Speech Recognition in daily Hospital practice: Human-Computer Interaction Lessons learned

*Andreas Holzinger[1], Siegfried Ackerl, Gig Searle, Erich Sorantin**

Institute of Medical Informatics, Statistics and Documentation (IMI)
*) Department of Radiology
Medical University of Graz
A-8036 Graz Austria
[1]andreas.holzinger@meduni-graz.at

Abstract

Dictation is the most common use of Automatic Speech Recognition (ASR) systems today. This includes medical transcriptions, legal and business dictation, as well as general word processing. In some cases special vocabularies are used to increase the accuracy of the system. Due to the continuously rising costs and stiffening competition in the health care industry, doctors and hospitals need to find effective ways of reducing spending and increasing efficiency whilst still providing high quality care for their patients. One of the main criteria in providing this care for patients, e.g. in a department of Radiology, is the speed with which the radiological finding can be supplied, together with the X-Ray. Therefore we implemented and realized a PC based software solution in the Radiology of the 2300 bed Graz University hospital and carried out experiments since 1998 using Philips SpeechMagic™ speech recognition technology, with the aim to eliminate the need to manually transcribe medical reports in the traditional way. According to Shneiderman (2000) spoken language is effective for human-human interaction but often has severe limitations when applied to human-computer interaction: Speech is slow for presenting information, is transient and therefore difficult to review or edit, and interferes significantly with other cognitive tasks. But most of all we found that most medical doctors were not willing to invest a lot of time and effort in this technology, which is absolutely necessary for getting satisfying results.

"In dictation, users may experience more interference between outputting their initial thought and elaborating on it." (Ben Shneiderman, 2000)

1. Introduction

The aim of this project was the implementation of a PC based Software Solution for the generation of digital medical texts, such as examination results (findings). The text can be dictated directly into the PC instead of onto tape. This is particularly important when a PACS system (Picture Archiving and Communication System) is installed in the clinics with corresponding satellites (BAUMAN et al., 1996).

The time between the dictation onto tape and typing the finding into the RIS (Radiology Information System) is particularly noticeable during night duty. Also, the release of radiological pictures is often linked to the RIS finding, causing occasional delivery delays. With digital dictation, the X-Ray can be released to the Medical doctor (MD) immediately dictation is complete. In this case the MD is able to listen to the spoken results while viewing the pictures on the PC (via PACSview). The same digital text can be typed, or merely corrected, and transferred into RIS at a more convenient time.

2. Speech recognition systems

Speech recognition (cf. e.g. (HOLZINGER, 2002a)) is the process by which a computer identifies spoken words. Basically, it means "talking" to a computer, and having it correctly recognize what was said. Although progress has been impressive, there are still many hurdles that speech recognition technology must clear before ubiquitous adoption can be realized (DENG and HUANG, 2004). The following definitions are the basics needed for understanding speech recognition technology (RUDNICKY et al., 1994), (HOLMES and HOLMES, 1999), (JURAFSKY and MARTIN, 2000).

Utterance
An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences.

Speaker Dependence
Speaker dependent systems (SDS) are designed around a specific speaker. They are generally more accurate with the speaker for whom they are designed, but much less accurate with other speakers. It is assumed that the speaker will speak in a consistent voice and tempo. Speaker independent systems (SIS) are designed for a variety of speakers. Adaptive systems usually start as speaker independent systems and utilize training techniques to adapt the program to the speaker in order to increase recognition accuracy.

Vocabularies
Vocabularies (or dictionaries) are lists of words or utterances that can be recognized by the SR system. Generally, smaller vocabularies are easier for a computer to recognize, while larger vocabularies are more difficult. Unlike normal dictionaries, each entry does not have to be a single word.

Accuracy
The ability of a recognizer can be examined by measuring its accuracy – or how well it recognizes utterances. This includes not only correctly identifying an utterance but also identifying if the spoken utterance is not in its vocabulary. Good ASR systems have an accuracy of 98% or more. The acceptable accuracy of a system really depends on the application.

Training
Some speech recognizers have the ability to adapt to a speaker. When the system has this ability, it may allow training to take place. An ASR system is trained by having the speaker repeat standard or common phrases and adjusting its comparison algorithms to match that particular speaker. Training a recognizer usually improves its accuracy. Training can also be

used by speakers having difficulty speaking, or pronouncing certain words. As long as the speaker can consistently repeat an utterance, ASR systems with training should be able to adapt.

