Forschungsberichte des
Instituts für Phonetik und
Sprachliche Kommunikation
der Universität München
(FIPKM) 35 (1997) 107-114

Comparison of Commercial Dictation Systems for Personal Computers

Susanne Burger and Hans G. Tillmann
Institut für Phonetik und Sprachliche Kommunikation
Ludwig-Maximilians-Universität
Schellingstr. 3
D - 80799 Munich

Abstract

Great progress in developing speech technology has made the provision of affordable speech recognition software with little memory requirements for the general user possible. The objective of a small project we have recently completed was to compare two different so-called "Dictation Systems" available also for the German language and to test their respective capabilities to recognize language as well as their usefulness when applied in a business world environment.
We conclude that both dictation systems can be expected to function at a word recognition rate of about 98% with vocabulary known to the system after a comparable training period.
On the other hand the effective application of the versions tested depends highly on the function desired by the user; in addition, the technique of speaking discretely needs getting used to.

Introduction

In the summer of 1996 a consulting company asked the institute to test quality and usefulness of commercially available dictation systems. For this task the producers of the dictation systems Dragon Dictate and IBM Voice Type provided the institute with their products free of charge. At the time these two systems were the only ones available for the German language at a comparable price and quality level. The paper gives a short summary of the work and the results of the project.

1 What are Dictation Systems?

A Dictation System offers to a PC user the opportunity for speaking text straight into a word processor instead of typing it by hand. Also, the user may control functions of his desk top per voice.
To do this both systems use a so-called engine operating on the basis of Hidden Markov Models and simple language models: our systems used trigram and respectively bigram statistics. The actual versions of these systems are speaker adaptive dealing with a speaker independent set of models containing a large vocabulary (the tested versions contained 30.000 words) which is adapted to the pronunciation of the individual user in a training period with small subsets prior to regular usage. In use the systems improve their models via the corrections the user makes in case of false recognitions. Both systems require a discrete speaking mode (e.g. small pauses between the words while speaking) for recognizing single words. They try to circumvent this when allowing fluent speech with common phrases like greeting phrases.
The advantage of this technology is a relatively low price and low consumption of storage capacity, the disadvantage is the need of the intensive learning phase before first efforts in using the system and a high capability of adaptation required on the part of the user.

2 Task of the Project

The basic question was whether the software at the present level of development can be of effective use for a big company, e.g. whether time and manpower may be economized.
The following aspects were considered in detail
- recognition rate
- dictation speed
- time requirements for learning and training.

3 Equipment and Test Data

All tests were carried out in a quiet office environment, in particular with continuous humming of appliances like PCs and background noise from other office staff.
All tests were carried out by a female test person (the first author of this report). She was familiar with the basic features of both systems as well as with the special discrete speaking mode. Also she was familiar with the most important commands to operate the systems.
The personal computer in use was a notebook Toshiba 610CT Protegé, with a 90 Hz Pentium CPU, 16 MB RAM and a 800 MB harddisc. The system was operated with Windows 95.
The following two dictation systems were tested:

1. Dictation System of **Dragon Dictate**, Version 1.2 by Dragon Systems GmbH (Referred to in the following also by Dd)
2. Dictation System of **IBM Voice Type**, Version 3.0 as a Not-For-Sale-Version by IBM (Referred to in the following also Vt)

The microphones included in the system packages not being satisfactory a comparable Sony microphone was used in the tests. This microphone could be used without need of any adapters and produced the best signal.
The original Dragon Dictate Shure microphone produced only a weak signal with the Toshiba notebook PC despite the adapter delivered with the package.
The original Voice Type Andrea microphone produced nearly the same quality of results as did the Sony, but only with the battery adapter delivered with the package.

3.1 Starting Basis

3.1.1 Standard Initial Start-up

- Dragon Dictate: Tuning of microphone, training (400 words, about 30 minutes)
- Voice Type: Tuning of microphone, testing of the language model (5 words to be spoken twice, about 3 minutes)

3.1.2 Dictation Application

Both systems may be used for dictating in combination with various word processors. For the tests those processors were chosen generating the least difficulties.
Dragon Dictate: Microsoft Word 7.0
Voice Type: IBM Voice Pad

3.1.3 Test Vocabulary

During the test a business-management oriented vocabulary was exclusively dictated. The example texts for the tests were taken from a business consulting environment.

