

Australian Standard[®]

**Interactive voice response systems user
interface—Speech recognition**



This Australian Standard® was prepared by Committee IT-022, Interactive Voice Response Systems User Interface. It was approved on behalf of the Council of Standards Australia on 13 February 2008.

This Standard was published on 19 March 2008.

The following are represented on Committee IT-022:

- Association of Consultants in Access Australia
 - Australia Post
 - Australian Bankers Association
 - Australian Electrical and Electronic Manufacturers Association
 - Australian Information Industry Association
 - Australian Telecommunications Users Group
 - Australian Teleservices Association
 - Small Enterprise Telecommunications Centre (SETEL)
 - Telstra Corporation Limited
-

This Standard was issued in draft form for comment as DR 05470.

Standards Australia wishes to acknowledge the participation of the expert individuals that contributed to the development of this Standard through their representation on the Committee and through the public comment period.

Keeping Standards up-to-date

Australian Standards® are living documents that reflect progress in science, technology and systems. To maintain their currency, all Standards are periodically reviewed, and new editions are published. Between editions, amendments may be issued.

Standards may also be withdrawn. It is important that readers assure themselves they are using a current Standard, which should include any amendments that may have been published since the Standard was published.

Detailed information about Australian Standards, drafts, amendments and new projects can be found by visiting www.standards.org.au

Standards Australia welcomes suggestions for improvements, and encourages readers to notify us immediately of any apparent inaccuracies or ambiguities. Contact us via email at mail@standards.org.au, or write to Standards Australia, GPO Box 476, Sydney, NSW 2001.

Australian Standard[®]

Interactive voice response systems user interface—Speech recognition

First published as AS 5061—2008.

COPYRIGHT

© Standards Australia

All rights are reserved. No part of this work may be reproduced or copied in any form or by any means, electronic or mechanical, including photocopying, without the written permission of the publisher.

Published by Standards Australia GPO Box 476, Sydney, NSW 2001, Australia

ISBN 0 7337 8600 6

PREFACE

This Standard was prepared by the Australian members of the Joint Standards Australia/Standards New Zealand Committee IT-022, Interactive Voice Response Systems User Interface. After consultation with stakeholders in both countries, Standards Australia and Standards New Zealand decided to develop this Standard as an Australian Standard rather than an Australian/New Zealand Standard.

This Standard is intended to complement AS/NZS 4263:2003, *Interactive voice response systems—User interface—Dual tone multifrequency (DTMF) signalling*.

The objective of this Standard is to provide telephone users with a user friendly interface to interactive voice response systems, which use speech as the mode of input, in order to promote the ease of use of Interactive Voice Response (IVR) systems in Australia. This Standard also provides an improved and consistent interface based on several years' experience with practical implementations of IVR systems which use speech recognition as the primary input mode.

Much of the technical content of this Standard has been reproduced from ETSI ES 202 076 V1.1.2, ETSI Standard Human Factors (HF); User Interfaces; Generic spoken command vocabulary for ICT devices and services.

In preparing this Standard, account was taken of:

Bednall, E.S. and Rogers, C.T. (2003). Telstra User Interface Guidelines for Speech Applications. *Telstra Research Laboratories, Report Number 8546*.

This Standard is within the framework of Standards for Open Systems Interconnection (OSI) as defined by ISO/IEC JTC1.

The term 'informative' has been used in this Standard to define the application of the appendix to which it applies. An 'informative' appendix is only for information and guidance.

