HANDBOOK OF MULTIMODAL AND SPOKEN DIALOGUE SYSTEMS

Resources, Terminology and Product Evaluation

Dafydd Gibbon
Inge Mertins
Roger K. Moore
(Eds.)

Microsoft - Exhibit 1017
Microsoft v. Dialect - IPR2025-00655
Page 1 of 10

HANDBOOK OF MULTIMODAL AND SPOKEN DIALOGUE SYSTEMS

Resources, Terminology and Product Evaluation

edited by

Dafydd Gibbon
University of Bielefeld

Inge Mertins
University of Bielefeld

Roger K. Moore
DERA and 20/20 Speech Limited

Kluwer Academic Publishers
Boston/Dordrecht/London

Microsoft - Exhibit 1017
Microsoft v. Dialect - IPR2025-00655
Page 2 of 10

HANDBOOK OF MULTIMODAL AND SPOKEN DIALOGUE SYSTEMS

Resources, Terminology and Product Evaluation

Microsoft - Exhibit 1017
Microsoft v. Dialect - IPR2025-00655
Page 3 of 10

Distributors for North, Central and South America:
Kluwer Academic Publishers
101 Philip Drive
Assinippi Park
Norwell, Massachusetts 02061 USA
Telephone (781) 871-6600
Fax (781) 871-6528
E-Mail <kluwer@wkap.com>

Distributors for all other countries:
Kluwer Academic Publishers Group
Distribution Centre
Post Office Box 322
3300 AH Dordrecht, THE NETHERLANDS
Telephone 31 78 6392 392
Fax 31 78 6546 474
E-Mail services@wkap.nl>

Electronic Services <http://www.wkap.nl>

Library of Congress Cataloging-in-Publication

Handbook of multimodal and spoken dialogue systems : resources, terminology, and product evaluation / edited by Dafydd Gibbon, Inge Mertins, Roger K. Moore.
p. cm. -- (Kluwer international series in engineering and computer science ; SECS 565)
Includes bibliographical references and index.
ISBN 0-7923-7904-7 (alk. paper)
1. Natural language processing (Computer science) 2. Automatic speech recognition.
I. Gibbon, Dafydd. II. Mertins, Inge, 1964- III. Moore, Roger, 1952- IV. Series.

QA76.9.N38 H362 2000
006.3'5--dc21

00-044780

LIBRARY OF CONGRESS
SEP 19 2000
COPY
COPYRIGHT OFFICE
90

Copyright © 2000 by Kluwer Academic Publishers.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, mechanical, photo-copying, recording, or otherwise, without the prior written permission of the publisher, Kluwer Academic Publishers, 101 Philip Drive, Assinippi Park, Norwell, Massachusetts 02061

Printed on acid-free paper.

Printed in the United States of America

Microsoft - Exhibit 1017
Microsoft v. Dialect - IPR2025-00655
Page 4 of 10

electrical characteristic
/ɪ'lektrɪkəl kærəkta'rɪstɪk/, /I'lektrIk@l k{r@kt@'rIstIk/, [N: [AJ: electrical][N: characteristic]], [plural: -s]. Domain: physical characterisation, speech recognition. Meronym. sup.: physical characterisation. Def.: An electrical characteristic is a parameter (post-production factor) in the physical characterisation of the properties of a speech signal.

embedded programming language
/em'bedɪd 'prəʊgræmɪŋ 'læŋgwɪdʒ/, /em'bedId 'pr@Ugr{mIN 'l{NgwIdZ/, [N: [AJ: embedded][N: programming][N: language]], [plural: -s]. Hyperonyms: programming language. Cohyponym: script language. Def.: Embedded (programming) languages run inside an application, e.g. languages such as Basic which are often used to program macros in order to automatise complex sequences of user operations in word processors and PC database systems, or such as Java or JavaScript in a web browser. Embedded languages may be integrated into an application, or additional modules sometimes called add-ons or plug-ins).

EMU
/'iːmjuː/, /'i:mju:/, [N: EMU], [plural: none]. Hyperonyms: software. Def.: EMU is a collection of software for the creation, manipulation and analysis of speech databases. At the core of EMU is a database search engine which allows the researcher to find various speech segments based on the sequential and hierarchical structure of the utterances in which they occur. EMU includes an interactive labeller which can display spectrograms and other speech waveforms, and which allows the creation of hierarchical, as well as sequential, labels for a speech utterance.

environment
/en'vaɪrənmənt/, /en'vaIr@nm@nt/, [N: environment], [plural: -s]. Hyperonyms: context of recognition. Def.: 1. The environment is the total context in which a system such as a recognition or interactive dialogue system is located. For example, a dashboard control system operates in an in-car environment. Environments may be characterised in many different ways. Most commonly, however, factors which might affect the performance of the system (such as high background noise) are singled out to describe environments. 2. The software and hardware within which a application is developed or used.

equivalence
/ɪ'kwɪvələns/, /I'kwIv@l@ns/, [N: equivalence], [plural: none]. Domain: multimodal systems. Hyperonyms: cooperation type. Cohyponym: complementarity, redundancy, specialisation, concurrency, transfer. Def.: 1. Formally, a relation between two expressions with the same meaning or function. 2. A relation between two or more modalities transmitting the same chunk of information , for instance an option to choose from a menu by either mouse or voice selection.

