Ambient Voice Control for a Personal Activity and Household Assistant

Niko Moritz, Stefan Goetze, and Jens-E. Appell

Fraunhofer Institute for Digital Media Technology (IDMT),
Project group Hearing, Speech and Audio Technology (HSA),
Marie-Curie-Str. 2, 26129 Oldenburg, Germany
{niko.moritz,s.goetze,jens.appell}@idmt.fraunhofer.de

Abstract. Technologies for ambient assisted living (AAL) are used to increase the quality of life of older or impaired persons. This contribution discusses the utilization of automatic speech recognition (ASR) as a natural interface for control of assistive technologies in everyday life situations. We focus on the use of hands-free systems, the technical challenges for the ASR software caused by this and the benefits for older persons. Moreover, state-of-the-art approaches for improving robustness of ASR systems are presented, discussed and demonstrated by an ASR experiment.

Keywords: Automatic Speech Recognition, Acoustic User Interfaces, Hands-Free, Distant Speech Recognition.

1 Introduction

Due to the demographic change that will lead to an increased amount of older persons [1-3], new solutions are necessary to cope with the challenges that will occur e.g., in the field of social care. Assistive technologies show great potential to support older people in their everyday life situations and, by this, allow for a longer independent period of life [4]. In this context, information and communication technologies (ICT) are of high social relevance. However, especially for older people this kind of technology is often not easy to use due to its inherent high complexity. For this reason the design of simple and intuitive user-interfaces is of great importance [4-6].

The use of speech is the most natural form of human communication. Therefore, the control of assistive technologies by speech recognition software is desirable and offers opportunities for an intuitive and user-friendly navigation of assisting devices in a home environment for example. Moreover, in situations in which the hands are not free to be used, speech input offers great advantages over conventional input methods, such as a remote control or mouse and keyboard [6, 7].

However, automatic speech recognition (ASR) in combination with hands-free equipment in adverse acoustic environments is still a challenging problem [8, 9]. The reliability (i.e. 'robustness') of current speech recognition systems still is far from reaching the performance of humans [10]. For this reason, in this contribution we will describe results of a user study that was conducted and in which the tolerance

threshold for erroneous recognitions was determined. Moreover, the user acceptance for such defective systems will be reviewed (cf. Section 2).

To improve recognition rates in acoustically adverse conditions several approaches exist. On the one hand concepts for improving the signal-to-noise ratio (SNR) of noisy and reverberant speech signals are applied (cf. Section 3.1 and references therein). These signal processing concepts can be used for pre-processing the speech input of ASR systems, to increase their robustness especially if hands-free equipment is used. However, it should be mentioned that these concepts may also cause distortions of the acoustic features an ASR system relies on, which can pose problems for the speech recognition algorithms even if the signal quality perceived by a human listener is improved [9]. Thus, benefits of possible pre-processing concepts must always be regarded in connection with the used speech recognition concepts.

On the other hand, in addition to an appropriate pre-processing also other concepts exist to improve robustness of the ASR system itself. Current methods are presented and discussed in Section 3.2. One common method is to use a structured dialogue for the ASR system instead of recognition of continuous speech, for example. The advantage is twofold. Firstly, the complexity of the speech recognition software can be kept small and secondly, the control of those systems by spoken commands is favoured by many users as it is shown e.g., in [11]. The main reasons that were given by users in [11] for preferring a structured dialogue are that this type of control is unambiguous and already known from other applications.

Section 4 of this contribution presents results of an ASR experiment, which was conducted in an AAL living lab that is built similarly to a conventional living room. In this living lab a distant ceiling microphone and a close-talk microphone were used for the experiments described in Section 4. The performance was compared in terms of word error rates and the used microphone for training and testing.

