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)

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

Page

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is entirely informative in nature and does not have to be reviewed until the data it provides are considered to be no longer valid or useful.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.

Correction management. in case of errors (by the user or system) a facility should be available to correct the error. This may be as simple as the “correction command” or as complex as recovery from an unexpected dialogue state.

4.5 Training (task-related training aspects)

Speaker dependent: a system trained for one or a limited group of speakers. For a word recognizer, this normally is accomplished for each speaker individually.

Speaker independent: a system trained with a large speech database. The database consists of speech samples from many speakers (up to 50-100 h of speech). This is normally performed in the factory.

Speaker adaptive: a system tuned for a specific speaker. Normally, the system starts as a speaker-independent-system and is adapted to a certain user by training for a specific individual. This feature is often used for dictation systems.

Type of speech: depending on the application, this may cover isolated words, connected words, continuous speech or spontaneous speech.

4.6 Environment (specification of the speech quality in a specific environment, for both input and output)

Noise: ambient noise may distort the speech signal. For automatic speech recognition, the effect of noise on the recognition performance is much larger than for human listeners. The noise level and spectrum should be determined. For speech synthesis, the ability of the human listener is responsible for the final intelligibility.

Reverberation: reverberating sounds will disturb the speech signal and reduce the recognition performance. In most cases, a noise-cancelling microphone at an optimal position near the mouth is required for acceptable automatic-speech-recognition performance.

Co-channel interference: cross-talk from other speech signals are generally more disturbing than stationary noise as the recognition algorithm cannot discriminate between the primary speech signal and the disturbing signal.

4.7 Input (specification of the transmission of the speech signal from the microphone to a recognizer input)

Microphone: an input microphone can have a great effect on the quality of the signal. Especially for telephone-network-operated systems, the microphone quality at the speaker side is uncertain. Training and testing a system with the same type of microphone is to be preferred but is not always feasible. Accurate microphone positioning is an important parameter.

Distortion: various distortions may appear if the system is integrated into a network. For a telephone network, a bandwidth limitation (300 Hz to 3 400 Hz) is normally found. Use with portable hand-held telephones may suffer from speech-coding algorithms with limited performance. Band-pass limitations, the overload response, echoes and system noise are major issues.

4.8 Specification of speech-technology modules

Recognizer: the system parameters of a recognizer are normally preset. In most cases, so many (often hidden) parameters are available that it is impossible to adjust these for optimal performance. It is important to specify the vocabulary and language model. If an adaptive system is used, the performance of the system may vary during use or the testing. It is therefore important to store the relevant parameters that are changed during use. If this is not possible, repeated use of the bootstrap system may be required.

Dialogue management: an accurate description of the dialogue structure is required in order to assess the effect of system, or user errors, on the task completion or error correction.

Speech output: system parameters of a text-to-speech system are, as for a recognizer, adjusted in the factory. Sometimes, some options are presented which may improve the speech quality for names, addresses, etc.

5 Assessment methods

5.1 General

The performance of speech-related services and technologies depends on many variables. Some of them are under control, others are affected by uncontrolled phenomena. The specification of the performance of a specific technology or system is normally restricted to a limited set of these variables with fixed parameter settings. For evaluation of a system in a given application, an assessment procedure representative of the application characteristics is required.

The spectrum of evaluation strategies and tests associated with these strategies is highly non-uniform. A number of factors contribute to this situation. First and foremost, spoken-language-system evaluation terminology is itself currently very varied. Common dimensions include: *assessment* vs. *evaluation*, *laboratory* vs. *field* methods, system transparency (*black box* vs. *glass box,* sometimes *white* or *grey box,* evaluation), *subjective* vs. *objective* testing. These dimensions are not completely independent, which implies that a set of meta-criteria is required in order to determine a useful and consistent terminology, which is likely to achieve wide acceptance. However, such criteria are not currently available and consequently the definition of *de facto* or binding standards in this area is a long way off, and even recommendations need to be made with care.

An important source of non-uniformity with regard to evaluation strategies and tests is the heterogeneity of the field itself. The dynamism of current research, development and product marketing and the increasingly wide variety of spoken-language-technology-related devices means currently that frequently an individual product demands a new and individual evaluation strategy with associated tests. The *naturalness* of speech synthesizers, for instance, is required in entertainment robotics or educational avatars. It is not hard to develop scenarios in which *naturalness* is not a primary criterion, and in which, on the contrary, a system must sound like an artificial system in order to allow the user to develop a confidence judgement. In cases like this, the twin criterion of *intelligibility* is always the more important, whether or not product marketing would like to have a human sound-alike voice.

