Computerized English Pronunciation Testing

Lim, Chang-Keun, Kang, Seung-Man*

(Keuk-Dong Information College,

* Korea National University of Education)

smkang@www.seowon.ac.kr

Abstract

The past decade has witnessed the abundant use of computer in testing language skills such as listening and reading. Compared with these language skills, we have experienced little use of computer in testing a speaking skill including pronunciation. This is largely due to limitations of the current computer technology. One of such limitations for testing pronunciation is to store and automatically evaluate what the learner utters. Due to this limitation, the computer simply stores what the learner utters and raters evaluate it afterward on a certain rating continuum. With the advent of voice recognition technology, however, the computer has been able to test pronunciation in a systematic way. This technology enables the computer to identify, visually show, and evaluate the learner's intonation pattern by means of autocorrection. The evaluation is expressed in terms of the degree in which the learner's intonation pattern overlaps with that of the native speaker of the target language. In particular, the degree is numerically displayed on the screen, and this numeral is considered as the score of the learner's utterance under our testing framework.

1. Introduction

As computer technology has highly developed since the 1970s, there have been a host of attempts to test language ability via computer. In the early stages of these

attempts, computer simply played a receptive role of grading and analyzing a traditional pencil-and-paper test, which was not so satisfactory for language testing. Late in the 1980s, however, a new attempt, called CLT (Computerized Language Testing), was made to use computer for language testing; that is, students directly respond to computer with regard to test items presented on the computer screen. This method was popularized widely among educators around the middle of the 1990s.

CLT was previously implemented by some representative CAT (Computer-Assisted Testing) programs such as PETA (Pitt Educational Testing Aids) and AIMS (Academic Instructional Measurement System), which were developed by Nikto & Hus (1986) and Psychological Corporation (1989) respectively. PETA was originally a computerized testing program used for Apple Basic. It was capable of testing approximately 500 students at a time and processing their test results in a straightforward way. PETA was popular among test administrators because its working procedures are quite simple and it enables the administrators to produce a variety of test items. However, it finds its limitation in producing highly complicated test items with a chart as well as a graphic and animation. Moreover, PETA employs a conventional method of analyzing test items.

AIMS, on the other hand, was able to produce test items including a picture and a chart with the help of Macintosh HyperCard. In addition, it had a great advantage of analyzing test items by IRT Rasch Model. But AIMS likewise could not produce test items with motion pictures, which are multimedia in nature.

Early in the 1990s, CAT enters into a new phase of testing, which is called a "computer-adaptive test." The test subsumes the notion of IRT (Item Response Theory) and utilizes the basic properties of computer such as computation and control. This in turn makes it possible to provide students with only the items within their ability, which is adaptive in nature. It follows that testing in this practice can be rather

individualized, resulting in an individualized adaptive test via computer. The test in question has been developed from an early tailored test such as Binet's intelligence test, which failed to spread widely because it basically rests on CTT (Classical Testing Theory) for item analysis and employs raw scores, as pointed out by Weiss (1985). The tailored test, however, gains its popularity in the 1980s as theoretical foundations of IRT are gradually established, and personal computers are popularized among people. Consequently, it develops into the form of CAT and gives rise to the birth of CALT (Computer-Adaptive Language Testing) in the field of language education.

It is true that CALT is generally used for language skills such as listening and reading, but it is rarely used for speaking due to some technological limitations. In this paper, we will briefly discuss such limitations especially with respect to testing pronunciation and then suggest some possible ways to apply the current computer technology to testing pronunciation.

2. Properties of CALT as a Testing Tool

CALT is presently considered to be one of new testing methods and thereby draws a lot of attention from researchers. It is credited especially with its ability to overcome some limitations of the classical testing theory mentioned above. In particular, ETS (Educational Testing Service), which has conducted a large scale of research and project on the use of CALT, comes to the conclusion that CALT proves to be of great use in language testing.

CALT finds its use in the following three aspects. First, in the affective aspect it helps to lessen learners' burden of testing by making them reply to the test items within their reach (Madsen 1991; Dunkel 1997; Brown 1997). As you might recall, in the traditional pencil-and-paper test the same test items are uniformly provided to all

students; thus, the students should answer all the items given with no exception. In CALT, however, a certain series of items are given to students first and a subsequent series of items follow on the basis of the results of the items previously given. If the scores of the first series of items are low, for example, the difficulty level of the subsequent series of items is accordingly adjusted. In other words, the presentation of items is highly optimized to the extent that students are given only the items compatible to their capacity. This function of CALT lends itself to lowering students' affective filter, virtually providing them with a high level of motivation for testing.