2 User Acceptance Study

In [11] a user study has been conducted to evaluate the user tolerance threshold and the acceptability of erroneous command recognitions. For that purpose a total of 12 subjects aged from 63 to 75 years were consulted. The experiment was made using a mock-up system, which simulated a calendar application as part of a personal activity and household assistant [12]. This means that the in- and output of the system was controlled by a human investigator, who responded to the commands of the subject. The investigator, thus, took over the duties of the ASR software, whereby previously planned failures of the system could be investigated. The output of the system was presented acoustically to the user by previously recorded speech sentences. By this approach three test phases were pre-defined to prove the users tolerance level. In each phase, the subject could enter a new appointment to the calendar application by spoken commands and after each phase the subject was asked about his or her impressions of the command controlled calendar application.

During the first test phase, the ASR system worked perfectly and didn't make any mistakes. By this, the subject could get used to the system and got an idea of how the system works in general. In the second test phase, one recognition error was introduced, which means that the subject had to correct the incorrectly recognized part

of his or her speech input by repetition of the corresponding commands. In the third and final test phase continuous recognition errors were introduced by the investigator. By this, an infinite loop of corrections had to be done by the user to measure his or her frustration tolerance. A yellow and a red card were handed out to the subjects. By using the yellow card a warning could be given to the system, which was considered as a first indication of frustration and the red card stopped the test. By this procedure the subjects' limit of tolerance for mistakes of the ASR software was determined.

After each phase the attributes "user-friendliness"," intuitiveness of use", "comprehensibility", "helpfulness" and "acceptance" of the presented device were evaluated. These properties were assessed on a scale between 1 ("not applicable at all") and 5 ("applies wholeheartedly"). The result after the first phase was that the user-friendliness and acceptance were rated very well (between 4 and 5 by all 12 test persons). The result regarding helpfulness of the system was also rated well (between 3 and 5). The assessments of the comprehensibility and intuitiveness were somewhat inhomogeneous. The interested reader is referred to [11] for a more detailed discussion of these results.

The assessment of the system after the second test phase (with one incorporated error) did not change significantly.

The evaluation of the frustration threshold within the third test phase leads to an interesting result. The amount of repetitions due to faulty command recognitions ranged between 0 and 17 until the subject showed the red card to abort the test. The median of repetitions was six. As stated in [11], the high tolerance threshold for this experiment was not expected by the investigators. Nine of twelve subjects stated that they would use such a system in the future, although the willingness to use such a system in current state was rather low. Reasons to be mentioned for these findings are that the mock-up system was still quite limited in features and in possible output modalities, i.e. only acoustic output was available for instance. Reference [7] and [12] provide a more detailed overview of the multi-modal outputs of the personal activity and household assistant.

Also many users assigned that they would like to activate the system by a keyword or by clapping. Reference [13] describes event detection methods in the context of AAL, which can be used for this purpose.

3 Challenges and Demands for Voice Controlled Ambient Assistive Technologies

Robust automatic speech recognition with non-close-talk recordings, i.e. with recordings captured by distant microphones ('hands-free'), is still a major challenge for research in ASR. This is due to the spatial distance between the used microphones and the source of the desired speech signal, which is generally the users mouth. Therefore, microphones do not only pick up the clean speech signal, but also ambient noise sources and delayed copies of the original signal (reverberation), which strongly depends on the environment in which the system is used.

Fig. 1 illustrates a user scenario of a system which can be installed as an ambient assistive device at home. The provided example shows a situation in which the microphones of the device capture several signals from different sources. On one

hand this is the desired signal of the user who provides speech commands to the system, and on the other hand these are undesired noise signals which interfere with the speech signal. The constituted noise sources in Fig. 1 are a vacuum-cleaner and also the loudspeakers of the device itself, that play back the acoustic output of the system. Those parts of the signal that are uttered by the loudspeakers and picked-up by the microphones again are commonly known as acoustic echoes [14]. Thus, the user's command input is heavily disrupted by (i) ambient noise, (ii) acoustic echoes and (iii) room reverberations, namely echoes introduced to the desired speech signal due to reflections at room boundaries as obvious e.g., from speech in churches.