error rate
/'erə 'reɪt/, /'er@ 'reIt/, [N: [N: error][N: rate]], [plural: -s]. Domain: speech recognition, speaker recognition. Hyperonyms: ratio. Hyponyms: word error rate, sentence error rate. Meronym. sup.: statistics. Def.: The fraction of errors made by a recognition system, i.e. the number of errors divided by the number of words to be recognised. Often expressed as a percentage.

error recovery
/'erə rɪ'kʌvəri/, /'er@ rI'kVv@ri/, [N: [N: error][N: recovery]], [plural: y/-ies]. Domain: speech synthesis, speech recognition, consumer off-the-shelf products. Hyperonyms: performance measure. Cohyponym: recognition accuracy, OOV-rejection, response time, situational awareness. Def.: Error recovery is property of a system based on a procedure by which a system returns to a defined state after the occurrence of an error, either directly or by undoing previous actions (user or system backtracking).

ESPS/waves+
/'iː 'es 'piː 'es 'weɪvz 'plʌs/, /'i: 'es 'pi: 'es 'weIvz 'plVs/, [N: ESPS/waves+], [plural: none]. Hyperonyms: software. Def.: ESPS/waves+ is widely used high end commercial suite of programs used for the analysis and display of speech signal data. It includes a library of signal processing programs to assist in computing spectra, analysing speech, converting data, and applying time-referenced labels, and synchronised windows for display of parallel data representations.

estimator
/'estɪmeɪtə/, /'estImeIt@/, [N: estimator], [plural: -s]. Domain: speech recognition, assessment methodologies. Hyperonyms: statistical value. Def.: A mathematical expression for estimating the value of a statistical property, such as the mean or the variance of a set of numerical data values.

evaluation methodology
/ɪvæljʊ'eɪʃən meθə'dɒlədʒi/, /Iv{ljU'eIS@n meT@'dQl@dZi/, [N: [N: evaluation][N: methodology]], [plural: y/-ies]. Domain: multimodal systems. Hyponyms: user-based evaluation, theory-based evaluation, expert-based evaluation, benchmarking. Def.: A set of well-defined methods for quantifying the properties of a system in terms of the system requirements and design specifications.

evaluation
/ɪvæljʊ'eɪʃən/, /Iv{ljU'eIS@n/, [N: evaluation], [plural: -s]. Hyperonyms: measuring. Cohyponym: validation. Def.: A process of measuring whether a given resource or the results of a given activity fit the requirements or design specifications.

event
/ɪ'vent/, /I'vent/, [N: event], [plural: -s]. Domain: language modelling. Hyponyms: unseen event, singleton event, doubleton event. Def.: 1. A state or change of state of a system, or a value or change of value of a parameter, at a point in time or over an interval in time. 2. In graphical user interface (GUI) design, one of a set of specified events such as mouse movements or keyboard inputs which may be used to influence the behaviour of the system.

event-dependent speaker recognition system
/ɪ'vent dɪ'pendənt 'spiːkə rekəg'nɪʃən 'sɪstəm/, /I'vent dI'pend@nt 'spi:k@ rek@g'nIS@n 'sIst@m/, [N: [N: event][AJ: dependent][N: speaker][N: recognition][N: system]], [plural: -s]. Domain: speaker recognition. Hyperonyms: text-independent speaker recognition system. Hyponyms: keyword spotting, concept spotting. Cohyponym: text-dependent speaker recognition system. Def.: A text-independent speaker recognition system for which test utterances must contain a certain linguistic event (or class of events) while the rest of the acoustic material is discarded. This approach requires a preliminary step for spotting and localising the relevant events.

exception vocabulary size
/ek'sepʃən və'kæbjʊlərɪ 'saɪz/, /ek'sepS@n v@'k{bjUl@ri 'saIz/, [N: [N: exception][N: vocabulary][N: size]], [plural: -s]. Domain: consumer off-the-shelf products, speech recognition. Hyperonyms: vocabulary size. Synonyms: user vocabulary size, extension vocabulary size. Cohyponym: active vocabulary size, passive vocabulary size. Def.: The number of words a user may add to the lexicon of a speech recogniser.

exchange
/eks'tʃeɪndʒ/, /eks'tSeIndZ/, [N: exchange], [plural: -s]. Domain: interactive dialogue systems. Meronym. sup.: dialogue. Def.: A pair of contiguous and related turns, one spoken by each party in the dialogue.

Microsoft - Exhibit 1017
Microsoft v. Dialect - IPR2025-00655
Page 5 of 10

iterative design
/ˈɪtərətɪv dɪˈzaɪn/, /’It@r@tIv dI’zaIn/, [N: [AJ: iterative][N: design]], [plural: none]. Hyperonyms: experimental technique. Synonyms: rapid prototyping. Cohyponym: benchmark evaluation, user study, simulation study. Def.: Iterative design has been widely adopted in the field of human-computer interaction, especially for product development. It is suitable for the development of multimodal applications, since many detail implementation issues can be explored rather quickly. The iterative design cycle includes (re)design of the application, implementation, and (informal) user testing. Iterative design is highly desirable from the HCI point of view but is difficult to reconcile with the pipeline or cascaded process organisation in software development which is currently still predominant, for reasons of cost control mainly.

