

2.2.2 Speech processing

Klaus Fellbaum and Diamantino Freitas

2.2.2.1 Introduction and state of the art

Communication is an essential part of human life. If communication is disturbed or impossible, the consequences are loneliness and isolation.

It is well known that speech plays a key role in communication and it explains why humans also want to have speech as a means of communication/interaction with computers. Although human-like speech dialogue with computers is still far off, even with current state-of-the-art technology, the benefits and potential of speech processing are obvious. As will be seen in the next sections, this is especially true in applications for persons with disabilities. Well-known examples are reading machines for blind people, voice control for wheel chairs or speech-based dictation systems for physically impaired computer users.

This chapter presents some new applications for speech-based systems that are (partly) still at the research or prototype stage. Since some of our readers may not be familiar with the principles of electronic speech processing and the state of the art, our presentation will start with some relevant basic definitions.

Speech recognition or equivalently voice recognition is the automatic recognition of spoken words or sentences by a machine. In many cases the result of the recognition is a displayed text and then the terms voice-to-text or dictation system are used. Other important areas for speech recognition are systems for the recognition of spoken commands and the control of basic functions of a personal computer.

There are three main modes for speech recognition.

a) Isolated word recognition up to a vocabulary in the order of 50 000 words and more is on the market. Most of the systems have to be trained before they reach a good level of reliability (up to 98 to 99% correct recognition in controlled environments) or they are speaker-adaptive, that means, at the beginning the recognition accuracy is very moderate, but after intensive use it continues to improve and the accuracy can also reach up to 98...99%.

b) Word spotting or key word recognition is another form of recognition with the aim of recognizing key words in continuous speech. Let us consider, for example, a

2.2. New technologies to help people with disabilities and elderly people

flight information dialogue system where a user wants to know when is the next flight to Brussels, he might ask in a different way like: 'next flight to Brussels' or 'when will be the next flight to Brussels?' or 'please give me the next flight to Brussels'. In all of these cases the key words are obviously 'next' and 'Brussels' and the rest of the words are not relevant. The advantage of word spotting is that the flight destination can be formulated as desired which makes the dialogue much more user friendly.

c) Continuous speech recognition has also reached market maturity but the recognition accuracy still leaves to be desired as regards robustness. The main applications for continuous speech recognition are dictation systems which can recognize more than 1 Million word forms. The term 'word forms' is not equivalent to words. It has to be noted that most words may appear in different forms (basic form, flexions, different tenses etc.) and each word form has to be considered as another word (pattern). That's why such a high number of word forms is needed for ordinary office vocabulary.

A serious problem of all speech recognizers is their sensitivity to noise. However, for certain applications in noisy environments (factory floor, aeroplane cockpit, cars in heavy traffic) very robust recognizers have been developed, but the vocabulary is of moderate size (in the order of some hundred words, isolated mode). This is, on the other hand, not very restricting because the vocabulary being used in such situations is rather limited anyway.

Speaker recognition tries to identify and/or verify the identity of the speaking person and is applied in many security-sensitive situations such as access control to secured areas or bank transactions. State of the art systems have an accuracy (correct recognition) of up to 98%.

Speech replay is the speech reproduction by a technical system (computer etc.). The speech being used was spoken in advance by a person and then stored in a fixed memory or disk. Typical applications are announcement systems (e.g. in public transportation) or system messages. A significant characteristic of a replay system is its limited vocabulary. The speech quality is usually good, in principle it can be increased to a high-quality level, this is only a question of the amount invested in the recording equipment and the storage capacity. It is important to mention that the adequate quality level strongly depends on the application [Jekosch, 2005]. For example, a user accepts a lower quality in a telephone conversation than in a radio announcement.

Speech synthesis has, in contrast to speech replay, an unlimited vocabulary. The speech is concatenated artificially from more or less short speech elements like

2.2. New technologies to help people with disabilities and elderly people

phonemes or diphones or even longer segments. Although speech synthesis has reached an advanced level of maturity, it still suffers from an audible 'machine accent' but since the intelligibility (not necessarily the naturalness!) of synthesized speech is comparable to natural speech, this kind of speech is usable in many practical applications. As a well-known example the screen readers for blind people can be mentioned.

A very important parameter which strongly influences overall speech quality (in both speech replay as well as speech synthesis) is intonation or, more generally, prosody. It is composed of several speech features such as intonation, speed and rhythm, pauses, intensity and is connected to other features such as voice quality (breathy, modal, creaky, etc). All these features, together as a whole multidimensional set, carry so-called supra-segmental information to the utterance that enriches the meaning and can make speech human-like and intelligent. Prosody is the underlying speech layer that conveys pragmatic information. It can also provide para-linguistic and non-linguistic information like intentional and emotional information, respectively [Botinis, 1997].