3.1.4 Dictation Procedure

All texts dictated were not spoken freely but read from paper.

3.1.5 Correction

The correction of misspelled or wrongly recognized words was undertaken using the keyboard. The correction by the use of voice is possible but may lead to further mistakes in speech recognition.

4 Description of the Test

Three types of tests were carried out:

Test A: Improvement of Recognition Rate
How does the recognition rate improve during the first six dictated pages?
In this test each page dictated contained different texts about business-management themes.

Test B: Learning Effects through Corrections
How does the recognition rate improve after the first six pages dictated?
In this test the same text (392 words) was dictated seven times, also containing business-management vocabulary, to test the system's ability to learn from corrections and to see which recognition rates may be reached with familiar vocabulary.

Test C: Time Factor
How many words per minute may be dictated?
The text used in Test B was typed by a professional secretary. The secretary's time required for typing, correction and verifying once was taken as a reference. In test B the time was taken from the beginning of the dictation with the system until the end of the correction process.

5 Results

5.1 Results of Test A

- Dragon Dictate: Despite the intensive training phase the system could recognize only 68,24% of the first text without errors. After the fifth page the system reached a recognition rate of 91,16%. The slight drop after the sixth page is due to the testing speaker, while all other tests with new texts remained at about 90 %(see figure 1).
- Voice Type: Although this system requires no training it already recognized 85,86% of the text of the first page. After the fourth page the system also reached a recognition rate of about 91% (see figure 1).

Summary:
After 4-5 pages the rate of speech recognition remains in the 90-94% range. As long as the vocabulary of a certain area of usage is not fully trained new words are continuously added causing recognition errors. All other errors are primarily caused by the speaker. After an extended period of dictating a minor drop of the recognition rate may occur because

a) the exact pronunciation suffers due to the return to normal speech rather than maintaining the initial effort to dictate carefully
b) the voice will be tired and less powerful after two to three hours of dictating.

Both systems will also become accustomed to these variants after a further couple of pages.

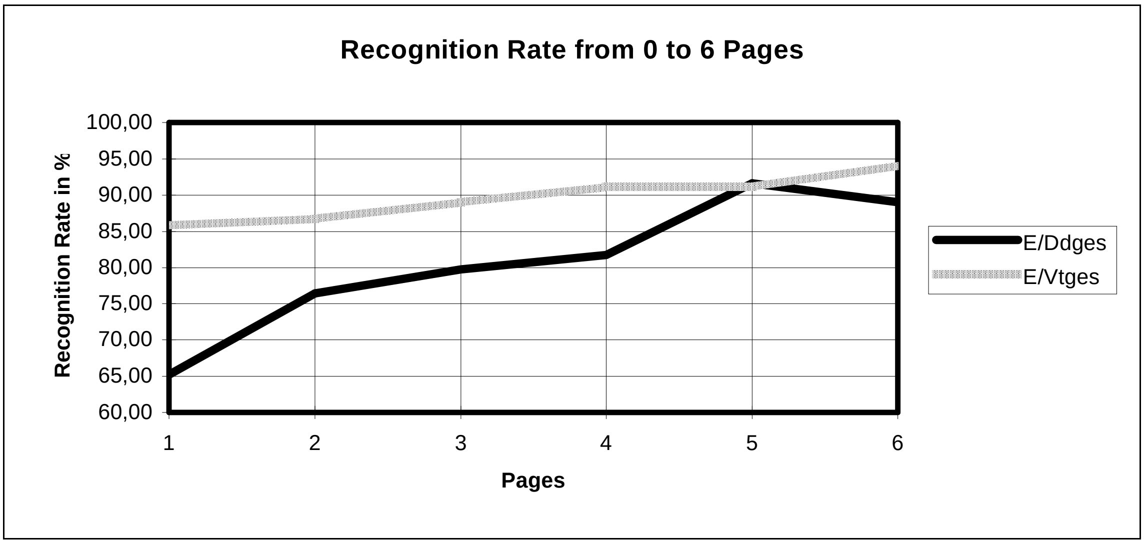

Figure 1: Recognition rate in % starting with zero pages dictated to six pages - each page containing a different text; E/Ddges = Dragon Dictate, E/Vtges = IBM Voice Type

