Univ Access Inf Soc (2009) 8:109–122
DOI 10.1007/s10209-008-0136-x

LONG PAPER

Attuning speech-enabled interfaces to user and context for inclusive design: technology, methodology and practice

Mark A. Neerincx · Anita H. M. Cremers · Judith M. Kessens · David A. van Leeuwen · Khiet P. Truong

Published online: 7 August 2008

Abstract This paper presents a methodology to apply speech technology for compensating sensory, motor, cognitive and affective usage difficulties. It distinguishes (1) an analysis of accessibility and technological issues for the identification of context-dependent user needs and corresponding opportunities to include speech in multimodal user interfaces, and (2) an iterative generate-and-test process to refine the interface prototype and its design rationale. Best practices show that such inclusion of speech technology, although still imperfect in itself, can enhance both the functional and affective information and communication technology-experiences of specific user groups, such as persons with reading difficulties, hearing-impaired, intellectually disabled, children and older adults.

Keywords Universal access · Speech technology · Multimodal interaction · User experience engineering

M. A. Neerincx (✉) · A. H. M. Cremers · J. M. Kessens · D. A. van Leeuwen · K. P. Truong
TNO Human Factors, P.O. Box 23,
3769 ZG Soesterberg, The Netherlands
e-mail: mark.neerincx@tno.nl

A. H. M. Cremers
e-mail: anita.cremers@tno.nl

J. M. Kessens
e-mail: judith.kessens@tno.nl

D. A. van Leeuwen
e-mail: david.vanleeuwen@tno.nl

K. P. Truong
e-mail: khiet.truong@tno.nl

M. A. Neerincx
Delft University of Technology, Mekelweg 4,
2628 CD Delft, The Netherlands

1 Introduction

Speech technology seems to provide new opportunities to improve the accessibility of electronic services and software applications, by offering compensation for limitations of specific user groups. These limitations can be quite diverse and originate from specific sensory, physical or cognitive disabilities—such as difficulties to see icons, to control a mouse or to read text. Such limitations have both functional and emotional aspects that should be addressed in the design of user interfaces (cf. [49]). Speech technology can be an 'enabler' for *understanding* both the content and 'tone' in user expressions, and for *producing* the right information with the right tone. For example, the right tone may help to motivate the so-far neglected user groups to use specific software applications (such as e-health services). Although vocal interaction has technological limitations, these might be overcome when used as a component of a multimodal user interface. Such interfaces seem to have higher levels of user preference, among other things, because people experience a greater degree of flexibility [47]. Designers of such interfaces should know how speech interacts with the other modalities, and how redundancies and complementarities can compensate for specific (combinations of) disabilities. In general, speech technology might help by adding value to the total user experience.

It should be noted that the user can have specific problems to access information in a specific context. For example, visual limitations may not be a problem in an optimal desktop setting, but a severe problem in a mobile context. Therefore, individual limitations should be addressed in combination with the contextual constraints in 'inclusive design' (cf. [57, 59]). Furthermore, these limitations have their own additional requirements for

personalization and support, to be included in the final user interface [44]. In many cases, accessibility guidelines have been 'only' applied after a web site has already been constructed and are consequently not applied for the development of the required personalization and support concepts. The question is how to integrate these guidelines into the design practices. This paper proposes a design approach in which accessibility is not a separate, additional aspect or objective of development processes, but is integrated into the design and test of personalization mechanisms for the user interfaces [15, 60]. This approach will extend the types of persons who can successfully use the resulting user interfaces, because they are attuned to the diverse user capacities and momentary work contexts, utilizing speech technology in a smart way (i.e., exploring new possibilities and taking care of the constraints).

To establish a thorough theoretical and empirical foundation, a situated user-experience engineering approach is further proposed in which state-of-the-art knowledge is used to distinguish specific *usage constraints* that may arise from an accessibility or a technological perspective (see Fig. 1) [43]. For example, an accessibility constraint can be a hearing disorder of some users, whereas a technological constraint can be the failure to automatically recognize speech in a noisy multiuser environment. Design is an iterative process, consisting of the generation, evaluation and refinement of interaction specifications. To acquire a valid and complete assessment, user experience sampling methods should measure both performance and affective aspects of the interaction in realistic usage contexts. The human–computer interaction (HCI) community brought forth an extensive and diverse set of evaluation methods (e.g. [38]). Such methods have to be combined in a smart way to get a concise, complete and coherent set of user experience data, such as performance, situation awareness, trust and acceptance. Which combination of methods is most appropriate for a specific project depends on the purpose of evaluation, development stage, complexity of the design, number and type of participants, usage setting, duration and cost of evaluation [61]. It should be noted that 'human-in-the-loop' evaluations can be done before complete technological implementation via the so-called Wizard of Oz method, in which a human experimenter simulates (parts of) the functionality of the technical system. Such an evaluation can be done in parallel and in combination with a technology assessment (see [36]). The technology can be integrated into the prototype incrementally, adding more and more advanced support functions—possibly agents—to the prototype.

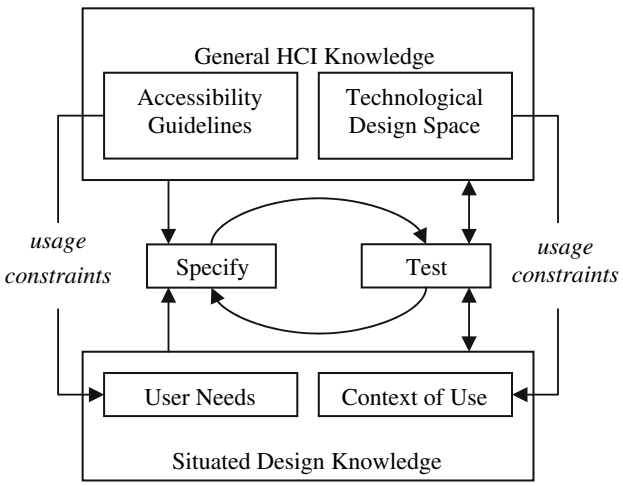

Fig. 1 Situated user experience engineering for inclusive design

Section 2 describes current accessibility guidelines that distinguish human disabilities, which might be compensated for by speech technology. Section 3 presents an overview of this technology and its combination with other modalities. Section 4 provides four best practices of 'beyond the desktop' user interfaces in which speech improved the experiences of 'nonstandard' user groups in specific usage contexts, and Sect. 5 contains general conclusions and discussion.

2 Accessibility guidelines

Different sets of 'Universal Accessibility' guidelines have been developed for people with a variety of disabilities. Being part of the HCI knowledge used in situated user experience engineering (see Fig. 1), they are an important source guiding inclusive user interface design and evaluation. Examples of guidelines are the Web Content Accessibility Guidelines and User Agent Accessibility Guidelines by the World Wide Web Consortium (W3C) [72] and 'Design for All' guidelines for information and communication technology (ICT) products and services by the European Telecommunications Standards Institute [24]. Inclusive design should include all relevant user groups as defined by these guidelines.

2.1 User groups that may benefit from speech-enabled interfaces

There is a large number of attributes that can be used to distinguish between people in a population. The ones that should be considered to have direct impact on the successful use of ICT products and services include [24] the following:

- Sensory abilities such as seeing, hearing, touch, taste, smell and balance.
- Physical abilities such as speech, dexterity, manipulation, mobility, strength and endurance.

- Cognitive abilities such as intellect, memory, language and literacy.
- Allergies can also be a significant factor in some products.

The individual user may have excellent ability in some areas and yet be poor in others. For the population as a whole, there can be a wide variability in any one attribute. The complexity of the problem increases dramatically as more attributes are considered. In general, attributes deteriorate with ageing, whereas the variability increases.

