

# The NativeAccent™ pronunciation tutor: measuring success in the real world

Maxine Eskenazi, Angela Kennedy, Carlton Ketchum, Robert Olszewski, and Garrett Pelton

Carnegie Speech Company

4615 Forbes Ave. Pittsburgh PA 15213 USA

{max, ack, carlton, bobski, gap} @carnegiespeech.com

## Abstract

This paper describes real user assessment of NativeAccent™, a pronunciation tutor using automatic speech recognition that is a commercial product. It describes the product and discusses the issues involved in assessments of real users in real situations, such as assessments based on the customer's own criteria instead of more academic measures, and the variations in the customers' measures. Results in one study show that subjects who used NativeAccent™ did more than twice as well as the control group while both groups had human instruction. The implications of these results are discussed in light of other measures and real world considerations.

**Index Terms:** pronunciation, error detection, tutoring system, assessment

## 1. Introduction

The past decade has seen a generalization of the development of algorithms to detect errors in non-native pronunciation [1] [2]. This wealth of work has led to higher precision compared to previous attempts which could at best detect that there was an error in a word, but could not point out where that error occurred. Developed first for English as a second language, these types of algorithms have been shown to be usable for other languages as well [3] [4].

While one could be interested in algorithmic development alone, there is much that lies beyond error detection, to approach the actual use of the algorithms in a language tutoring system, and for research on learning. Many of those who have worked on the algorithms have made this transition. This has brought the state of the art from the simple algorithm to the demonstration of feasibility. Using a system centered on the detection algorithm, researchers can now show how a system's interface interacts with students and can give an idea of the materials it can teach.[5]. While some have already shown success [3] [6] [7] [8], one study has expressed reserved caution about the validity of the use of error detection for learning [4]. This demonstrates a move of one branch of automatic speech processing research out of the laboratory and into the real world. Another branch making this move is the work done on automatic assessment of non-native speech [9] [10].

Yet, bringing a system from proof-of-concept to useful product is a complicated task, expensive and labor-intensive. The software must be robust (consistent high performance for the majority of users, useful on many computer platforms, etc). A realistic market must be found (people who will actually pay money to have the product, how potential users there are, etc.). The software also must be appealing. All of these issues are not

high on the average academic language technology researcher's agenda. Thus, much of the systems based on pronunciation error detection have so far remained in the laboratory. One of the exceptions is the work of Yamada [11]. It has become a complete pronunciation training system for Japanese speakers who want to speak English and has been commercialized by ATR in Japan. NativeAccent by Carnegie Speech is another exception. With its origins in research at Carnegie Mellon [1], the Fluency project was spun off into a company, Carnegie Speech, in 2001 [12]. At present, there are users of NativeAccent in several branches of industry on several continents. Since, up to now, the product has been sold to companies for their employees rather than to individual end users, the companies/clients are the ones who test the efficacy of the product in order to determine if their employees are showing real improvement in their spoken English as used in their jobs. Results of some of these tests are given in below.

## 2. Background

To go from automatically detecting errors in pronunciation to tutoring students in the pronunciation of a target language, some of the main components that need to be developed are: leveled corrective feedback information; a full curriculum; a student model; a strategy on how to proceed through the curriculum for different learners (fast and slow, for example); a reporting mechanism for the teacher (to follow individual and grouped student progress). NativeAccent™ has been endowed with these features as well as others, requested by the customers, that are less essential to the basic system.

