University of Southern Denmark · Syddansk Universitet

NISLab document

Natural Interactive Systems Laboratory

Pronunciation Trainer Status

8 June 2004

Authors

Thomas K. Hansen and Niels Ole Bernsen

1 Introduction

This report describes the results achieved by June 2004 with a Danish Pronunciation Training application developed at the Natural Interactive Systems Laboratory, University of Southern Denmark, Odense. The report describes the goals of the project, the system's functionality, hardware and software, and the set-up and results of the user tests made so far.

2 Project goals

The goal of the project in its first phase, reported on below, is to iteratively build and test with representative users a first series of prototypes of a Danish pronunciation trainer which can teach Danish *single-word* pronunciation to learners of the Danish language. The core development and test objectives have been to:

- build a stable system prototype version which can be used effectively by second-language learners of Danish in a laboratory environment;
- evaluate the system's performance to verify that it behaves in a way which is close to that of a Danish language teacher who listens to language learners, judges how close their pronunciation is to correct and fully intelligible Danish pronunciation, and advices the students on particular issues in their pronunciation which merit additional training;
- baseline the system with native speakers of Danish in order to identify a "gold standard" for the second-language learners; and
- measure the progress in Danish language pronunciation made by representative Danish second-language learners who work with the system by themselves, outside of but in conjunction with, their class-based Danish course.

The last point mentioned above - assuming the positive results we have obtained concerning the three previous points - has, obviously, been of paramount importance in this first phase of the project. The steeper the average curve of improvement in Danish pronunciation by the test subjects, the more promising is the pronunciation trainer for accelerating the mastery of Danish single-word pronunciation through self-training with the computer.

Following the successful completion of the first set of project goals, the goal for the second phase of work is two-fold. We aim to:

- move use of the system outside the laboratory and into the Language Schools environments, and possibly elsewhere as well, in order to be able to take into account all manner of issues to do with system installation on different platforms and by different people, actual use of the system in conjunction with Language School teaching, teachers' experiences with using the system for supplementing classroom teaching, the variety of students at the schools as contrasted with the more uniform laboratory student population, teachers' use of student skill measurement software, etc.;
- conduct proof-of-concept experiments in the laboratory with an enhanced system prototype version able to teach pronunciation of Danish at the *sentence level,* including prosody. The methodology will be similar to the methodology for testing single-word learning presented in the present report.

Throughout, it is important to make the system versions on trial as pedagogically sound and motivating as possible, making the user interface simple and intuitive to use, enabling informative feedback on the students' input, making them comfortable during training and, if possible, motivated if not entertained when using the system, and, ultimately, making them fast learners of the Danish skills they need to become valuable workers and Danish citizens.

It is important to emphasise that the learning of Danish pronunciation at the word- and sentence levels is far from tantamount to mastering the Danish language. In particular, the pronunciation trainer described in this report falls short of teaching a really large Danish vocabulary and a comprehensive Danish grammar to the students, and it does not strongly support spontaneous Danish language production. As to the former, we are making increasing efforts as we move from the single-word level to the sentence level, to have the students pronounce "useful Danish" in terms of useful words, useful phrases, and useful sentences for daily life purposes. Still, the pronunciation trainer is primarily useful for students who are being taught Danish grammar and semantics in parallel, as well as for the considerable number of foreigners living in Denmark who already have substantial mastery of Danish grammar and semantics but who were never properly trained in Danish pronunciation.

3 The system

3.1 Functionality

The current functionality of the Danish pronunciation trainer is to present the student with a relatively large, phonetically balanced set of 450 Danish words which the student is encouraged to train to pronounce correctly. The words are presented orthographically on the screen one-by-one and the user can navigate the word library to focus on specific words. To help prepare and support correct pronunciation, the student can select to view the spelled word, hear the word pronounced by a native Danish speaker, or both listen to and view a native Danish speaker pronounce the word on video, cf. Figure 3.1.