IVR
/ˈaɪ ˈviː ˈɑː/, /’aI ’vi: ’A:/, [N: IVR], [plural: -s] . Domain: interactive dialogue systems, system design. Hyperonyms: dialogue system. Synonyms: interactive voice response, interactive dialogue. Def.: Interactive Voice Response (IVR) is what the commercial world calls interactive dialogue. As such, its scope encompasses certain kinds of simple interaction which research scientists do not normally think of as dialogues. For example, a telephone caller calling a weather forecasting Audiotex service might be asked to say one of the words “today”, “tomorrow” or “weekend”. In the basis of what the system recognises, a canned weather forecast will be played. This is an example of IVR which is also widely known as Voice Response (VR), and a system which supports VR is usually known as a Voice Response Unit (VRU).

jacknife method
/ˈdʒæknaɪf ˈmeθəd/, /’dZ{knaIf ’meT@d/, [N: [N: jacknife][N: method]], [plural: -s]. Hyperonyms: testing method. Def.: A method of testing a speech recognition system, for example, in which part of the data is kept aside for testing and the rest used for training, and the part kept aside is rotated until all the data has been used for testing.

Java
/ˈdʒɑːvə/, /’dZA:v@/, [N: Java], [plural: none]. Hyperonyms: object-oriented programming language. Def.: An object-oriented programming language developed by Javasoft of SUN Microsystems. It has become the de facto standard programming language for applets, i.e. programs which are distributed over the WWW to run inside a WWW browser. Java features a large class library including classes for graphical display, and audio data access. The Java Speech API specification supports voice command recognition, dictation, and text-to-speech synthesis.

jitter
/ˈdʒɪtə/, /’dZIt@/, [N: jitter], [plural: -s]. Domain: physical characterisation. Hyperonyms: physical characteristic; frequency modulation. Cohyponym: wow. Def.: Jitter is a form of frequency modulation of a signal by noise. The term is used as a measure of the average perturbation of a speaker’s fundamental frequency and of its magnitude.

judgment testing
/ˈdʒʌdʒmənt ˈtestɪŋ/ , /’dZVdZm@nt ’testIN/ , [N: [N: judgment][N: testing]] , [plural: none] . Domain: speech synthesis. Hyperonyms: testing procedure. Synonyms: opinion testing. Cohyponym: functional testing. Def.: Procedure whereby a group of listeners is asked to judge the performance of a speech output system along a number of rating scales.

J_ToBI
/ˈdʒeɪ ˈtəʊbi/, /’dZeI ’t@Ubi/, [N: [N: J_ToBI]], [plural: none]. Domain: corpora. Hyperonyms: ToBI . Cohyponym: E_ToBI, GlaToBI, G_ToBI. Def.: J_ToBI is a variant of ToBI developed for the transcription of Standard (Tokyo) Japanese.

keyword detection
/ˈkiːwɜːd dɪˈtekʃən/, /’ki:w3:d dI’tekS@n/, [N: [N: keyword][N: detection]], [plural: -s]. Syn-

knowledge-based speech recognition system
/nɒlɪdʒˈbeɪst ˈspiːtʃ rekəgˈnɪʃən ˈsɪstəm/, /nQlIdZ’beIst ’spi:tS rek@g’nIS@n ’sIst@m/, [N: [N: knowledge][AJ: based][N: speech][N: recognition][N: system]], [plural: -s]. Domain: speech recognition, corpora. Hyperonyms: speech recognition system. Cohyponym: stochastic speech recognition system. Def.: A knowledge-based speech recognition system specifies explicit acoustic-phonetic rules that are robust enough to allow recognition of linguistically meaningful units and that ignore irrelevant variation in these units. (Gibbon et al. 1997, p. 94)

labiodental consonant
/leɪbɪəʊˈdentəl ˈkɒnsənənt/, /leIbI@U’dent@l ’kQns@n@nt/, [N: [AJ: labiodental][N: consonant]], [plural: -s]. Hyperonyms: consonant. Cohyponym: bilabial consonant, dental consonant, alveolar consonant, postalveolar consonant, retroflex consonant, palatal consonant, velar consonant, uvular consonant, pharyngeal consonant, glottal consonant. Def.: A labiodental consonant is a consonant classified on the basis of its place of articulation: a sound in which one lip is actively in contact with the teeth. (Crystal 1988, p. 172)

laboratory room
/ləˈbɒrətri ˈrʊm/, /l@’bQr@tri ’rUm/, [N: [N: laboratory][N: room]], [plural: -s]. Domain: physical characterisation. Hyperonyms: recording room. Cohyponym: soundproof booth, recording studio, anechoic chamber. Def.: Speech recordings in typical laboratory environments are sometimes made in a kind of workbench situation when no special recording facility is available. Recordings made in a laboratory environment are often used to test speech recognition systems, as lab speech recordings seem to reflect best natural speech recognition situations, without requiring too much effort concerning the recording setup. (Gibbon et al. 1997, p. 309)

laboratory testing
/ləˈbɒrətri ˈtestɪŋ/, /l@’bQr@tri ’testIN/, [N: [N: laboratory][N: testing]], [plural: none] . Domain: speech synthesis. Hyperonyms: testing procedure. Cohyponym: field testing. Def.: Speech output test procedure entirely run in a laboratory, either abstracting from in vivo complications or trying to simulate real-life situations.