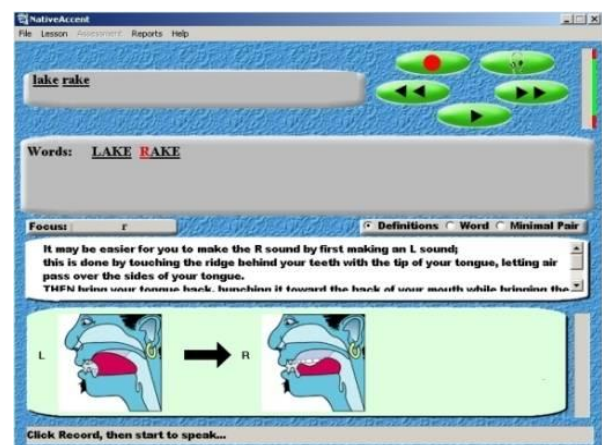


Figure 1 Main screen of NativeAccent showing feedback

Since the place where an error has occurred in the string of target sounds (for example, in a sentence) and the exact incorrect target sound (for example, a TH) is known, it is possible to offer specific corrective information to the student. It can be argued that self-discovery would be an alternative and valid method of learning here (“listen to what you said and find your own errors”). Yet without being trained on how to listen discriminatively, the task is extremely difficult and errorful for the student. Training in self-discovery requires that the student have enough time to devote to both pronunciation training and the basics in discriminative listening and pronunciation of the target language. This is not always possible in the real world. The ATR [11] system may offer hope in this direction; it has shown that training on listening alone can translate into gains in pronunciation for some students. NativeAccent on the other hand uses immediate corrective articulatory help for each type of phonetic or prosodic error that the student may make, as seen in Figure 1. Recent work has shown that at least some explicit instruction is more effective than implicit instruction alone [13] and so Native Accent’s strategy can furnish the explicit training. By calling attention to the precise item that is incorrect and giving several forms of targeted specific instructions immediately (while the student still remembers how they produced the erroneous sound), the system increases the probability that: 1) the student will find helpful information (one form of feedback will perhaps mesh more than another with a student’s style of learning), and 2) the student will know exactly what was pronounced incorrectly, and thus maximize the chances of success. Corrective help in NativeAccent is language-specific and presently covers 28 different native languages and includes an “other” category.

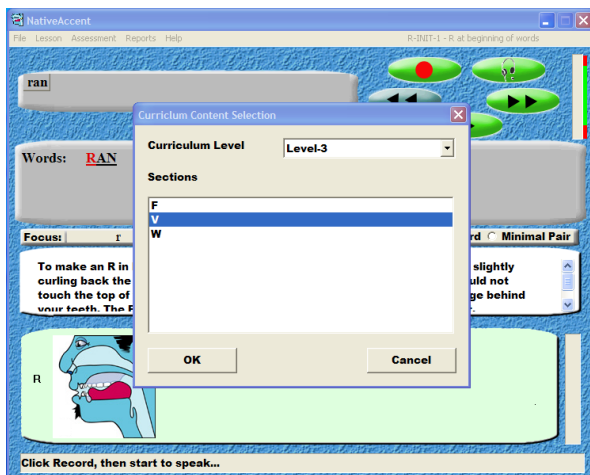


Figure 2 Selecting a lesson and overriding the intelligent tutoring mechanism

The NativeAccent curriculum has about 800 exercises that cover all of the sounds in English as well as aspects of duration and pitch. Some of the exercises are specific to one or more of the 28 native languages (for example, the /TH/ exercises are not the same for Japanese learners and for Russians). The completeness of the curriculum allows us to select user group-specific exercise subsets. This is essential for many of Carnegie Speech’s clients since their employees may not have the 100

hours of training time needed to proceed through the whole curriculum. By providing a smaller, very focused subset the students can improve their pronunciation within the amount of time they can afford to devote. The real user tests we have conducted are with clients who have taken this subset option.

By modeling student progress, the software can estimate how well it believes that each student knows each sound/skill and how well they should do the next time they have to pronounce it. It can then proceed through the curriculum, spending less time on some sounds/skills that the student performs well on and more time on poorly mastered ones. Moreover, if the student wishes to practice some skill again, they can override this intelligent tutoring mechanism by using a pull-down menu (see Figure 2) that allows them to bring up an exercise at any level and on any skill that they want. To service client needs and the needs of the teachers, NativeAccent also generates summary reports of student progress based on this knowledge.

Endowed with a very flexible curriculum, NativeAccent goes beyond the one-size-fits-all solution. It is adaptable to the variable needs of the clients/employers. The end measure is whether the student, who usually needs to speak English clearly in their job, can achieve better job performance after NativeAccent training. Thus, unlike other assessments of language learning software, the measure used to assess NativeAccent here is not created by the authors of the software, or by some academic domain standard, but by the individual clients. Their judgment becomes the gold standard, and that standard changes from client to client. The numbers of users at any one site may vary considerably, so expected sample size is sometimes sufficient to conclude statistical significance, but many times it is not. Since a comparison across sites is not possible due to the change in criteria, while they do reflect real world use, most of the client results can only be used reveal a trend in the data.

### 3. Experimental setup

In this section we will describe one assessment of NativeAccent that took place in a company in India. The goal of the assessment was to show that employees trained with NativeAccent perform better, according to their employer’s criteria, than employees who did not use NativeAccent.

There were 43 subjects, 25 in the test group and 18 in the control, divided into groups randomly. All of the subjects were recent university graduates in their first real job and all had passed the company’s entrance assessment. Both groups had training with the company’s usual teacher. during this time The assessment was conducted over a three week period. The test group had a total of 20 hours of NativeAccent training time in addition to time with the teacher. The control subjects did not have any additional training. All subjects continued to use English in their jobs when they were not training.