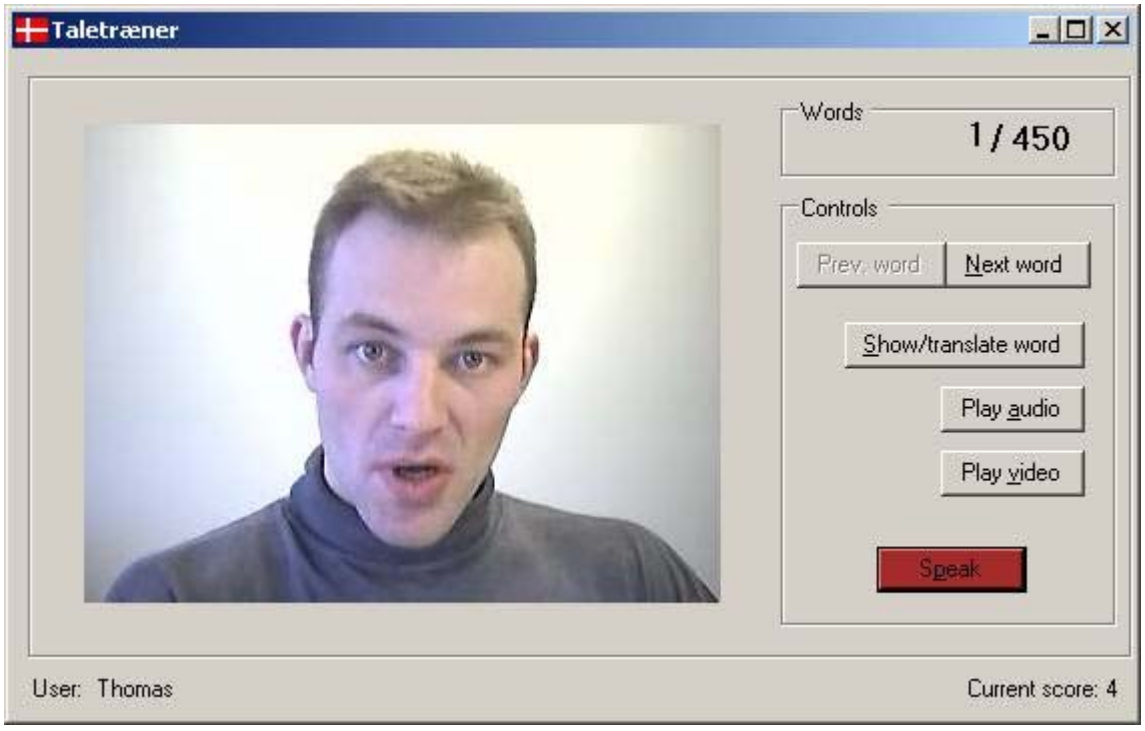

Figure 3.1. Pronunciation on video.

To support proper pronunciation and point out up-front major pitfalls of spelling versus pronunciation, words whose spelling-pronunciation relationships are irregular, as is often the case in the Danish, such as "bryllup"/"brølåp", are presented in square brackets on the screen in simplified and easily recognisable phonetic transcription using the ordinary alphabet. To support learning of word meaning, the word's

English translation is presented on the screen as well, cf. Figure 3.2. Once the learner is ready to pronounce the word, the Speak button is pressed, which activates the speech recogniser and informs the user that it is time to speak the word. When the user has spoken the word, the system evaluates the correctness of the pronunciation and feeds back to the user a score which reflects the quality of the pronunciation. In student test mode, the user is only allowed to pronounce each word once. In training mode, the student can train pronunciation ad libitum.

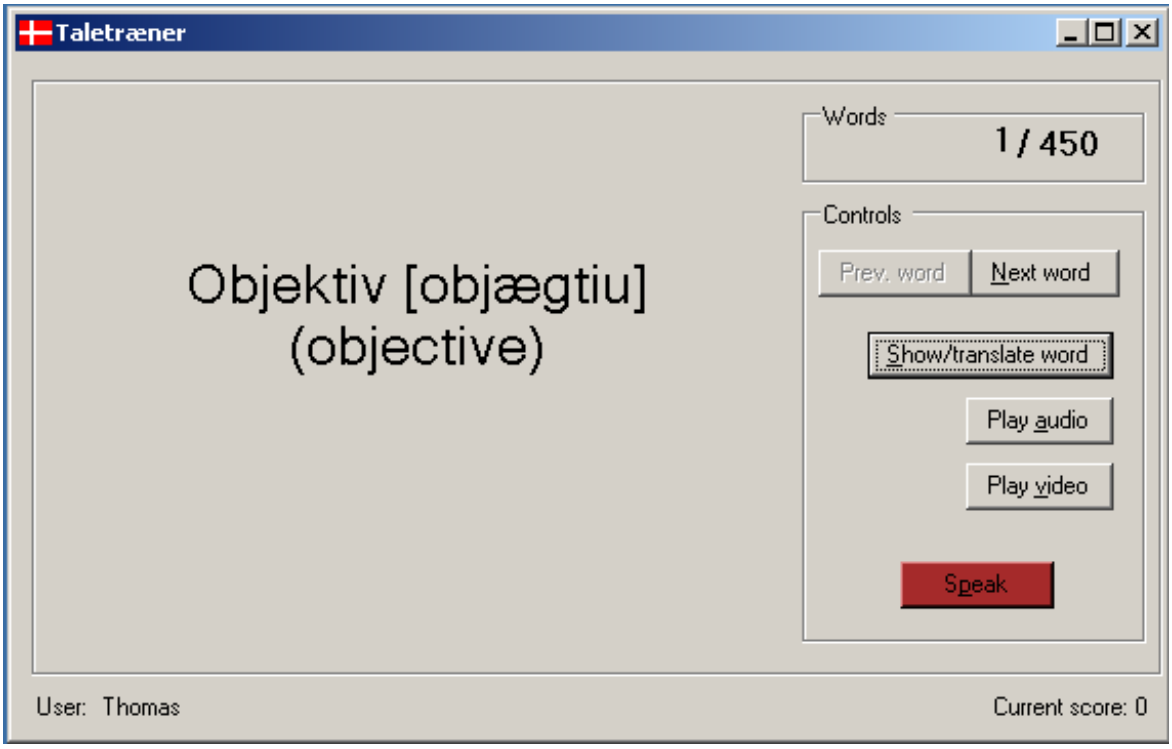

Figure 3.2. Graphical user interface.

The word spoken by the user is recorded in a separate file for purposes of later evaluation and comparison. In addition, the application logs the number of times the user presses the audio or video button in order to gather information on the learning strategy of the user.