CONTENTS

	<i>Page</i>
FOREWORD.....	4
 SECTION 1 SCOPE AND GENERAL	
1.1 SCOPE	6
1.2 APPLICATION	6
1.3 REFERENCED DOCUMENTS	6
1.4 DEFINITIONS	6
 SECTION 2 NAMING CONVENTIONS	
2.1 GENERAL	12
2.2 TERMINOLOGY IN SYSTEM PROMPTS AND SPOKEN COMMANDS	12
 SECTION 3 SYSTEM INPUT	
3.1 GENERAL	19
3.2 BARGE-IN	19
3.3 DTMF INPUT IN SPEECH RECOGNITION APPLICATIONS	19
 SECTION 4 SYSTEM OUTPUT	
4.1 GENERAL	20
4.2 PROMPT CHARACTERISTICS	20
4.3 OUTPUT OF DIGIT STRINGS	21
4.4 CALL RECORDING	21
4.5 MENUS	21
4.6 NON-SPOKEN OUTPUT	22
 SECTION 5 ERRORS AND TIMEOUTS	
5.1 GENERAL	24
5.2 TIMEOUTS	24
5.3 TRANSFER TO AN OPERATOR	24
5.4 UNAVAILABILITY OF SERVICE	25
5.5 ERROR HANDLING	25
 SECTION 6 HELP AND TUTORIALS	
6.1 GENERAL	27
6.2 HELP	27
6.3 TUTORIAL	27
 APPENDIX A TIPS FOR DESIGNERS	
	28

FOREWORD

Automated Telephone services can be classified as shown in Figure 1.

Interactive Voice Response (IVR) systems allow callers to interact with a computer by using their telephones as terminals. While this Standard focuses on IVR functionality which may be incorporated in any of the applications shown in Figure 1, it does not cover those functions which are specific to voice messaging.

