TECHNICAL REPORT

ISO/TR 19358

First edition
2002-10-01

Ergonomics — Construction and application of tests for speech technology

Ergonomie — Élaboration et mise en œuvre des tests des systèmes de technologie de la parole

Reference number
ISO/TR 19358:2002(E)

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.ch
Web www.iso.ch

Printed in Switzerland

Contents

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 3.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

In exceptional circumstances, when a technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example), it may decide by a simple majority vote of its participating members to publish a Technical Report. A Technical Report is entirely informative in nature and does not have to be reviewed until the data it provides are considered to be no longer valid or useful.

Attention is drawn to the possibility that some of the elements of this Technical Report may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO/TR 19358 was prepared by Technical Committee ISO/TC 159, *Ergonomics*, Subcommittee SC 5, *Ergonomics of the physical environment*.

Introduction

This Technical Report advises on methods for determining the performance of speech-technology systems (automatic speech recognizers, text-to-speech systems and other devices that make use of the speech signal) and on selecting appropriate test procedures.

Human-to-human speech communication is not included in this Technical Report but is covered by ISO 9921.

Ergonomics — Construction and application of tests for speech technology

1 Scope

This Technical Report deals with the testing and assessment of speech-related products and services, and is intended for use by specialists active in the field of speech technology, as well as purchasers and users of such systems.

Advanced users are referred to the detailed evaluation chapters of the *EAGLES Handbook of Standards and Resources for Spoken Language Systems* (Gibbon et al. 1997) and the *EAGLES Handbook of Multimodel and Spoken dialogue Systems*. EAGLES was a research project partly sponsored by the European Community.

2 Terms and definitions

For the purposes of this Technical Report, the following terms and definitions apply.

2.1
Automatic Speech Recognition
ASR
ability of a system to accept human speech as a means of input

2.2
dialogue
interactive exchange of information between the speech system and the human speaker

2.3
dialogue management
control of the dialogue between the speech system and the human

2.4
Natural Language Processing
NLP
automatic processing of text originating from humans

2.5
objective assessment
assessment without direct involvement of human subjects during measurement, typically using prerecorded speech

2.6
performance measures
means used to assess the system performance, typically by diagnostic or relative performance methods

2.7
speaker-dependent system
need of a speech-recognition system to be trained with the speech of the specific user

2.8
speaker identification
identification of a particular speaker from a closed set of possible speakers

2.9
speaker-independent system
system not trained for a specific user but applicable for any user of a selected group (native speakers, adults, etc.)

2.10
speaker recognition
general term for technology which identifies or verifies the identity of a speaker

2.11
speaker verification
verification of the identity of a person by assessment of specific aspects of his/her speech

2.12
speaking style
speech may be isolated or continuous, read or spontaneous, or dictated

2.13
speech communication
conveying or exchanging information using speech, speaking, and hearing modalities

NOTE Speech communication may involve brief texts, sentences, groups of words, isolated words, hums and parts of words.

2.14
speech recognizer
process in a machine capable of converting spoken language to recognized words

NOTE This is the process by which a computer transforms an acoustic speech signal into text.

2.15
speech synthesis
generation of speech from data

2.16
speech understanding
technology that extracts the semantic contents of speech

2.17
subjective assessment
assessment with the direct involvement of human subjects during measurement

2.18
text-to-speech synthesis
generation of audible speech from a text

2.19
vocabulary
set of words used in a particular context

2.20
vocabulary size
number of words in a vocabulary of the speech recognizer

3 Description of speech technologies

3.1 Introduction

Speech technology includes the automatic recognition of speech and of the speaker, speech synthesis, etc., Natural Language Processing (NLP) includes the understanding of text items and the management of a dialogue between a human speaker and a machine. Modern technologies are mostly based on algorithms, which make use of digital-signal processing embedded in a digital-signal processor or a (personal) computer system. The algorithms produce near real-time responses. The performance depends on the application. For example, a speech-recognition system designed for use with a small vocabulary and trained with speech from a single user (e.g., control of a personal hand-held telephone) will generally perform (for this particular user) much better than a system designed for a domain with a large vocabulary and generally for a large group of unknown users (e.g., information services through a public telephone network).