A second advantage of CALT lies in the effectiveness of test and the accuracy of measurement. It is widely noted that CALT helps to obtain a much higher level of validity and reliability even with a small number of items (Henning 1991; Larson 1996; Moon 1997). It is generally estimated that half or one-third of the items for the classical test are needed in CALT to take the same effect on establishing validity and reliability. Another great advantage of CALT in measurement is that testing can be readily individualized. In general, it is extremely difficult to obtain correct evaluation results of each individual with a certain number of items uniformly given to the whole group of students. In CALT, however, items are adapted and tailored to each individual according to his/her ability, resulting in individualized testing. This in turn enables us to correctly find out each individual's shortcomings in learning and thus provide him/her with immediate feedback on them.

Shohamy (1993) suggests another advantage of CALT in terms of variety with respect to items. He illustrates that we can prepare a variety of items with the help of computer technology, which are virtually characterized as multimedia in nature. These items are treated as authentic; they accordingly lead to authentic and integrative testing.

Finally, CALT has an advantage in administrative aspects. The administration of

CALT is not limited in time and space. As mentioned earlier, CALT is fully individualized, and thus each individual can take a test at his/her convenient time and space. This practice is unimaginable in the classical test environment, in which the whole group of students are uniformly given a certain set of items and asked to complete them within the limited time. A further advantage of CALT is that testing can be administrated on the Internet, completely eliminating the limit of time and space for testing. Testing on the Internet also makes it possible to provide students with real-time advice online.

3. Computer and Pronunciation Test

Just as the teaching of pronunciation is a little-studied area of research, so does pronunciation testing suffer from a lack of serious attention. This is due to the fact that the pronunciation test suffers from low reliability and thereby low feasibility in the educational setting where accuracy is highly valued. Nevertheless, Koren (1995) indicates the need of the pronunciation test for the following reasons.. First, we need a means for measuring or testing pronunciation in order to establish any kind of relationship between pronunciation and other variables such as musicality, age, and personality factors. Second, the pronunciation test is inevitable for a language laboratory class designed mainly to measure the learner's pronunciation.

Recently, however, some new attempts have been made to measure the learner's pronunciation ability using computer technology. Thus far, we have experienced little use of computer in testing a speaking skill, including pronunciation, compared with other language skills. This is largely due to the limited functions of the current computer technology. One of the most serious limitations of the current computer technology for testing speaking is to store and automatically evaluate what the learner speaks. For this reason, computer simply stores what the examinee utters and raters

evaluate it afterward on a certain rating continuum. In particular, the testing of such delicate skills as pronunciation, stress, and intonation, which requires highly advanced computer technology, has been so limited that raters, simply by listening, make a rough check on them.

With the advent of voice recognition technology, however, the computer has been able to test pronunciation. This technology enables the computer to identify, visually show, and evaluate the learner's intonation pattern by means of autocorrection. Strictly speaking, however, this is not for testing pronunciation; rather it is for the comparison of the learner's utterance with the native speaker's.

In general, testing pronunciation has been carried out indirectly. That is, indirect testing of pronunciation attempts to measure the abilities which underlie the skills in which we are interested. Perhaps the main appeal of indirect testing is that it seems to offer the possibility of testing a representative sample of a finite number of manifestations of them. The main problem with indirect tests, as Hughes (1989) points out, is that the relationship between performance on them and performance of the skills in which we are usually interested tends to be rather weak in strength and uncertain in nature. In other words, indirect testing raises a serious problem in terms of validity. Nevertheless, it is widely preferred by test administrators because of its high feasibility.

Now computer has paved the way for direct tests of pronunciation. The voice recognition technology enables the computer to read the examinee's oral production and measure it based on the given criteria set by the rater such as the pitch of each phoneme and the intonation of the whole sentence. The measurement in this practice is represented in terms of the acceptance rate of the examinee's utterance by the computer. If fourteen sentences out of twenty are accepted, for example, the acceptance rate is 70 percent, and the examinee's oral production is scored as much.

A difficulty that arises here, though, is to determine how accurate oral production should be in order to be accepted.

4. Toward Computer-Adaptive Pronunciation Testing

4.1. Developing a Computerized Testing Tool

In order to perform computer-adaptive testing, we need to develop a computerized testing tool as a preliminary step. Some testing tools are currently available such as MicroCAT (1997) and MultiCat (1999). One can also develop a testing tool that can serve the need of one's own. When it comes to developing a tool for testing pronunciation, it requires highly advanced technology because it should be able to store and analyze the examinee's utterance.