Speech-enabled interfaces (input and/or output) may compensate for other poor abilities. Below, an overview of the sensory (Sect. 2.1.1), physical (Sect. 2.1.2) and cognitive (Sect. 2.1.3) disabilities that may be relieved by speech is provided, based on relevant parts of the ETSI 'Design for All' guidelines [24]. However, note that, using speech may pose serious disadvantages to people with other or multiple disabilities related to hearing and speech. These are described in Sect. 2.2.

2.1.1 *Sensory disabled: seeing and touch*

Sight (or vision) refers to the ability to sense the presence of light and to sense the form, size, shape and color of visual stimuli. Visual disabilities vary from refractive errors, diminished ability to adapt to changes of ambient illumination, color blindness, opacities of the crystalline lens of the eye caused by cataract, diabetic retinopathy causing local loss of visual function, macular degeneration of the photosensitive cells at the center of the retina, to blindness or loss of central vision. Some people cannot perceive light at all, some can distinguish between brightness and darkness, and others can perceive slight movement or some images. Loss of sight can involve one eye, leading to a deterioration of depth perception and field-of-view, or can involve both eyes. Any form of visual disability makes activities such as reading or writing very difficult if not impossible. Also, there is a significant memory load both to read large text and to locate information on a screen. Thus, visually disabled cannot effectively use ICT products and services that rely on visual displays.

The sense of touch refers to the ability to sense surfaces, their texture or quality and temperature. As people age, they lose tactile sensitivity and may no longer be able to rely on touch and pain to give early feedback on temperature or injury. In conjunction with changes in fine motor control, this means that any manipulation that requires very fine adjustment or touch discrimination will be compromised. Those who lack touch sensation, particularly those with prostheses, may not be able to use touch-sensitive screens or touch-pads on computers. Some people have hypersensitive touch and are hurt by stimuli, for example by sharp points and edges, which might only cause discomfort to others.

2.1.2 *Physical disabled: dexterity, manipulation, mobility, strength and endurance*

Dexterity is defined as the skill of manipulation, but can also refer to right-handedness. It implies coordinated use of hand and arm to pick up and handle objects, manipulating and releasing them using the fingers and thumb of one hand. Manipulation relates to activities such as carrying, moving and manipulating objects and includes actions using legs, feet arms and hands. Mobility is the ability to move freely from place to place. Mobility problems can extend from minor difficulties in movement, to being confined to a wheelchair or being bedridden. Some people with impaired mobility have difficulty with control, where muscles are tense and contracted (spasms). They may also have small or missing limbs. Strength relates to the force generated by the contraction of a muscle or muscle group exerted on a specific object. It depends on endurance or stamina (the capacity to sustain such a force) and can be related to heart and lung function.

Dexterity impairment may cause more complex operations, such as simultaneous push and turn, hold down multiple keys simultaneously or using a mouse, to be painful or impossible. Manipulation can be impaired by the inability to use both hands (or feet) or move joints when carrying out some function. Diverse mobility problems may result in difficulties in carrying out controlled and coordinated movement. People with involuntary movements or spasms have problems with tasks that require precision. Reduction in strength can make it difficult to operate a device against significant resistance or torque. A weak grip may make it difficult to hold an object, such as a telephone, particularly for extended periods.

2.1.3 *Cognitive disabled: language and literacy, intellect*

Language and literacy are the specific mental functions of recognizing and using signs, symbols and other components of language. Language impairment can be caused by a stroke or dementia. Sufferers may be able to think as before, but be unable to express their thoughts in words. Literacy refers to the ability to read and write. People of all ages with dyslexia have difficulty with reading and writing. It is therefore very important to keep the wording of signs and instructions as simple and short as possible. Most prelingually deaf (people who are born deaf or have lost their hearing before they learnt to speak) and some postlingually deaf have poor or no reading abilities.

Intellect is the capacity to know, understand and reason. People with intellectual impairment will typically not have the necessary reading skills to comprehend written instructions. They can often recognize simple icons and abbreviations and may be able to follow graphic instructions. They can often function well in a familiar environment but can easily be confused when required to respond quickly.

2.2 Hearing and speech disabled

2.2.1 Hearing

There is a fairly wide spread in hearing ability, such that a deviation of ±20% about the nominal is considered within the normal range of hearing. Most middle-aged people start experiencing impairment in some of the more demanding listening situations such as hearing faint sounds, listening with excessive background noise, hearing with multiple sources (e.g. picking out a single voice in a din of voices, the 'cocktail party' effect). People who are moderately hard of hearing may have difficulty in hearing warning tones. People who are severely hard of hearing generally use hearing aids. Profoundly deaf people generally rely on sign language and lip-reading.

2.2.2 Speech production

Speech production occurs in the mouth and larynx and depends on the coordinated action of many muscles. Effects of diminished breathing ability also affect sound production, the ability to control voice volume and the precision of pronunciation and intonation. Problems of stammering can be accentuated by excessive echo or sidetone, the effect of hearing one's own speech in the earphone. Hearing impairments may affect speech due to changes in the perceived feedback. Prelingually, deaf people typically have no speech or poor speech intelligibility and poor or no reading abilities. Postlingually, deaf people may have retained anything from intact and fully intelligible speech to very unintelligible or no speech at all. Their reading abilities are normally also retained, but some of them may not be able to read or not read very well.

2.2.3 Intellect, memory and language

People with diminished intellect have difficulty in concentrating and paying attention to a task; they require more time to perform tasks and memory for new information deteriorates. Impairment of intellect leads to difficulties in perception and problem solving and can include difficulty in taking in information. Failing memory affects people's ability to recall and learn things and may also lead to confusion. Either, or both, short-term and long-term memory can be affected. People with short-term memory problems can forget where they are in a sequence of operations. Further, loss of language ability may cause sufferers to be unable to express their thoughts in words.

2.3 Conclusions

Knowledge of user groups with special needs is to a large extent documented in accessibility guidelines. To effectively apply speech technology in user interfaces, it should *not* be used for user groups with disabilities in the exact same modality. Instead, it should be used for user groups experiencing difficulties in other modalities, and designed to compensate for these difficulties. So, for people hard of hearing, speech synthesis is not going to be a very useful interface, but speech transcription *can* be effective, as we will see in Sect. 4.2. In addition, care should be taken to avoid shortcomings of applying speech technology (see Sect. 3). Accessibility guidelines form valuable input for the specification and testing of user interfaces. However, Situated Design Knowledge in the form of additional user needs and context of use should also be taken into account (see Fig. 1).

3 Technological design space

Speech technology opportunities to compensate for the disabilities discussed in Sect. 2, and the corresponding usage constraints, are an important source guiding inclusive user interface design and evaluation (see Fig. 1). In addition to providing redundant or alternative ways for input and output, speech technology can help to automatically acquire knowledge about the user, which in turn can help to attune the user interface to this user. This section gives an overview of the capabilities of speech technology, and describes how this technology is to be integrated into multimodal user interfaces.

3.1 Speech technology

In addition to techniques for automatically recognizing and producing speech, technologies are available to sense-specific user and interaction characteristics, such as the user's language, language proficiency level, identity and emotional state, through speech. This section discusses the state-of-the-art of speech technology with potential to improve universal accessibility.

3.1.1 Automatic speech recognition

Automatic speech recognition (ASR) is the automatic conversion of human speech to a sequence of words. The

aim of ASR is to recognize automatically *what* has been said. Nowadays, ASR systems are generally based on the hidden Markov models (HMM) for modeling the acoustics of speech and use either statistic language models (*n*-grams) or rule-based grammars to model the language component. Furthermore, there are many techniques that normalize the acoustic features of the signal and adjust the acoustic models to a specific speaker or different recording conditions.