For speech products and services, we can identify four main categories:

a) **Command and Control**. The interface between a user and a system is accomplished by automatic speech recognition (ASR). ASR is normally used in a multimodal design, in which the control of a system by speech is one of the possible modalities (i.e., a keyboard, mouse, touch screen, etc. may be an alternative modality). Control by an ASR system may be essential in “hands busy” situations.

b) **Services and Telephone Applications**. Services such as an information kiosk normally require a combination of speech recognition, understanding, speech synthesis and dialogue management in order to control the unsupervised dialogue between user and system. Present state-of-the-art systems cover relatively simple dialogue structures such as travel-information systems (day, time and “from-to”), and call centres (selection of the required information).

c) **Document Generation**. Dictation systems trained for many languages are presently on the market. These systems can be linked to standard word-processing systems. Simple applications include data entry for a specific user domain (e.g. medical reports), more complex systems allow dictation of full documents and the control of the text processing system. These more complex systems are often trained for a large vocabulary and speaker-dependent use. However, for acceptable performance, the system has to be familiarized with the user and the domain of the use. This is often accomplished in two steps: by an (adaptive) acoustical training session in which the user has to read a predefined text, and by presentation of a number of documents written for the user, which are used to extend the vocabulary and to modify the language model.

d) **Document Retrieval**. Retrieval of complete documents (from a spoken-document archive), information retrieval of specific passages from a document or utterances from a specific speaker are of interest for archive documentation and management and the compilation of overviews. Various technologies are used for labelling of the speech utterances such as ASR, word spotting and speaker recognition. Specific search algorithms are used to retrieve the required information.

3.2 Available technologies

3.2.1 Speech recognition

Automatic speech-recognition systems are capable of producing a transcription (text string) from a speech signal. For this purpose, trained systems are used. Modern systems, for use with a large vocabulary, extract specific spectral parameters that identify sub units (phonemes) from the speech signal. Words are described in terms of strings of these phonemes. The recognition architecture may require various levels related to models of the phonemes (phone models), words (vocabulary) and the statistically description of word combinations (language model). Phone models are normally trained for a large number of speakers resulting in statistically based representation. The statistical approach is normally based on a Hidden Markov Model (HMM) or a Neural Network (NN). The vocabulary and the language model are obtained from digitally available text that are representative for the application domain.

3.2.2 Speaker identification and verification

Automatic speaker *identification* is the capability to identify a speaker from a group of known speakers. It answers the question “To whom does this speech sample belong?” This technology involves two steps: modelling the speech of the speaker population (training) and comparing the unknown speech to all of the speaker models (testing).

Speaker verification is a method of confirming that a speaker is the person that he or she claims to be. The heart of the speaker-verification system is an algorithm, which compares an utterance from the speaker with a model built from training utterances gathered from the authorized user during an enrolment phase. If the speech matches the model within some required tolerance threshold, the speaker is accepted as having the claimed identity. In order to protect against an intruder attempting to fool the system by making a recording of the voice of the authorized user, the verification system will usually prompt the speaker to say particular phrases, such as sequences of numbers which are selected to be different each time the user tries to gain entry. The speech verification system is combined with a recognition system to assure that the proper phrase was spoken.

3.2.3 Speech synthesis

For speech synthesis two methods are used: the first, generally known as “canned speech”, is generated on the basis of prestored messages. The coding techniques to compress the messages are normally used in order to save storage space. With this type of synthesis, high-quality speech can be obtained, especially for quick-response applications that make use of a number of standard responses. The second method, “text-to-speech synthesis,” allows the generation of any message from a written text. This generally involves a first stage of linguistic processing, in which the text-input is converted into an internal representation of phoneme and prosodic markers, and a second stage of sound generation on the basis of this internal representation. The sound generation can be made either entirely by rule, typically using complex models of the speech production mechanism (formant synthesis, intonation), or by concatenating short prestored units (concatenate synthesis). The speech quality obtained with concatenate synthesis is generally considered higher.