A computerized testing tool needs an item bank module, a computerized testing module, and an item and test analysis module. These elements are kept not in a single execution file but in separate files, and they are integrated only when executed. One of the most fundamental components of CALT can be said to be an item bank storing a great number of items analyzed by IRT. Items in the item bank should be selected by the same standards and arranged in advance by IPE (Item Parameter Estimation). The size of the item bank varies according to the purpose and/or scope of testing.

The computerized testing module should have a function to determine the difficulty level of the first item provided, to branch items, to keep updating the examinee's ability estimate, and finally to finish testing according to pre-specified stopping rules. Accordingly, decisions should be made concerning the following steps:

(1) to select an initial ability estimate for the examinee

(2) to score the ability estimate based on the examinee's response and pre-calibrated item parameter value

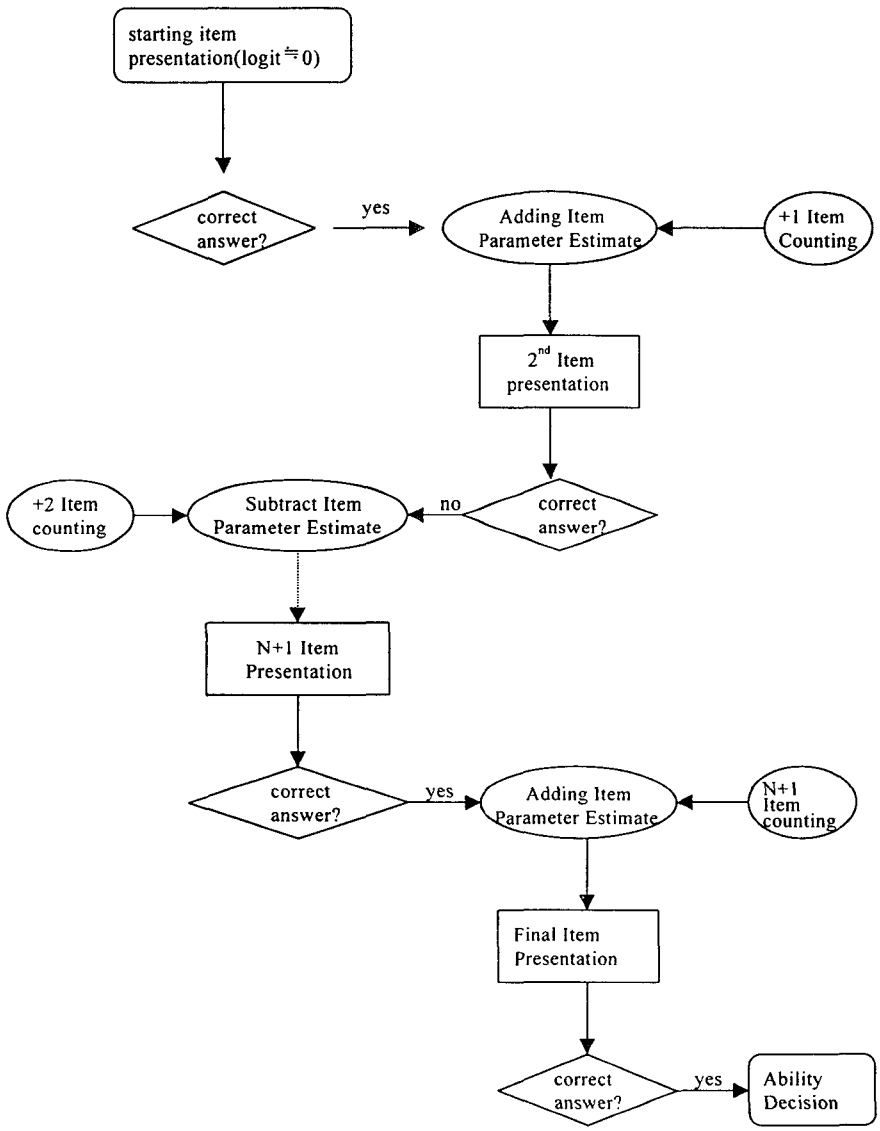
(3) to select the next item given the updated ability estimate

(4) to continue the CAT process until a designated stopping rule is met

An initial estimation of ability enables the examinee to determine with which item in the item bank he should start testing. If there is no prior information concerning the examinee's ability estimate, an initial item of moderate difficulty had better be set to zero as a starting point. Updating ability estimation based on the examinee's response and pre-calibrated item parameter value can be done by the following three estimation rules: a) maximum likelihood estimation, b) Bayseian estimation, and c) Bayes modal estimation. Item selection, on the other hand, can be made by either MLAE (Maximum likelihood ability estimation or BAE (Bayseian ability estimation) if items in the item bank are already calibrated by IRT (Item Response Theory).