The first speech recognizer dates from 1952 and consisted of the recognition of spoken digit recognition [17]. What started as digits spoken in isolation by a single speaker has now evolved to speaker-independent, large-vocabulary recognition of fluent spontaneous speech. One can appreciate the techonological advances made over the past decades from observing the speech recognition performance over time. In speech recognition research, the most widely used performance measure is word error rate (WER), which is defined as the percentage of incorrectly recognized words, determined using a specific test set. The performance of ASR systems are dependent on the task, as can be seen in Fig. 2, which shows the WER of the best performing ASR systems for a number of tasks/application domains [48].

If a speech recognizer is part of an application, usually it is used as an input technology or technique [29]. Various other performance measures can augment the WER, depending on the application. For instance, van Leeuwen et al. [66] mention the following issues that are important to consider when estimating the performance of an ASR application: the type and quality of feedback, error correction strategy, dealing with out-of-domain words, speed and response time, the user's situational awareness in the dialog structure, dialog/task success rate, and subjective impression of the overall performance of the system (cf. [25, 65, 67]).

In going from speaker-dependent, isolated digit recognition to real-life speech, like the meeting domain, the difficulty of the speech recognition task is gradually increasing. ASR systems have to model a wide variability of realizations of speech sounds, and be able to search efficiently in a virtually unlimited space of word sequence hypotheses, which makes the task very challenging. The following aspects of real-life speech make ASR difficult:

- Modeling of different speakers
- Size of the vocabulary
- Continuity of the speech signal, which has to be segmented
- Occurrence of nonspeech sounds (music, applause, paper rustling)
- More spontaneous speaking style, which causes variations in pronunciation and the occurrence of filled pauses, hesitations, false starts and repetitions
- Deterioration of the speech signal (effects of acoustics such as reverberation, noise, transmission channel, cross talk)

Nowadays, speech recognition technology is used in many commercial applications. In these applications,

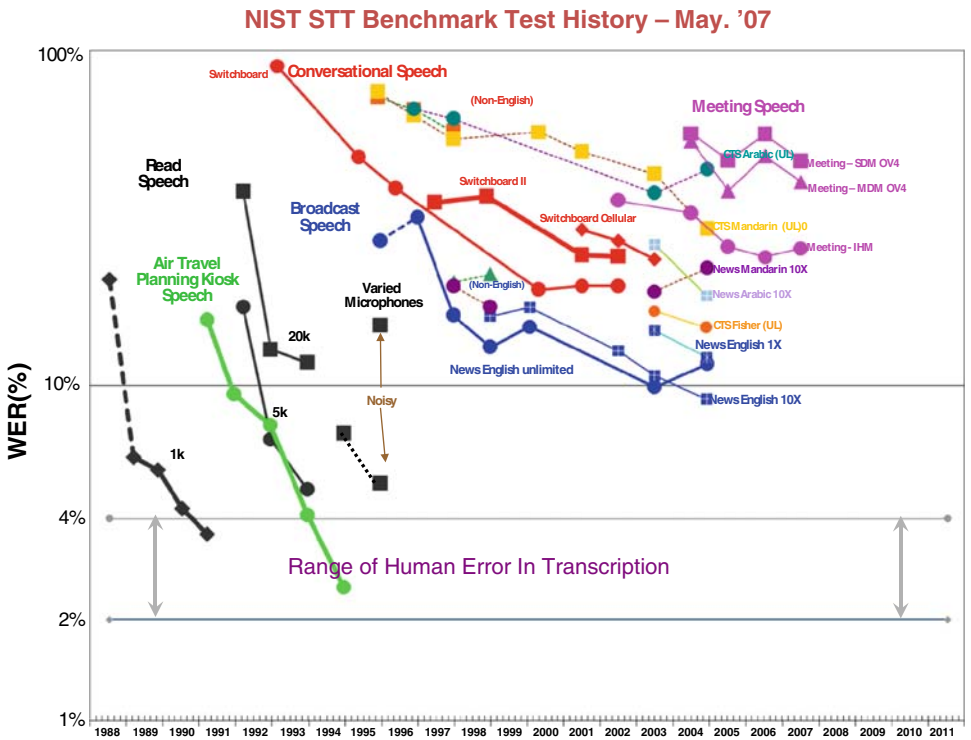

Fig. 2 NIST Benchmark Test History [48] showing the general decrease in word error rate (logarithmic scale), as determined in formal evaluations, as a function of date

shortcomings of the technology are reduced by tailoring the interaction to the individual user and context, for instance, by making the system speaker-dependent (e.g., a dictation systems), by limiting the vocabulary (e.g., voice dialing) or by using a simple and predictable syntax (e.g., command and control, data-entry). In some applications, recognition accuracy does not need to be perfect, e.g., for spoken document retrieval a word recognition accuracy of 70% produces similar retrieval results as manually generated speech transcripts [46].

3.1.2 Speech synthesis

Speech synthesis is the artificial production of human speech. Usually, text is converted into speech; therefore, speech synthesizers are also known as text-to-speech systems. Early speech synthesizers sounded robotic and were often difficult to understand (e.g., Dudley's VOCODER [21]). However, the quality of synthesized speech has steadily improved.

The three basic methods for speech synthesis are formant, articulatory and concatenative synthesis. Formant synthesis is based on the modeling of the resonances in the vocal tract and is perhaps the most commonly used in the last decades. Articulatory synthesis tries to model the human vocal organs as perfectly as possible, but is less popular, as it is difficult to implement, the computational load is high and the level of success is lower than for other synthesis methods [34]. Concatenative speech synthesis is becoming more popular. For this approach, prerecorded speech is used; during synthesis, pieces of prerecorded speech are concatenated. Diphone synthesis is the concatenating of small fixed-size speech units called diphones, or the transitions between two subsequent sounds. At runtime, the target prosody of a sentence is superimposed on the diphones, using techniques such as LPC, PSOLA ([12] or MBROLA [22]). Diphone synthesis is generally more natural-sounding than the output of formant synthesizers, but the best quality systems extend the concept of diphones to longer duration units. In these so-called 'unit selection systems,' the appropriate variable length units are selected at runtime. This type of synthesis provides the greatest naturalness, because it applies only a small amount of speech processing. A disadvantage of unit-selection synthesis, however, is that a large database with prerecorded speech is needed for each synthetic voice.

Speech synthesis systems are usually evaluated by measuring speech intelligibility and speech quality using objective or subjective testing methods (e.g., in the Blizzard Challenge[1] [4]). Nowadays, commercial speech synthesis systems with a very high intelligibility are available in a number of languages. However, artificially produced speech still does not sound natural. Although during the last decades attempts have been made to add emotion effects or the right 'tone' to synthesized speech, emotional speech synthesis is not yet applicable in many real life settings. One of the problems is that there is a trade-off between flexibility of acoustic modeling and perceived naturalness. To express a large number of emotional states with a natural-sounding voice, either the rule-based techniques need to become more natural-sounding or the selection-based techniques must become more flexible [53].

The main advantage of synthetic speech over natural speech is that any text can be converted to speech, therefore allowing spoken access to any information that is stored in writing. Another advantage is that the speaking rate can be controlled. Listeners can understand speech, which is artificially time-compressed to two to three times the original rate, but the maximum speech rate that speakers can attain is lower than that [31]. Foulke [27] showed that for visually impaired listeners a playback of 1.4 as fast as normal is the preferred rate for listening to speech.

3.1.3 User profiling (identity, language)

In many situations, additional information of the user may be beneficial to the interaction process. For instance, in a call center, basic information about the caller, such as age and sex, may be of help to route the call to the most appropriate agent. For known customers who have identified themselves, this information may be obtained from a database, but this identification process may not be desirable in all situations. Information such as the caller's sex can quite easily be retrieved from the speaker's voice. The automatic and unobtrusive techniques to obtain this kind of metadata from the speaker is called *speaker characterization* or *classification* [42]. Closely related to these techniques are the recognition of the speaker's identity (speaker recognition) and the speaker's language or accent [39, 66].