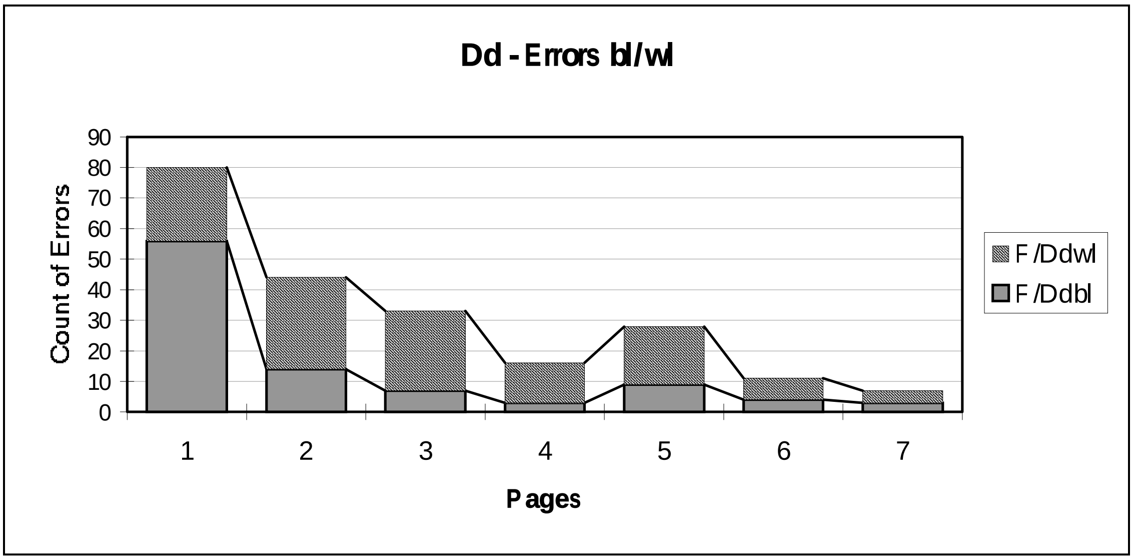

Figure 2: Dragon Dictate - Number of recognition errors per page; F/Ddwl=correctly spelled word in word list, F/Ddbl=correct new word entered by keyboard

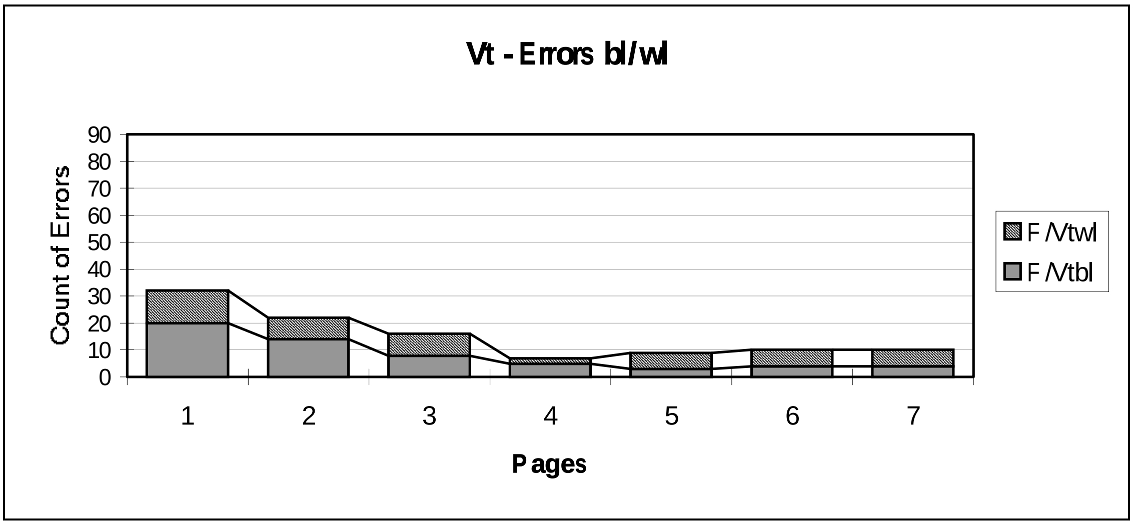

Figure 3: Voice Type - Number of recognition errors per page; F/Vtwl=correctly spelled word in word list, F/Vtbl=correct new word entered by keyboard

5.2 Results of Test B

5.2.1 Minor and Major Errors in Speech Recognition

In general in both systems two levels of recognition errors may be distinguished:
wl (Word List): (minor recognition error) A word dictated is not recognized but appears in a word list from which it may be chosen by either clicking it (Voice Type), or using an oral command (Dragon Dictate).
bl (Spelling List): (major recognition error) A word dictated is not recognized and also does not appear in a word list. In this case the word has to be typed by hand. (This may also be done by voice, but in the test only the keyboard enhancement of vocabulary was exercised to avoid further recognition errors.)
In the following diagrams (fig. 2 and 3) there occur a change in the ratio of the two error types to the benefit of error correction on the basis of word lists as dictation progresses.