For increased user support, we are currently implementing foreign-language-specific advice to students. The idea is to provide the system with knowledge about the specific problems speakers of a particular foreign languages have in pronouncing Danish. Each Danish word to be pronounced is tagged with the specific pronunciation issues that may arise, if any. When the word is not pronounced correctly, possibly after several failed attempts, the system offers dedicated pronunciation training exercises to that particular student based on its diagnosis of the problem the student has. The corrective feedback functionality just described is planned to be completed by the end of July 2004.

3.2 Hardware and software set-up

The pronunciation trainer prototype includes an automatic speech recogniser (ASR), a graphical user interface (GUI), logfile facilities, and kernel software for managing user-system interaction.

The ASR chosen is Philips' (now Scansoft) SpeechPearl2000, due to the quality of its Danish recognition. An entirely new GUI was created which allows the user to select between several word pronunciation presentation options which affect language perception and production at multiple levels in the training

process, cf. Section 3.1. The GUI was designed with simplicity in mind. It should be intuitively clear how to use the application.

The application is implemented in C++. All output training audio files have been recorded using the **audiomagic** application which is freeware found on the Internet. Recordings were saved as .wav files in 16 bit 44100 Hz stereo. Video output training files have been recorded using an ordinary webcam and saved in .avi format. The program used is **virtualdub**, also available from the Internet as freeware.

The complete application runs under Windows 2000 with soundblaster live/AW64 soundcards. Headsets and microphones are of average/good quality, ensuring that, to a reasonable extent, these devices do not pick up environmental noise that may damage word recognition. The entire setup is designed to mimic standard home-owned equipment. Furthermore, the use of software for recording audio/video files lends itself to easy manipulation by teachers in a class-based setting.

In addition, when using the system described in this report, the students can also use, in parallel, the listening trainer developed elsewhere in the project. The listening trainer develops the student's acuity of perception of phonetic differences in the Danish, which, we argue, is a necessary step in learning correct Danish pronunciation.

4 Baselining the application

In order to test the reliability of the ASR and in view of the fact that no current speech recognition system, including the human system, has a zero per cent word error rate, eight native speakers of Danish (4 men and 4 women) from different parts of the country pronounced 100 words to measure how many of these the ASR would recognise. The overall mean score was **90,25%** correct recognitions.

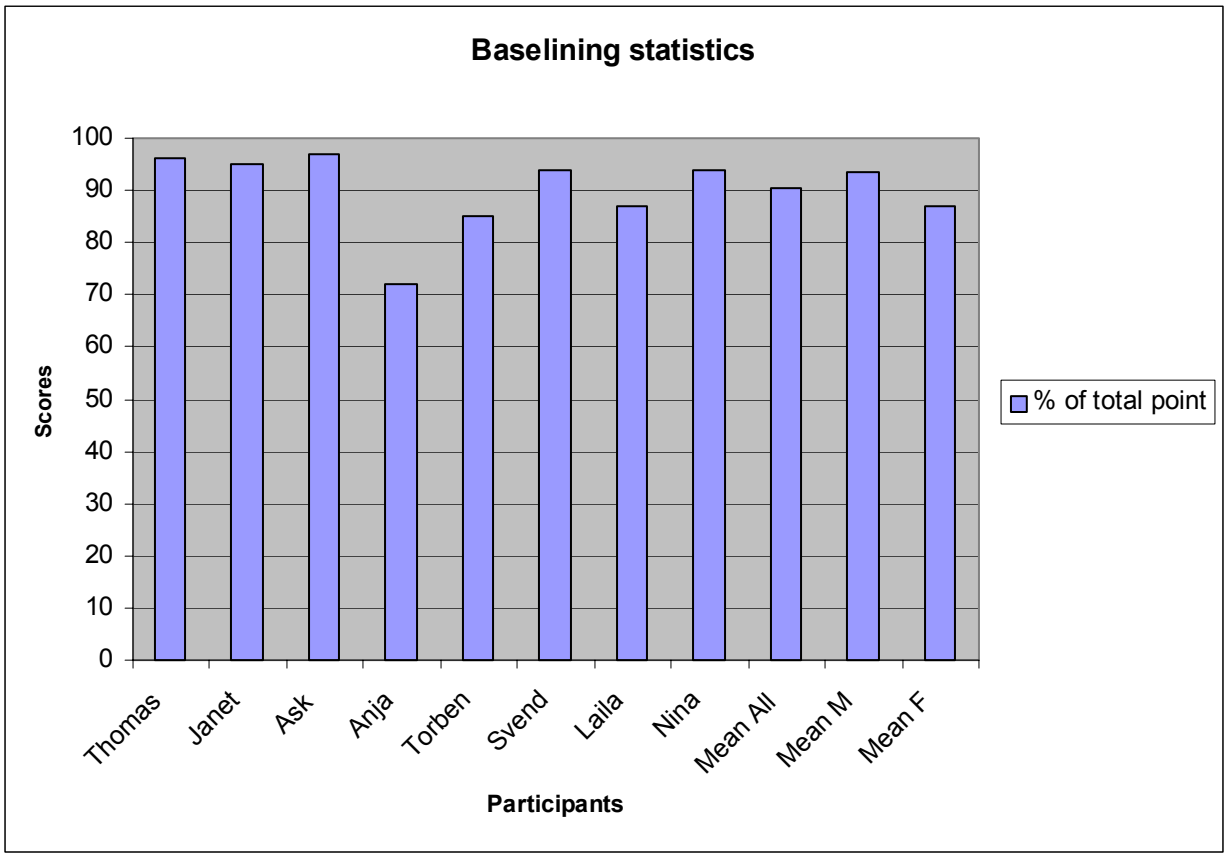

Figure 4.1. Baselining with Danish speakers. Points scoring is explained in Section 5.4.