Recognition systems can be broken down into two main types. Pattern Recognition systems compare patterns to known/trained patterns to determine a match. Acoustic Phonetic systems use knowledge of the human body (speech production, and hearing) to compare speech features (phonetics such as vowel sounds). Most modern systems focus on the pattern recognition approach because it combines nicely with current computing techniques and tends to have higher accuracy. Most recognizers can be broken down into the following steps:

1. Audio recording and Utterance detection
2. Pre-Filtering (pre-emphasis, normalization, banding, etc.)
3. Framing and Windowing (chopping the data into a usable format)
4. Filtering (further filtering of each window/frame/freq. band)
5. Comparison and Matching (recognizing the utterance)
6. Action (Perform function associated with the recognized pattern)

Although each step seems simple, each one can involve a multitude of different (and sometimes completely opposite) techniques (PICKETT, 1999; HOLZINGER, 2002b).

(1) Audio/Utterance Recording: This can be accomplished in a number of ways. Starting points can be found by comparing ambient audio levels (acoustic energy in some cases) with the sample just recorded. Endpoint detection is harder because speakers tend to leave "artifacts" including breathing/sighing, teeth chatters, and echoes.

(2) Pre-Filtering: This is accomplished in a variety of ways, depending on other features of the recognition system. The most common methods are the so called "Bank-of-Filters" method that utilizes a series of audio filters to prepare the sample and the Linear Predictive Coding method that uses a prediction function to calculate differences (errors). Different forms of spectral analysis are also used.

(3) Framing/Windowing involves separating the sample data into specific sizes. This is often rolled into step 2 or step 4. This step also involves preparing the sample boundaries for analysis (removing edge clicks, etc.)

(4) Additional Filtering is not always present. It is the final preparation for each window before comparison and matching. Often this consists of time alignment and normalization.

There are a huge number of techniques available for (5), Comparison and Matching. Most involve comparing the current window with known samples. There are methods that use Hidden Markov Models (HMM), frequency analysis, differential analysis, linear algebra techniques/ shortcuts, spectral distortion, and time distortion methods. All these methods are used to generate a probability and accuracy match.

3. The system used: Philips SpeechMagic™ 4.0

This system includes a Software Development Kit (SDK), which allows the integration of the Philips´ professional speech recognition engine into any application through traditional C style APIs or ActiveX controls (including user interface modules). It is designed for integration into third-party professional market-specific solutions for the medical, legal and insurance markets. As input devices we used the standard Microphone SpeechMike Pro (SM6174). SpeechMagic™ supports English US, English UK, French, German (including the new German orthography), Austrian, Dutch and Flemish. Professional ConTexts are available with common terms and abbreviations for a particular profession (medical, legal, insurance). SpeechMagic provides high sound compression (CELP – 19.2 kBit/s) to transfer sound data over band-limited channels with high recognition rates. It can process the following sound file formats: PCM 16khz / 16bit - 256 kb/s (PC), CELP 16khz / 16bit – 19 kb/s (PC default format), PCM 11khz / 16bit – 176 kb/s (mobile), DSS Standard Mode- 13 kb/s (mobile), PCM 8khz / 8bit and 16bit – 64/128 kb/s (telephone) and CCITT μ-law, 8khz / 8bit – 64 kb/s (telephone).

Microphones

A quality microphone is the key to utilizing ASR. In most cases, a desktop microphone just won't do the job. It tends to pick up ambient noise, giving ASR programs a hard time.

Hand held microphones are not the best choice as they can be cumbersome to hold continually. While they do limit the amount of ambient noise, they are useful in applications that require changing speakers often, or when little speaking is done (e.g. when wearing a headset is not appropriate).

The best choice, and by far the most common is the headset style which allows the ambient noise to be minimized, while allowing to have the microphone at the tip of the tongue all the time.

The integration into the existing system

Integration via ActiveX controls or C/C++ API SpeechMagic™ provides three different C/C++ interfaces for integration into a professional solution: SpeechMagic™ API (SmApiV3) – Standard workflow API to pass dictations to speech recognition and administration SNCAPI – Standard dictation API to integrate into your professional solution Correction API (CspApi) – to integrate correction facilities into any word processor. It also provides standard ActiveX Controls for Digital dictation, standard workflow and Dictation list for starting correction.