lamb
/ˈlæm/, /’l{m/, [N: lamb], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker. Synonyms: vulnerable speaker. Cohyponym: resistant speaker. Def.: A speaker with a high mistrust rate. (Gibbon et al. 1997, p. 433)

language disorder
/ˈlæŋgwɪdʒ dɪsˈɔːdə/, /’l{NgwIdZ dIs’O:d@/, [N: [N: language][N: disorder]], [plural: -s]. Domain: corpora. Hyperonyms: speech disorder. Synonyms: aphasia. Cohyponym: articulation disorder, resonance disorder, voice disorder, rhythm disorder. Def.: A language disorder is one which does not affect only the production of the speech message, but rather its content. (Gibbon et al. 1997, p. 115)

Microsoft - Exhibit 1017
Microsoft v. Dialect - IPR2025-00655
Page 6 of 10

MVIP
/'em 'viː 'aɪ 'piː/, /'em 'vi: 'aI 'pi:/, [N: MVIP], [plural: -s]. Domain: system design. Hyperonyms: bus. Synonyms: Multi-Vendor Integration Protocol. Def.: MVIP is a multiplexed digital telephony highway for use within one computer chassis. It provides standard connection for digital telephone traffic between individual circuit boards. It supports telephone circuit-switching under direct computer control, using digital switch elements distributed amongst circuit boards in a standard computer. MVIP software standards allow system integrators to combine MVIP-compatible products from different vendors. The communication technologies that are supported include call management, voice store and forward, speech recognition, text-to-speech, Fax, data communications, and digital circuit-switching. The objective of an MVIP bus is to carry telephone traffic. It allows the interface to the telephone network to be separated from voice processing resources so that the telephone interface may be obtained from one vendor while the voice processing resources are obtained from others. A single MVIP bus has a capacity of 256 full-duplex telephone channels.

n-gram grammar
/'engræm 'græmə/, /'engr{m 'gr{m@/, [N:[N: n-gram][N: grammar]], [plural: -s]. Domain: . Cohyponym: null grammar, unigram grammar, bigram grammar, trigram grammar. Def.: A probabilistic grammar based on transition probabilities of words, predicting the probability of a word in a given context from the product of the a priori probability of the word and the probability of its (n-1) predecessors. The transition probabilities of words are calculated from their distribution in a corpus. Analogously, the probability of a word in a bigram or trigram grammar is calculated using the probabilities of the preceding word or the preceding two words.

n-gram language model
/'engræm 'læŋgwɪdʒ 'mɒdəl/, /'engr{m 'l{NgwIdZ 'mQd@l/, [N: [N: n-gram][N: language][N: model]], [plural: -s]. Domain: language modelling. Hyperonyms: stochastic language model. Synonyms: n-gram language model. Cohyponym: bigram language model, trigram language model. Def.: An n-gram language model is a stochastic language model that is based on a n-gram grammar, i.e. on conditional probabilities depending only on the (n-1) immediate predecessor words. (Gibbon et al. 1997, p. 243)

narrow phonetic transcription
/'nærəʊ fə'netɪk træn'skrɪpʃən/, /'n{r@U f@'netIk tr{n'skrIpS@n/, [N: [AJ: narrow][AJ: phonetic][N: transcription]], [plural: -s]. Domain: corpora, lexicon. Hyperonyms: phonetic transcription. Cohyponym: broad phonetic transcription. Def.: Phonetic transcription which is relatively detailed and represents more than the minimum number of phonetic features which are needed to represent phonemes. (Crystal 1988, p. 313)

nasal
/'neɪzəl/, /'neIz@l/, [N: nasal], [plural: -s]. Hyperonyms: consonant; manner of articulation. Cohyponym: plosive, trill, tap, flap, fricative, lateral fricative, approximant, lateral approximant. Def.: A nasal consonant or vowel is classified on the basis of manner of articulation: a sound produced while the soft palate is lowered to allow resonance of the nasal cavity and escape of air through the nose. (Crystal 1988, p. 203) E.g. [m, n, N].

Natural Language Processing
/'nætʃərəl 'læŋgwɪdʒ 'prəʊsesɪŋ/, /'n{tS@r@l 'l{NgwIdZ 'pr@UsesIN/, [N: [AJ: Natural][N: Language][N: Processing]], [plural: none]. Synonyms: NLP. Cohyponym: SLP, spoken language processing. Def.: The parsing of written language into semantic concepts.

natural language
/'nætʃərəl 'læŋgwɪdʒ/, /'n{tS@r@l 'l{NgwIdZ/, [N: [AJ: natural][N: language]], [plural: -s]. Hyperonyms: language. Cohyponym: formal language. Def.: The forms of language used by humans for communication, as opposed to artificial or formal languages such as logics, algebras, programming languages. The controlled language used between people and invented systems can be also termed "natural" if it is what users spontaneously produce in response

near end echo
/nɪər'end 'ekəʊ/, /nI@r'end 'ek@U/, [N: [AJ: near][N: end][N: echo]], [plural: -es]. Hyperonyms: echo. Cohyponym: far end echo. Def.: The electrical echo of the output of a telephone system generated at the point where the system is connected to the network. Near end echo is always an electrical phenomenon.