Based on prior experience, it was suspected that female voices would be harder to recognise than their male counterparts. The average score for men was **93,5%** and **87%** for women, tending to confirm our hypothesis. However, the 6,5% difference in male/female recognition success is mainly attributed to a single female speaker, Anja, who scored significantly lower than the other Danish test subjects. There is no apparent reason why this was the case. The individual and average scores are shown in Figure 4.1.

5 Testing the system

5.1 Training material

450 (mostly) common Danish words were transcribed using a slightly modified version of the Speech Assessment Methods Phonetic Alphabet (SAMPA) for Danish. The words were selected on the basis of several criteria: the lexicon should, as far as possible, reflect words that are commonly used in ordinary conversation. The website www.korpus2000.dk was consulted to examine which words are used most commonly. Function words were largely disregarded and emphasis was on form words, primarily (proper)-nouns, adjectives and verbs. To ensure a phonetically balanced word test set, the following measures were taken:

- some close minimal pairs are present in the corpus to see how the user and recogniser handle these;
- all phonemes in the Danish inventory are present a number of times and in different combinations;
- monosyllabic as well as polysyllabic single words have been used to create diversity. Due to the limitations of the ASR, the lexicon mainly consists of polysyllabic words. Unless specifically trained on short words, such as the cardinal numbers, current ASR technology is known to perform better with longer words than with two- or three-character words. The predominant use of longer words does not matter for our purposes since the primary objective is to teach Danish word pronunciation rather than to re-confirm known recogniser limitations;
- some words were chosen purely because these contain phones or phone combinations that typically represent a problem for the speakers of Chinese who participated in the test;
- conversely, some words were chosen because, theoretically, they should not present any problems for the Chinese test subjects.

5.2 Speakers

Ten native speakers of Mandarin Chinese were recruited from Studieskolen to spend a period of 3 months practising pronunciation with the pronunciation trainer for 2-3 hours a week. Eight students were selected to constitute the experimental group while the remaining two served as the control group. The control group was tested at the beginning and end of the project to measure potential increase in pronunciation quality without having been subjected to specific training, while the experimental group was evaluated at regular one-month intervals. Finally, the results obtained at the beginning and end of the test period were compared to measure increase in pronunciation quality.

All students were between 23 and 31 years of age. Most students had been in Denmark between 6 to 18 months already. They had all received, or were currently receiving, some form of Danish language teaching for an average of 5-6 months. They are all students at the University of Southern Denmark, Odense.

5.3 Control group

The results of the initial and final testing of the control group are shown in Figure 5.1.

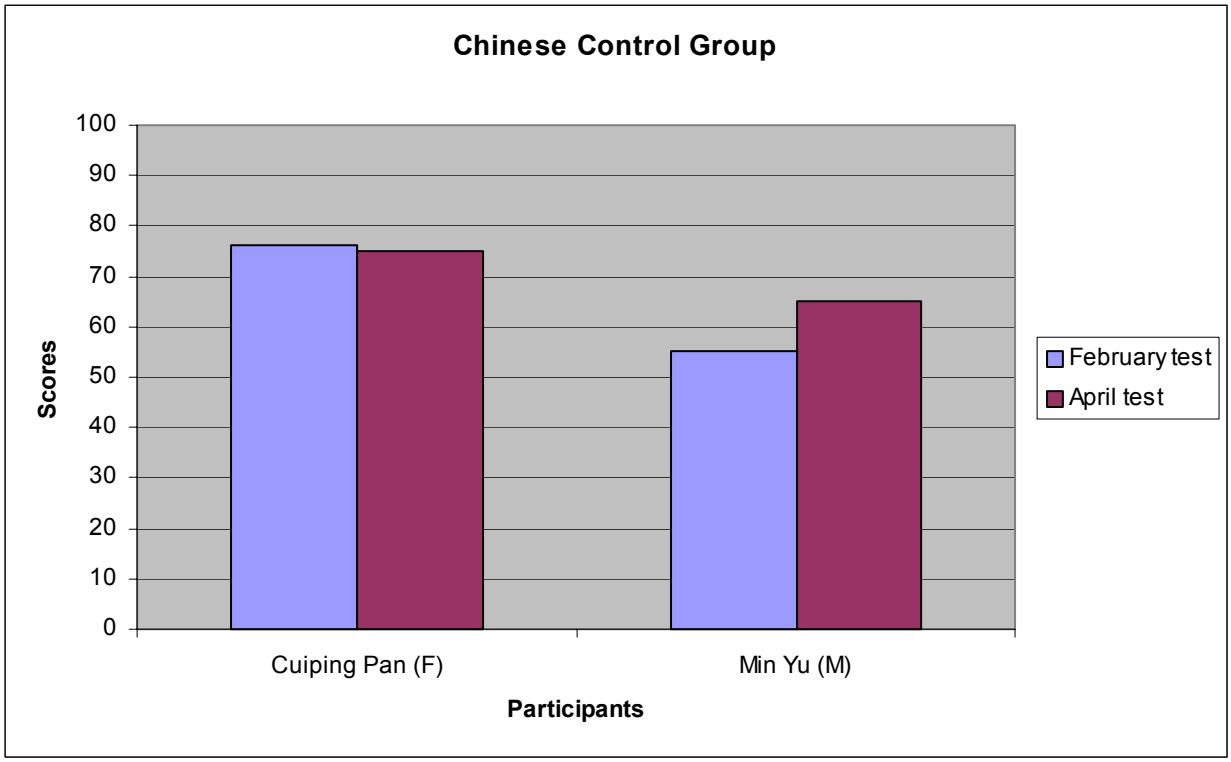

Figure 5.1. Initial and final testing of the Chinese control group. F is female, M is male.