But these two estimation rules are not used for pronunciation testing. Rather, IPE (Item Parameter Estimation) is used for the purpose, since pronunciation testing does not need a large number of items and it requires a minimized adaptive function. MicroCat (1999), a computerized testing tool, makes use of IPE, in which the examinee provided with an initial item responds to the item and the computer, based on the examinee's response, estimates an approximate item parameter value and presents the next item with this value. As shown in the following flow chart in Figure 1, an examinee signs up his/her name and press the start button. The computer will then display one randomly chosen item with logit scale zero from the item bank to start. The examinee responds to the item, and the response is immediately evaluated against the answer key and counted for or against the examinee's proficiency record. The number of items at the certain level shown to the examinee is counted and serves as a determining factor along with the number of correct responses whether or not to show the next level of items. The student ability level counter is to check whether or

not enough tries have been made at the certain level to set the level as the student's proficiency level. The test ends if it satisfies one of the following two conditions. First, an examinee administers a fixed number of items (i.e., 20). Second, an examinee administers a test until the standard error of θ estimate reaches the pre-specified level.



<Figure 1> Flow Chart of Examinee's Ability Estimation

The item and test analysis module is designed to analyze the results of the examinee's response to items according to the pre-specified test item characteristics and then provide the evaluator with basic information for feedback and the improvement of items.

4.2. Constructing Item Bank

For computerized language testing it is necessary to construct a large item bank consisting of lots of items. Test items must be stored in a format for examiners to increment additional items easily and for examinees to retrieve any appropriate level of items to their language ability efficiently. Henning(1991) suggests that the item bank must include the following five characteristics:

(1) Items included in an item bank have been pre-administered and analyzed for use the same intended population of examinees.

(2) The items in a CAT items bank are stored together in a system that permits ready retrieval and rapid presentation for testing purposes.

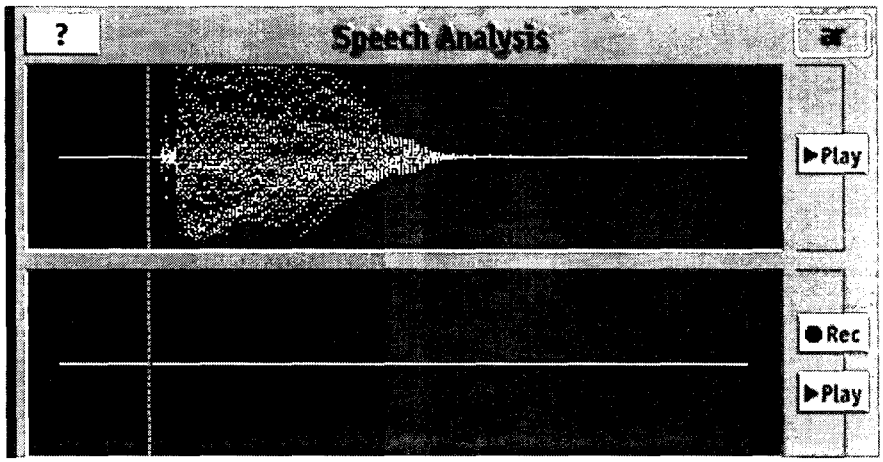
(3) Items in an item bank have been calibrated or placed on some measurement continuum according to a uniform scalar metric.

(4) The possibility exists of increasing or decreasing the number of items in the item bank without destroying the usefulness of the prior remaining items or changing the meanings of the item classifiers.

These criteria can be divided into two groups, the validation of test items and the implementation of efficient interactive design.

There are two types of items in the item bank for pronunciation tests. The first type of items is one for measuring the accuracy of the examinee's utterance, which is numerically displayed on the screen. The second type of items is one for measuring the degree that the computer, using voice recognition technology, recognizes the

examinee's utterance based on the pre-specified criteria. Computerized pronunciation testing, in particular, requires particular peripheral devices and software that can catch the pitch of a sound, display its pitch pattern on the screen, and finally analyze it. In order to measure the accuracy of the examinee's utterance with the first type of items mentioned above, we need a D/A board, DSP board, and A/D board, which are all internally equipped with the high-speed sound analysis system analyzing the waves, pitches, and stress of a sound. In this system, it is possible to visually identify and aurally measure the examinee's intonation patterns of the target language, as shown in Figure 2.