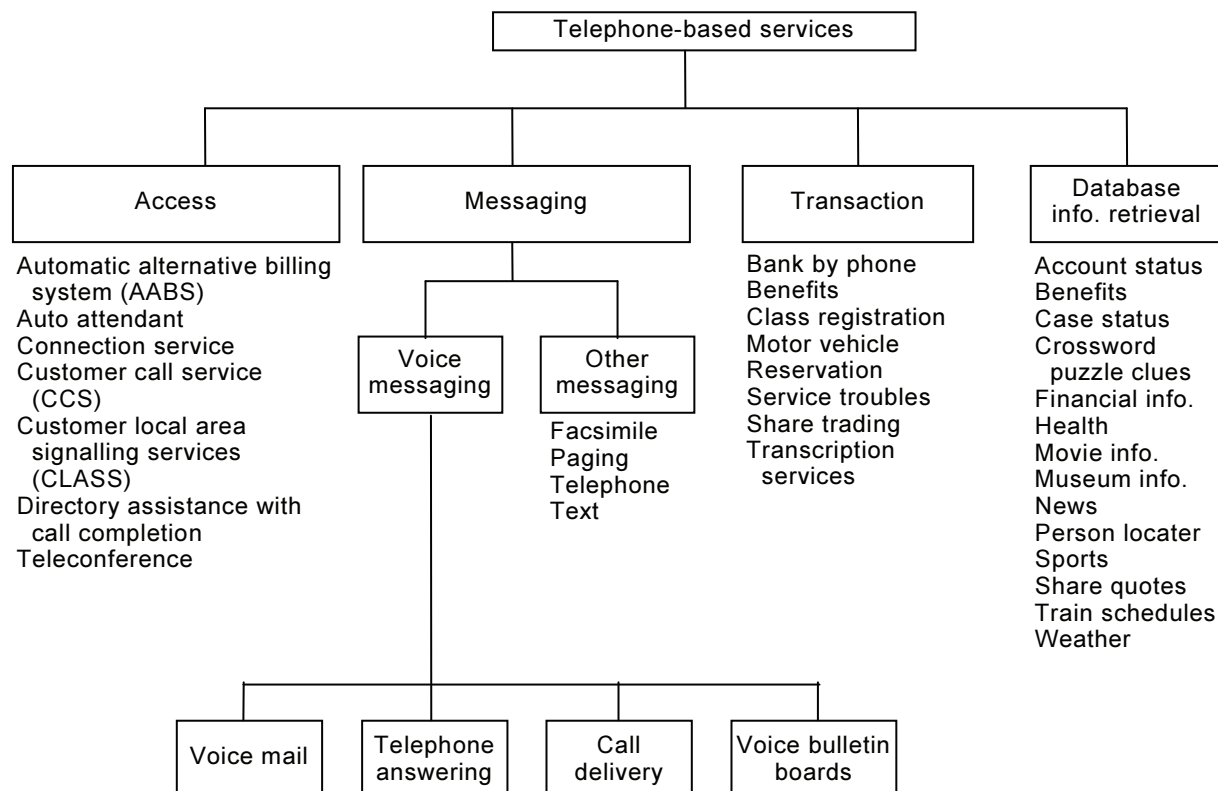


FIGURE 1 CLASSIFICATION OF USER-SYSTEM APPLICATIONS

DTMF signalling is used to send switching information to a telephone exchange. Once a connection is established, the automated telephone service interprets DTMF tones or human speech as commands or system input.

The two input modes addressed in this Standard are—

- (a) DTMF tones; and
- (b) speech.

Systems can be configured to use both DTMF and speech recognition in parallel. This Standard is a companion document to AS/NZS 4263, which deals with DTMF-based IVR services. This Standard addresses applications which also incorporate speech recognition input.

Experience has shown that consistent and predictable human interfaces benefit users. The benefits can include faster learning, greater productivity, fewer errors and greater satisfaction.

Consistent interfaces also benefit the industry by promoting greater acceptance of products and services. Standardization of the user-to-system interface can deliver these benefits.

Standardization of IVR systems is particularly important because callers do not have the opportunity to read an instruction manual each time they access a different IVR service.

As IVR services incorporating speech recognition continue to become more prevalent, users will become extremely frustrated if they encounter systems with varying design principles.

In practice, a speech recognition system has a limit on the number of words that it can recognize. The design of a good user interface is therefore dependent on how well the designer has predicted the words a user is likely to utter.

In human-to-human communication, people tend to mirror the vocabulary of the person with whom they are speaking. In a similar way, IVR systems can guide users to use certain preferred expressions by including those expressions in system prompts.

This Standard contains a core vocabulary of natural and intuitive words and phrases. The vocabulary has been derived from international guidelines, and takes account of the Australian context and vernacular. A key benefit of this common set of terms is that it can be adopted across diverse services and domains, thereby minimizing the time to learn and use new speech-based applications.

STANDARDS AUSTRALIA

Australian Standard

Interactive voice response systems user interface—Speech recognition

SECTION 1 SCOPE AND GENERAL

1.1 SCOPE

This Standard specifies requirements for the design of the user interface in Interactive Voice Response (IVR) systems where the primary user input is by speech.

User input by Short Message Service, Wireless Application Protocol and General Packet Radio Service is outside the scope of this document.

NOTE: Where the user interface allows input by means of both DTMF and speech, the requirements of AS/NZS 4263 should apply to DTMF input except where this Standard includes specific requirements for multi-mode input.

While this Standard does not specifically address issues concerned with accessibility by people with disabilities, attention is drawn to the *Australian Disability Discrimination Act 1992* which makes it unlawful for a person or organization who provides IVR services, to discriminate against another person on the grounds of the other person's disability.

Of importance is that any service accessible by the public should be able to be accessed without discrimination unless making the service available would impose unjustifiable hardship on the person or organization providing the service, would be technically impossible, would impose major difficulties or involve unreasonable costs.

1.2 APPLICATION

This Standard is intended for use by user interface designers and developers of IVR systems which incorporate speech recognition.

1.3 REFERENCED DOCUMENTS

The following documents are referred to in this Standard:

AS/NZS

4263 Interactive voice response systems—User interface—Dual tone multifrequency (DTMF) signalling

Australian Disability Discrimination Act 1992

ETSI ES

ITU-T Rec.

E.183 Guiding principles for telephone announcements

ACIF

G516 Industry Guideline—Participant monitoring of voice communications

1.4 DEFINITIONS

For the purposes of this Standard the definitions below apply.

1.4.1 Application

Computer software which is designed for a specific business area or purpose.