For more details about the principles of electronic speech processing, the interested reader is referred to the literature; recommended are for example [Furui, 2001], [Gardner-Bonneau, 1999], [Vary, 2006] and, for an extended description of the mathematic principles of speech processing, [Deller, 2000].

2.2.2.2 Speech-based applications for persons with disabilities

Advances in synthetic speech

Multilingual speech synthesis

We are living in a multilingual world. Especially in Europe, different languages are closely related and usually we are trained from school to speak different languages. The same situation exists with written documents or websites. It is thus obvious that most speech synthesis applications (for example enquiry systems or reading machines for blind persons) have to be multilingual.

There are several multilingual systems on the market. One of these was produced by the Bell Laboratories (AT&T, Murray Hill, New Jersey). It functions as a synthesizer for English, French, Spanish, Italian, German, Russian, Romanian,

2.2. New technologies to help people with disabilities and elderly people

Chinese and Japanese. Interestingly, the underlying software for both linguistic analysis and speech generation is identical for all languages, with the exception of English. However, it is clear that the acoustic elements, used for the concatenation into continuous synthetic speech, must be spoken from a native speaker and thus this part of synthesis is language-dependent. The same holds for the base data of linguistic analysis. However, these components are stored externally in tables and parameter files and they can be loaded when needed in real-time. A detailed description of the AT&T Synthesis can be found in the book of Sproat [Sproat, 1998]; the synthesis of different languages is demonstrated on [Synthesis testsite AT&T].

Another system which became very popular in the speech synthesis society is the MBROLA system. *“The aim of the MBROLA project, initiated by the TCTS Lab of the Faculté Polytechnique de Mons (Belgium), is to obtain a set of speech synthesizers for as many languages as possible, and provide them free for non-commercial applications. The ultimate goal is to boost academic research on speech synthesis, and particularly on prosody generation, known as one of the biggest challenges taken up by Text-To-Speech synthesizers for the years to come.”* More details and demos are presented on the home page of MBROLA [MBROLA].

Emotional speech

Emotional speech can remarkably extend the content and expression of spoken information. Moreover, sometimes the way how items are expressed is more important than what is expressed. The key parameter which determines the emotional content is the prosody as discussed before.

A great deal of work has been done in the recognition and production of emotional speech; among others, there was the EU FP6-IST project HUMAINE (Human-Machine Interaction Network on Emotion). For more information visit the home page which is under [HUMAINE].

In a man-machine communication, let's consider a speech-based dialogue system, emotional speech can be used in two directions:

a) The user speaks with emotions and the system has to recognize these emotions in addition to the 'pure' speech recognition. As an example, a situation might occur where the system does not sufficiently recognize the user and reacts in an unsatisfying manner. This is very often annoying and leads to an angry voice. If this anger is recognized by the system, then it might be wise for it to react with excuses and/or an explanation why the recognition failed [Lee, 2002].

2.2. New technologies to help people with disabilities and elderly people

b) If the system produces speech (be it stored or synthetic speech), it can in principle be used to express emotions. Everyone has a need to transmit emotions. But if we think of deaf persons or those with severe speech disorders or people suffering from muscular dystrophies and cerebral diseases that often have also aphasia along with body paralysis, these persons are unable to express their emotions through speech although they have a strong desire to do so.

Several research groups have investigated emotional speech. Concerning the speech quality and, above all, the naturalness of the recognizability of the emotions, the results are encouraging; see for example [Burghardt, 2006].

Lida, Campbell and Yasumura [Lida, 1998] describe an application concept of an affective communication system for people with disabilities who cannot by themselves express their emotions. They get help from some buttons for the selection of emotions. These 'emoticons' are very helpful and they can be easily added to an ordinary text-to-speech synthesis (figure 2.6).

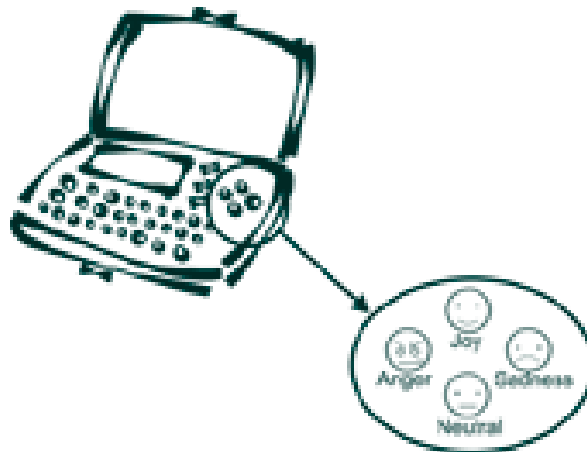


Figure 2.6 EMOTICONS (Emotion keys).