We used the Windows NT 4.0 operating system, (The installation of SP5 is absolutely necessary) Microsoft Word 2000, Windows NT 4.0 Network operating system (file server), CPU Intel Pentium® III with 500 MHz, 256 MB RAM and 4 GB hard disk space.

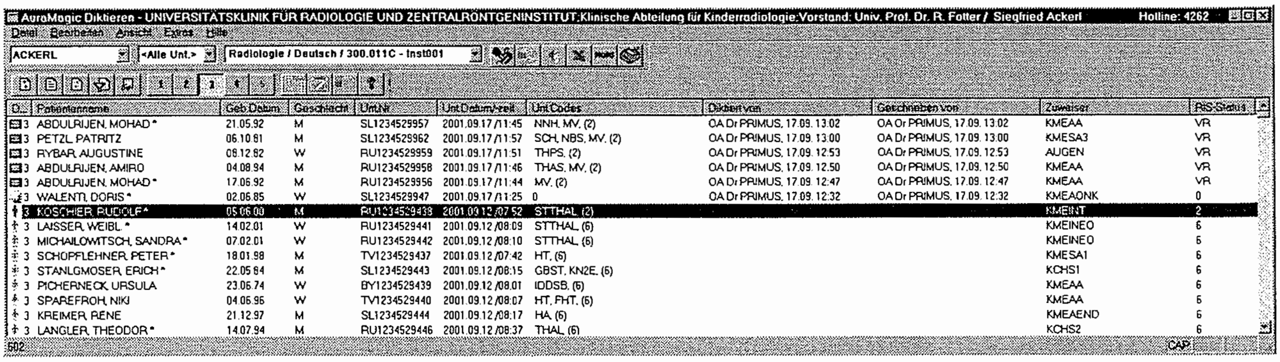

Figure 1: A look at the user interface (shown data anonymized), explanation below

Explanation of symbols (from upper left to lower right):

Titelleiste:	University Department, Head, current user, Hotline phone number
[ACKERL]:	name of the current medical doctor
[<Alle Unt>]:	list of certain examination types
[Radiologie / Deutsch...]:	context for speech recognition
Mikrofon:	starting the speech recognition
Druckersymbol:	printing of a result
Lautsprechersymbol:	listening to created result
Excelsymbol:	output of the shown list in MS Excel
[HOLD]:	system lock, logging in of the medical doctor is necessary
Druckersymbol:	default printer setting
Symbol mit zwei „+“:	notes to the patient (already examined)
Blattsymbol:	new patient
Pfeilsymbol:	updating list
Filter mit Blatt Symbol:	show filtered list
Brillensymbol:	showing preliminary results
[1 ... 5]:	priority of the result
weiteren Symbole:	no function
Fragezeichen:	help function

Symbols within the list:

- Priority with symbol showing the status of the result (dictated, corrected, in the RIS, ...)
- Patient name
- Patient birthdate
- Patient sex
- Patient examination number
- examination date and time
- examination codes
- Dictated by: name of medical doctor
- Written by: medical doctor or secretary
- Assigned by
- RIS-Status: Status of the result within the RIS

4. Lessons learned

User Expectations:

The expectations on the part of the doctors testing the system were too high, causing disappointment at the results of the first dictation.

User Discipline:

During dictation it is necessary for the radiologist to divide their attention between the console showing the x-rays and the speech recognition PC. This results in an involuntary movement towards and away from the microphone and a relatively unsteady level of sound. This results in a correspondingly low recognition rate.

Solution: We propose to use another type of microphone to solve this problem, a possible alternative is the type used for simultaneous translation on television (small and attached to the shirt/top).

User Acceptance:
The systems testers were unwilling to invest a great deal of their time in "fine tuning" the system. According to Philips, a minimum of 15 minutes is necessary and an hour is recommended to adapt the system to the voice and speech mannerisms of the user. Time should also be expended on context adaptation and text correction. This job can be completed by secretaries or typists.