5.2.2 Recognition Rate in Test B

- Dragon Dictate: The rate of recognition was 79,60% after the first page and reached more than 95% after the fourth. After the seventh attempt the result of 98,22% was even better then the best result of Voice Type (see figure 4).
- Voice Type: The rate of recognition reached 91,84% after the first page and remained stable at 97,5% after the fourth page (see figure 4).

General remarks: After the fifth page the settings for the speech rate were altered to a higher speed which resulted in a minor drop of the recognition rate with Dragon Dictate. With Voice Type also the rate droppped slightly. A 100% recognition rate was never reached with either system, even though texts and vocabulary were known to both systems. This is due to small variations in pronunciation. From the sixth attempt both systems functioned at about the same recognition rate.

5.3 Results of Test C

As a reference for comparing the time involved the text used in test B was typed by a professional secretary and the time was taken (the Z/gerwmin line in figure 5). For reasons of comparability this was done without any special formatting and the time for verification and corrections was added to the time used for typing. The reference time was 26,13 words per minute.

As figure 5 shows both systems reached only 21,77 words per minute beginning with the fourth page dictated due to more corrections necessary on the first three attempts of test B. In the fifth attempt this value dropped again due to the change of the speed settings and the resulting higher number of errors. Despite the higher speed setting which actually should result in a higher rate of words dictated per minute this rate stabilized and remained at about 21 words per minute. In this context the individual speech rate of the person dictating has to be considered.

5.4 Summary of Test Results concerning Recognition Rates

In general it may be stated that Voice Type reaches high recognition rates in shorter time and without training in comparison to Dragon Dictate. But both systems reach about the same recognition rate after 10 pages if errors are consistently corrected.

The final result of 98% with a familiar text is extremely high; only by exercising a precise pronunciation an even higher rate could be reached. With unfamiliar texts with vocabulary from a familiar domain the level rate of 90% may even be exceeded because with time less and less words will remain unfamiliar and the systems will keep adapting to the speech of the individual user.

5.5 Summary of Test Results concerning Time Requirements

In view of the fact that Voice Type requires no training of speakers with no strong dialects, the words per minute ratio is slightly better in the beginning than with Dragon Dictate. The time requirement for one page after the tenth dictation remains constant with both systems at about 18 minutes for 392 words or 21 words per minute. After further use the systems will adapt to higher speaking rates so that more words may be dictated per minute. Both systems are slower than the secretary; on the other hand misspellings are practically impossible because every word is automatically compared with the wordlists and the spelling-checker.

6 Dictation and Correction in Comparison

6.1 Voice Type using IBM VoicePad

Dictating and correcting are two separate sequential steps with IBM Voice Type. This implies dictating in the dictating mode and correcting afterwards when this mode is turned off. The system is based on a language model based on trigram statistics. It always decides three dictated words later which word is to be written. By this method words may appear wrong at first and then be corrected later on. In the beginning this appears confusing to the user.

6.1.1 Dictation

Texts must be dictated discretely. Minimal pauses must be made between words so that the system may identify on the one hand words and their boundaries and on the other hand system commands always consisting of two words spoken without a pause.
By separating the two steps dictation and correction, entire pages may be dictated without interruptions, nevertheless, the system needs a pause every six to eight words for processing. As the discrete dictating is unfamiliar to the user at first, it appears tiring.
Generally it is not necessary watch the screen while dictating.