In many cases, a user (who cannot speak as well as a normal speaking person) needs a synthesizer to produce a specific voice from a selected person or with specific features. The underlying concept which fulfills this requirement is called voice personalisation. This facility is very interesting when there is the need to transmit synthetic speech from a text given by a specific person. Voice personalisation is nowadays available at a constantly decreasing cost with the advent of statistical speech-model-based speech synthesizers [Barros, 2005].

Support of a speech conversation for hard of hearing or deaf persons

In this application two persons have, for example, a telephone conversation. One person has normal hearing, the other has a severe hearing impairments. The idea is now to support the hard of hearing person with additional visual information, either in the form of an animated face or as text or in both forms which are presented on a screen (figure 2.7).

The technical implementation works as follows. The speech of the normal hearing person is automatically recognized by a high-level speech recognition system. The result is a text which can be displayed. In the next processing step the text is converted into control parameters for a talking head. At least the person with hearing problems can receive the message in three versions: as original speech, as text and as an animated face. It is assumed here that the hard of hearing person speaks normally, which is quite common.

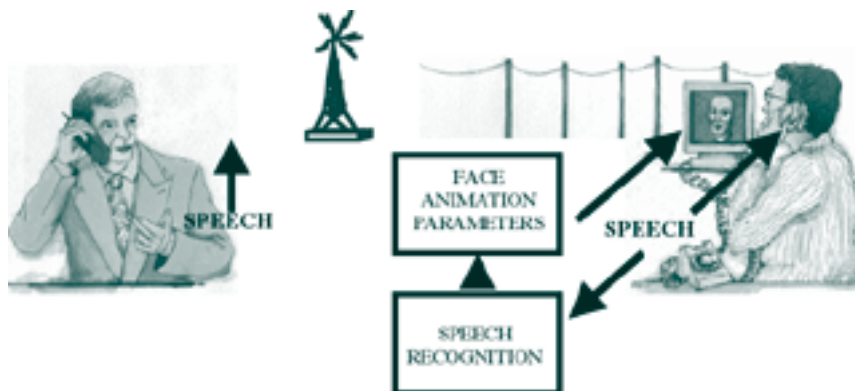


Figure 2.7 Telephone conversation, the partner on the right is hard of hearing.

If the person is deaf, he or she will not have serious problems to understand the message by reading the text and watching the animated face. But problems arise when the deaf person wants to respond to the normal hearing person. This problem will be discussed in the next section.

There are several research projects dealing with speech to text or speech to animated faces. One of it is SYNFACE which was developed at the KTH in Stockholm until 2004 [SYNFACE, 2005]. In the meantime it has become a commercial product. The speech recognition is based on phoneme recognition and a speech synthesizer activates the talking head, mainly the lips. The movements of the talking head are synchronized with the telephone speech and thus the listener

2.2. New technologies to help people with disabilities and elderly people

can directly complete the part of the information which he or she does not hear.

A similar system that is on the market is iCommunicator [icommunicator]. The system aims mainly at the group of deaf persons, but also at those who are hard of hearing. The kernel of the system is the Dragon Naturally Speaking Professional Engine [DRAGON, 2006], at the moment one of the best and most powerful speech-to-text systems on the market. iCommunicator runs on a higher end laptop computer. Among other features, iCommunicator converts in real-time, speech to text, speech to video sign language, speech to computer generated voice, text to computer-generated voice or to video sign language.

A third system, which can be mentioned here, was developed in a project called MUSSLAP at the University of West Bohemia in Pilsen, Czech Republic. One of the outcomes was a real-time recognizer which presents its results as text on the screen. As a very impressive example, an ice hockey match is shown on a tv screen and the system automatically recognizes the comments of the reporter and displays the result as text in real-time [MUSSLAP].

Speech processing for the communication of a deaf person

If deaf persons communicate over a distance (telecommunication), a very common method since a long time is text telephony or fax which also has the advantage that the communication between deaf and normal hearing persons is possible without any problem. For several years, SMS has also served as a cheap and widespread communication tool. Above all, the Internet with its many services (for example Web and email) has dramatically widened the communication in general and specially between deaf and normal hearing persons.

On the other hand, text communication has some drawbacks: text information is rather impersonal and the typing procedure is laborious and time consuming and not all deaf people have a sufficiently high level of understanding of written language to be able to access text.