Figure 5.1 shows that, upon initial testing, the mean score for the Chinese control group is, as expected, somewhat below that of the Danish baseline. After the test period of three months for the experimental group, no major progress had been made by the control subjects. Min Yu shows an increase of 10% in recognition from February through April whereas Cuiping Pan shows a slight decrease in the otherwise impressive pronunciation score. Clearly, we would have wished the control group to have been larger than two subjects. However, we had to make do with the Chinese subjects we could recruit and, as the results to be presented below illustrate, the experimental group did show far more substantial improvement in Danish pronunciation than did the subjects in the small control group.

5.4 Evaluation method for the experimental group

Evaluation of user progress is both qualitative and quantitative. In the following, we show the results of the quantitative evaluation. In the qualitative evaluation (in progress), the recorded sound files from the start and end of the experimental group tests are being evaluated by two phoneticians using a specifically tailored scoring method. The final result of the qualitative evaluation, comprising 1600 files, is expected to be completed by mid-July 2004.

As regards quantitative evaluation, the ASR's recognition score is logged for each pronounced word. The ASR provides a numerical score (Confidence Score, CS) ranging from 0-1000 with 1000 being the best score possible. It should be noted that the CS expresses confidence in terms of similarities in recogniser vocabulary items (acoustic models) rather than confidence in user pronunciation per se. Hence, the more similar words can be found in the vocabulary, the more likely it is that the CS will decrease without reflecting any corresponding decrease in the subject's abilities to pronounce Danish words correctly. It is therefore desirable and fortunate that a phonetically representative and rich selection of Danish words can be made by choosing only 450 words as we have done. A vocabulary this size does not appear to

significantly distort the recogniser's ability to evaluate Danish word pronunciation, especially since we have mostly avoided monosyllabic words, as explained above. However, the limitations of the recogniser just noted call for supplementing the quantitative scores by qualitative scorings made by human experts, at least this first time of using the recogniser.

The ASR presents an N-Best list which presently includes the three topmost Hidden Markov Model path choices of words that the ASR believes the user to have spoken (see example below). The words are listed in a most-to -least likely order.

User: Thomas

Word 1 loaded: **syde**

Word seen: 1 times

Audio played: 0 times

Video played: 0 times

Recognition result, path choices:

Path choice #0 (confidence = 777): syde (start = 0, end = 580, confidence = **912**)

Path choice #1 (confidence = 135): øde (start = 370, end = 580, confidence = **135**)

Path choice #2 (confidence = 87): cykel (start = 0, end = 580, confidence = **87**)

In the above case, showing the internal feedback expression, the target word is **syde** (sizzle) and the 3-best list provides this word as its topmost choice which is indicative of pronunciation success.

All students were initially tested on the same 100 words as were the Danish speakers to see how much of a difference exists. Similar tests were conducted every following month. By the end of the experiment, initial and final student tests were compared to assess pronunciation improvement. The ultimate goal is for the student to achieve a score similar to that of native Danish speakers.

In conformance with the simplicity and intuitiveness design requirements for the system's user interface, the above logfile format is translated into a different feedback output for the user. All listings of path choice and confidence are removed and replaced by a list showing the recognized word(s), a smiley which depends on the recognition success achieved, and a simple numerical score. 0 points are awarded if the spoken word is not in the N-best list; 1 point is awarded if the word is included in the list but not at its top, and 2 points are awarded if the word is either the topmost choice or the only choice. The user's goal is to achieve as high a total score as possible. Although the application will continue to display the achieved score if the word is repeated several times, only the first score will be calculated in the GUI. Subsequent attempts and scores will still be incorporated in the logfile. The three scoring stages are illustrated below.

- The target word is the only listed word = [big smiley] 2 points.
- The target word is on the list = [smaller smiley] 1 point.
- The target word is not on the list = [sad smiley] 0 points.

Figure 5.2 illustrates a top score achievement.

5.5 Improvement through constructive feedback strategy

In order to enable the application to provide constructive feedback to the student, the following strategy has been devised and is currently being implemented.

Based on the most common pronunciation errors known to be made by Chinese speakers of Danish, a list of *tags* has been created. The tags are being associated with the individual words in the existing system vocabulary and, when a pronunciation error is detected, the application will inform the student what the

problem most likely is. Consequently, the student will be presented with a list of words specifically designed to address the problem area in question. After completing the list, the student will be taken back to the training data to continue training.