In multicultural and multilingual environments, the automatic detection of the spoken language can be the first step in a spoken human–machine interaction. Current techniques concentrate on *speaker-* and *text-independent* approaches [39]. Text-independent means that the language can be identified without requiring specific words to be used in the interaction. If the system would have to request the preferred language of interaction, a very lengthy and undesirable verification process would occur.

If the language can be identified unobtrusively by monitoring some of the user's conversation with others, a system may use this information to select the right

[1] http://festvox.org/blizzard/.

language for ASR and speech synthesis. In an even more enhanced scenario, automatic detection of dialect or even socio-geographic accent may select specific extensions to the vocabulary and choice of synthesized voice.

3.1.4 User state (emotion) assessment

Besides information on the speech content and user profile, speech also contains information on the user state or emotion. The acoustic–phonetic correlates of emotional speech have been exhaustively investigated, but no reliable acoustic–phonetic voice profiles for emotions that can also be used for the automatic assessment of vocal emotional expressions have been unraveled yet.

Often, statistics of fundamental frequency, intensity, duration and speech rate are used to characterize emotional speech [2, 40, 70]. Further, spectral features as used in ASR are increasingly used in automatic assessment of emotion. They usually prove to be very powerful but have as a disadvantage that they are less insightful. The energy distribution in the spectrum, jitter and shimmer are also among the features that are frequently used.

Traditionally, studies in emotional speech have focused on typical, full-blown emotions that are expressed on demand by actors. The most famous set of full-blown emotions is composed of six so-called basic, universal emotions ([23], note that the origin of this set of emotions is based on facial expressions rather than vocal expressions): anger, disgust, fear, joy, sadness and surprise. The classification accuracies reported in emotion recognition studies are hard to interpret and to compare to each other. For example, on a set of seven basic acted emotions, Schuller et al. [54] achieved an accuracy rate of 88%. Banse and Scherer [2] obtained an accuracy rate between 25 and 53% on a set of 14 emotions. On a set of five emotions, Ververidis and Kotropolos [68] achieved 51% correct, while with six emotions, Nwe et al. [45] had a correct rate between 77 and 89%. Because of many dependencies and diversity of emotions used in these studies, it is difficult to obtain a global view on the performance of emotion detection in general based on these figures.

Recent research has shown that an approach using acted and full-blown emotions to automatic recognition of human emotion is perhaps too simplified for a complex phenomenon as emotion. First of all, there is increasing evidence that there are significant differences in the production and perception of emotional speech [71]. Therefore, a growing number of spontaneous emotional speech databases are being made available to enable researchers to investigate real emotions instead of acted emotions (e.g., [9, 20, 41, 58], HUMAINE[2]). An intermediate form of collecting spontaneous emotional speech data is by using a Wizard-of-Oz (WOZ) setting where the user thinks he/she is interacting with a machine, while, in reality, a human is controlling the machine. Reported classification accuracies of studies using real or WOZ elicited speech lie between 75 and 83% (e.g., [1, 3, 69]).

Secondly, emotions in daily life are not always that extreme and full-blown [8, 19]; so, there is a need for another emotion-labeling scheme that is less rigid. The arousal (ranging from active to passive) valence (ranging from positive to negative) model seems to be flexible enough: many emotions appear to have a place in this two-dimensional space [6, 13, 50, 51] and can be described in terms of arousal and valence. A third dimension called dominance (ranging from submissive, weak to dominant, strong) can be added but is not often used. The advantage of this model is that labels are no longer necessary and that gradations of emotions or less extreme emotions can be better described in this space. Recent automatic emotion recognition technologies are therefore aiming at prediction of emotion in the arousal-valence space. Grimm et al. [28] used Support Vector Regression to predict arousal, valence and dominance values in spontaneous emotional speech and found relatively good recognition results for activation and dominance, while valence scored moderately. Valence has been known to be difficult to characterize with speech features [52, 73]. Several studies have pointed out that some emotions were better recognized in the visual domain than the auditory domain (e.g., [18]). The use of facial expressions in addition to vocal expressions may therefore improve the performance of automatic assessment of emotion (see Sect. 3.2).

Obstacles in this relatively new research area include the sparseness of real emotional speech data, the lack of evaluation standards that hamper comparison between performances of systems and the complexity of measuring emotion in general. It is fair to say that the automatic assessment of human emotion based on vocal expressions is still in development. The current performances seem to be comparable to the ASR performances 55 years ago where ASR systems were able to recognize spoken digits if spoken very clearly and in isolation [17].

3.2 Speech in combination with other modalities

Speech technology can be integrated into multimodal user interfaces that include interaction in other modalities, such as visual (face, gestures), tactile and nonspeech audio. The term ‘multimodal’ is frequently used as a vogue word in user interface research to stress the innovativeness of project plans and results. However, often multimodality does not go beyond allowing the user to interface with a system through different modalities, e.g., speech and touch screen (cf., alternate multimodality; [10]). The modality

[2] http://www.emotion-research.net/download/pilot-db/.

appropriateness framework can be used to include the best set of (alternate) modalities in the user interface [64]. 'Weak multimodal interaction' takes place when there is no semantic interaction between modalities. Contrarily, 'strong multimodal interaction' occurs when the *combination* of modalities opens up new semantics to the system. A strong multimodal interface would be able to deal with spoken references to words like 'there' or 'him,' where by face, head or hand gestures, the place or person can be resolved (cf. synergic multimodality; [10]). When the gestures of virtual characters in a user interface duplicate pieces of information conveyed by speech (redundancy over modalities), users' verbal information recall and subjective experience are improved [7]. Implementation of a 'strong' combination of modalities does not necessarily have to be hard; a simple example is the combination of 'shift-click' keyboard–mouse interaction, which can have different semantics from the individual modalities. Still, before the power of strong multimodal interaction can be utilized, the individual modalities need to have an acceptable level of performance.

Another form of interaction between modalities is the combination of recognition of the individual modalities to improve the performance. For instance, person identification can be improved when both visual (face) and speech (voice) characteristics are combined. Contrary to the term 'interaction,' the term 'fusion' is used to indicate such combination. A second example is the recognition of emotion via fusion of these two modalities, which may carry complementary information and could therefore lead to higher classification accuracies. In a literature study, Truong et al. [62] observed that emotion classification accuracies increase when audiovisual information (AV) is used instead of individual audio (A) or video channels (V). They made a distinction between fusion on feature-level and decision-level. On feature-level, features from different modalities can be concatenated to each other to form one large N-dimensional feature vector. Feature selection techniques may then be used to remove redundant features. Fusion on decision-level means that the features of the different modalities are processed separately, and are fused when the separate classifiers give outputs/scores, which are usually in terms of posterior probabilities or likelihoods. These scores are then subsequently fused by summing, or taking the product of the scores, etc. Fusing classifiers and data streams is not straightforward. Other studies have not only used speech and facial expressions, but also other physiological measures such as skin response, heart rate, etc.

3.3 Conclusions

Shortcomings in speech technology can be reduced by tailoring the interaction to the individual user and context, and by combining modalities. The latter can have the purpose to improve effectiveness (fusion), in terms of better recognition performance, and attractiveness for the user (weak multimodal interaction). When context and modalities are both used, there is the opportunity to open new ways of interaction by making the interpretation of recognized speech depend on the context and other modalities (strong multimodal interaction). Adaptation to the user can be based on absolute identity (obtained actively from an identification interaction, or passively from voice or face), or on more general characteristics such as sex, age, spoken accent, dialect or language. Sensing of the expression of emotion, which is possible using multiple modalities, can be used to monitor the effectiveness of the interaction, and to adapt the interface to suit the user better—specifically when additional user characteristics are available.