High administrative expenditure:
Our results confirm that a high administrative expenditure is necessary during the initial phase, including:

- Dictation training; (once per person) – preferably one hour.
- Adjustment; of the audio level for the microphone and loud speakers (once per speaker).
- Context adaptation; dictionary adaptation, integrating unknown words into the system, so these can be used, and recognized, in future.

The more often the system is used, the more words that are "learnt" by the system. However at the beginning the dictionary must be frequently expanded (during the initial phase at least once a day) depending on the amount dictated. The extended dictionary is available for all uses. SpeechMagic continuously adapts the vocabulary from the documents the user dictates. It adds any unknown words and learns the way the user dictates. To further improve the initial performance of the speech recognition system SpeechMagic also allows you to "tune" your vocabulary (ConText) by analyzing existing documents.

Corrections:
A 100% recognition rate is extremely unlikely. This means that every text dictated must be corrected. Although no exact figures are available, it is to be supposed that any time saving is minimal.

Recognition Rates:
Despite high expenditure of time, this is not really satisfactory. During the pilot phase in the Paediatric Radiology a recognition time of between 50% and 100% was achieved with an average of 90%. Only short dictations of 1 or 2 lines achieved a zero error rate. Word endings were frequently lost due to the fact that lots of people swallow the ends of words. A further difficulty was caused by the difference in language customs. When dictating in German the user found that the system frequently split up words, even after a number of dictations, following the American English, rather than the German custom, which does not allow adjectives and nouns to be combined.

5. Experiences

Radiologists in the University Clinic of Pediatric Radiology (Dr. SORANTIN) have employed voice recognition for the dictation transcription of radiological findings and diagnoses, within in a limited sphere, since April 2001.

The system runs on a PC (Pentium III, 800 Mhz, 512 RAM MB). Philips "SpeechPro" and Philips "SpeechMike" is installed with 3 demonstration licenses for the radiologists Dr. LINDBICHLER, Dr. PILHATSCH and Dr. SORANTIN.

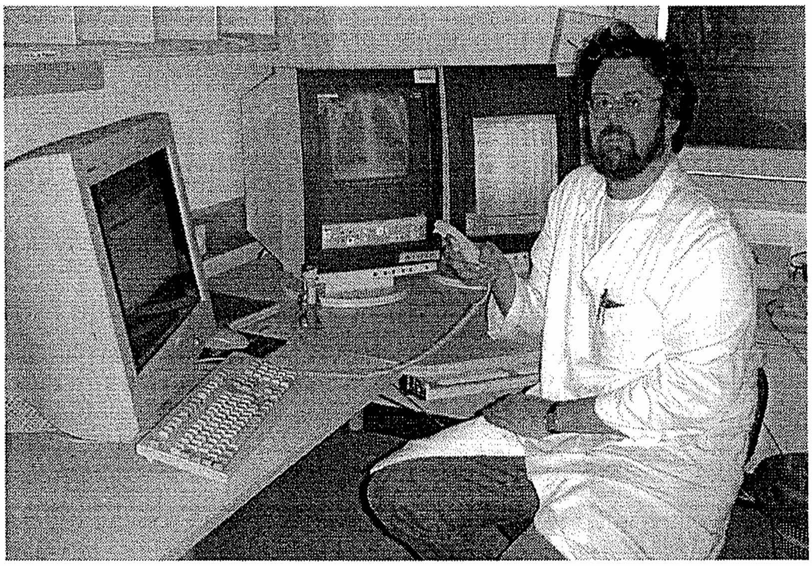

Figure 2: Prof. Dr. ERICH SORANTIN within the experimental setting

The Radiology Information Systems (RIS) uses a VMS-VAX mainframe software for the permanent storage of written patient documentation. The User Interface Application (UIA), designed and developed by the Institute of Medical Informatics (Zeilinger, Stückler, Searle, Ackerl), includes two-way interfaces to the documentation systems, the user authorization files and the previous diagnosis records.

The recognition software was mainly used to dictate short diagnoses and findings (Brief). Some longer dictations, for Sonography and CT-Scans, were also attempted with this system

The Radiologists' appraisal occurred in three parts:

- Functionality of Philips speech recognition software "SpeechPro"
- The implementation
- Collective appraisal

Functionality:
After the initial difficulties in adjusting and calibrating Philips "SpeechPro" with the PC's sound card, a second initial training was carried out. The results achieved with short dictations were satisfactory. However, the longer texts required for the CT and Sonography examinations resulted in unacceptably high recognition errors.