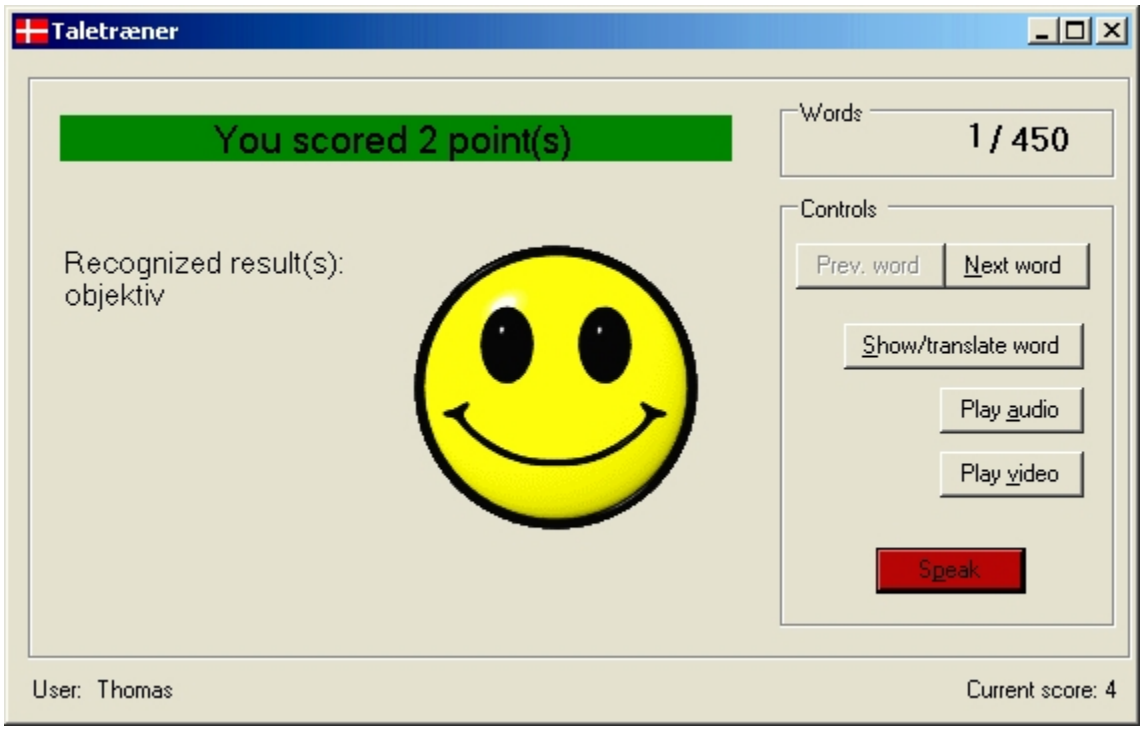

Figure 5.2. A top score of 2 points achieved.

6 Test results for the experimental group

Although the idea was initially that each student would use the pronunciation trainer for 2-3 hours a week, this proved to be very hard to achieve due to the fact that the students had their regular university course schedules to attend to. All in all it was expected that the students would practice approximately 24-30 hours for the full period of three months.

Figure 6.1 shows the total amount of time actually spent by each student. The figure shows that the amount of time spent on pronunciation training by each student varied heavily, with no student reaching the minimum target time initially expected.

All students in the experimental group were asked to pronounce the first 100 words the first time they used the application, thereby establishing a foundation which could be compared to the baselining of the Danish subjects. Subsequently, tests including 100 words were carried out every month to monitor progress.

Most students chose to address the task in a cyclical fashion by going from word 1 through 450 and then starting over. Only one student, Geng Tian, focused on a small number of words for each session, striving to achieve mastery of each word before progressing.

The monthly test results for all experimental group participants are presented in Figure 6.2.

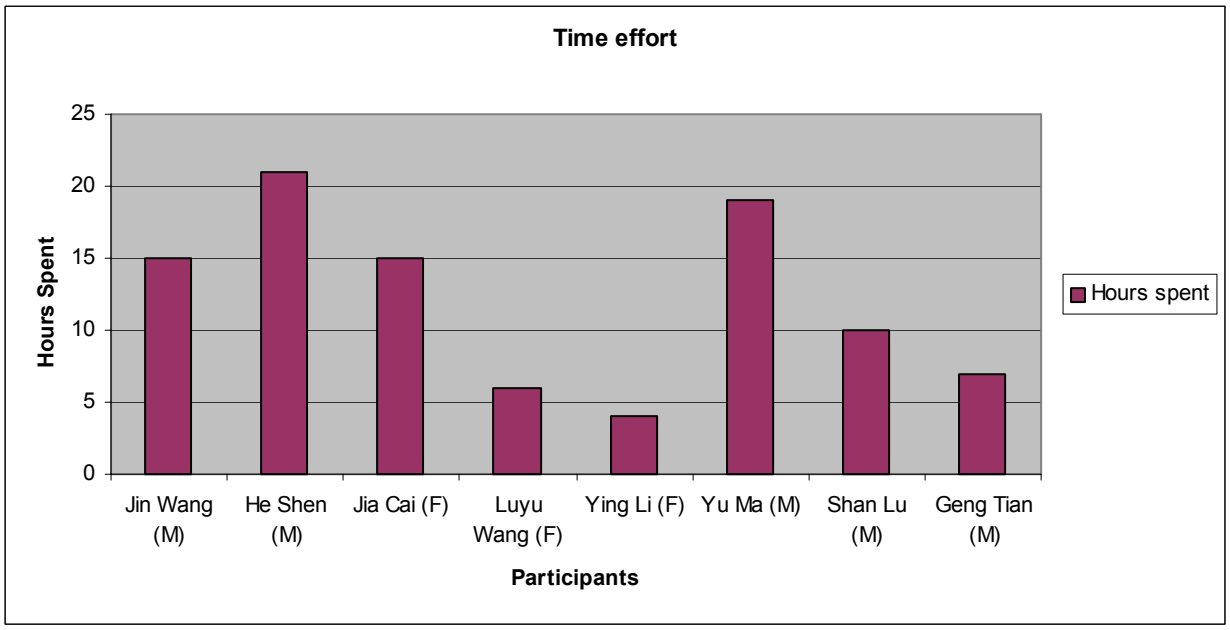

Figure 6.1. Total mount of training time spent per student. F is female, M is male.