1.4.2 Audio icon

Natural sounds which are commonly associated with contexts and actions. An example of an audio icon might be a 'breaking glass' sound, to represent destroying an object.

1.4.3 Automatic speech recognition

See 'speech recognition'.

1.4.4 Barge-in

(Noun) A mechanism that enables a speech recognition application to detect and recognize audio input while a prompt is playing. As soon as the input is detected the prompt should stop playing.

(Verb: To barge in) To start speaking while a prompt is playing.

1.4.5 Confirmation

A special case of feedback, this occurs where a number string or other data input is read back to the user and there is opportunity for the user to cancel or change the data before proceeding to the next step.

1.4.6 Correct accept (CA)

An outcome of recognition, where the system correctly deems the user's utterance has matched an item in its grammar, i.e. it has successfully and correctly recognized the user's utterance.

1.4.7 Correct reject (CR)

An outcome of recognition, where the system correctly deems that the user's utterance does not match any item in (the system's) currently loaded grammar.

1.4.8 Dialogue state/module

A single conversational interchange between the system and the user, typically consisting of a prompt and the user's response.

1.4.9 Directed dialogue

Dialogue between a system and a user that is guided by prompts that ask the caller to make a specific response; for example, 'Please tell me what you'd like: say [Option A], [Option B] or [Option C]'.

1.4.10 Disability

The *Australian Disability Discrimination Act* (DDA) uses a broad definition of 'disability' that includes:

Physical, intellectual, psychiatric, sensory, neurological, and learning disabilities, as well as physical disfigurement, and the presence in the body of disease-causing organisms.

1.4.11 Disambiguate

To remove ambiguity from a situation, where the speaker's response can have more than one meaning and the system needs further information to resolve the ambiguity.

1.4.12 DTMF

Dual Tone Multi Frequency. The combination of high and low frequency tones generated by pushbutton or touchtone telephones.

For more information about DTMF and its uses in IVRs, refer to AS/NZS 4263:2003.

1.4.13 Earcon

Abstract musical tones and cadences which can assist the user with context or system status. See also 'audio icon'.

1.4.14 End-of-speech timeout

A period of time allowed after a user has finished speaking.

1.4.15 Enrolment Session

For voice recognition systems where the system recognises the user's voice as well as their speech, a session where the user provides information to the system to enable the system to recognize the user's voice in subsequent sessions.

1.4.16 False accept (FA)

An outcome of recognition, where the system incorrectly matches a user's 'in grammar' utterance to the wrong grammar item, (FA In) or incorrectly matches a user's 'out-of-grammar' utterance to a grammar item (FA out).

1.4.17 False reject (FR)

Where the system incorrectly deems that the user's utterance did not match any item in the grammar, despite the utterance actually being an expected grammar item.

1.4.18 Feedback

Information given to the user to indicate that their request has been received. Feedback may be in the form of prompts or sounds.

1.4.19 Global error count

A running total of the number of no matches and no speech timeouts the user has made during the session.

1.4.20 Grammar

A collection of items that are anticipated to be recognised by the system at any particular stage. Sometimes known as 'vocabulary'. Responses can be either 'in grammar' (IG) or 'out of grammar' (OOG).

1.4.21 Impairment

Any loss or abnormality of psychological, physiological or anatomical structure or function. E.g. 'vision impairment', 'hearing impairment'.

1.4.22 'In grammar' utterance

A caller response that is contained within the current set of grammar items. See 'out of grammar'.

1.4.23 Interactive voice response (IVR) system or service

A telephone-based system or service that allows users to enter information and to make menu selections using speech or DTMF, and to receive audible information.

NOTE: Sometimes the term 'IVR' is inappropriately used to refer to DTMF-only systems, when in fact the term applies to any telephone-based voice output system irrespective of input mode.

1.4.24 Key phrase

A word or phrase that a speech recognition system is listening for in word-spotting mode. Also known as a 'hotword' or a 'magic word'.

1.4.25 Keyword

See 'key phrase'.