Features:

- continuous dictation of entire pages
- no visual contact necessary while dictating
- punctuation
- extra mode for numbers
- extra mode for letter spellings
- macros may be applied

Problems:

- no cancel command for already dictated words while dictating (like "cross out the last word")

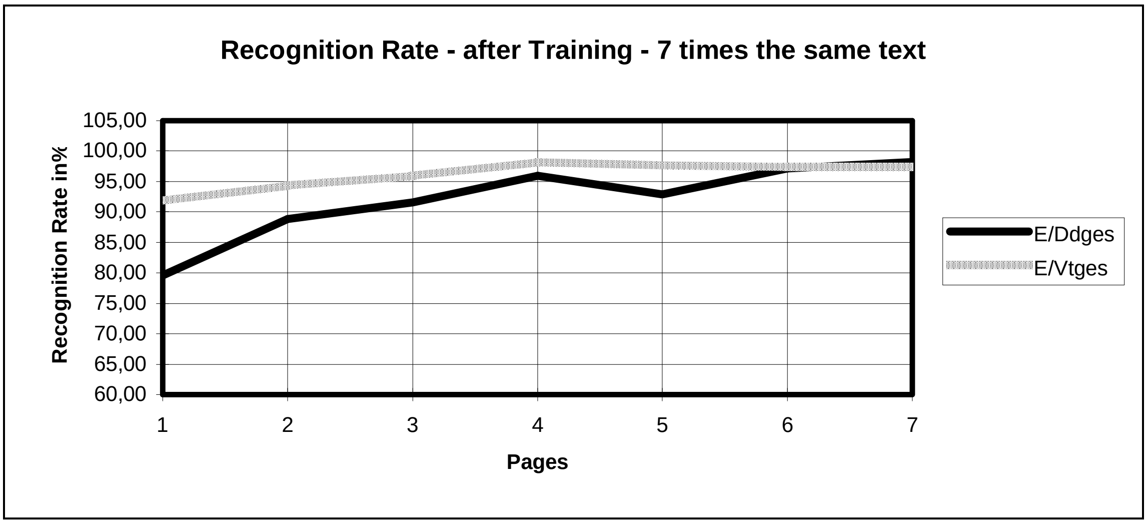

Figure 4: Recognition rate in % after six pages dictation and seven dictations of one and the same page; E/Ddges=Dragon Dictate, E/Vtges= IBM Voice Type

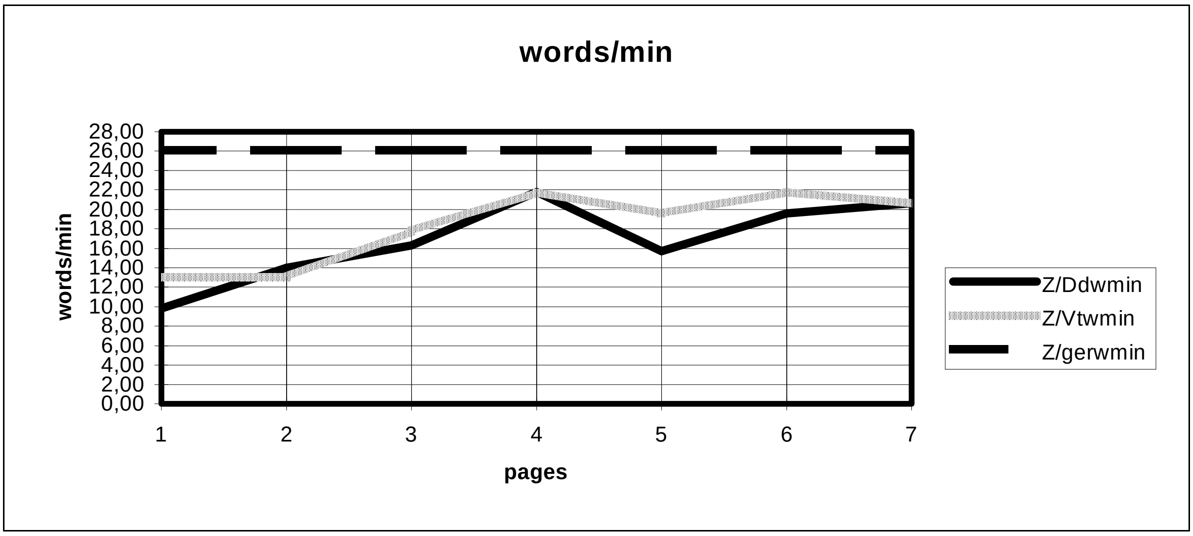

Figure 5: Number of words dictated per minute per page; Z/Ddwmin=Dragon Dictate, Z/Vtwmin= IBM Voice type, Z/gerwmin =secretary

- unfamiliar composition of compound words by use of macros: "Binde-s", "Bindewort"
- complex formats have to be defined in macros
- all other noise e.g. produced by breathing, coughing, throat clearing creates words.