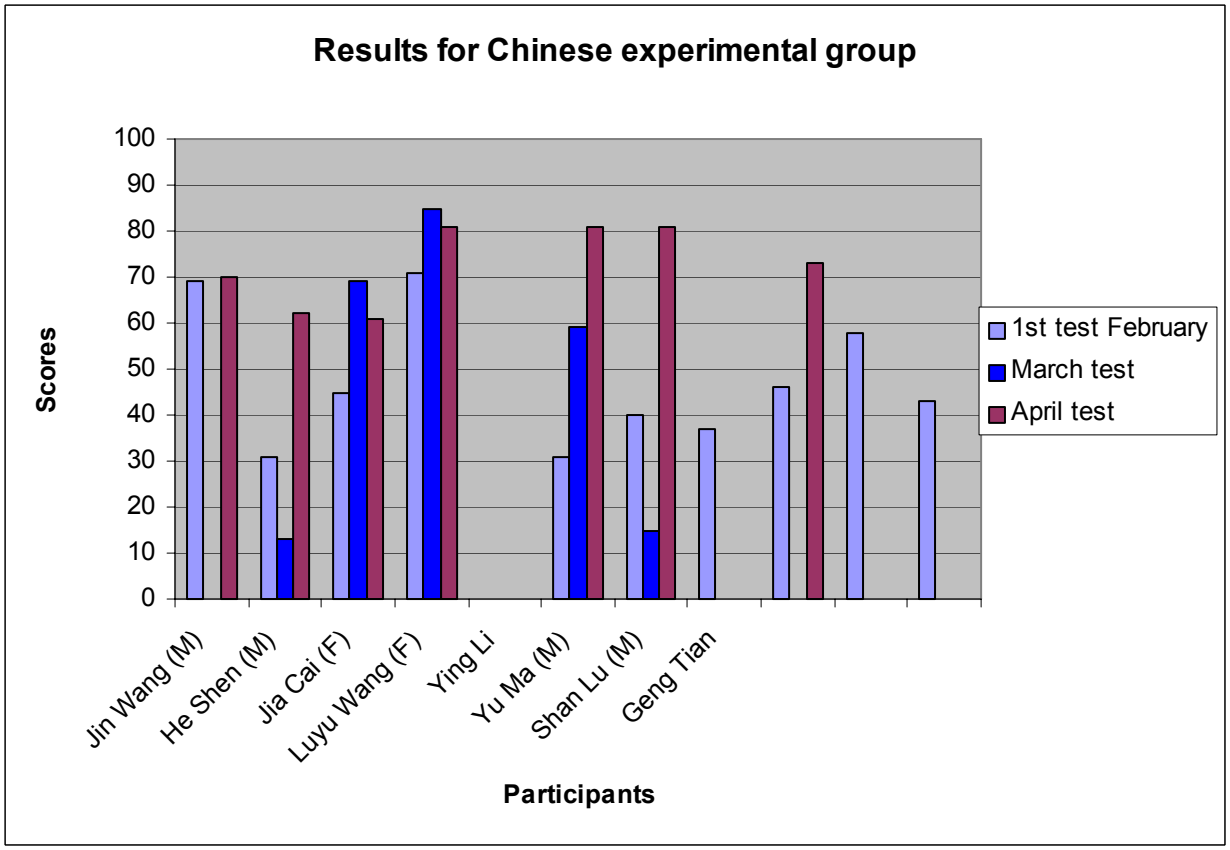

Figure 6.2. Monthly test results for the experimental group. F is female, M is male.

As can be seen from Figure 6.2, not all students were tested equally often. They were all informed of what to do when they arrived on the days of testing, but not everyone did as instructed, and some misinterpreted the instructions. Thus, too little data was collected for Ying Li to be of any use and only one set of data from Geng Tian proved to be of value. Furthermore, the second test of Jin Wang is missing.

Part of the problem encountered with testing was the duration of the test (approx. 1+ hour) and the amount of time that the students could spend on each visit to the lab. Sometimes the students had to leave before finishing the test and did not return for several days, thereby partly invalidating the final result of that particular test.

Still, when comparing the results of the experimental group (EG) with those of the control group (CG), it becomes immediately apparent that the EG shows substantial increase in recognition score from Test 1 to Test 2. Examining the results of Tests 2 and 3, the progress continues for some students, flatlines in some cases, or fall slightly. The results are displayed in Table 6.1.

Student	**Test 1**	**Test 2**	**Test 3**	**Total increase**
Jin Wang (M)	69%	N/A	70%	1%
He Shen (M)	31%	13%	62%	31%
Jia Cai (F)	45%	69%	61%	16% (max 24%)
Luyu Wang (F)	71%	85%	81%	10% (max 14%)
Ying Li (F)	N/A	N/A	N/A	N/A
Yu Ma (M)	31%	59%	81%	50%
Shan Lu (M)	40%	15%	81%	41%
Geng Tian	37	N/A	N/A	N/A

Table 6.1. Scoring rates for the individual tests.