neural network based approach
/'njuːrəl 'netwɜːk 'beɪst ə'prəʊtʃ/, /'nju:r@l 'netw3:k 'beIst @'pr@UtS/, [N: [AJ: neural][N: network][AJ:based][N: approach]], [plural: -es]. Domain: multimodal systems. Hyperonyms: template matching. Cohyponym: Principal Component Analysis, PCA, geometric template matching, deformable template matching, optical flow technique. Def.: Another variant of template matching are neural-network based approaches to, for instance, speech or face recognition, for example applying multi-layer perceptrons or Kohonen self-organising maps.

neural network
/'njuːrəl 'netwɜːk/, /'nju:r@l 'netw3:k/, [N: [AJ: neural][N: network]], [plural: -s]. Hyponyms: multi-layer perceptrion, back-propagation network, Kohonen map. Def.: An artificial neural network is a pattern recognition system modelled on the human nervous system, in which input values are mapped via a network of weighted functions (nodes) into output classes.

newsgroup
/'njuːzgruːp/, /'nju:zgru:p/, [N: newsgroup], [plural: -s]. Def.: Moderated or unmoderated Internet information service; a newsgroup is focused on a specific subject, e.g. Spoken Language Processing (SLP). Users subscribe to a newsgroup to read and post mailings.

NIST-SPHERE header
/'nɪst 'sfɪə 'hedə/, /'nIst 'sfi:@ 'hed@/, [N: [N: NIST-SPHERE][N: header]], [plural: -s]. Hyperonyms: audio signal header. Def.: Standard header for audio signals proposed by the National Institute of Standards and Technology (NIST), consisting of a simple ASCII formatted text information describing the signal following the header; the header options can be user-specified.

NLP
/'en 'el 'piː/, /'en 'el 'pi:/, [N: NLP], [plural: none]. Synonyms: Natural Language Processing. Cohyponym: SLP, Spoken Language Processing. Def.: A theoretical and applied discipline, involving computational linguistics and artificial intelligence, which is concerned with natural language understanding (parsing and interpretation) and natural language generation (production of sentences and texts).

noise-cancelling microphone
/nɔɪz 'kænsəlɪŋ 'maɪkrəfəʊn/, /nOIz 'k{ns@lIN 'maIkr@f@Un/, [N: [N: noise][V: cancelling][N: microphone]], [plural: -s]. Hyperonyms: microphone. Synonyms: close-talking microphone. Def.: A noise-cancelling microphone is more sensitive to nearby sound sources (i.e. the mouth) than to distant sound sources (e.g. noise sources).

non-adaptive language model
/nɒnə'dæptɪv 'læŋgwɪdʒ 'mɒdəl/, /nQn@'d{ptIv 'l{NgwIdZ 'mQd@l/, [N: [AJ: non-adaptive][N: language][N: model]], [plural: -s]. Domain: language modelling. Hyperonyms: language model. Cohyponym: adaptive language model. Def.: A non-adaptive language model does not depend on the test data, but remains unchanged as trained on the original training data. (Gibbon et al. 1997, p. 257)

Microsoft - Exhibit 1017
Microsoft v. Dialect - IPR2025-00655
Page 7 of 10

segmental quality
/seg'mentəl 'kwɒlɪti/, /seg'ment@l 'kwQlIti/, [N: [AJ: segmental][N: quality]], [plural: y/-ies]. Domain: speech synthesis. Cohyponym: prosodic quality. Def.: The quality of the phoneme-synthesis factors in a speech synthesiser.

segmentation
/segmen'teɪʃən/, /segmen'teIS@n/, [N: segmentation], [plural: -s]. Domain: corpora, speech synthesis, assessment methodologies. Hyponyms: manual segmentation, automatic segmentation, semi-automatic segmentation, morphological segmentation, phonological segmentation. Def.: The procedure of isolating minimal distinctive temporal phonetic segments (phones). (Gibbon et al. 1997, p. 206)

semantic fusion
/sə'mæntɪk 'fjuːʒən/, /s@'m{ntIk 'fju:Z@n/, [N: [AJ: semantic][N: fusion]], [plural: -s]. Domain: multimodal systems. Hyperonyms: fusion. Synonyms: semantic-level fusion. Cohyponym: signal-level fusion. Def.: Semantic fusion performs the combination of multimodal input at the meaning level. Semantic fusion of multimodal input proceeds in two steps. First, input events in different modalities are combined in a low-level interpretation module by grouping input events in different modalities to multimodal input events. Next, the multimodal input event is passed on to the high-level interpretation module to derive the meaning of multimodal input events by extracting and combining the information chunks. Thus, the high-level interpretation module determines what type of action the user wants to trigger, and what its parameters are. This parametrised action is then passed to the application's dialogue manager that can initiate the execution of the intended action.