For these reasons most deaf persons prefer sign language communication. This form of communication has remarkable advantages:

- *Sign language is fast and its speed is comparable to speech because sign language is produced 'multidimensional'. The 'speaker' can express several items at the same time, for example using in parallel both hands and face expressions. There are, however, some exceptions, which make sign language slower, among others, finger spelling or some complicated words can be mentioned, but these exceptions do not significantly affect the average speed*

2.2. New technologies to help people with disabilities and elderly people

- *Sign language is individual. Persons can (beside the 'pure' information) express their personal accentuations and emotions.*

The adequate tool for a sign language communication is obviously video telephony, mostly using a standard like H.320 which is also compatible to ISDN. With the advent of UMTS (3G) and WLAN, a mobile video communication became reality. In both cases usually relay services are applied to connect deaf users, but also, with the aid of a sign language interpreter, deaf and normal hearing subscribers can (indirectly) communicate.

Several projects exist which work on sign language transmission. One is the European IST project WISDOM (Wireless Information Services for Deaf people On the Move, lifetime from 2000 to 2004), in which several advanced wireless services for the Deaf were developed and evaluated [WISDOM].

The situation is different when a direct (face-to-face) situation between a normal hearing and a deaf person is considered. As a first observation it comes out that the communication is obviously much easier from the side of the deaf person because he or she has learned to understand a speaking person by lip reading and watching face expressions and gestures. Although this special form of 'human speech recognition' is never perfect (among other reasons because some sounds are invisibly produced inside the mouth), often fragmented utterances can be completed by the context. It is interesting to state that the recognition of emotions works rather well by watching the face movements and gestures.

For an additional support of the communication process for the deaf person, a speech to text and/or a speech to sign language transformation, as described in the previous section, might be useful. The result of such a transformation can be presented on a display or, more advanced; it could be beamed to little mirrors in the spectacles of the deaf person.

But looking at the other direction: what about the normal hearing person who does not understand sign language?

If we imagine this situation, we can state that - even without any knowledge of sign language - valuable information is transferred about the intention of the deaf person and his/her emotions when we watch gestures, mimic, body movements and other kinds of visual information. In this respect, the situation is similar to those of the other communication direction (from the speaking to the deaf person). The key problem is the recognition of the objective, content-carrying part of the message. For this we can come back to the relay service solution. The deaf person has a camera (maybe as a part of a mobilephone) which records the gestures to

2.2. New technologies to help people with disabilities and elderly people

the interpreter who translates them into speech, which is then audible for the hearing person. This procedure works well as several projects (also the WISDOM project) have shown, but the problem here is the availability of the interpreter and the fact that a face-to-face situation often happens unforeseen.

Obviously a better solution would involve an automatic gesture recognition which transforms gestures into synthetic speech. In this case the normal hearing person receives the information of the deaf person twice: as gestures and as voice and both forms of information complete each other. There is no need for emotional synthetic speech because emotions are optimally expressed by gestures and the face, as mentioned before.

It is important to state that automatic gesture recognition or, more extended, automatic sign language recognition, is probably one of the most difficult research tasks in the area of communication aids. Difficulties are:

- *the detection and triggering of face and hands*
- *the ambiguity and individual variations of the movements, above all from the hand movements and*
- *the structure of the sign language itself which has no strict one-to-one relation to text or speech syntax.*

The first systems for sign language recognition were based on the data glove(s). These gloves are well-known tools, mostly used in the Artificial Intelligence research and in entertainment applications. The advantage of such a glove is the precision with which hand positions are recognized. But for many situations in the daily life, the use of gloves might be too uncomfortable.

A better (but much more complicated) alternative are video-based systems. The deaf person uses sign language, a video camera recognizes gestures and facial movements (above all lip movements) and as result of the video processing, the sign language is transformed into text which can be displayed somewhere and/or the text can be transformed into synthetic speech. A very detailed description of problems and solutions in that area are presented in a recently published book on human interaction with machines [Kraiss, 2006].

We will now briefly mention some research projects.

In the framework of the European IST research program ARTHUR, the Lab. of Computer Vision and Media Technology, Aalborg University Denmark investigated the automatic recognition of hand gestures used in wearable Human Computer

2.2. New technologies to help people with disabilities and elderly people

Interfaces [Moeslund, 2003]. Different gesture detection devices are described, among others the 'classical' data glove and reduced versions of it (index finger tracker with a wired or wireless connection to the receiver), a 'Gesture Wrist', a 'Gesture Pendant' and, of course, camera solutions.

A famous researcher, Christian Vogler, who is deaf himself, has made his PhD in automatic recognition of American Sign Language (ASL). He describes the problem of simultaneous events in sign language (for example, the handshape can change at the same time as the hand moves from one location to another, or hand(s) and face express signs simultaneously). Another important aspect is the segmentation of the continuous stream of movements into discrete signs and the breaking-down of signs into their constituent phonemes. If this works satisfactorily, the next steps, namely transformation of signs into text and then into synthetic speech, are relatively easy to manage. For more information see [Vogler, 2000].