3.2.4 Speech understanding

Speech-understanding systems can be divided into two broad categories. The first set of problems addresses human-machine interactions. In this case, the person and the machine are working jointly to solve a particular problem. The interactive nature of the task gives the machine a chance to respond with a question when it does not understand the intentions of the user. In turn, the user can then rephrase the query or command. In the second type of problem, the machine has to extract some desired information from the speech without the opportunity for feedback or interaction. This is the case with a summarization of spoken documentation.

3.2.5 Dialogue management

A dialogue is usually considered to be an interaction between two cooperating partners during which some information is passed from one to the other. It may be better to treat the concept differently, recognizing that one of the partners has initiated the dialogue for a certain purpose. The two partners in a dialogue should be considered asymmetrically, one being the originator of the dialogue, the other being the recipient. The dialogue itself is successfully concluded when at least the originator believes that the recipient is in the state for which the dialogue was intended. The intended state may be that the recipient now has some information, or that the recipient has provided some information, or that the recipient is performing some task on behalf of the originator. In effect, a single one-way message has passed between the originator and recipient, and has had a desired effect observable by the originator.

4 Description of relevant variables related to speech technology

4.1 Introduction

Various factors influence the suitability of speech and language systems. Therefore, the optimal use of a system may be related to a certain application. For this purpose, the task-related characteristics and specification of the required performance are required prior to the design of a probable assessment activity. The relevant characteristics include a specification of the speech type, speaker, task, training, environment, input and system. Each of these characteristics covers various variables that are described in 4.2 to 4.8.

4.2 Speech type

Isolated words: a string of words spoken separately, often used for a command and control task or simple data entry. Short pauses indicate the word boundaries.

Connected words: a string of connected words spoken contiguously, often used for a command and control or data entry as number strings. These systems are usually trained with isolated words.

Read speech: speech read continuously, such as from a textbook, without pauses.

Dictation speech: speech read continuously but at a controlled speed and with extra attention for proper pronunciation. The speaker is aware that automatic recognition is taking place.

Spontaneous speech: conversational speech, including all types of discontinuities such as coughs, hesitation, interruptions, etc. Usually the speakers are not aware that recognition is taking place.

4.3 Speaker (specification of speaker-dependent aspects)

Speaker dependency: speaker dependency relates to a system trained for one speaker or a small group of speakers, speaker independency relates to a system trained for many speakers, normally for use with speakers who were not in the training set.

Gender: speech obtained from male and female speakers normally differs with respect to the fundamental frequency (pitch) and spectral contents. This may have an effect on the performance of a recognizer if the system is not trained for the corresponding gender.

Age: the age of a speaker has, as does the gender, an influence on pitch and spectral components. Classification by age may cover 12-18 years, 19-22 years, 22-65 years. However, within each group a large variation may be observed. Below 12 years and above 65 years, very large individual variations may occur.

Vocal effort: the level of the speech signal depends on the vocal effort of the speaker. The vocal effort is expressed by the equivalent continuous sound-pressure level of speech measured at a distance of 1 m in front of the mouth.

Speaking rate: number of speech items spoken in a certain time slot. Number of words per minute or number of syllables per second. A normal rate is 3-5 syllables per second.

Native language, accent: a reduced recognition performance may be obtained for non-native but fluent speakers of a second language or speakers who have a strong accent.

4.4 Task (application-specific description of relevant recognition parameters)

Vocabulary size: the vocabulary size is task related. For a command and control application, 15 to 100 words may suffice. For large vocabulary recognition, 50,000 words or more may be used. In the latter case, the use of words not present in the vocabulary may occur (so-called OOV's, out-of-vocabulary words).

Syntax complexity: for a tree-structured command, in a (nested) menu, a limited selection set may be needed. The number of alternatives available at a given level corresponds to the complexity.

Dialogue structure: the start position in a dialogue and the sequence to follow should be identified. In case of recognition errors, the system may arrive in an unexpected state. The way back requires situational awareness of the (untrained) user.