6.1.2 Correction

The correcting phase begins after the end of the dictation. The defectively recognized words are clicked at and the dictated words can be heard again.

There are two different levels of recognition errors.

1. word list:
   The correct word appears in a listing of choices and may be selected by mouse click or by voice.
2. entry of a new word:
   The correct word does not appear in the listing and must be entered by the user. Voice Type will now try to find the word in its vocabulary:

- If the word is found in the vocabulary the system gives a message.
- If the word is not contained in the vocabulary or the pronunciation differs strongly from the pronunciation familiar to the system it will try to add the word to its vocabulary. This requires an entry of a description of the pronunciation, which often fails because orthography in some cases does not describe the sound correctly and it may also be unclear which letter Voice Type uses for which sound. Again the system will try to add the word to the vocabulary.
  Under some circumstances the system may not be able to interpret the description of a pronunciation. A second attempt may solve the problem or the addition of the new word has to be left out.

It seems impossible to operate the system completely without the use of the keyboard. The correction of exceptional words or abbreviations is complicated and may even lead to system failures. One solution may be the application of especially designed macros for certain abbreviations. The correction of faultily recognized format commands or punctuation marks has so far proved impossible and always resulted in system failures. The reference manual also gave no explanations on how to deal with this.

Features:

- by clicking at a dictated word the user can listen to what he had said earlier
- alternate words may be selected by mouse click
- deleting of entire words or noises is possible
- a special training of single words is possible

Problems:

- words changed by word processor action cannot be corrected by the dictation system
- completely unrecognized words must be corrected using the keyboard
- words not recognized by the system must be enhanced by a description of pronunciation
- the description of one's own pronunciation is by no means trivial
- the correction of words easily leads to system failures with mistyping, abbreviations, extraordinarily long words, punctuation marks or system commands.

6.2 Dragon Dictate using Word 7.0 for Windows

Dictating and correcting are integrated into the same process with Dragon Dictate. This implies that the user must correct a faultily recognized word immediately or a couple of words later. Therefore it is necessary to watch the screen while dictating. Dragon Dictate decides which word to choose after the following word is dictated (bigram statistic).

6.2.1 Dictation

Here too, words are to be dictated discretely while system commands are spoken contiguously. The system is not quite as sensitive towards breathing noise or throat clearing. Because of the necessity of watching the screen while dictating the discrete pronunciation appears not as strange as with Voice Type.

Features:

- formats like "new line"/"new paragraph" while dictating
- extra mode for numbers
- extra mode for letter spellings
- many format commands
- application of macros is possible
- delete command for already dictated words during the dictatation ("Streich das")

Problems:

- the user must watch the screen
- frequent switching between dictation and spelling mode
- unfamiliar composition of compound words by use of macros ("Wort-Beginn", "Wort-Ende")

6.2.2 Correction

With every word dictated a list of alternative words is presented. Correction is necessary if the correct word does not top the listing, which implies that it was not recognized. There are two levels of recognition errors:

1. word list:
   The correct word appears in a listing of choices and may be selected by mouse click or by voice.
2. entry of a new word:
   The correct word does not appear in the listing. The user gives the command to enter the spelling mode and may spell the correct word either by voice or keyboard.
   - cases where the correct word is in the listing, generally only the first three letters must be given until the correct word tops the listing.
   - cases where the word is not contained in the vocabulary, it must be entered completely. Here Dragon Dictate offers assistance by giving a similar word which needs altering.
   - cases where the pronounced commands for selection, correction, spelling or format are not recognized, these have to be corrected, too. In comparison to Voice Type this is fairly easy and also improves further recognition of these commands.
   - Dragon Dictate also offers the possibility to correct errors which occurred earlier during dictation. By means of the "hoppla" command the system enters a separate line mode. This is only recommended for up to ten words, especially because Dragon Dictate does not have a facility for listening to the dictated text and often the user will not remember what he had said.

Dragon Dictate may function completely without the use of the keyboard after only a short time. The first corrections should be performed by using the keyboard, because in the beginning there may be frequent incorrect recognition of commands as well. Correction in the various modes is relatively confusing and time intensive.

In this system the correction of unusual words, abbreviations and other punctuation is easy. All entries are taken over into the system's vocabulary. On the other hand, because of this misspellings may also easily be adopted.