Thad Starner and his group from Georgia University of Technology, Atlanta USA, are working on several projects in American Sign Language recognition. They use multiple sensors for the recognition, among others a hat-mounted video camera and accelerometers with three degrees of freedom mounted on the wrist and torso to increase the information of the video camera. For control reasons, the deaf user has a head-mounted display which shows what the camera captures [Brasher, 2003]. The aim of the activities is a flexible mobile system for the output of text or speech, depending on the application. Figure 2.8 shows the head-mounted camera and a recorded gesture.



*Figure 2.8 Base-cab-mounted camera and a recorded gesture
(with kind permission of Thad Starner, Media Lab, MIT).*

Visual and audio-visual speech recognition based on face or lip reading

A methodology which is quite similar to gesture recognition, mentioned before, is automatic facial reading or lip reading. The result is a text sequence which represents the content of the utterance. Figure 2.9 shows the region that is investigated for lip reading.

2.2. New technologies to help people with disabilities and elderly people



Figure 2.9 The region of interest of the video facial image.

The automatic recognition of facial images has been used for a number of years for the improvement of a (spoken) speech recognition under noisy conditions and it has been proved to be very successful [Kraiss, 2006], [Moura, 2006], although the accuracy, obtained with purely visual speech recognition, is not as high as in audio speech recognition. There are a number of reasons for this; one is that visual speech is partially phonetically ambiguous.

Nevertheless, for the communication between deaf and normal hearing persons, facial or lip reading is a very valuable help and, as previously mentioned, the human face can optimally express emotions and this information is detectable for the visual recognizer.

Small-vocabulary preliminary trials have been reported [Moura, 2006] to obtain word recognition rates of about 65% for a one speaker lip-reading task with grammar correction. Interestingly, the performance of professional observer was in the range of 70%-80% for the same corpus. Figure 2.10 shows the situation under remarkable noise conditions and it demonstrates the advantage (in terms of recognized words error rate – WER) of a simple combination in a multi-stream recognition approach [Moura, 2006].

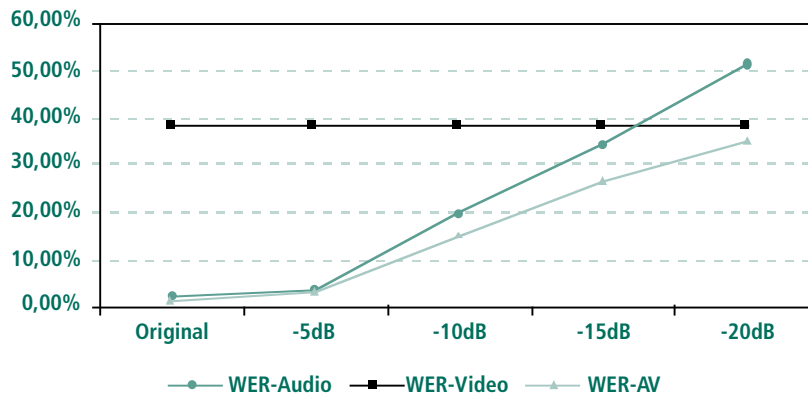


Figure 2.10 Variation of the total word error rate in function of the signal-to-noise ratio.

Correction of speech defects, unintelligible speech

If a person is unable to speak 'normally' resulting in unsatisfactory intelligibility, a speech recognition and synthesis system can be a valuable aid. The impaired speech is the input for the recognizer, which converts it into text and the text is then converted into clean synthetic speech.

It is very important to state that even totally unintelligible speech or any acoustic utterance can be recognized, the only prerequisite is the ability of the 'speaker' to reproduce utterances with sufficient similarities and to train the recognizer with this kind of 'vocabulary'. As a matter of fact, even emotions can be expressed, using emotional speech synthesis. Finally, visual speech recognition, as mentioned before, can significantly contribute to better speech recognition.

A system for speech therapy

It is well known that many deaf persons have fully functioning speech organs but the problem is that they cannot control articulation because they do not have acoustic feedback through the ears.

When the deafness occurred after the complete language/speech acquisition, the deaf person can maintain (with restrictions) his/her speaking ability with the help of a speech therapist. But there is the necessity of a permanent training with a therapist which is obviously not always possible.

Many attempts have been made to develop systems which perform a visual control of a spoken utterance. The time signal or the spectrum of the speech are not very suitable because the relation between the sound production and the resulting signal is rather complicated and abstract.

A better solution is obviously a face animation showing two speaking faces: the 'reference' face and the (deaf) speaker's face. Thus the deaf person can directly see deviations between the two faces and he or she can try to adapt. Since some sounds are produced invisibly inside the mouth, as mentioned earlier, a useful help is a transparent mouth region (figure 2.11).