Another source of non-uniformity lies in the increasing use of spoken language input/output devices in embedded systems, sometimes with safety-critical functions. This implies a rapid rise in the complexity of human-machine interfaces with which many current types of evaluation are not designed to cope and which must be approached with extreme caution and expert understanding of the limits of current spoken-language technology. Two examples of areas of this kind are in real-time voice-control of safety critical systems, and in automatic warning systems.

The determination of the assessment method is also conditioned by the purpose behind the evaluation which can be for

a) comparing different systems or different versions of the same system (*progress evaluation* as defined by EAGLES in King, 1996),

b) validating the use of a system for a given task or with respect to a given standard (*adequacy evaluation* as defined by EAGLES in King, 1996),

c) diagnosing dysfunction and their origins (*diagnostic evaluation* as defined by EAGLES in King, 1996), or

d) predicting the future behaviour of a system in a given environment (*predictive evaluation*, as applied to Spoken Language Dialog Systems in Walker et al. 2000).

5.2 Field vs. laboratory evaluation

In laboratory evaluation, some of the specific aspects of the application deployment environment are abstracted from the assessment set-up, while field evaluation considers the actual performance of a system in the context of a specific deployment environment, considered to be representative of the intended deployment environment. Thus, it may happen that a system shows good performances under laboratory conditions which are not repeated under field conditions, since the latter is more concerned with the overall usability of the system being assessed. The key problem when going from field evaluation to laboratory evaluation is to abstract enough aspects to get rid of the noise introduced in the measures by the specificity of the deployment environment, while still remaining faithful to the real issues at stake, i.e. the assessment of the system in respect of its intended use. Field evaluation tends to take into consideration the attributes of the system that are essential for usability but which are not necessarily directly related to the underlying technology performance (provided these are located beyond an acceptable threshold). Some measures then become irrelevant for the assessment of the performance of the underlying technology itself, but are more related to ergonomics or even marketability.

The assessment set-up can be located anywhere between the two extreme types of approach: field evaluations and laboratory evaluations, each with its own properties:

Field Evaluation	Laboratory Evaluation
Application representative	Application generic
Uncontrolled conditions	Reproducible conditions
Expensive	Inexpensive
Large-sized variable set	Small-sized variable set
Usability testing	Technology performance testing
Extrinsic criteria	Intrinsic criteria

With respect to the assessment of speech-based systems, a combination of both methods can be applied. By calibrating a representative database (e.g., recorded in field trials) the parameter values can be determined and controlled laboratory experiments can be conducted. Many calibrated databases, mainly for telephone-oriented applications, are available through specific consortia (ELRA, LDC).

Because natural language is close to the human psyche, the behaviour of the users and their reaction to technology have a significant influence on the measured performance in actual field conditions. For instance, when testing a phone server for booking train tickets under laboratory conditions, it was found that the measured transaction success rate and the average transaction length were significantly higher than those measured under field conditions, since in the first case the testers were paid to interact with the system and did not always react to dialog inconsistencies and repetitions, while in the second case the real users were hanging up at the first missed command in the reservation procedure. When both laboratory assessment and field assessment are performed, the different results should be correlated. This last remark offers a simple way to validate, at a very coarse grain, the evaluation procedure itself.

5.3 System transparency

The system transparency can be located anywhere in the range between the two extreme types of methodology which are respectively termed: white box methodology and black box methodology (Sparck and Jones, 1995). When assessing a system under white box conditions, the investigator has full access to the inner working of the system and associated documentation (when the documentation is not available, the condition is often termed glass box condition). He has the possibility to arbitrarily choose his measurement points, i.e. the points between which he will perform the measurement of a selected parameter, chosen as representative of the performance of a given system function. In black box evaluation, the investigator considers only the system input-output relation without regard to the specific mechanism linking input to output.