Implementation:

Due to the complexity of the converging systems the Microsoft Office applications, Excel (for patient control) and WinWord (as text editor) were integrated into the UIA.

The X-Rays are viewed on a Magic View screen next to the PC.

Wide-ranging consequences result from the operation of these two systems: The patient must be *separately* selected on each. Although the RIS provides each examination with a unique identification and this is clearly shown on the UIA patient list; the X-Ray picture and the Win Word Text screen, it is still *possible* to select non-corresponding patient data. Given the quantity of findings and diagnoses made daily, the risk of irrelation increases. This could be solved by the inclusion of an interface to the Magic View screen, configured to load the correct pictures, thereby removing the element of human error. Due to the high resolution of the Magic View screen, a complete integration with the patient record software could only be detrimental. Within the framework of the telemedicine college practical training, Dr. SCHIPFER and Dr. SORANTIN measured the mean time necessary for *dictating* a medical finding. The results shown on table below only take the time required by the radiologist into consideration. Attention: Tape delivery delays and the reiteration of secretarial duties are ignored.

	Mean Value (s)
Voice Recognition	139.02
Dictation to Tape	30.32
Hand written report	107.97

Figure 3: The amount of work is noticeably increased for the diagnostician

Collective appraisal:

The amount of work is noticeably increased for the diagnostician. Whereas, with the conventional systems, only the results of the patient examination needed to be selected for display and one tape could be used to dictate a number of findings, Voice Recognition requires the doctor to concentrate, equally, on unsimilar computer applications; patient selection; handling the voice recognition software controls (start, stop, adaption, understanding, spelling etc.); patient data conformity *and* the medical examination.

Prof. Dr. E. SORANTIN summed up: "*This certainly does not increase the vigilance of the diagnostician. The above mentioned factors make a wide-area, routine implementation of Voice Recognition for all examination findings appear problematical.*"

However, the radiologists feel that the application of the voice recognition would be acceptable as an alternative to hand written documentation. Given the equal time factor, the advantage of a typed, and therefore more legible, permanent documentation, would outweigh the disadvantages described.

The formatting of the WinWord documentation for the German language, in particular the hyphenation at the end of a line causes some problems, which would have to be investigated more closely. The existence of a finished document tends to diminish the reliability of controlling the automatically transferred records in the RIS – to the possible detriment of the permanent documentation.

A system-wide implementation of this hybrid system – using three platforms – was a difficult task and we would never recommend this as a basic design. Further tests should be made after the completion of the planned roll-out of the new Hospital Information System, which uses the same platform as Speech Magic. Given a fully implemented system, on a single platform, combined with a modern viewing system and supported by a database, the acceptance would be more positive [cf. also with (LANGER, 2002)].

Suggested optimization:
Implementation of a "Remote Folder Opens" function, whereby the selection of a patient in the UIA automatically triggers the opening of the relevant Magic View and Voice Recognition files would increase both reliability and efficiency.

Implementation of the new XSL standard documentation library, with a pointer stored in the RIS application. The use of hyphenation in the text conversion files transferred to the permanent RIS would reduce the time required for control by the secretarial staff. Marking all documents created by Voice Recognition for later identification. Opportunities for medical staff to increase their acquaintance with the use of an integrated RIS/voice recognition system would serve to minimize the difficulties in handling and optimize the acoustic adaptation.