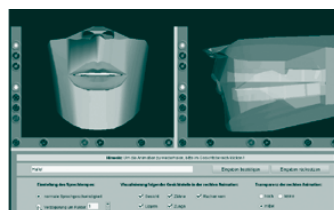


Figure 2.11 Face animation with a transparent area of the mouth region [Pritsch, 2005].

Screen readers for blind or partially sighted persons

The usual computer desktop metaphor practically leaves blind persons out because it is a Graphical User Interface (GUI), based on a more or less rich graphic display of icons, windows, pointers and text. Since blind persons require non-visual media, the alternative is, among tactile information (Braille), primarily an aural interface which can be called, analogous to GUI, Aural User Interface (AUI), based on the terminology supported by many authors including T.V.Raman [Raman, 1997].

Since the early 80's, after some trials with special versions of self-voicing software, capable of driving a speech synthesizer and so providing access for blind persons, a more general concept appeared and a family of applications, called screen-readers, was initiated with the purpose of creating a vocal rendering of the contents of the screen under user control through the keyboard, using a text-to-speech converter [Wikipedia]. In this way, properly installed screen reader software stays active in the operating system and operates in the background, analysing the actual contents of the screen. From the initial command-line interface (CLI) to the now existing ubiquitous graphical user interface (GUI) screen reader software has evolved much in 2,5 decades.

Screen readers can also analyse many visual constructs like menus and alert or dialogue boxes and transform them into speech to allow interaction with a blind user.

Navigation in the screen is possible as well, to allow a non-linear or even random exploration and acquisition of the depicted information. Control of the produced speech is normally given to the user so that quite fast navigation becomes possible when the user works with shortcuts. A simulation of a screen reader is available at the WebAIM website [WebAIM].

Although many screen reader applications exist, there are many limitations that current screen readers cannot overcome per se, for instance those related to images and structured text (tables etc.). Screen readers cannot describe images, they can only produce a readout of a textual description of these and the user has problems to realize how the page is organized.

The basic requirement in terms of speech processing for screen reader applications is a robust text-to-speech converter with the possibilities of spelling and reading random individual characters and all kinds of text elements that may appear like numeric expressions, abbreviations, acronyms and other coded elements. Punctuation is also spoken in general, besides being determinant in introducing some prosodic manipulation in the synthetic voice.

2.2. New technologies to help people with disabilities and elderly people

Following this idea, the World Wide Web Consortium (W3C) in 1998, with the issue of the Cascading Style Sheet 2 (CSS2) recommendation, has introduced the Aural Cascading Style Sheet (ACSS); a chapter respective to the acoustical rendering of a web page is presented in [WDAC].

Auditory icons, sometimes also called earcons, are made audible to the user by means of a loudspeaker or earphone system that should have advanced acoustic features (high quality, stereo etc.). The acoustic elements contain voice properties like speech-rate, voice-family, pitch, pitch-range, stress, and others that are used as command parameters to the speech synthesizer.

An extended investigation of spatial acoustic features as a component of a screen reader was performed in the GUIB (Graphical User Interfaces for the Blind) project in the framework of the European TIDE initiative [Crispien, 1995]. The idea was to generate an acoustic screen in front of the user on which windows, icons and other graphic elements are audible on different places, and the mouse position is also audible when the mouse is moving.

In a former project (AudioBrowser, 2003-2005, see [Repositorium]), developed for Portuguese, but applicable for most other languages, the structure or outline of a web page can be discovered and used as a table of contents, and it was implemented successfully. The user in this application can freely navigate inside the contents of each window or jump between windows from contents to tables of contents or vice-versa in order to scan or navigate through the page in a more structured and friendly way. The blind or low-vision user is constantly helped by the text-to-speech device that follows the navigation accurately.

The W3C consortium, through its Web Accessibility Initiative (WAI) has been issuing a relevant set of Web contents accessibility guidelines (WCAG), now in version 2. These guidelines are greatly helpful in orienting web page design for accessibility [WAI]. Authoring Tool Accessibility Guidelines (ATAG), nowadays in version 2.0, are also important for developers of authoring tools.

Reproduction of complex documents for blind persons

Complex documents like mathematical and other scientific, technical or even didactic documents are usually equipped with graphical representations. Above all, equations and other mathematical expressions have posed a substantial barrier to the access by visually impaired persons. Most representations and charts may also be included in this group.

2.2. New technologies to help people with disabilities and elderly people

Representation in special Braille codes of complex mathematical elements can almost totally solve the problem for blind persons. The LAMBDA project [LAMBDA, 2005] has produced a mathematical rendering package using such a system.

In the case of more lengthy mathematical objects, more refined solutions might be preferable using audio rendering of the mathematical expressions through synthetic speech. Using the codification of the expression in MathML, a browsable textual description of the expression can be automatically derived from the MathML code by means of a special lexicon and a grammar. Both must be specially designed for the purpose according to the mathematical conventions and concerns of non-ambiguity of the textual description. This work has been carried out in the AUDIOMATH project [Ferreira, 2005] carried out at the Faculdade de Engenharia da Universidade do Porto. A demonstration page is available at [Ferreira].