In practice, most of the time, the investigator has little or no control over the system transparency, and the assessment method performed is driven by whatever possibilities are offered by the system. In some cases, a sort of *grey box* evaluation is possible, since intermediary information-taping points are provided in the system, e.g. when tracing or debugging functionality are available or when reusable modules from a toolkit are considered. In that case, it may happen that the investigator will have to hypothesize on the function performed by the system between the measurement points, since he may have only a partial description of it. Note also that there is not necessarily a one-to-one correspondence between the actual modules composing a system and the set of functionality that may be subjected to assessment. For example, in any spoken language dialog system, *dialog management* is an essential functionality which may be distributed over different modules involved at different stages of the processing of the input information.

5.4 Subjective vs. objective methods

Assessment methods can be further categorized depending on whether they use *subjective* methods (assessment with the direct involvement of human subjects during measurement) or *objective* methods (assessment without direct involvement of human subjects during measurement, typically using prerecorded speech), or a combination of both. Objective methods have the advantage of producing reproducible results and of being automated by nature, thus they are also cheaper. The problem of objective methods for speech and language application assessment is that they cannot be made to cope easily with the complexity required for understanding natural language or for speech interaction. On the other hand, subjective methods are more suited for evaluating applications with higher semantic or dialog content but they suffer from the fact that a human being cannot reliably perform measurement and that he cannot either handle fine-grained measurement scales (on average, one uses gradation scales with 5 to 10 levels, no more). The use of smoothing statistical techniques like the kappa-statistics to assess inter-annotator agreement (Cohen, 1960, 1968, Krippendorf, 1980) can help, but they do not bring a definite answer to the problem. In addition, their use generally requires more testers, thus increasing the cost of the evaluation.

5.5 Speech recognition systems

There are many parameters which define speech recognition systems. However, since spoken-language input devices are driven by statistical training, the objective testing of many types of system requires a prerecorded, well-defined large corpus (data set), which is partitioned into a training set and a test set, the ratio of training to test data being up to 9:1, often with multiple tests based on different partitions of the data set. Clearly, testing a system on the test data should define an upper bound to performance which is not valid in practice. Adequate training must be performed before testing; a general recipe for "adequate" cannot be given; for a given product, the manufacturer will specify the procedure. Training on specific products will often not be via a prerecorded corpus, however, but by direct microphone input. For particular contexts, noise may be added, either well-defined noise signals or the noise from relevant environments, such as offices or vehicles. Not only the acoustic decoder components but also linguistic factors, such as vocabulary size and the language model employed by the system, are critical parameters which have a strong effect on results. In embedded systems, such as dictation software, there are many other parameters, including error recovery, which cannot be tested exhaustively.

5.6 Speech synthesis systems

The wide range of evaluation dimensions mentioned above also applies to spoken language output (speech synthesis) systems. During development, objective tests may be applied, but for typical embedded applications, judgmental scaling (e.g. of naturalness, pleasantness, appropriateness) and functional testing (e.g. of intelligibility or identification of sounds) may be used. These are types of subjective test which involve human subjects. As with speech input systems, noise levels and types are important factors, and voice adaptation to the specific task requires careful attention: a warning spoken softly in an attractive voice may be not only inappropriate but ineffective. For example, in a case like this, not only the intelligibility must be examined, but also the triggering of appropriate responses by listeners; a hard task to simulate.

5.7 Speaker identification and verification

For speaker identification and verification systems, which may now be considered as one particular type of biometric system, the main parameters pertain to error types, i.e. false rejection and false acceptance, and to the handling of different speaker roles: *applicant speaker* (a given user), *registered speaker* (an authorized user),

genuine speaker (an applicant speaker who is a registered speaker, *impostor* (an applicant speaker who is not a registered speaker). In false rejection, a genuine speaker is not accepted, and in false acceptance, an impostor is accepted. Population size and environment are highly critical factors both in training and in testing. Unlike spoken language input/output devices, a system may have to *unlearn,* as when a speaker's registration is cancelled. Both real-time and non-real-time applications of speaker identification and verification are likely to be security-sensitive, and biometric system technology is currently developing very rapidly and increasingly coming onto the market, and becoming highly complex; consequently reference should be made to the standard handbooks for defining test procedures.

5.8 Corpora

Three types of speech and language corpora are typically of interest:

— "analytic-diagnostic" material which is of primary importance to progress in basic science and which is specifically designed to illuminate specific phonetic and linguistic behaviour;

— "general purpose" material which includes vocabularies which are either common or which are typical of a wide range of applications (for example, alpha-numeric words or standard control terms);

— "task-specific" material which reflects different levels of formalized spoken monologue/dialogue within constrained discourse domains.

