extent in the Reading Fluency subscore. An analysis of the length of utterances was performed to assess the effect of reading speed on score result. Since the reading part of the test consists of sentences from a limited bank of material, many examinees read a number of the same sentences, whose length can be compared.

The lowest score shown in Figure 1 was Mrmfa's second trial. The highest result was Tmmro's second trial. Coincidentally, these two tests contained three sentences in common. Examinee Tmmro, with a score of 6.9, read these sentences twice as quickly as Mrmfa, who had a score of 3.5. Table 1 shows the sentences and the speed at which they were read.

Table 1. Reading speed for selected utterances, best and worst scoring students.

Sentence	Tmmro2 (6.9)	Mrmfa2 (3.5)
"It's really expensive, but his friends eat there a lot."	4.06 seconds	7.92 seconds
"He gives them a pretty big discount."	2.05 seconds	3.89 seconds
"And they, in turn, always leave him a generous tip."	3.66 seconds	7.42 seconds
Total	9.77 seconds	19.23 seconds

3.2 Pronunciation

Part A of the PhonePass test also provides data for the Pronunciation subscore. Students' before and after results on the pronunciation subscore are shown in Figure 2.

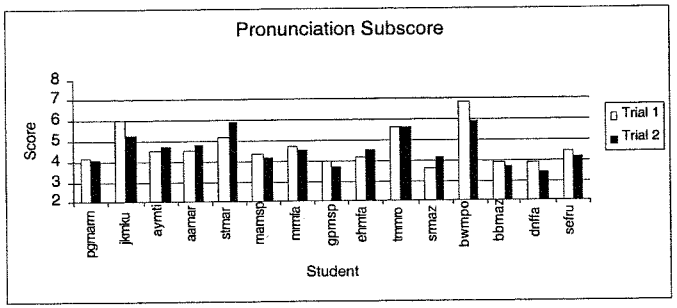

Figure 2. Pronunciation subscore, before and after, by student.

Ordinate claims that the standard error of measurement is .4 for the PhonePass test (Bernstein, personal communication). According to this measure, only two students have achieved significant improvement in their pronunciation, and three have achieved significantly worse scores. Does this reflect their production? The high scorer, Bwmpo, received a full point lower on the second trial. An analysis of the phonetic errors he made in both recordings was performed to see whether his pronunciation had in fact degraded during the course.

A native speaker of Polish, Bwmpo's main problem was with devoicing the consonants /z/, /d/ and /v/ in final position. The sentences he read in the first trial contained 11 words with one of these final phonemes, and he mispronounced 9 of these (Table 2). In the second trial, there were 15 words containing these phonemes, and he made errors in 13 of them. Another basic problem he has is with initial /h/, which he sometimes realizes as [x]. While there were 5 places where this phoneme appeared in the first trial, there were twice as many in the second trial due to the nature of the sentences. He mispronounces 2 out of 5 in the first trial, and 7 out of 10 in the second trial.

Bwmpo's Pronunciation score can have been pulled down by the fact that the second set of sentences contained more words that he had particular problems with. The positive effect of the more natural prosody present in the second trial is not reflected in this score, though it can be seen in Figure 1 that his Reading Fluency score increased by more than half a point.

Table 2. Bwmpo errors

Error	First trial (score 6.9)	Second trial (score 5.9)
devoicing final /z/, /d/, /v/	9 out of 11 possible	13 out of 15 possible
/h/>[x]	2 out of 5 possible	7 out of 10 possible
Total (for these phonemes)	11 out of 16 possible	20 out of 25 possible

4 Discussion

This study is part of a year-long project to examine the effect of alternative means of computer-assisted pronunciation instruction. Students in the spring of 2001 are being taught with software that can be run on students' home computers, using a dialogue system and providing feedback on the sentence, word, and phoneme level. Pre- and post- trials using the PhonePass test will determine whether this is beneficial to the students' spoken language development. This paper has shown how one part of the PhonePass test uses speech recognition to provide a measure of rate of speech and phonetic accuracy, and also reveals some potential problems with the test, such as the absence of a way to measure the effects of prosody and the randomness of the presence of particular phonetic pitfalls.

References

Bernstein, Jared. 1999. 'Validation Summary for PhonePass SET-10'. <www.ordinate.com/technology.jsp?reports>.

Öster, Anne-Marie. 1999. 'Strategies and results from spoken L2 teaching with audio-visual feedback.' Stockholm: STL-QPSR 1-2/99.

Rivers, Wilga & Temperley, Mary. 1978. *A Practical Guide to the Teaching of English.* New York: Oxford University Press.

Townshend, Brent; Bernstein, Jared; Todic, Ognjen & Warren, Eryk. 1998. 'Estimation of spoken language proficiency.' In *Proceedings of ESCA Workshop on Speech Technology in Language Learning*. Stockholm: KTH Department of Speech, Music and Hearing.