As mentioned above, for clients whose employees have a very limited amount of training time, Carnegie Speech generates a mini-curriculum that includes a mixture of phonetic and prosodic pronunciation training. For this test, 20-hour mini-curricula for each of several native Indian languages were used. All of the subjects were pretested on the phones in English. The pretest, which was furnished by Carnegie Speech, automatically examines a speaker’s phone production of every sound in

English. Phones were ranked for overall scores from worst to best (for example, the combined scores of all 43 subjects could have been the lowest on IH as in “hɪp”). The test subjects worked on 4 phones per day, some days concentrating on the worst scoring phones and on alternate days, the best scoring ones. Prosody exercises were interleaved according to their usual order in NativeAccent.

At the end of the training period, both the test and control subjects were posttested via an individual face-to-face interview, which is the company’s usual manner of assessing their employees’ fluency in English. This assessment comprises many areas of company performance such as typing. The area we are interested in is “accent training”. As mentioned above, a Carnegie Speech assessment was used as a pretest to determine the order of the phones to learn and the client company also gave each student a pretest with a live interviewer. The interviewer assessed performance based on a 5-point scale in several areas including pronunciation and degree of accent. A final composite score was obtained, ranking each criterion evenly. Thus, unlike academic assessment procedures, the pretest and posttest measures are not the same due to the client’s request. But the goal here is to make more employees competent in the use of English in their jobs. Thus we cannot measure how much has been learned. Rather we examine how much more competent the employees have become.

#### 4. Results

The client company’s pretest shows that, at the outset, the control group was stronger overall than the test group. On a 5-point scale (1 is worst, 5 is best), the control group scored higher in all categories. Those results are shown in Table 1.

Group	Accent	Pronunciation
Test	1.60	1.60
Control	2.61	2.72

Table 1 Average pretest scores for control and test groups on client company-internal test.

The results on the client company’s posttest show that the test group improved more on the accent training assessment part of the test and passed the course (and was thus judged to be capable of using English in their jobs) at a rate of 68% as opposed to the 33% of the test group, who were a stronger group at the outset. Other individual skills that were tested, such as typing, did not show as great a difference in results between the two groups. This seems to indicate that adding NativeAccent to the curriculum to complement classroom work boosts the success rate. Even though the test group had more room for improvement, both groups performed relatively poorly on the pretest for accent and had ample opportunity to improve during training.

Group	% Increase in Accent Score	Course pass rate
Test	76.00%	68.00%
Control	44.44%	33.33%

Table 2 Posttest pass rates for control and test groups on client company’s internal test: accent training is only one of several criteria.

#### 5. Discussion of results

Results from the above test show that a tutoring system based on automatic detection of pronunciation errors does help users in the real world. Yet these results leave room for much discussion.

First, the number of subjects was relatively small. While this is true for this one test, it should be noted that the company is constantly hiring new employees and training them, so if the test is carried out for new groups of incoming employees, we should see more statistically significant results.

Second, the interviews had some different scoring on for the pre and posttests, but the “accent training” scoring was the same. Basically, the pretraining test determines how well the employee speaks at that point (and thus how willing the company will be to expend time and money to train this subject) and that posttest determines how well the employee is now able to handle their work. We also note that if we are able to increase the abilities of a group that was weaker at the outset, then NativeAccent seems to be useful for students at a variety of levels and not just for those who already have a good command of the language.

Linked to this concern, it has been noted in general, across companies, that the pretest pass rates for groups of new employees over time has decreased. Companies have a very high rate of employee turnover and are thus constantly hiring and training new employees. As they move through more and more people, they are obliged to hire less and less well-trained folks. Thus our results imply that NativeAccent may be capable of dealing with students who are less well-prepared.

There is also the issue of tests being specific to each client. Having the flexibility to perform well on a wide range of client criteria is important for acceptance and use. To provide contrast to the above study, the following is another study, carried out in a company in Latin America. In this company there were two groups of employees, each at a different site. For each group, the managers designated 20 employees who they felt needed to improve their English to perform better in their jobs. Each group of 20 was given the NativeAccent pretest that initializes the settings of the student model. The 10 employees in each group who had the lowest overall scores on the pretest were retained for this test. Each set of 10 subjects was then divided into two groups of 5 persons each so that there was a balance in proficiency between the two groups (according to our pretest). One group of 5 at each site used NativeAccent and the other group used another product commonly used in large companies for English training. Each of the four groups (5+5 at each site) used their designated software for a total of 13 hours. The subjects using NativeAccent were trained on 16-20 of the most difficult sounds as well as on prosody. It should be noted that not all of the subjects used the software as they were instructed.