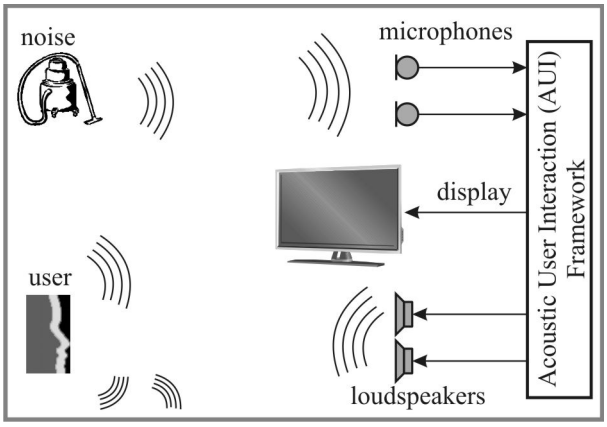

Fig. 1. Schematic of the technical structure of the personal activity and household assistant. The acoustic user interaction framework contains the ASR system and the intelligence to control the in- and outputs depending on the specific application.

Different signal processing strategies to enhance the desired signal and to suppress others are described in Section 3.1. Section 3.2 presents a common strategy for improving the recognition performance of an ambient ASR system used in adverse acoustical conditions. Furthermore, an insight to a current field of research for enhancing robustness in ASR is provided.

3.1 Pre-processing Concepts for Speech Quality Enhancement

As depicted in Fig. 1 multiple microphones as well as multiple loudspeakers can be used for sound pick-up and play back in hands-free systems. The microphone signals are processed by the signal processing unit (acoustic user interaction (AUI) framework in Fig 1), which enhances the captured signal and analyses the content of the signal by means of ASR. Feedback is given by the system acoustically via its loudspeakers.

Fig. 2 shows the signal processing unit AUI for the single-channel case in more detail. Here, the desired signal part, e.g., a spoken command, is denoted as $s_n[k]$.

Ambient noise $n[k]$ and acoustic echoes $\psi[k]$ are disturbances for the ASR system that are superimposed to the desired signal part in the microphone signal. The acoustic output of the system $s_f[k]$ is played back by the loudspeaker. Due to the acoustic coupling between loudspeaker and microphone, parts of the signal are picked up by the microphone again. Numerous reflections of the signal at the room boundaries (walls, floor and ceiling) lead to a reverberated version of the system output at the position of the microphone. Mathematically these reflections are characterized by the so-called room impulse response $h[k]$ as depicted in Fig. 2. Since the system output may contain speech information, this could heavily disturb the ASR system if no suppression filter is used. So-called acoustic echo cancellation filters $c_{\mathrm{AEC}}[k]$ estimate the signal part stemming from the loudspeaker signal contained in the microphone signal, whereby its cancellation is principally possible [15, 16]. Highly reverberant environments and multiple loudspeakers pose particular challenges to such acoustic echo cancellers. The interested reader is referred e.g., to [14-18] for a more detailed discussion of the technical challenges.

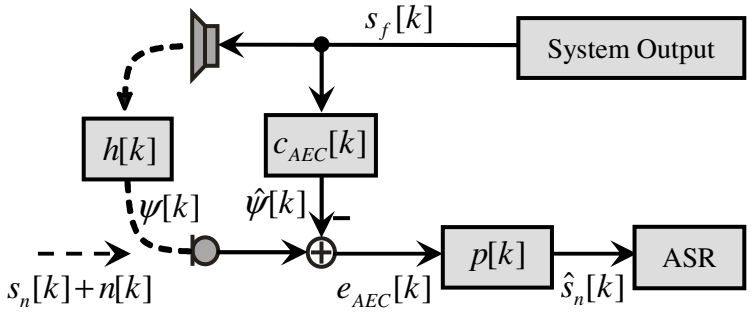

Fig. 2. Suppression of ambient noise and acoustic echoes.

Residual echoes that may remain after the compensation point of the acoustic echo canceller $c_{\mathrm{AEC}}[k]$ as well as additional disturbances $n[k]$ that are picked-up by the microphone are suppressed by the succeeding suppression filter $p[k]$ before the enhanced signal is analysed by the ASR software [9, 19]. Such suppression filters generally do not only suppress disturbances but may also introduce distortions to the desired part of the signal. Although these distortions may be small in amplitude they might heavily distort the acoustic features an ASR system relies on. Therefore, distortions of the desired part of the signal have to be kept as small as possible [9].