<Figure 2> Speech Analysis by Computer

Take a sample sentence "this is my car" and compare its different intonation patterns produced by the examinee and the native speaker. As noted earlier, they will be displayed in the two layered tiers on the screen. The degree is numerically displayed in which the two intonation patterns overlap with each other, and it serves as a determining factor of the examinee's accuracy rate of pronunciation. To be more specific, the accuracy rate of the exmainee's pronunciation depends on how close its

intonation pattern is to that of the native speaker of the target language. The closer it is, the higher score it obtains.

With the second type of items illustrated above, we need to set criteria with which the computer will take an utterance as accepted. The determining criteria can vary according to the examinee's grade level. For example, they can be set to the extent that the native speaker can readily understand the utterance in a real communication. When it comes to the selection of items for utterance, the examinee is not allowed to select any item for an open communication, which is due to the limitations of the current voice recognition technology. Instead, the examinee is obliged to utter only the items presented by the computer, which now plays a role of measuring the accuracy rate of his/her utterance.

The number of items to be included in the item bank for pronunciation tests can vary according to the examinee's grade level. It is difficult to determine the size of the item bank. But suffice it to say that it is around a third of the items needed for testing other language skills such as listening and reading, since computerized pronunciation testing needs the least adaptive function. It is generally acknowledged that the optimal size of the item bank for pronunciation achievement tests is around 150. However, pronunciation diagnostic tests need around 200 items, which is 30 % higher than the former.

5. Conclusion

Pronunciation testing has been problematic due to the lack of validity and reliability. The appearance and use of computer, however, have provided us with a valuable momentum for working out such problems and greatly improving the quality of testing. The rapid development of multimedia computer technology speeds up the activation of the high-speed sound analysis system which analyzes the pitches, waves,

and stress of a sound. We are now experiencing the application of this technology partly in the testing as well as in the teaching of pronunciation. Furthermore, the recent voice recognition technology shows great possibilities in testing oral ability, including pronunciation, with an outstanding degree of validity and reliability as in other language skills.

The current computer technology is still far from perfect in testing oral ability. But it is expected that such technological difficulties will be reduced in the near future. Then the focus of our research of pronunciation testing will be shifted to the following item-related issues; 1) what types of items should be used for the item bank; 2) how the achievement level and proportion of each type of items should be determined; 3) how the criteria for the item parameter estimate are set.

<References>

Bachman, L. F. (1996). What does language testing have to offer? In D. Brown & S. Gonzo, (Eds.). *Reading on second language acquisition.* Englewood : Prentice Hall Regents.

Baker, F. (1992). *Item response theory : Parameter estimation techniques.* New York: Marcel Decker, Inc.

Brown, J. D. (1997). Computers in language testing : Present research and some future directions. *Language learning & technology,* 1(1).

Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language testing,* 11(3).

Dunkel, P. (Ed.). (1997). Computer-Adaptive test of listening comprehension: A blueprint for CAT development. Document URL:http://langue.hyper.chub.ac.jp/jalt/pub/tlt/97/oct/ dunkel.html.

Henning, G. (1991). Validating an item bank in a computer-assisted or computer

adaptive test. In P. Dunkel (Ed.), *Computer-assisted language learning and testing*. New York: Harper Collins.

Hughes, A. (1989). *Testing for language teachers*. Cambridge : Cambridge University.

Kim, J. R. & Lim, C. K. (1997). Computer adaptive English testing. *English teaching,* 52(3).

Koren, S. (1995). Foreign language pronunciation testing: A new approach. *System,* 23(3).

Larson, F. (1996). An argument for computer-adaptive language testing. *Multimedia-Assisted Language Learning*, 1(1).

Madsen, H. S. (1991). Computer-adaptive testing of listening and reading comprehension: The Brigham Young University approach. In P. Dunkel (Ed.). *Computer-assisted language learning and testing*. New York: Harper Collins.

Moon, O. H. (1997). Improving the measurement of English language ability with computerized adaptive language test. *English teaching,* 52(3).

Nitko, A. & Hus, T. (1984). *Pitt educational testing aids. User manual*. Pittsburg: University of Pittsburg.

Shohamy, E. (1993). A collaborative/diagnostic feedback model for testing foreign language. In C. Chapelle & D. Douglas (Eds). *A new decade of language testing research*. Alexandria : TESOL inc.

Spolsky, B. (1995). *Measured words*. Oxford : Oxford University Press.

Weiss, D. (1982). Improving measurement quality and efficiency with adaptive testing. A*pplied psychological measurement.* 6.