Features:

- word alternatives may be selected by voice or mouse click
- refusal of mispronunciations or noises is a possible system feature
- simplifications of correction procedure; often the entry of the first three letters is enough to select the correct word
- the correction of abbreviations, extremely long words, punctuation, system commands is straightforward
- the training of single words is possible

Problems:

- long text passages may not be edited supported by speech recognition
- it is not possible to listen to the dictated text
- switching between modes requires a lot of "talking" and patience.

7 Comparison of Installation and other Functions

7.1 Installation

IBM Voice Type:	Dragon Dictate:
CD-ROM, good documentation. Installation of microphone: the documentation by the manufacturer is confusing.	7 disks, a dongle secures copyrights, good documentation. Installation of microphone: the documentation by the manufacturer is incomplete.

For the test we received the CD-ROM version from IBM and the disk version from Dragon Dictate. Both systems are also available in the other version.

Dragon Dictate only functions when the dongle for copy security contained in the package is hooked up with the parallel interface of the PC.

The first installation and tuning of the microphones takes time and requires knowledge of the Windows sound system. The installation guides are a bit cryptic, so that some experimentation is involved.

Before the system may be started, various settings and tests must be done with the sound system of the PC. Both system packages provide various adapters to amplify or damp the microphone signal. Whether and which adapter has to be applied can only be found out by trial and error. For the ideal tuning of the microphones the Windows Audio settings have to be modified for recording mode. Voice Type provides a little program which adjusts the system to the microphone settings. With Dragon Dictate the signal is either too strong or too weak and the Windows Audio settings must be modified. It is highly recommended to either save the ideal settings or memorize them precisely.

7.2 Manual

IBM Voice Type:	Dragon Dictate:
Learning Program "Erste Schritte" (30 minutes), Parts 1 and 2 of the reference guide	Learning Program "Drache Martin" (30 minutes), reference guide 1 "Introduction"

Both systems provide an introductory presentation. These presentations are helpful during the first encounters with systems. Further introductory advice may be found in the first chapters of the reference guides or, for Dragon Dictate, the special introduction book. Also both systems provide an extensive on-line help.

In Voice Type's presentation a speaker explains the first steps and system features. Various sections may be selected by mouse click, e.g. the positioning of the microphone, dictation or correction. The complete demonstration takes approximately 30 minutes.

Both introductions, reference guide and on-line, are quite detailed. Nevertheless it is essential to read both for better understanding, because some late entries in the online references are not yet printed in the book version. This may be due to the Not-for-Sale-Version used in the test.

Dragon Dictate is introduced on screen by an animated cartoon figure ("Martin, the dragon"). Here too, it is possible to choose from various chapters of the introductory presentation. The presentation is interactive in the sense that the user is asked to repeat and speak commands. The complete demonstration also takes about 30 minutes. The ideal preparation for work with the system is achieved by studying the printed reference guide intensively.

7.3 Training

IBM Voice Type	Dragon Dictate
- a real training is not really necessary - short tests for installing an individual speech model - optimizing by constant corrections - optional training: a minimum of 50 out of 252 sentences have to be read, duration about 30 minutes	- training: three different groups of words, about 400 words have to trained - the more intense the training the better the recognition rate - intensive training takes about one hour. - optimizing by constant corrections.

With Voice Type intensive training is not really necessary. The system offers every user a little program for individual initialization, which asks the user to speak five different words for tuning the microphone and the same five words for the system to adapt to the speaker. An intensive training - called registration - may also be executed. Then at least a minimum of 50 or for an intensive training the complete set of 252 sentences must be spoken.

This registration is necessary for users
- speaking strong dialect or with an accent
- with husky or hoarse voice
- with articulatory handicap

In Dragon Dictate's Version 1.4 intensive training is still required. The company announced that the new version 2.0 would need no more intensive training.

The training may be done with about 400 words and commands offered by the system, which have to be spoken from one to five times. The best recognition rates were reached by repeating every single word until the system recognized them without difficulties. This type of training requires about one hour. New vocabulary may be added by this procedure.