semantic-level fusion
/sə'mæntɪk 'levəl 'fjuːʒən/, /s@'m{ntIk 'lev@l 'fju:Z@n/, [N: [AJ: semantic][N: level][N: fusion]], [plural: -s]. Domain: multimodal systems. Hyperonyms: fusion. Synonyms: semantic fusion. Cohyponym: signal-level fusion. Def.: Semantic fusion performs the combination of multimodal input at the meaning level. Semantic fusion of multimodal input proceeds in two steps. First, input events in different modalities are combined in a low-level interpretation module by grouping input events in different modalities to multimodal input events. Next, the multimodal input event is passed on to the high-level interpretation module to derive the meaning of multimodal input events by extracting and combining the information chunks. Thus, the high-level interpretation module determines what type of action the user wants to trigger, and what its parameters are. This parametrised action is then passed to the application's dialogue manager that can initiate the execution of the intended action.

semantics
/sə'mæntɪks/, /s@'m{ntIks/, [N: semantics], [plural: –]. Cohyponym: phonology, phonetics, morphology, syntax, pragmantics. Meronym. sup.: linguistics. Def.: Semantics is the study of the meaning in language. (Crystal 1988, p. 273)

semivowel
/'semivaʊəl/, /'semivaU@l/, [N: semivowel], [plural: -s]. Hyperonyms: consonant. Def.: A sound functioning as a consonant but lacking the phonetic characteristics normally associated with consonants (such as friction or closure); instead, its quality is phonetically that of a vowel, though, occurring as it does at the margins of a syllable, its duration is much less than that typical of vowels. (Crystal 1988, p. 276)

sentence accent
/'sentəns 'æksənt/, /'sent@ns '{ks@nt/, [N: [N: sentence][N: accent]], [plural: -s]. Domain: corpora, speech synthesis. Hyperonyms: accent. Synonyms: contrastive accent (Crystal 1988, p. 2). Cohyponym: word accent. Def.: Sentence accent is the emphasis which makes a particular word or syllable stand out [in a sentence]. (Crystal 1988, p. 2)

sentence error rate
/'sentəns 'erə 'reɪt/, /'sent@ns 'er@ 'reIt/, [N: [N: sentence][N: error][N: rate]], [plural: -s]. Domain: speech recognition. Hyperonyms: error rate. Cohyponym: word error rate. Def.: Proportion of utterances/sentences that contain at least one recognition error.

sentence syntax
/'sentəns 'sɪntæks/, /'sent@ns 'sInt{ks/, [N: [N: sentence][N: syntax]], [plural: none]. Domain: lexicon. Hyperonyms: syntax. Def.: Sentence syntax defines the structure of a (generally unlimited) set of sentences.

sex identification
/'seks aɪdentɪfɪ'keɪʃən/, /'seks aIdentIfI'keIS@n/, [N: [N: sex][N: identification]], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker classification task. Cohyponym: age identification, health state identification, mood identification, accent identification, speaker cluster selection. Def.: If the goal is to decide whether a given speech utterance was uttered by a male speaker or a female speaker, this particular problem of speaker classification can be referred to as sex identification. (Gibbon et al. 1997, p. 408)

SGML
/'es 'dʒiː 'em 'el/, /'es 'dZi: 'em 'el/, [N: SGML], [plural: none]. Hyperonyms: formal language. Hyponyms: eXtended Markup Language (XML). Synonyms: Standard Generalized Markup Language. Def.: An ISO standard (ISO 8879:1986) for the description of text by its structure.

sheep
/'ʃiːp/, /'Si:p/, [N: sheep], [plural: –]. Domain: speaker recognition. Hyperonyms: registered speaker. Synonyms: dependable speaker. Cohyponym: goat, unreliable speaker. Def.: A registered speaker with a low misclassification rate. (Gibbon et al. 1997, p. 432)

shimmer
/'ʃɪmə/, /'SIm@/, [N: shimmer], [plural: -s]. Domain: physical characterisation. Def.: Shimmer is a measure of the average perturbation of someone's fundamental frequency and of its magnitude.

sign language
/'saɪn 'læŋgwɪdʒ/, /'saIn 'l{NgwIdZ/, [N: [N: sign][N: language]], [plural: -s]. Domain: multimodal systems. Meronym. sub.: handshape, hands placement, hands orientation, hands movement, facial expression, body gestures. Def.: Exclusively visual representation of words by (sequences of) hand and face gestures. Also the environment, e.g. people present in the scene, are relevant. Sign language is used by hearing-impaired people to communicate.

signal detection theory
/'sɪgnəl dɪ'tekʃən 'θiːəri/, /'sIgn@l dI'tekS@n 'Ti:@ri/, [N: [N: signal][N: detection][N: theory]], [plural: y/-ies]. Domain: assessment methodologies. Hyperonyms: theory. Def.: A model that may be used for studying speech recogniser performance. The basic idea behind signal detection theory is that errors convey information concerning how the system is operating (in this respect, it is an advance on simple error measures).

signal-level fusion
/'sɪgnəl 'levəl 'fjuːʒən/, /'sIgn@l 'lev@l 'fju:Z@n/, [N: [N: signal][N: level][N: fusion]], [plural: -s]. Domain: multimodal systems. Hyperonyms: fusion. Synonyms: lexical fusion. Cohyponym: semantic-level fusion, semantic fusion. Def.: Signal-level fusion performs the combination of multimodal input at the level of the input signal. Signal-level fusion has to date been tried for audio-visual speech recognition, combining speech as audio signals and lip movements as visual signals. Other types of signal-level fusion have been explored in the robotics field (e.g. combining image data with other sensor input, such as laser ranger finders, or infrared sensors.