4 Best practices

Based on the current technological design space, as described in Sect. 3, speech-enabled interfaces have been developed that address specific aspects of the human disabilities that were distinguished in Sect. 2 (i.e., taking account of the accessibility- and technology-based usage constraints of Fig. 1). Users with specific difficulties to interact with standard user interfaces and experts of the user groups were involved in the development process to define the specific user needs and context of use (bottom part of Fig. 1). The involvement of these experts—such as medical health professionals—is crucial to generate adequate interaction designs and to perform the user experience tests in an appropriate way [14]. The following four 'best practices' show that speech technology can improve accessibility from simple automatic teller machine (ATM) machines to more advanced robot-mediated services.

4.1 Vocal interaction in an automatic teller machine for illiterates

Design objective UNESCO estimates that there are 771 million functional illiterate adults in the world (20% of the world's adult population). This percentage of illiteracy is already high, but it is even climbing. This is caused by a more demanding definition of literacy, including being able to work with machines and understand the instructions, and using ICT, for example being able to operate an ATM. It is necessary to develop models, methods and guidelines to address the illiterate population to make ICT more accessible to them. As a case study, the user interface of the ATM will be adapted, making it more accessible for

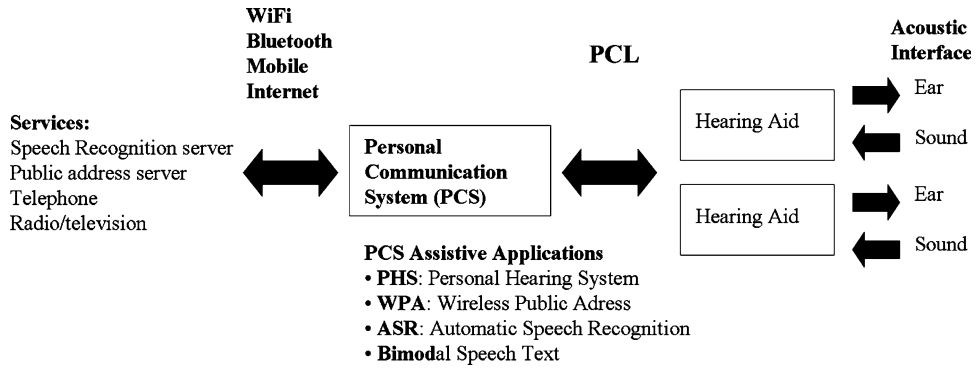

Fig. 3 A personal communication system (PCS) to support hearing-impaired persons during communication

illiterates. One of the possible investigated changes is to add vocal interaction to the ATM interface.

Accessibility Illiterates have difficulty with reading and writing. It is therefore very important to keep the wording of signs and instructions as simple and short as possible. For persons with reduced reading skills, people with dyslexia, but also people with visual disabilities, reading aloud text can improve the accessibility of several interfaces and information sources.

Technology Vocal interaction is currently applied in Talking ATMs. This type of ATMs provide audible instructions, so that persons who have difficulties in reading an ATM screen can independently use the machine. All spoken information is delivered privately through a headphone either through prerecorded sound files or via speech synthesis. Similar tools already exist for reading aloud text from common PC-applications like word processors, Internet browsers, e-mail and pdf readers[3] and websites.[4] The reading aloud functionality might even be visualized as a person who helps the user on navigating the site (e.g., http://www.steffie.nl). In several studies, the results with ASR prototypes were often disappointing, as the implemented speech recognition systems still lacked the performance that is necessary to use it as a viable input interface for ATMs [11, 30, 32]. However, speech recognition can be improved by applying multiple early evaluations of the speech recognition system in the design process (see Fig. 1).

Prototype and evaluation In a literature study, the target group (illiterates) and the application domain (ATMs) were analyzed. This analysis forms the basis for the modeling of the illiterate target group and the subsequent user-centered design process. A couple of interviews, focus groups and a participatory design session were held with a group of six illiterates with varying degrees of illiteracy. An experiment will also be conducted to compare different types of multimodal interfaces including speech output, in which the task performance of the illiterates is compared to their cognitive abilities. Two types of speech output modalities will be tested in the experiments. Firstly, the user can optionally get help and instructions by actively clicking on a specific function, thus focussing on the interface functionality. Secondly, a talking avatar will be used to illustrate the use of the ATM and to give instructions to the user, thus focussing on the task. The experiment with a group of illiterates and a norm group of nonilliterates will reveal to what extent the vocal interaction is beneficial to illiterate users. Based on the results of the experiments, a 'best practice' ATM user interface for illiterates will be designed combining speech with other modalities (e.g. icons/text).

Conclusions The conducted investigations show that speech output (speech synthesis or prerecorded speech) can be a promising interaction type for improving the accessibility of an ATM for illiterate users.

4.2 Automatic speech recognition as an assistive tool for hearing-impaired persons

Design objective Within the EU-funded project 'Hearcom,' a personal communication system (PCS) on a handheld device is being developed.[5] The PCS will support hearing-impaired persons during communication. The system will include several assistive applications and will be connected to hearing aids and other communication systems (e.g., phone, Internet, public address, etc.) by wireless technology (Fig. 3).

Accessibility Some applications have already been developed that support accessibility for hearing-impaired persons, e.g., text telephone systems (e.g., CapTel in the US) and subtitles of television programmes. However, a disadvantage of these tools is that a trained human operator is needed for transcribing the telephone speech and for

[3] e.g. AspireREADER, ReadPlease, TextAloud, ClaroRead, DeskBot and Spika.

[4] e.g. 'Browse aloud' and 'ReadSpeaker'. By selecting text on the website and clicking on the 'read aloud' icon, the text is read aloud.

[5] http://www.hearcom.org/about/Audio-VisualAssistiveTools.html.

making the subtitles. Therefore, these services are costly and will not always be accessible 24 h a day.

Technology One of the assistive applications in the PCS is an automated speech recognition (ASR) application. The goal of this ASR application is to improve speech comprehension in situations that the hearing-impaired does not see the face of the person he/she is listening to, for instance during telephone conversations. The ASR system 'listens' to the speech and automatically produces subtitles for the speech. The hearing-impaired might benefit from subtitling as it provides information on the speech content that the hearing-impaired is missing. Advantages of automatic subtitling is that it can be used 24 h a day, and that it is fast, cheap and easy to implement. A possible disadvantage of ASR is that the subtitles are not error free.

Prototype As a first step in the development of the ASR application of the PCS, it was investigated whether the subtitles produced by the ASR system improves speech intelligibility for hearing-impaired. To that end, speech intelligibility tests were performed in two conditions: with and without automatic subtitling, for a group of young normal listeners and in a subsequent study also for two groups of elderly listeners with normal hearing and with a hearing loss.

Evaluation The first set of experiments demonstrated that normal hearing listeners were able to use partly incorrect and delayed presented ASR output to increase their comprehension of speech in noise [74]. This supports the further development and evaluation of an assistive listening system that visually displays automatically recognized speech to aid communication of hearing-impaired listeners. In the future, the benefit from ASR subtitling will be evaluated in a more realistic communication setting. To this end, a normal hearing person and a hearing-impaired will be forced to communicate with each other by playing the card game 'Black Jack' on a computer, while they do not see each other.

Conclusions The results of the experiments support the hypothesis that automatic subtitling provided by an ASR system can help hearing-impaired persons in communicative situations, thus improving the accessibility of all kind of services (e.g., telephone, radio/television).