7.4 Additional Application Options

Both systems have a central menu offering the key functions training, settings, command and vocabulary listings.
Voice Type offers four different dictation features:

- Fast dictation: a simple text editor is used for the dictation of short notes and messages or numbers for a spread sheet, which may be transferred to other applications. The editor has no word processing functions.
- Voice Pad is the IBM equivalent to Windows WordPad, which has integrated word processing functions.
- Dictation using MS-Word: In this mode the dictation is written in MS-Word. All Word system and format commands may also be spoken. The problem was that system failures occurred frequently in Word, when it was used with Voice Type (This may be a problem with WinWord 7.0)
- Voice Type direct: Dictations may be made directly to most of the Windows applications.

Dragon Dictate: In general the system may be operated with any Windows application. Mouse control and navigation may be exercised by voice commands.

7.5 Vocabulary

Both systems were tested in a version containing 30.000 active and 200.000 passive words. New words overwrite seldom used ones in the active vocabulary.

The size of the active vocabularies seemed totally satisfactory: Both systems knew surprisingly many words of the business-management vocabulary used in the tests (Although it may be stated that Dragon Dictate's vocabulary seemed to fit this domain better than Voice Type's). The addition of new words was only necessary at an acceptable volume (about three to four words per page containing unknown text, mostly special business abbreviations). A larger active vocabulary might be necessary for areas of specific terminology as in the domains of medicine, law or prose writing.

7.6 Overview of the most important Features of the two Systems

IBM Voice Type	Dragon Dictate
- microphone with good performance (when operated with a laptop a battery adapter is necessary)	- microphone with poor performance
- training is not required but possible	- training required
- oriented to word passages	- oriented to single words
- dictation may be replayed	- no replay of the dictation is possible
- a lot of keyboard involved, less voice required	- little keyboard involved, more voice required
- no visual screen contact required while dictating	- visual screen contact at least in the beginning phase is essential
- correction mode is straightforward	- correction mode is complicated
- correction is susceptible to errors	- correction mode provides assistance
- generalized vocabulary (also convening business management vocabulary)	- vocabulary oriented to a business management environment
- new vocabulary may be created but not entered directly	- new vocabulary may easily be created and added
- system requirements: Pentium 90 and up, 16 MB RAM minimum, 60 MB harddisc space, Win95 and up	- system requirements: Pentium 90 and up, 16 MB RAM minimum, 30 MB harddisc space, Win3.1 or Win95

8 Summary

In conclusion, a double training is required with either system: First the user has to be trained to handle the systems and learn the discrete speaking mode. On the other hand the systems have to be trained to the individual pronunciation of its user. With both systems the recognition rates are surprisingly good. Both systems recognized about 93% when occasionally confronted with unknown words from a common vocabulary after the first five pages had been corrected consistently. When using only known words the systems reached about 98%.
With their dictation speeds both systems can almost keep up with a professional secretary at about 21 words per minute (including time necessary for corrections).
The term "dictation" is a bit misleading; actually the user only reads out aloud a prepared text to the computer. The field of possible applications was limited to input of prepared texts. The client company instructing the project decided to wait for further software and system development. The preparations required prior to system usage asked for too much patience and time. Learning the discrete speaking mode appeared non-efficient. The main disadvantage however is that dictation during the creative process of writing, with e.g. deleting and rephrasing, still remains too strenuous with the dictation systems tested.
Both systems seem to be a useful alternative to typing prepared texts by hand. The original version of this project report was dictated with IBMVoiceType; for editing and alterations Dragon Dictate was used, because here format and control commands are easier to apply.
A final decision on which system to choose is a question of taste and individual preferences. Regarding their performances both systems appear to be more or less of equal quality.

Literature

[1] H. Mangold (ed.), "Sprachliche Mensch-Maschine-Kommunikation", R. Oldenbourg Verlag, München, Wien, 1992

[2] D. Tynan, "What You Say Is What You Get", PC World Online, Januar 1995
http://www.pcworld.com/reprints/kurzweil.htm

[3] M.H. Vincken (ed.), "Special Issue on Speech Technology", Philips Journal of Research, Vol. 49, No. 4, pp. 315-494, Philips International BV, Eindhoven, Netherlands, 1995

[4] T.J. West, "Implementing and Evaluation COTS Voice Recognition Software", Naval Postgraduate School, Monterey, USA, 1996
http://dubhe.cc.nps.navy.mil/~fargues/research/tjwest/thesis~1.htm

[5] Dragon Systems Homepage,
http://www.dragonsys.com

[6] IBM Voicetype Homepage,
http://www.software.ibm.com/is/voicetype