Noteworthy items in Table 6.1 are that the students who started with low scores in Test 1 show substantial progress in the final result, ranging from 31 to 50 percent increases. The students who started with fairly high scores still show some increase in pronunciation recognition. This is to be expected as it seems intuitively correct to assume that going from high to higher is more demanding than going from low to high.

It is also noteworthy, but inexplicable at this point, that two students showed a significant decrease from Test 1 to 2, eventually achieving a very high score at the end of the testing period.

When the results are plotted and viewed in terms of hourly progress, the results look as outlined in Figure 6.3.

Although none of the subjects reached the results of the Danish native-speaker baseline, all curves are increasing in pronunciation recognition to a degree which clearly surpasses that of the control group. In total, the initial average score for the Chinese experimental group was 46% (horizontal line from 46 percent) which increased to 73% (horizontal line from 73 percent) after a maximum of 21 practice hours.

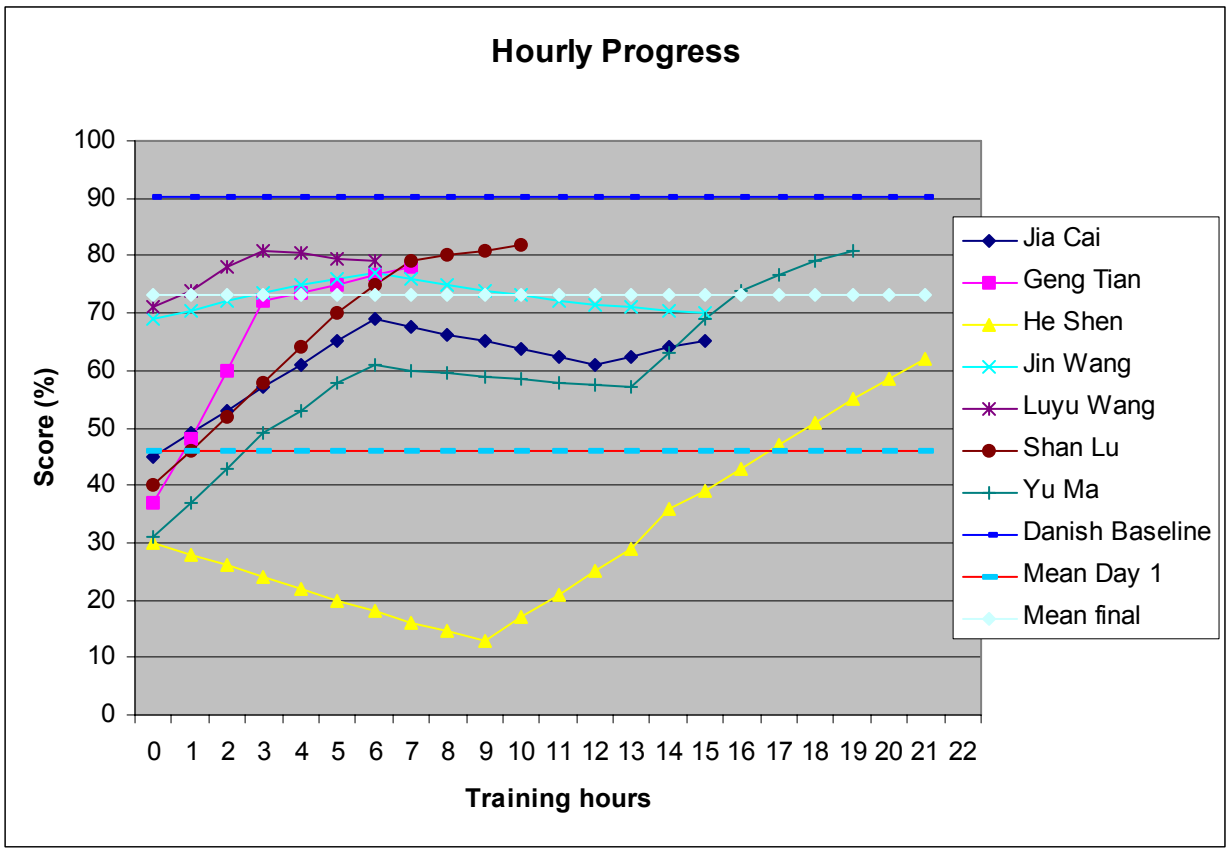

Figure 6.3. Hourly progress. Note that only the major curve break points are informative.

7 Conclusion

We have presented the status of the Danish pronunciation trainer prototype by June 2004 and results of testing the system with a group of Chinese learners of Danish. The results suggest that second-language learners of Danish can use the system through self-training to rapidly achieve very substantial progress in their pronunciation of Danish words. The quantitative results reported will be supplemented shortly by a qualitative evaluation of the students' progress. In addition, the system will be used for training 30 students from Finland in Danish pronunciation in the summer of 2004, providing valuable data for the continued evaluation of the system.

8 Acknowledgements

The work described in this report was supported by grants from Kommunernes Landsforening/Momsfonden, the Ministry of Immigration and Integration, and the Center for Immigration, the city of Odense. We gratefully acknowledge the support. We would also like to thank Laila Dybkjær, NISLab, who kindly provided her comments on an earlier draft.