Acoustical cues, contributing to the clearness of the speech rendering, are also important. Previous authors have used, for instance, prosodic modifications such as raising or lowering the pitch of the synthetic voice to signal upper or lower parts of the expression, respectively. In the work of AUDIOMATH the influence of pitch movements as well as of pauses during description of expressions was studied and rules were extracted. An intra-formula navigation mechanism was designed in order to allow the user to explore the formula at her/his own will thereby not putting too much stress on audio memory in the case of longer formulas.

2.2.2.3 Conclusions and future developments

The aim of this chapter was to show how electronic speech processing works and how persons with disabilities can benefit from it.

Since speech is man's most important form of communication, all efforts must be done to make speech communication possible, and if the speech channel is disturbed, technical solutions have to be found to overcome the obstacles.

The accuracy and quality of modern speech recognition systems as well as synthesis systems has reached a state of maturity which allows the development of very powerful support systems for persons with disabilities and to bridge the gap between these persons and those without disabilities, as was shown, for example, between deaf persons and the rest of the world.

Looking into the future of speech technology, some important research areas can be identified as follows:

2.2. New technologies to help people with disabilities and elderly people

- *Improving the robustness of speech recognition systems. Although the robustness has been remarkably improved over the last years, the systems are still far behind human capabilities. Noise, especially non-stationary noise, background speakers or music can still reduce the recognition reliability well below an acceptable error rate. Improvement is expected (and partly proven, as has been seen) from a multimodal recognition which includes also visual information (above all, mimics, facial and hand gestures)*
- *A more extended use of semantic and pragmatic information. When the system (recognizer or synthesizer) 'knows' what the speaker wants to express, which covers both, the content and the emotion, then the recognizer can usefully complete a spoken message which has recognition errors. A synthesizer could automatically generate the right accentuations and emotional 'colouring' of the speech. For the sake of completeness it has to be mentioned here that the permanent improvement of the quality of synthetic speech also includes multilinguality as well as speaker-specific synthesis and will remain within the scope of research. Audio rendering of complex documents through synthetic speech is also a very important development area where document description strategies, their conversion into full text form and intra-document navigation or browsing are the crucial steps*
- *A challenge and wide field of research is sign language recognition. As mentioned earlier, there are several research activities but much more work has to be done. More needs to be known about structures of sign languages (and there are very many and all are different!) and their relations to spoken and written languages. Automatic translations should be possible in different directions (sign language into speech and vice versa, sign language into another sign language, speech into a foreign sign language and vice versa, for example German speech into American sign language). Also the technical part of the problem is challenging. Using the Ambient Intelligence (Aml) approach, we can expect micro cameras in the clothes or in a pendant as well as position sensors in finger rings etc., and the environment will have enough intelligence to take on most of the processing activities needed for recognition and translation*
- *For blind persons, screenreaders and the automatic recognition of graphics, pictures and the environment are a never ending research area. As a matter of fact, for blind persons a verbal (spoken) description of the recognition result is, in many cases, the best solution. As before, Aml will be of crucial importance here.*

It should be mentioned here that the enumeration given in this chapter from being complete. Further examples will be given in other chapters, showing that speech technology and speech applications will play a dominant role whenever communication is discussed.

2.2.2.4 References

BARROS M.J., MAIA R., TOKUDA, K. RESENDE, F.G., FREITAS, D., (2005). HMM-based European Portuguese TTS System, artigo apresentado e publicado nas actas da Interspeech'2005 - Eurospeech — 9th European Conference on Speech Communication and Technology, Lisboa.

BOTINIS (ed.) et al., (1997). Intonation: Theory, Models and Applications. Proceedings of the ESCA Workshop Sept. 18-20 Athens, Greece.

BRASHER, H., STARNER, T. et al., (2003). Using Multiple Sensors for Mobile Sign Language Recognition. ISCW White Plains, WA,
Also: http://www-static.cc.gatech.edu/~thad/031_research.htm

BURGHARDT, F. et al., (2006). Examples of synthesized emotional speech
<http://emosamples.syntheticspeech.de/>

CRISPIEN, K., FELLBAUM, K. (1995). Use of Acoustic Information in Screen Reader Programs for Blind Computer Users: Results from the TIDE Project GUIB. In: Placencia Porrero, I., & de la Bellacasa, R. P., (Eds.): The European Context for Assistive Technology - Proceedings of the 2nd TIDE Congress, Paris, , IOS Press, Amsterdam.