4.3 A mobile travel assistant for persons with an intellectual disability

Design objective The Electronic Travel Companion, which runs on a PDA, allows persons with an intellectual disability (ID) to travel independently by public transport [63].

Accessibility The main problems of persons with ID concern reading, memory, concentrating and problem solving. Also, they get easily confused in unfamiliar environments when required to respond quickly. Often, this user group suffers from multiple disabilities, such as a poor sight, dexterity and speech production. Most can recognize simple icons and follow simple graphic instructions. The design was based on existing usability guidelines for PDAs, criteria for websites for ID persons, as well as requirements formulated by representatives of the target group.

Technology Accessibility options applied in the concept include the following: setting of reading level of the texts, choice of symbol set, read aloud on/off, adjust font size, contrast, color, response time and representation of the clock (analog/digital). The PDA can be operated by means of a pen or a finger, suitable for people with limited dexterity. If read aloud is on, text lines, icons and functions can be read aloud by touching them, applying speech synthesis technology. The prototype envisages the use of a built-in global positioning system (GPS) to know where the user is located, to provide location-based information. Finally, it includes a telephone to get in touch with familiar persons.

Prototype The concept application consists of two parts. The first part is a desktop application (Fig. 4a) for the ID person's caregiver, where a user profile (including accessibility options) can be set and the trip can be planned. The second part is the PDA application (Fig. 4b), which, based on the current location of the user, automatically presents each new travel step on the screen, accompanied by an auditory alert.

Evaluation The concept was evaluated with three members of the target group in a real public transport travel setting (by bus and train). The ID person was accompanied by a care giver. At a short distance, the test leader also traveled along, holding a laptop connected to the PDA via a wireless network. The laptop allowed the test leader to simulate GPS by sending commands to the PDA to present new travel steps, relative to the current location. All three participants were able to perform the trip more or less independently. The number of steps was sufficient, the auditory alerts were appreciated and participants did not face major problems navigating through the screens. Icons were considered to be sufficiently large and clear, but their functions were not always obvious. All participants relied on speech output as an alternative to reading the texts, which was judged positively.

Conclusion Although results of the limited evaluation suggest acceptance of the technology by the target group, clearly more testing is necessary. The speech technology in the form of the 'read aloud' function could be improved by a better quality speech synthesis, a possibility to adjust the volume to surrounding noise and by using an earphone for

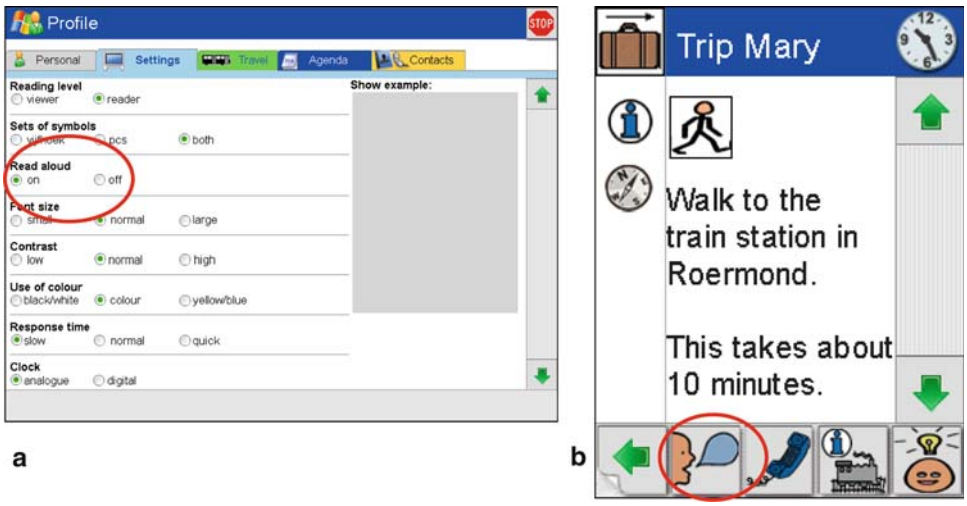

Fig. 4 **a** Desktop user interface settings (including settings for accessibility); 'read aloud' can be set on or off. **b** PDA with one step of the trip; during interaction, 'read aloud' can be switched on or off

Fig. 5 **a** iCat as virtual health assistant (VHA) for older adults. **b** iCat as VHA for children

a better perception and a lower risk of overhearing by other people, without missing relevant ambient sounds.

4.4 A virtual or robot character for self-care of children and older adults

Design objective In modern Western society, there is a major increase in patients with chronic diseases, such as obesities and diabetes. The virtual health assistant (VHA) applies persuasive technologies [26]) to support the daily health-related activities of such patients by educating, informing, motivating and instructing patients at home, and mediating the communication with remote specialists [5].

Accessibility Because of their health illiteracy, older adults and children need this kind of support particularly. Both user groups have relatively limited skills (partly due to limited sensory, physical and cognitive abilities), and motivation to use a standard Windows, Icons, Menu and Pointing device (WIMP). A substantial group of older adults have negative experiences when trying to use such interfaces [16], whereas children's willingness to spend effort in the use is limited [56].

Technology A speech-based robot is being developed that reduces such accessibility bottlenecks and improves the motivation for disease self-management, based on Philips' iCat (see Fig. 5a, b). Automatic speech synthesis was implemented, and speech recognition was simulated via the Wizard of Oz method in the first prototype. In the evaluation of this prototype, speech data were recorded that will be used to train the speech recognizer (to be implemented in the next prototype). Following the proposed incremental design approach of Sect. 1, user profiling and state assessments will be implemented subsequently.

Prototype For the older adults, the VHA educates and motivates patients to adhere to a healthy life-style by conducting a limited form of motivational interviewing, and by guiding the maintenance and assessment of an electronic diary containing health-related information [37]; cf. [55]. For children in particular, the interaction with the health service should be engaging and the VHA a buddy for both serious and entertainment activities (see Fig. 5b). All dialogs are restricted to the specific usage context: a motivational interviewing questionnaire, a diary, educational health videos and games. In such settings, automatic dialog management is feasible, and speech synthesis can be easily implemented. Furthermore, the restricted context-of-use makes it possible to automatically determine important social aspects of the communication (such as who is

talking, and what the emotional impact of a message is). Based on this information, iCat can show the required eye-movements and facial expressions.

Evaluation Older adults experience iCat as empathic; children are more engaged when interacting with iCat compared to a standard WIMP interaction. The communication proves to be more efficient and pleasant among other things, because it requires less specific computer usage skills like reading and typing [33]. For the social behavior, it proved to be important for iCat to show the right expressions consistently in all modalities: the speech content, facial expressions and eye movements. In general, similar results, although less extreme, were acquired for a virtual iCat that is shown on a computer screen.

Conclusion So far, it can be concluded that *a listening and talking robot*—embodied or virtual—can help to improve the accessibility of health-related information and services for older adults and for children.

5 General conclusions and discussion

It is human's nature to communicate via speech, and most people have the appropriate skills to easily exchange information in this way. For such people, speech-enabled user interfaces seem to have the potential to enhance ICT's accessibility. It might decrease specific difficulties to use applications or services, which may have a sensory, motor, cognitive and/or emotional basis. However, it should always be checked that the speech-enabled interfaces do not introduce or increase usage difficulties of specific user groups, such as people who have limited skills for producing or hearing speech.