Spatial distortions from different directions can be reduced by the use of multiple microphones as in Fig. 1. Like humans are able to focus on a spatial direction by exploiting information from their two ears and suppress acoustic sources stemming from other directions, this is also possible with signal processing strategies based on multiple microphones [9, 19, 20].

Another signal processing concept in ASR is to spread several microphones in the room and then selecting that one, which provides the best voice quality for the recognition process [21]. This normally is the microphone which is closest to the

speaker's mouth, since it generally provides the best SNR. However, the usage of more sophisticated approaches is worthwhile.

3.2 Concepts in Automatic Speech Recognition to Improve Recognition Performance for Acoustically Adverse Conditions

Besides an appropriate signal pre-processing additional concepts for improving ASR performance for acoustically adverse conditions exist. Two major concepts regarding the ASR algorithms itself are presented in the following.

The first concept is to use training data that was recorded in the same conditions as those in which the speech recognition device will be used later. By this, the acoustic models, such as the hidden Markov models (HMM) for instance, learn the noise characteristics, which may disturb the speech signal [8, 23-25]. Also convolutional noise like channel characteristics and reverberation can be learned to a certain extent by the acoustic models. However, if the acoustic environment changes (e.g., the amount of ambient noise, spectral noise characteristics or the reverberation time), benefits of this condition-adopted training can get reduced or even lost [8, 23]. Therefore, to obtain a reliable improvement, this method is only applicable if the acoustic conditions, in which the ASR system will operate, are precisely known. If the system is used in changing environmental conditions, but which are known in advance, then the ASR device may switch between different acoustic models that were trained under these different conditions. For this case a reliable detection of the current acoustic condition is needed, which can be a challenging problem as well [24].

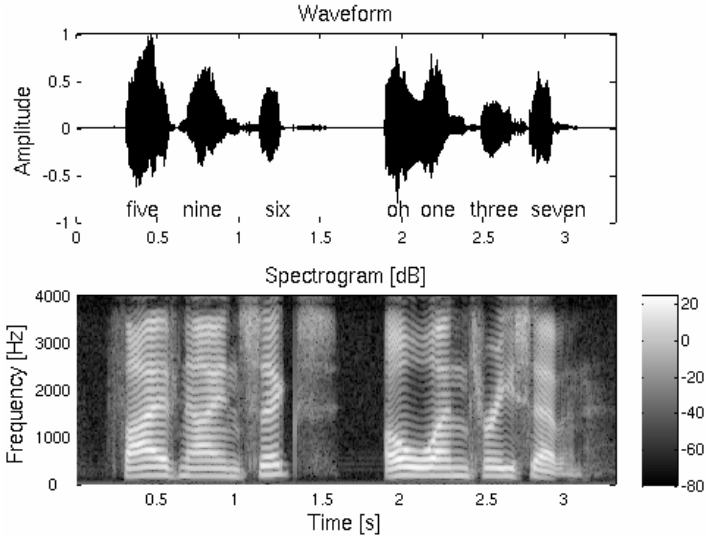

Fig. 3. Time-domain representation of a speech sequence (upper panel) and its spectro-temporal representation, i.e. spectrogram (lower panel).

Another concept to increase recognition rates in acoustically adverse conditions is a proper selection of the acoustic features that are used to represent the characteristics

of speech. These have to match two important requirements. Firstly, the properties that allow for distinguishing between the decisive characteristics of the different units of speech must be mapped as well as possible. Secondly, disturbances of the acoustic signal, such as noise and reverberation, should influence the properties of interest in feature domain as little as possible.

Most feature extraction methods in ASR are based on the spectral representation of the signal's waveform. To obtain this, the acoustic signal is first divided into short overlapping segments of typically 20-30 ms duration that are then transformed to the frequency domain by a Fourier transformation. This procedure is called short-time Fourier transformation (STFT). The result is a time versus frequency representation, which is commonly known as spectrogram (as depicted in the lower panel of Fig. 3).