Microsoft - Exhibit 1017
Microsoft v. Dialect - IPR2025-00655
Page 8 of 10

speech pathology
/'spiːtʃ pə'θɒlədʒi/, /'spi:tS p@'TQl@dZi/, [N: [N: speech][N: pathology]], [plural: y/-ies]. Domain: corpora, speaker recognition. Def.: Speech pathology is the study of the various types of pathological speech. (Gibbon et al. 1997, p. 92) E.g. hoarseness, aphasia.

speech recogniser training
/'spiːtʃ 'rekəgnaɪzə 'treɪnɪŋ/ , /'spi:tS 'rek@gnaIz@ 'treInIN/ , [N: [N: speech][N: recogniser][N: training]], [plural: -s]. Domain: speech recognition. Def.: Data for speech recogniser training is generated by means of collecting and analysing samples of real spontaneous speech. (Gibbon et al. 1997, p. 578)

speech recognition system
/'spiːtʃ rekəg'nɪʃən 'sɪstəm/, /'spi:tS rek@g'nIS@n 'sIst@m/, [N: [N: speech][N: recognition][N: system]], [plural: -s]. Domain: speech recognition. Hyponyms: discrete speech recognition system, continuous speech recognition system. Synonyms: speech recogniser. Cohyponym: speech understanding system, speech synthesis system. Def.: System that automatically recognises speech. Speech recognition systems support a restricted finite vocabulary, bounded by the limitations of the current technology. (Gibbon et al. 1997, p. 579)

speech synthesis
/'spiːtʃ 'sɪnθəsɪs/, /'spi:tS 'sInT@sIs/, [N: [N: speech][N: synthesis]], [plural: speech syntheses]. Domain: speech synthesis. Synonyms: production of speech sounds. Cohyponym: speech recognition, speech understanding. Def.: Speech synthesis is the name given to the production of speech sounds by a machine. Most speech synthesisers take a text string as input and produce a spoken version of the text as output. Some systems allow the text string to be annotated with prosodic markers which result in changes to the intonational pattern of the speech produced. (Gibbon et al. 1997, p. 92)

speech technology
/'spiːtʃ tek'nɒlədʒi/, /'spi:tS tek'nQl@dZi/, [N: [N: speech][N: technology]], [plural: none]. Hyperonyms: technology, human language technology. Hyponyms: speech input technology, speech output technology. Synonyms: spoken language technology, spoken language processing, SLP. Cohyponym: text technology, natural language processing, NLP. Def.: The discipline concerned with the research and development of spoken language input and output systems, using contributions from the neighbouring disciplines of acoustics, electrical engineering, statistics, phonetics, natural language processing, and involving system requirements specification, design, implementation and evaluation, corpus and linguistic resource processing, and consumer oriented product evaluation. (Gibbon et al. 1997, p. 578)

speech understanding system
/'spiːtʃ ʌndə'stændɪŋ 'sɪstəm/, /'spi:tS Vnd@'st{ndIN 'sIst@m/, [N: [N: speech][N: understanding][N: system]], [plural: -s]. Domain: speech recognition, consumer off-the-shelf products. Cohyponym: speech recognition system, speech recogniser. Def.: A system that not only recognises speech, but also interprets the words. One could call these systems 'speech-to-concept' systems as opposed to 'concept-to-speech' systems. In dialogue systems that ask open questions ("What do you want?") speech understanding plays an important role. If the questions are closed ("Do you want information about trains?") or specific ("Where do you want to go by train?") the system relies on speech recognition to a greater extent or word spotting alone. Generally, however, any speech interactive system that reacts to spoken input sensibly could be called a speech understanding system.

speech-to-speech translation
/'spiːtʃ tə 'spiːtʃ træns'leɪʃən/ , /'spi:tS t@ 'spi:tS tr{ns'leIS@n/ , [N: [N: speech][PREP: to][N: speech][N: translation]], [plural: -s]. Def.: Translation from spoken utterances in one language directly to spoken utterances in another without intervening textual representations.

spelling alternation
/'spelɪŋ ɒltə'neɪʃən/, /'spelIN Qlt@'neIS@n/, [N: [N: spelling][N: alternation]], [plural: -s]. Domain: lexicon. Synonyms: orthographic alternation. Def.: The differences between spellings of parts of composite words and the spellings of corresponding parts of simplex words. (Gibbon et al. 1997, p. 216) E.g. 'y' - 'i' - 'ie' in English 'fly', 'flier', 'flies'.

spelling rule
/'spelɪŋ 'ruːl/, /'spelIN 'ru:l/, [N: [N: spelling][N: rule]], [plural: -s]. Domain: lexicon. Hyperonyms: rule. Def.: Rule which describes spelling alternations. (Gibbon et al. 1997, p. 216)

spoken language corpus
/'spəʊkən 'læŋgwɪdʒ 'kɔːpəs/, /'sp@Uk@n 'l{NgwIdZ 'kO:p@s/, [N: [AJ: spoken][N: language][N: corpus]], [plural: spoken language corpora]. Domain: corpora. Hyperonyms: corpus. Synonyms: collection of speech sound recordings. Cohyponym: written language corpus. Def.: Any collection of speech recordings which is accessible in computer readable form and which comes with annotation and documentation sufficient to allow re-use.