1.4.26 Landmarking

Explicit confirmation of the user's location in a call flow. For example, if the user requests *Accounts and Payments*, the next prompt might say: '*Accounts and Payments*. Now what would you like?: say [Option A], [Option B] or [Option C].'

1.4.27 Menu

A list of options offered to the user that includes the actions required to select each option.

1.4.28 Mixed initiative

A style of application where at times the control is in the hands of the user (that is, there appears to be a natural dialogue) and at other times, the system appears to control the interaction—where there are highly structured prompts and options to which the user responds. It is not uncommon to mix directed dialogue (structured menus) with a more natural language approach in the one application.

1.4.29 Multi-mode

Describes an application that allows more than one method of input; for example, speech and web. Multi-modal applications can allow either or both methods at each point in the call flow.

1.4.30 n-best

A function that returns a set number of speech recognition results (n) in order of confidence (from highest to lowest).

1.4.31 No speech timeout

A period of time where no speech has been detected, which if exceeded is interpreted as no input from user. This is often interpreted as a 'no-input' or 'NSTO' error.

1.4.32 Non-speech audio

Sounds or music used to inform the user of the system state, or for providing feedback on the status of system actions. Examples include: a fragment of music played while waiting for a system response; sounds that indicate success or incorrect input; and sounds that identify a company (audio logos).

See also 'audio icon' and 'earcon'.

1.4.33 Open dialogue

Dialogue between a system and a user that is based on open-ended questions; for example, 'Please tell me what I can help you with today'.

1.4.34 'Out-of-grammar' (OOG) utterance

A call response which would be deemed by the system as not matching any of the known items contained within the current grammar.

1.4.35 Natural (language) speech recognition

See 'speech recognition'.

1.4.36 Pre-processing

Fine-tuning adjustments to the raw text inputs so that it sounds more natural when it is read out via text-to-speech. This is especially important for email addresses, message headers and URLs.

1.4.37 Prompt

An audible system output, usually, but not always, in the form of a sentence that instructs or guides the user. A prompt may consist of recorded voice, synthesized speech, sounds, or any combination of these.

1.4.38 Recognition threshold

A parameter (confidence score) set by the application designer below which an utterance will be treated as not recognized.

1.4.39 Prosody

Speech characteristics which include speed, pitch, intonation, pauses and stress.

1.4.40 Self-service

Resolution of an enquiry or completion of a transaction within an automated system, without recourse to an operator.

1.4.41 Side-speech

Speech that is not directed at the system but at person or thing external to the system.

1.4.42 Skiplist

A function used during confirmation of N-best speech recognition results, whereby a caller is not asked to confirm a result that they have already rejected. The system retains a list of rejected results and ‘skips’ them when asking for confirmation.

1.4.43 Speaker dependent

Involves the recognition of a particular speaker’s voice where the application is trained with speech samples from that speaker.

1.4.44 Speaker independent

Speech recognition based on a generalized model where the application is trained on many speech samples from a range of people of different genders, accents and ages. No enrolment session is needed; that is, a user is not required to identify themselves to the system or make voice recordings in order to be understood.

1.4.45 Speech recognition

Recognition by a computer of spoken words and phrases. The words and phrases recognized may be responses to specific prompts (directed speech) or to ‘open’ prompts (‘natural language’).

NOTE: The term ‘voice recognition’ is ambiguous because it can apply to either speech recognition or to speaker verification. Therefore this term is not used in this Standard.

1.4.46 Statistical language model

A statistical language model is a probabilistic description of the likelihood of a given word occurring in an utterance.

1.4.47 System

An interactive voice response (IVR) system.

1.4.48 Text-to-speech (TTS)

Synthesized speech generated by a computer, which is formulated based on text.

NOTES:

- 1 This contrasts with digitized speech, which is a digital recording of human speech.
- 2 A special case of TTS is concatenated speech, which uses small segments of recorded human speech that are spliced together to form words and phrases. TTS is often used in telephony applications which require spoken output. TTS is useful and economical where information must be frequently updated or large chunks must be presented (where pre-recording is too expensive).