Furthermore, a filter bank motivated by the human auditory system, such as the Mel or Bark filter bank, is usually applied to the spectrogram [8, 22, 23]. Such filter banks group certain frequency sections to frequency bands, which shall approximate the frequency resolution found in the inner human ear. For today's feature extraction methods further computational steps are necessary, however detailed descriptions are beyond the scope of this paper. Here, the interested reader is referred to [8, 22, 23].

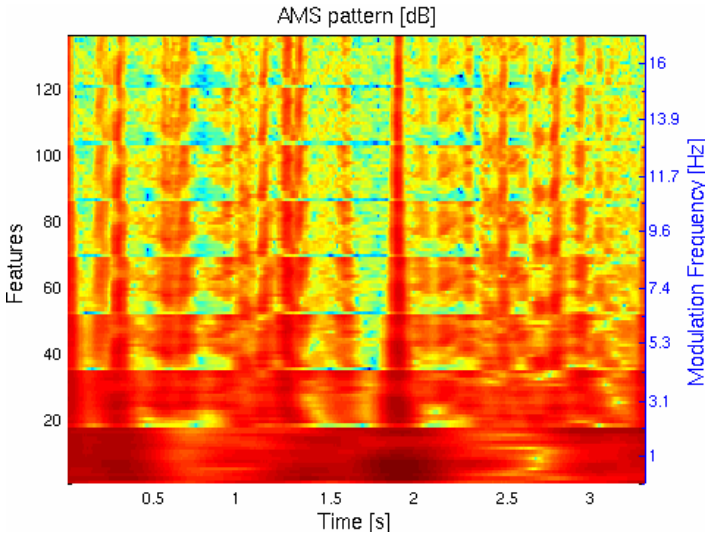

Fig. 4. AMS pattern of the signal depicted in Fig. 3. In order to make all three dimensions visible, patterns of time versus acoustic frequencies are stacked for each modulations frequency.

Instead, a feature extraction method is exemplarily presented in the following, to demonstrate which kind of approach is investigated in current research for improving robustness of ASR systems. Current research tends towards analysing longer time trajectories of spectral envelopes, which were taken from the spectrogram representation. As an example the so-called amplitude modulation spectrogram

(AMS) is introduced [26]. This feature type analyses about 300 to 350 ms long time trajectories taken from a spectrogram, which was previously half-wave rectified, squared and frequency decomposed into Bark bands [27]. By this, the amplitude modulations are analysed from the acoustic signal. Subsequently the modulation frequency domain gets filtered by a band pass filter to ignore modulation frequencies lower than 1 Hz and greater than 16 Hz. Thus, only the relevant modulation frequencies for speech are passed, whereby non-speech influences can already be reduced [28-30].

Fig. 4 shows an example of an AMS pattern for the signal depicted in Fig. 3. Since AMS is a three-dimensional representation of a signal, namely time versus acoustic frequency and modulation frequency, it is non-trivial to generate an easily interpretable graphical representation. For this reason modulation frequencies and acoustic frequencies are stacked in Fig. 4. This can be seen by the axis labels on the right, which indicates that patterns of time versus acoustic frequencies are stacked for each modulation frequency. For the example presented in Fig. 4 this results in a 136 dimensional feature vector (17 Bark band frequencies times 8 modulation frequencies) for each time frame.

It should be mentioned that dimension reduction methods (such as the principal component analysis (PCA)) are usually applied before this representation is used as an acoustic feature [8, 25]. This is done since classification algorithms, such as HMMs, are more effective if the dimensionality of feature vectors is low and if the feature components are decorrelated [8, 25].

4 An Automatic Speech Recognition Experiment in an AAL Scenario

In this section an ASR experiment is presented to demonstrate the influence of channel distortions and room reverberation on the performance of a speech recognizer. To do so, recordings of eight male speakers were collected in an AAL living lab. A room decorated like a living room was equipped with almost invisible ceiling microphones. These ceiling microphones represent the hands-free equipment for the conducted ASR experiment. In addition, a close-talk microphone was situated in front of the speaker's mouth. For the recordings the speaker was sitting on an armchair in the middle of the room turning the head towards the television screen (cf. Fig. 5).