Finally everyone took the NativeAccent pretest again as a posttest.

Here we have a comparable pre and post test, but small numbers of subjects. Results here show that, for site R, 100% of the subjects using NativeAccent showed improved posttest scores while 80% of the subjects using the other software showed a *decrease* in performance on the posttest. For the other site, E, the subjects complied with instructions less than at site R. They only sat through training from 4 – 9 hours, not the 13 requested from them. In this case, subjects trained on NativeAccent improved by an average of 75% on the posttest while the subjects using the other software showed no significant change between pre and post test. We should note that using the NativeAccent assessment scheme does bias the results toward NativeAccent. However, the performance of the group using the other product certainly should not have decreased, even in light of this bias.

## 6. Conclusions

We have shown that, using criteria that differ amongst sites, but that correspond to real world needs, NativeAccent has been shown to help improve employee speaking performance. We can see that spending more time on pronunciation training definitely results in better language learning. NativeAccent is one tool that can achieve this.

These results are encouraging. As more tests are carried out (and as Carnegie Speech obtains permission to cite their results), it will be possible to form a more complete picture of the effectiveness of NativeAccent. We will be able to map its success across continents and across the domains that client companies do business in.

This paper also shows how difficult it is to obtain meaningful results in the real world. Many companies do not want results made public, many test in ways that satisfy their own needs, but do not correspond to the manner in which systems are assessed in academia. One way to deal with this issue would be to have a group of paid subjects, who are not employed by any specific company, come to Carnegie Speech to assess the effects of training with NativeAccent. Although this might furnish a clearer measure of learning, most companies would not use this result as a measure of effectiveness since the only convincing results for them are tests in the real world with real subjects – their own employees. This appears to be typical of the general education arena and is something that those who produce CALL software will have to deal with in the future.

## 7. References

- Eskenazi, M., “Detection of foreign speakers’ pronunciation errors for second language training - preliminary results”, Proc. International Conference on Spoken Language Processing, Philadelphia, 1996.
- Neumeyer L.; Franco H.; Digalakis V.; Weintraub M., Automatic scoring of pronunciation quality, Speech Communication, Elsevier, Volume 30, Number 2, February 2000, pp. 83-93(11), 2000.
- Tsubota, Y., Dantsuji, M., Kawahara, T., “Practical use of autonomous English pronunciation learning system for Japanese students” in ICALL-2004, 2004, paper 033.
- Neri, A., Cucchiari, C., Strik, H., “ASR corrective feedback on pronunciation: Does it really work?”, Proceedings of ICSLP2006, Pittsburgh, USA, 2006, pp. 1982-1985
- Eskenazi, M., Hansma, S., The Fluency Pronunciation Trainer, Proc. STiLL Workshop on Speech Technology in Language Learning, Marhollmen., 1998.
- Eskenazi, M., Hansma, S., The Fluency Pronunciation Trainer, Proc. STiLL Workshop on Speech Technology in Language Learning, Marhollmen, 1998.
- Neumeyer L.; Franco H.; Digalakis V.; Weintraub M., 2000, “Automatic scoring of pronunciation quality”, Speech Communication, Elsevier, Volume 30, Number 2, February 2000, pp. 83-93(11)
- Mayfield Tomokiyo, L., Wang, L., Eskenazi, M., “An Empirical Study of the Effectiveness of Speech-Recognition-based Pronunciation Training”, Proc. ICSLP 2000, Beijing, 2000.
- Eskenazi, M., Pelton, G., “Pinpointing pronunciation errors in children’s speech: examining the role of the speech recognizer”, Proc. Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology Workshop, Sept 2002, Colorado
- Bernstein, J. “New uses for speech technology in language education” Proc, ISCA Still Workshop, Marhollmen, 1998.
- Hincks, R., “Using speech recognition to evaluate skills in spoken English”, in Lund University Dept of Linguistics, Working Papers 49 (2001), p. 58-61.
- Akahane-Yamada, R., Kato, H., Adachi, T., Watanabe, H., Komaki, R., Kubo, R., Takada, T., Ikuma, Y., “ATR CALL: A speech perception/production training system utilizing speech technology”, The 18th International Congress on Acoustics, Proc. ICA 2004, III 2319-2320.
- <http://www.carnegiespeech.com>
- Ellis, N. ed., Implicit and Explicit Learning of Languages, Cambridge University Press, 2000.