6. Conclusions

Simple but true: an increasing quantity of dictations increases the quality of the speech recognition. Obvious physical problems observed include fatigue from speaking continuously and the disruption in the office filled with people speaking, which is almost unavoidable in a hospital. The problems encountered coincide perfectly with the current theories considering human acoustic memory and processing (NOUSAK et al., 1996; BARD et al., 1988). Short-term and working memory is some-times referred to as acoustic or verbal memory. The part of the human brain that transiently holds chunks of information and solves problems also supports speaking and listening (HOLZINGER, 2002c). Therefore, working on difficult problems is best done in quiet environments – without speaking or listening to anyone. However, because physical activity is handled in another part of the brain, problem solving is compatible with routine physical activities like walking and driving. In short, humans speak and walk easily but find it more difficult to speak and think at the same time (SHNEIDERMAN, 2000). Similarly when operating a computer, most humans type (or move a mouse) and think but find it more difficult to speak and think at the same time. Hand-eye coordination is accomplished in different brain structures, so typing or mouse movement can be performed in parallel with problem solving. In accordance with SHNEIDERMAN (2000) our results may also support the explanation why after 30 years of ambitious attempts to provide military pilots with speech recognition in cockpits, aircraft designers persist in using hand-input devices and visual displays (SHNEIDERMAN, 2000). Some resistance from the staff is to be expected, since this method of working could eventually imply a complete transference of the secretarial responsibilities to the medical staff. The UIAs designed for secretarial staff are not necessarily suitable for medical staff and vice versa, but further research is necessary.

7. Future Outlook

Talking is an activity that most humans learn at a very early age and carry out fairly effortlessly for the rest of their lives. Although no one would dispute the naturalness of speech in human-to-human communication (NOYES, 2001), we have to carry out much research to improve human-computer interaction (NASS and GONG, 2000), (DENG and HUANG, 2004). Consequently there are great future possibilities to improve human-computer interaction by using speech recognition. We work towards the motto of (SHNEIDERMAN, 2002): "The old computing is about what computers can do; The new computing is about what people can do". The new radiology platform is expected to provide interesting opportunities for further investigation.

References

Bard, E. G., Shillcock, R. C. and Altmann, G. T. M. (1988), The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context, *Perception and Psychophysics,* 44, 5, 395-408.

Bauman, R. A., Gell, G. and Dwyer, S. J. (1996), Large picture archiving and communication systems of the world, *Journal of Digital Imaging: the Official Journal of the Society for Computer Applications in Radiology,* 9, 4, 172-177.

Deng, L. and Huang, X. (2004), Challenges in adopting speech recognition: Multimodal interfaces that flex, adapt, and persist, *Communications of the ACM,* 47, 1, 69-75.

Holmes, J. N. and Holmes, W. J. (1999), *Speech Synthesis and Recognition,* Taylor & Francis, London.

Holzinger, A. (2002a), *Multimedia Basics, Volume 1: Technology. Technological Fundamentals of multimedial Information Systems,* Laxmi Publications, New Delhi (see: www.basiswissen-multimedia.at, available in Hungarian: A multimédia alapjai, Budapest: Kiskapu Kiadó (http://kiado.kiskapu.hu).

Holzinger, A. (2002b), *Multimedia Basics, Volume 2: Learning. Cognitive Fundamentals of multimedial Information Systems,* Laxmi, New Delhi (www.basiswissen-multimedia.at).

Holzinger, A. (2002c), *Multimedia Basics, Volume 3: Design. Developmental Fundamentals of multimedial Information Systems,* Laxmi Publications, New Delhi (www.basiswissen-multimedia.at).

Jurafsky, D. and Martin, J. H. (2000), *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall.,* Prentice-Hall, London.

Langer, S. (2002), Radiology speech recognition: Workflow, integration, and productivity issues, *Current Problems in Diagnostic Radiology,* 31, 3, 95-104.

Nass, C. and Gong, L. (2000), Speech interfaces from an evolutionary perspective, *Communications of the ACM,* 43, 9, 36-43.

Nousak, J. M. K., Deacon, D., Ritter, W. and Vaughan Jr., H. G. (1996), Storage of information in transient auditory memory, *Cognitive Brain Research,* 4, 4, 305-317.

Noyes, J. (2001), Talking and writing: how natural in human-machine interaction?, *International Journal of Human-Computer Studies,* 55, 4, 503-519.

Pickett, J. M. (1999), *Acoustics of Speech Communication, The: Fundamentals, Speech Perception Theory, and Technology,* Allyn & Bacon.

Rudnicky, A. I., Hauptmann, A. G. and Lee, K.-F. (1994), Survey of current speech technology, *Communications of the ACM,* 37, 3, 52-57.

Shneiderman, B. (2000), The Limits of Speech Recognition, *Communications of the ACM,* 43, 9, 63-65.

Shneiderman, B. (2002), *Leonardo's Laptop: Human Needs and the New Computing Technologies,* MIT Press, Boston (MA).