Clearly general-purpose corpora are easy to collect and are useful in a general sense but, of course, they have only limited practical value. On the other hand, although task-specific corpora can be time-consuming to collect and are only relevant to a specific domain, they are obviously directly useful for the purposes of practical applications. Diagnostic corpora are time-consuming to design, but they are extremely useful for research purposes.

The availability of standard corpora is of great importance for the speech community and a number of national and international bodies have been responsible for coordination, distribution and production of appropriate databases.

5.9 Related sources of information

A valuable resource is a series of handbooks of speech and language standards and resources produced by the Expert Advisory Group on Language Engineering Standards, sponsored by the EU. The initiative covers a wide range of topics including methodologies for the creation and interchange of electronic language resources such as text and speech corpora, computational lexicons and grammatical formalisms, and the evaluation and quality assessment of language-processing systems and components.

The standard handbooks also cover the appropriate scoring methods; in the older handbook (Gibbon et al., 1997), an overview chapter with the appropriate scoring methods and statistical measures is included. In both this handbook and in Gibbon, 2000, the discussions of evaluation methods and tests include discussion of the appropriate scoring and statistics. For many purposes, the most basic statistics (mean, standard deviation, standard error) are sufficient, with analysis of variance (ANOVA) being desirable in some cases. Correlation measures and confusion measures are also common. One of the greatest mistakes one can make with statistical measures is that of interpreting too much into an overly complex or otherwise inappropriate method, for example when nominal and numerical data are confused. For complex questions, statistics experts should be consulted and brief overviews are likely to be misleading.

The Linguistic Data Consortium (LDC) in the US provides a mechanism for large-scale development and widespread sharing of resources for research in linguistic technologies. The consortium distributes previously created databases, as well as funding and coordinating the funding of new databases. The LDC is closely tied to the evolving needs of the community it supports and has helped researchers in several countries to publish and distribute databases that would not otherwise have been released.

In Europe, the European Language Resources Association (ELRA) was established with the goal of creating an organization to promote the creation, verification, and distribution of language resources. Eventually, ELRA will serve as the European repository for EU-funded language resources and interact with similar bodies in other parts of the world (such as the LDC).

Annex A
(informative)

Example of assessment

A.1 Command and control: voice-controlled dialling for GSM

Two voice-controlled dialling systems were compared in a car-oriented simulation. For this purpose, a mock-up of a car was used in which a hands-free set-up of a GSM phone was mounted. The acoustical environment included background noise with a frequency spectrum that is representative of cabin noise. The hands-free system was equipped with a special noise-cancelling microphone, which was placed at a representative position (50 cm from the mouth). The car telephone system was connected to a real telephone network that included the voice-dialling service. The performance of two different networks was determined and compared. In order to exclude unwanted interaction between the network and a specific GSM phone, two different GSM phones from two different manufacturers were included in the test.

For the test 20 subjects were used and all the subjects were inexperienced in using a voice-controlled dialling system. The selection of the subjects was balanced with respect to gender and age (age between 18 and 60 years).

Before starting the test, the subjects were instructed by giving them the official user manual as supplied by the network provider who offered the voice-dialling service. The subjects were asked to read the manual within 10 min. After the instruction, the subjects were asked to select five persons from their personal relations. The names of these persons were used for the voice-controlled dialling. This procedure has the advantage that the subjects may pronounce the names without hesitation and that a representative sample of names is used that is related to potential users. Each trial of the test was performed in two steps:

a) training of the system with the five selected names according to the directions of the user manual;

b) performing a test run per condition, each run consists of a dialling sequence for all five chosen names in a random order.

The variables in the test for each user were

— two voice-dialling systems,

— two GSM hands-free telephone systems,

— two noise conditions of car noise (80 km/h, 110 km/h),

— two groups of ten subjects (male, female).

The sequence of the test conditions per subject (dialling system, GSM phone, noise condition) was balanced in order to avoid any effect of learning in the comparison of the dialling systems (see Gibbon, 2000).