This paper provided a methodology to design and evaluate speech-enabled interfaces taking these issues into account, systematically applying general HCI-knowledge on accessibility and technology for the identification of user needs and usage contexts. Current accessibility guidelines, containing knowledge of user groups with special needs that can be relieved by speech, form valuable input for specification and testing of user interfaces. Current speech technology can process both the content of communication and other types of information such as speaker's identity, language and emotion. Although far from perfect, this technology can improve the ease and pleasantness of ICT-use for a diverse set of users, if it is appropriately integrated into the overall—often multimodal—user interface and if current usage constraints are taken into account adequately. Four best practices showed how this methodology can help to identify accessibility difficulties and speech technology opportunities to reduce these difficulties. The resulting user interfaces were evaluated via user experience sampling methods, showing the specific strengths and weaknesses of the design. These results will be used to further improve the applications and to refine a general and situated 'knowledge-base' for inclusive design (cf. [35, 43]).

Acknowledgments This paper is based on the results of several projects, among others: the BSIK project MultimediaN, the IOP-MMI project SuperAssist and the EU-project Hearcom.

References

1. Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in Human–Computer Dialog. In: Proceedings of the ICSLP International Conference on Spoken Language Processing, Denver, Colorado, September 2002, pp. 2037–2040
2. Banse, R., Scherer, K.R.: Acoustic profiles in vocal emotion expression. J Pers Soc Psychol **70**, 614–636 (1996)
3. Batliner, A., Steidl, S., Hacker, C., Nöth, E., Niemann, H.: Tales of tuning—prototyping for automatic classification of emotional user states. In: Proceedings of Interspeech, Lisbon, Portugal, September 2005, pp. 489–492
4. Bennett, C.L., Black, A.W.: Blizzard Challenge 2006: Results. Blizzard Challenge 2006 Workshop, Pittsburgh, PA, September 2006
5. Blanson Henkemans, O.A., Neerincx, M.A., Lindenberg, J., van der Mast, C.A.P.G.: SuperAssist: supervision of patient self-care and medical adherence. In: Proceedings of the 16th Triennial Congress of the International Ergonomics Association (CD-rom), pp. 3637–3643. Elsevier, Amsterdam (2006)
6. Bradley, M.M., Lang, P.J.: Affective reactions to acoustic stimuli. Psychophysiology **37**, 204–215 (2000)
7. Buisine, S., Martin, J.-C.: The effects of speech-gesture cooperation in animated agents' behavior in multimedia presentations. Interact Comput **19**, 484–493 (2007)
8. Campbell, N.: A language-resources approach to emotion (2006). Corpora for the analysis of expressive speech. In: Proceedings of the 5th International Conference on Language Resources and Evaluation LREC, Genoa, Italy, May 2006
9. Campbell, N.: The JST/CREST Expressive Speech Processing project, introductory web pages at: http://feast.atr.jp
10. Carbonell, N.: Ambient multimodality: towards advancing computer accessibility and assisted living. Universal Access Informat Soc **5**(1), 96–104 (2006)
11. Chan, F.Y., Khalid, H.M.: Is talking to an automated teller machine natural and fun? Ergonomics **46**, 1386–1407 (2003)
12. Charpentier, F., Moulines, E.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis. Proc ICSLP **2**, 13–19 (1989)
13. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schroder, M.: FEELTRACE: an instrument for recording perceived emotion in real time. In: Proceedings of ISCA ITRW Workshop on Speech and Emotion, ISCA, Belfast, Northern Ireland, 2000, pp. 19–24
14. Coyle, D., Doherty, G., Matthews, M., Sharry, J.: Computers in talk-based mental health interventions. Interact Comput **19**, 545–562 (2007)

15. Cremers, A.H.M., Neerincx, M.A.: Personalisation meets accessibility: towards the design of individual user interfaces for all. In: User-Centered Interaction Paradigms for Universal Access in the Information Society. Lecture Notes in Computer Science, pp. 119–124. Springer, Berlin (2004)
16. Czaja, S.J., Lee, C.C.: The impact of aging on access to technology. Universal Access Informat Soc **5**, 341–349 (2007)
17. Davis, K.H., Biddulph, R., Balashek, S.: Automatic speech recognition of spoken digits. J Acoust Soc Am **24**(6), 637–642 (1952)
18. De Silva, L.C., Miyasato, T., Nakatsu, R.: Facial emotion recognition using multi-modal information. In: Proceedings of the ICICS International Conference on Information, Communications and Signal Processing, Singapore, 9–12 September 1997, pp. 397–401
19. Douglas-Cowie, E., Devillers, L., Martin, J., Cowie, R., Savvidou, S., Abrilian, S., Cox, C.: Multimodal databases of everyday emotion: facing up to complexity. Proceedings of Interspeech, Lisbon, Portugal, September 2005, pp. 813–816
20. Douglas-Cowie, E., Cowie, R., Schröder, M.: A new emotion database: considerations, sources and scope. In: Proceedings of the ISCA Workshop on Speech and Emotion, ICSA, Belfast, Northern Ireland, 2000, pp. 39–44
21. Dudley, H.: The vocoder. Bell Labs Rec **17**, 122–126 (1939)
22. Dutoit, T., Pagel, V., Pierret, N., Bataiile, F., van der Vrecken, O.: The MBROLA project: towards a set of high quality speech synthesizers of use for non commercial purposes. In: Proceedings of ICSLP'96, Philadelphia, October 1996
23. Ekman, P., Friesen, W.V.: Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto (1978)
24. ETSI EG 202 116 V 1.2.1: Human Factors (HF); Guidelines for ICT Products and Services; "Design for All". Sophia Antipolis Cedex, ETSI (2002)
25. Fiscus, J., Ajot, J., Garofolo, J.: The Rich Transcription 2007 Meeting Recognition Evaluation, The Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops, Lecture Notes in Computer Science, vol. 4625. Springer, Berlin (2007)
26. Fogg, B.J.: Persuasive technology: using computers to change what we think and do. Morgan Kaufmann Publishers, Amsterdam (2003)
27. Foulke, E.: A survey of the acceptability of rapid speech. New Outlook Blind **60**, 261–265 (1966)
28. Grimm, M., Kroschel, K., Narayanan, S.: Support vector regression for automatic recognition of spontaneous emotions in speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007), Honolulu, 2007. Publication Date: 15–20 April 2007
29. Hinckley, K.: Input technologies and techniques. In: Jacko, J.A., Sears, A. (eds.) The Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, Chapter 7. Lawrence Erlbaum, Mahwah (2003)
30. Hone, K.S., Graham, R., Maguire, M.C., Baber, C., Johnson, G.I.: Speech technology for automatic teller machines: an investigation of user attitude and performance. Ergonomics **41**, 962–981 (1998)
31. Janse, E.: Production and perception of fast speech. PhD Thesis, University of Utrecht, The Netherlands (2003)
32. Johnson, G.I., Coventry, L.: "You talking to me?" Exploring voice in self-service user interfaces. Int J Hum Comput Interact **13**, 161–186 (2001)
33. Lange, de V.: iCat as personal assistant for diabetic children. MSc Thesis, Delft University of Technology, Delft, The Netherlands (2007)
34. Lemmetty, S.: Review of speech synthesis technology. Master's Thesis, University of Helsinki, Electrical and Communications Engineering (1999)
35. Lindenberg, J., Neerincx, M.A.: The need for a 'universal accessibility' engineering tool. In: ACM SIGCAPH Computers and the Physically Handicapped, Issue 69, pp. 14–17. ACM Press, New York (2001)
36. Lindenberg, J., Pasman, W., Kranenborg, K., Stegeman, J., Neerincx, M.A.: Improving service matching and selection in ubiquitous computing environments: a user study. Pers Ubiquit Comput **11**, 59–68 (2007)
37. Looije, R., Cnossen, F., Neerincx, M.A.: Incorporating guidelines for health assistance into a socially intelligent robot. In: Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication (Ro-Man 2006), University of Hertfordshire, Hatfield, UK, 6–8 September 2006, pp. 515–520
38. Maguire, M.: Methods to support human-centred design. Int J Hum Comput Stud **55**, 587–634 (2001)
39. Martin, A.F., Le, A.N.: Current state of language recognition: NIST 2005 evaluation results. In: Proceedings of Odyssey 2006 Speaker and Language Recognition Workshop, San Juan (2006)
40. McGilloway, S.,Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S.: Approaching automatic recognition of emotion from voice: a rough benchmark. In: Proceedings of the ISCA Workshop on Speech and Emotion, ICSA, Belfast, Northern Ireland, 2000, pp. 207–212
41. Merkx, P.A.B., Truong, K.P., Neerincx, M.A.: Inducing and measuring emotion through a multiplayer first-person shooter computer game. In: van den Herik, H.J., Uiterwijk, J.W.H.M., Winands, M.H.M., Schadd, M.P.D. (eds.) Proceedings of the Computer Games Workshop 2007, Amsterdam, The Netherlands, MICC Technical Report Series, ISSN 0922-8721, number 07-06, pp. 231–242 (2007)
42. Müller, C. (ed.): Speaker Classification I & II. Lecture Notes in Computer Science, vols. 4343 and 4441, ISBN 978-3-540-74186-2. Springer, Heidelberg (2007)
43. Neerincx, M.A., Lindenberg, J.: Situated cognitive engineering for complex task environments. In: Schraagen, J.M.C., Militello, L., Ormerod, T., Lipshitz, R. (eds.) Natural Decision Making & Macrocognition, pp. 373–390. Ashgate Publishing Limited, Aldershot (2008)
44. Neerincx, M.A., Lindenberg, J., Grootjen, M.: Accessibility on the Job: Cognitive Capacity Driven Personalization. HCII 2005, vol. 7: Universal Access in HCI: Exploring New Interaction Environments (10 pages). MIRA Digital Publishing, St Louis (2005)
45. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden Markov models. Speech Commun **41**, 603–623 (2003)
46. Ordelman, R.J.F., de Jong, F., Huijbregts, M.A.H., van Leeuwen, D.A.: Robust audio indexing for Dutch spoken-word collections. In: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC2005), Amsterdam, The Netherlands, 14–17 September 2005, pp. 215–223
47. Oviatt, S.: User-centered modeling and evaluation of multimodal interfaces. Proc IEEE **91**(9), 1457–1468 (2003)
48. Pallett, D.: A look at NIST's benchmark ASR tests: past, present, and future. In: Proceedings of ASRU, St Thomas, US Virgin Islands, 2003, pp. 483–488
49. Picard, R.W.: Affective Computing. MIT Press, Cambridge (1997)
50. Russell, J.A.: A circumplex model of affect. J Pers Soc Psychol **39**(6), 1161–1178 (1980)
51. Schlossberg, H.: Three dimensions of emotion. Psychol Rev **61**, 81–88 (1954)
52. Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., Gielen, S.: Acoustic correlates of emotion dimensions in view of speech synthesis. In: Proceedings of Eurospeech 2001, Aalborg, September 2001, pp. 87–90