spoken language dialogue system
/'spəʊkən 'læŋgwɪdʒ 'daɪəlɒg 'sɪstəm/ , /'sp@Uk@n 'l{NgwIdZ 'daI@lQg 'sIst@m/ , [N: [AJ: spoken][N: language][N: dialogue][N: system]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: interactive dialogue system. Cohyponym: question/answer system. Def.: A variety of interactive dialogue system in which the primary mode of communication is spoken natural language. Spoken language dialogue systems take human-human conversation as their inspiration, though differences are bound to persist into the forseeable future by virtue of the character of such systems as constrained designed artifact. Spoken language dialogue systems support a much more natural kind of dialogue than Interactive Voice Response systems.

spoken language dialogue
/'spəʊkən 'læŋgwɪdʒ 'daɪəlɒg/, /'sp@Uk@n 'l{NgwIdZ 'daI@lQg/, [N: [AJ: spoken][N: language][N: dialogue]], [plural: -s] . Domain: interactive dialogue systems. Hyperonyms: dialogue. Synonyms: oral dialogue. Def.: A complete spoken verbal interaction between two parties (a system and a human being), each of whom is capable of independent actions. A dialogue is composed of a sequence of steps which are, in some way, related and build on each other. Dialogue systems are thus more sophisticated than question/answer systems, in which one agent may pose a succession of unrelated queries to the other agent.

spoken language identification
/'spəʊkən 'læŋgwɪdʒ aɪdentɪfɪ'keɪʃən/ , /'sp@Uk@n 'l{NgwIdZ aIdentIfI'keIS@n/ , [N: [AJ: spoken][N: language][N: identification]] , [plural: -s] . Domain: speaker recognition. Hyperonyms: decision-making process. Cohyponym: spoken language verification. Def.: Any decision-making process that uses some features of the speech signal to determine what language is spoken in a given utterance.

spoken language lexicon
/'spəʊkən 'læŋgwɪdʒ 'leksɪkən/, /'sp@Uk@n 'l{NgwIdZ 'leksIk@n/, [N: [AJ: spoken][N: language][N: lexicon]], [plural: spoken language lexica, -s]. Domain: lexicon. Hyponyms: system lexicon, lexical database. Cohyponym: written language lexicon. Def.: 1. A spoken language lexicon may be a component in a system, a system lexicon, or a background resource for wider use, a lexical database, in each case containing information about the pronunciation, the spelling, the syntactic usage, the meaning and specific pragmatic properties of words; lexica containing subsets of this information may also be referred to as spoken language lexica, though the simpler cases are often simply referred to as wordlists. (Gibbon et al. 1997, p. 184) 2. A spoken language lexicon is defined as a list of representations of lexical entries consisting of spoken word forms paired with their other lexical properties such as spelling, pronunciation, part of speech (POS), meaning and usage information, in such a way as to

Microsoft - Exhibit 1017
Microsoft v. Dialect - IPR2025-00655
Page 9 of 10

LIBRARY OF CONGRESS
0 007 095 307 9

HANDBOOK OF MULTIMODAL SPOKEN DIALOGUE SYSTEMS

Resources, Terminology and Product Evaluation

Dafydd Gibbon, Inge Mertins and Roger K. Moore (Eds.)

Dictation systems, read-aloud software for the blind, speech control of machinery, geographical information systems with speech input and output, and educational software with "talking head" artificial tutorial agents are already on the market. The field is expanding rapidly, and new methods and applications emerge almost daily. But good sources of systematic information have not kept pace with the body of information needed for development and evaluation of these systems. Much of this information is widely scattered through speech and acoustic engineering, linguistics, phonetics and experimental psychology.

THE HANDBOOK OF MULTIMODAL AND SPOKEN DIALOGUE SYSTEMS presents current and developing best practice in resource creation for speech input/output software and hardware. This volume brings experts in these fields together to give detailed "how to" information and recommendations on planning spoken dialogue systems, designing and evaluating audiovisual and multimodal systems, and evaluating consumer off-the-shelf products.

In addition to standard terminology in the field, the following topics are covered in depth:

- How to collect high quality data for designing, training and evaluating multimodal and speech dialogue systems.
- How to evaluate real life computer systems with speech input and output.
- How to describe and model human-computer dialogue precisely and in depth.

Also included:

- A fully searchable CD-ROM containing a hypertext version of the book in HTML format for fast look up of specific points, convenient desktop use, and lightweight mobile reference, and
- The first systematic medium scale compendium of terminology with definitions.

This handbook has been especially designed for the needs of development engineers, decision-makers, researchers and advanced level students in the fields of speech technology, multimodal interfaces, multimedia, computational linguistics, and phonetics.

ISBN 0-7923-7904-7

Microsoft - Exhibit 1017
Microsoft v. Dialect - IPR2025-00655
Page 10 of 10