During the test, the response time of the system, from the moment of the action to dial a selected name till the moment of the actual connection, was established. Also the number and the type of errors was administered. The scoring procedure was based on a penalty system. For each dialling sequence, no penalty was given when the required connection was obtained after speaking the required command string. When extra user interaction was required, the following penalties were used: system asks for name confirmation = 1, deletion of a name = 2, substitution = 5, identified training error = 15. The mean penalty was calculated per subject and test condition.

The results of these tests are given in Table A.1. The major point of interest was the comparison of the performance of the two voice-dialling systems. As system A provided a mean penalty of 3,1 and system B of 5,1,

the conclusion should be that system A performs better. However, to prove this a analysis of variance was performed in order to analyse the significance between the two scores. For this purpose, a so-called ANOVA test was performed. The result of this analysis showed that, for the given number of trials, the two systems are significantly different in performance with respect to successful trials on a probability level p = 0,03.

The second performance measure was the time for completion of a trial. The average completion time was 27,7 s and 17,4 s for system A and B respectively. It was tested with an ANOVA that this difference is significant, p = 0,001. Hence, with respect to this performance measure, system B is better.

Also the effect on the performance for all independent variables was analysed (gender, GSM set, and noise level). These results are not given here as the fall outside the scope of this example.

Table A.1 — Response time and performance of voice-controlled dialling for two systems

System	Gender	GSM device	Noise level	Response time s	Mean penalty	Significance (p-value)
A				24,7	3,1	0,03
B				17,4	5,1	
	Male			19,9	3,3	0,08
	Female			22,2	4,9	
		A		21,2	4,0	0,84
		B		20,9	4,2	
			80 km/h	20,6	3,5	0,07
			110 km/h	21,4	4,7	

A.2 Dictation: Multilingual comparison of a dictation system

This example concerns a comparison of a dictation system for large vocabulary applications and isolated words. The system was developed for five languages (German, Spanish, Italian, French and English). The test was conducted by the manufacturer. The results were published in the open literature (Barnett et al., 1995).

The language-dependent dictation systems consisted of the same basic software but different vocabularies and language models. It is always difficult to compare systems based on different languages as the language-dependent variables can hardly be controlled; language-dependent issues and the different training materials might affect the performance.

Test protocol. The system was tested with text material obtained from widely translated authors. Hence, the texts for the different languages consisted of the same subjects translated into the five different languages. Also a section from the user manual of the dictation system was included.

For each language, four native speakers were used (two male and two female). The dictated speech signals were recorded and stored with the reference transcription in order to perform the test in an automatic computer-controlled test scheme. It also allows repeating the test with the same recognizer but with different parameter settings. This was the case with the adaptation facility. Adaptation is a feature for which the recognizer is additionally trained with the speech of a specific user. The system is made partly user dependent. The test included two modes, with and without adaptation. For the non-adaptation condition, the test was run with the initial factory settings of the system for each language and for the four, language-dependent, native subjects. In the adaptation mode, the recognizer was trained with four texts from different authors after which a test was run for a different fifth author. This was repeated for each author delivering the test material and the other four delivering the training material for the adaptation.

Scoring was performed for four conditions: without adaptation, with adaptation, for words that are included in the vocabulary, and the homophone error rate. It is obvious that adaptation should improve the performance, however, it requires some effort to train the system for each individual. It is also of interest to determine the performance for words that are included in the vocabulary; this indicates how well the system works for the trained material. Homophones (words that have the same phonological form but differ with respect to their orthographic form, hence they sound similar but are spelled differently) are to be detected by the language model.

Some of the results of this study are given in Table A.2. It is clear that the adaptation feature provides a significant improvement of the performance. The effect of the language on the word error rate is also significant (p = 0,01).

German is much more difficult to recognize than Italian or English. In fact, for German, coverage with the same vocabulary size is much smaller than for English because of the presence of many inflected word forms in German.

Table A.2 — Word error rate for some conditions of the experiments on language and training dependency of a dictation system

Language	German	Spanish	Italian	French	English
No adaptation	82	86	89	84	87
With adaptation	87	89	92	87	91
With adaptation, in vocabulary	91	92	94	88	91
Homophone errors	22	25	17	73	25

Annex B
(informative)

Performance measures

The term-recognition system includes systems designed for words, spoken commands, text strings, speakers and languages.