53. Schröder, M.: Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis. PhD Thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University (2004)
54. Schuller, B., Müller, R., Lang, M., Rigoll, G.: Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In: Proceedings of Interspeech, Lisbon, Portugal, September 2005, pp. 805–809
55. Scopelliti, M., Giuliani, M.V., Fornara, F.: Robots in a domestic setting: a psychological approach. Universal Access Informat Soc **4**, 146–155 (2005)
56. Sim, G., MacFarlane, S., Read, J.: All work and no play: measuring fun, usability, and learning in software for children. Comput Educ **46**, 235–248 (2006)
57. Stary, C., Stephanidis, C. (eds.): User-Centered Interaction Paradigms for Universal Access in the Information Society. Lecture Notes in Computer Science. Springer, Berlin (2004)
58. Steiniger, S., Schiel, F., Dioubina, O., Raubold, S.: Development of user-state conventions for the multimodal corpus in SmartKom. In: Proceedings of the Workshop on Multimodal Resources and Multimodal Systems Evaluation, Las Palmas, Spain, 2002, pp. 33–37
59. Stephanidis, C. (ed.): User interfaces for all: concepts, methods, and tools. Lawrence Erlbaum, London (2001)
60. Stephanidis, C.: Adaptive techniques for universal access. User Model User-Adapt Interact **11**, 159–179 (2001)
61. Streefkerk, J.W., van Esch-Bussemakers, M.P., Neerincx, M.A., Looije, R.: Evaluating Context-Aware Mobile User Interfaces. In: Lumsden, J. (ed.) Handbook of Research on User Interface Design and Evaluation for Mobile Technology, Chapter XLV, pp. 756–776. IGI Global, Hershey (2008)
62. Truong, K.P, van Leeuwen, D.A., Neerincx, M.A.: Unobtrusive multimodal emotion detection in adaptive interfaces: speech and facial expressions. In: Schmorrow, D.D., Reeves, L.M. (eds.) Foundations of Augmented Cognition, 3rd edn., LNAI 4565 proceedings, pp. 354–363, ISBN 978-3-540-73215-0 (2007)
63. Van der Pijl, D.J., Cremers, A.H.M., Soede, M.: Personalized PDA accessibility for intellectually disabled persons: concept guidelines based on the development of an Electronic Travel Companion. In: HCII 2005, vol. 7: Universal Access in HCI: Exploring New Interaction Environments. MIRA Digital Publishing, St Louis (2005)
64. Van Erp, J.B.F., Kooi, F.L., Bronkhorst, A.W., van Leeuwen, D.L., van Esch, M.P., van Wijngaarden, S.J.: Multimodal interfaces: a framework based on modality appropriateness. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, October 2006, pp. 1542–1546
65. van Leeuwen, D.A.: Consumer off-the-shelf (COTS) product and service evaluation. In: Gibbon, D., Mertins, I., Moore, R. (eds.) Handbook of Multimodal and Spoken Dialogue Systems, Resources, Terminology and Product Evaluation, Chapter 3. Kluwer, Boston (2000)
66. van Leeuwen, D.A., Martin, A.F., Przybocki, M.A., Bouten, J.S.: NIST and TNO-NFI evaluations of automatic speaker recognition, *Computer*. Speech Language **20**, 128–158 (2006)
67. Van Wijngaarden, S.J., Smeele, P.M.T., Steeneken, H.J.M.: A new method for testing communication efficiency and user acceptability of speech communication channels. In: Proceedings of Eurospeech 2001, Aalborg, September 2001, pp. 1675–1678
68. Ververidis, D., Kotropolos, C.: Automatic speech classification to five emotional states based on gender information. In: Proceedings of Eusipco, Vienna, Austria, 6–10 September 2004, pp. 341–344
69. Vidrascu, L., Devillers, L.: Detection of real-life emotions in call centers. In: Proceedings of Interspeech, Lisbon, Portugal, September 2005, pp. 1841–1844
70. Williams, U., Stevens, K.N.: Emotions and speech: some acoustical correlates. JASA **52**, 1238–1250 (1972)
71. Wilting, J., Krahmer, E., Swerts, M.: Real vs. acted emotional speech. In: Proceedings of Interspeech, Pittsburgh, September 2006
72. World Wide Web Consortium. Web Content Accessibility Guidelines 1.0; User Agent Accessibility Guidelines 1.0. http://www.w3.org
73. Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A.: An acoustic study of emotions expressed in speech. In: Proceedings of ICSLP, Jeju Island, Korea, 4–8 October 2004, pp. 2193–2196
74. Zekveld, A., Kramer, S.E., Kessens, J.M., Vlaming, M.S.M.G., Houtgast, T.: The Benefit obtained from visually displayed text from an automatic speech recogniser during listening to speech presented in noise. Ear Hear (2008, accepted for publication)