DELLER, J.R., (2000). Discrete-time processing of speech signals.
New York : Institute of Electrical and Electronics Engineers.

DRAGON Naturally Speaking Professional Engine, (2006). NUANCE communications <http://www.nuance.com/naturallyspeaking/>.

FERREIRA, H., FREITAS, D., (2005). AudioMath—Towards Automatic Readings of Mathematical Expressions”, 11th International Conference on Human Computer Interaction, Las Vegas, EUA.

FERREIRA. <http://lpf-esi.fe.up.pt/~audiomath>

FURUI, S., (2001). Digital speech processing, synthesis, and recognition
2nd ed., rev. and expanded. New York : Marcel Dekker.

2.2. New technologies to help people with disabilities and elderly people

GARDNER-BONNEAU, D., (1999). Human Factors and Voice Interactive Systems. Kluwer Academic Publishers, Boston.

HUMANE, Network of Excellence. <http://emotion-research.net/aboutHUMAINE>. iCommunicator homepage. <http://www.mycommunicator.com/>].

IIDA, A., CAMPBELL, N., YASUMURA, M. (1998)., Emotional Speech as an Effective Interface for People with Special Needs, *apchi*, p. 266, Third Asian Pacific Comp. and Human Interaction.

JEKOSCH, U., (2005). Voice and Speech Quality Perception. Springer-Verlag Berlin, Heidelberg.

KRAISS, K.F., (ed.), (2006). Advanced Man-Machine Interaction. Springer Berlin Heidelberg, New York.

SYNFACE project research page <http://www.speech.kth.se/synface/>.

LEE, C.M., PIERACCINI, R., (2002). Combining Acoustic and Language Information for Emotion Recognition. Proc. of the International Conference on Speech and Language Processing (ICSLP 2002). Denver, Co.

LAMBDA (2005). <http://www.lambdaproject.org/>.

MBROLA website <http://tcts.fpms.ac.be/synthesis/>.

MOESLUND, T., NORGAARD, L., (2003). A Brief Overview of Hand Gestures used in Wearable Human Computer Interfaces. Technical Report CVMT 03-02, Computer Vision and Media Technology Lab., Aalborg University, DK.

MOURA A., PÊRA V., FREITAS, D., (2006). (in Portuguese) Um Sistema de Reconhecimento Automático de Fala para Pessoas Portadoras de Deficiência", artigo publicado nas actas da conferência IBERDISCAP'06, realizada em Vitória-ES, Brasil.

MUSSLAP. University of West Bohemia, MUSSLAP website <http://www.musslap.zcu.cz/en/audio-visual-speech-recognition/>.

PRITSCH, M., (2005). Visual speech training system for deaf persons. Proceedings of the 16th Conference Joined with the 15th Czech-German Workshop "Speech Processing, Prague, Sept. 26-28, 2005. TUD press Dresden, Germany.

RAMAN, T.V., (1997). Auditory User Interfaces, Kluwer Academic Publishers, August.

2.2. New technologies to help people with disabilities and elderly people

RAMAN, T.V., (1998). Conversational gestures for direct manipulation on the audio desktop, Proceedings of the third international ACM SIGACCESS Conference on Assistive Technologies, Marina del Rey, California, United States, pgs 51 – 58. ISBN:1-58113-020-1.

REPOSTIRORUIM. <https://repositorium.sdum.uminho.pt/bitstream/1822/761/4/iceis04.pdf#search=%22audiobrowser%22> SPROAT, R. (ed.) (1998).: Multilingual Text-to-Speech Synthesis. Kluwer Academic Publishers. Dordrecht, Boston, London.

SYNFACE - Synthesised talking face derived from speech for hard of hearing users of voice channels
<http://www.speech.kth.se/synface/> and <http://www.synface.net/>.

SYNTHESIS TESTSITE, AT&T. <http://www.research.att.com/~ttsweb/tts/demo.php>.

VARY, P., MARTIN, R., (2006). Digital Speech Transmission. Enhancement, Coding and Error Concealment. J. Wiley&Sons.

VOGLER, C. et al. A Framework for Motor Recognition with Applications to American Sign Language and Gait Recognition.
<http://www.cis.upenn.edu/~hms/2000/humo00.pdf>
see also Vogler's homepage <http://gri.gallaudet.edu/~cvogler/research/>.

WAI. Web accessibility homepage. <http://www.w3.org/WAI/>

WDAC (1999). Aural Cascading Style Sheets (ACSS), W3C Working Draft
<http://www.w3.org/TR/WD-acss>.

WebAIM Screen Reader Simulation.
<http://www.webaim.org/simulations/screenreader.php>

Wikipedia about screenreader http://en.wikipedia.org/wiki/Screen_reader

WISDOM project page. <http://www.bris.ac.uk/news/2001/wisdom.htm>.