The technical assessment (i.e. laboratory assessment) of recognition systems normally uses the recognition rate as a figure of merit. Related to this, the error rate can be used. The accuracy is a measure that accounts for the type of errors that were made (rejections, insertions and false alarms). Rather than a general figure of merit for the performance of a recognition system, more selective measures may be used such as, for speech recognition, the OOV rejection. An out-of-vocabulary word (OOV words) is a word that is spoken by a user but not included in the trained vocabulary of a system. An OOV can therefore not be recognized correctly.

At the application-oriented assessment, in general a complete system assessed by potential users, performance measures are related to the task, i.e., number of successful trials, response time, and error recovery. Systems with a tree-type input structure may cause disorientation of the user with respect to the position in the task completion. The user must be aware of the status of the system. This situational awareness is essential for successful completion of a task or for, in the case of errors, the error recovery.

The word error rate is obtained by:

$$w = \frac{i + d + s}{N}$$

where

w is the word error rate;

i is the number of insertions;

d is the number of deletions;

s is the number of substitutions;

N is the number of words.

The word error rate can also be given as an percentage. An estimate of the standard deviation of w (s_w) is obtained by:

$$s_w = \sqrt{\frac{w(1-w)}{N}}$$

For a detailed description of performance measures, see chapter 3 in Gibbon et al.

Bibliography

[1] ISO 9921:—[1)], *Ergonomics — Assessment of speech communication*

[2] COHEN J., A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, **20**, pp. 37-46, 1960

[3] COHEN J., Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin*, (**70**)4, pp.213-220

[4] BARNETT, J., BAMBERG, P., HELD, M., HUERTA, J., MANGANARO, L. and WEISS, A. (1995), *Comparative performance in large vocabulary isolated word recognition in five European languages*. Proc. Eurospeech '95 Madrid, Spain, pp. 189-192

[5] ELRA (European Linguistic Resources Association), ELRA/ELDA, "http://www.icp.grenet.fr/ELRA/home.html"

[6] GIBBON, DAFYDD, Inge MERTINS & Roger MOORE, eds. (2000). *Handbook of Multimodal and Spoken Language Systems: Resources, Terminology and Product Evaluation*. Boston, Dordrecht, London: Kluwer Academic Publishers

[7] GIBBON, DAFYDD, Roger MOORE & Richard WINSKI, eds. (1997). *Handbook of Standards and Resources for Spoken Language Systems.* Berlin: Mouton de Gruyter

[8] KING, M. et al., *Evaluation of Natural Language Processing Systems - EAGLES Final Report*, EAG-WEG-PR.2, (October 1996), ISBN-87-90708-00-8

[9] KRIPPENDORF, K., *Content Analysis: An Introduction to Its Methodology*, Sage Publications, Beverly Hills, CA, 1980

[10] LDC (Linguistic Data Consortium), "http://www.ldc.upenn.edu"

[11] LEEUWEN, D.A. van, and STEENEKEN, H.J.M., *Handbook of Standards and Resources for Spoken Language Systems,* Chapter Assessment of recognition systems, pp. 381-407. Mouton de Gruyter, Berlin, New York (1997)

[12] LEEUWEN, D.A. van, and STEENEKEN, H.J.M., *Handbook of Multimodel and Spoken Dialogue Systems,* Chapter: Consumer off-the-shelf (COTS) speech technology product and service evaluation, pp. 204-239. Kluwer academic publisher. Berlin, New York (2000), ISBN 0-7923-7904-7

[13] SPARCK Jones, K., GALLIERS, J. R, *Evaluating Natural Language Processing Systems*, Springer-Verlag (1995), ISBN-3-540-61309-9

[14] STEENEKEN, H.J.M. *Digital Speech Processing*, Chapter 6, Quality evaluation of speech processing systems. Kluwer Academic Publishers Boston/Dordrecht/London (1992)

[15] WALKER, M., Kamm, C. and Litman, D., *Towards Developing General Models of Usability with PARADISE*, Natural Language Engineering, Best Practice in Spoken Language Dialogue System Engineering, Special Issue, Volume 6, Part 3, October 2000

[16] *Potentials of speech and language technology systems for military use: an application and technology-oriented survey*. Ed. H.J.M. Steeneken, NATO-RTO, Neuilly sur Seine, (1996)

1) To be published.

ICS 13.180

Price based on 15 pages