The synchronous recordings of spoken commands with the close-talk microphone and with one ceiling microphone that was situated in the middle of the room (c.f. Fig. 5) were then used to train and test a speech recognition system. 25 different command words have been selected and recorded for the purpose to manage a calendar application of the personal activity and household assistant. Each of these commands was recorded ten times from each speaker in a silent environment. The testing was done with the so-called cross validation procedure [25]. For this, the recordings of seven speakers were used for training and the recordings of the remaining speaker were used for testing. By this, all eight possible combinations were tested and an average word error rate was determined.

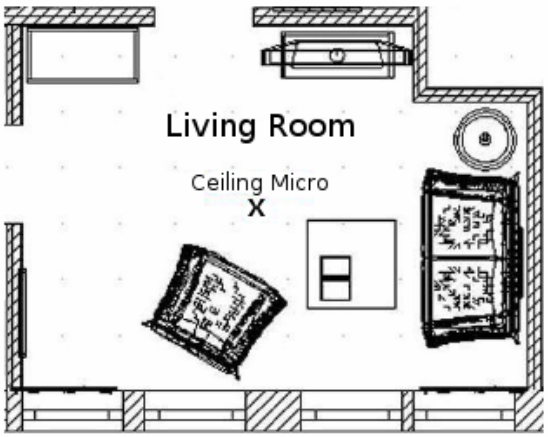

Fig. 5. Layout of the acoustic environment in the living lab. The position of the ceiling microphone, which is used for the experiment, is indicated by an 'X'. The armchair in the bottom centre of the room indicates the position of the speaker.

The most commonly used acoustic features, namely the Mel-Frequency Cepstral Coefficients (MFCCs - here including the 0^{th} cepstral coefficient plus dynamic features such as deltas and double deltas) [8, 22, 23], are also used for this experiment. The classifier relies on the statistical description by linear whole-word hidden Markov models (HMM) with 14 states each (including the two non-emitting states).

Three test scenarios were defined:

a) The recordings of the close-talk microphone are taken for training and the ceiling microphone for testing.
b) The ceiling microphone is used for training as well as for testing.
c) The close talk microphone is used for training as well as for testing.

Table 1. ASR results for the different test scenarios. SD: Standard Deviation.

Test scenario	**a)**	**b)**	**c)**
WER (SD) in %	56,9 (26,5)	2,2 (2,3)	1,1 (2,1)

Table 1 shows the results in terms of word error rates (WER). The WER is calculated by the ratio of mistakes made by the ASR system and the total amount of spoken words. Mistakes that may occur are deletions (*DEL*), substitutions (*SUB*) and insertions (*INS*) of words. Expressed in a formula that is

$$WER = \frac{\#DEL + \#SUB + \#INS}{N} \tag{1}$$

where N is the total amount of spoken words to be recognized.

The results show that the performance strongly depends on the training data. For the case that the recordings for training and testing are produced in the same way (see results of scenario b) and c)), the system achieves very low WERs. As soon as the recordings of the close-talk microphone are taken for training, which can be considered to be free of reverberation, and reverberated speech taken by the ceiling microphone is used for testing, then strongly increased WERs can be observed (see results of a)). In addition to room reverberation also channel distortions and internal noise due to different types of microphones used in the ceiling and for close-talk plays a role for the increased WERs observed in test scenario a).

The results demonstrate that it is necessary to account for the specific environmental conditions and known distortions when preparing training data for speech recognition software. This is especially a very effective concept, if hands-free equipment is used, which is permanently installed in the room since changes of the spatial conditions do not have to be expected. However, it should be mentioned that in addition other difficulties, which are not considered by this experiment, are present when using hands-free equipment. These problems result from the fact that non-close-talk microphones do not only pick up sound from the desired speaker, but also from noise sources in the vicinity and other talking people, for instance.

5 Summary and Conclusion

In this paper an overview of current challenges and demands for acoustic user interfaces for AAL technologies with a focus on automatic speech recognition (ASR) is provided. In addition, the discussed results are supported by an ASR experiment in an AAL living lab environment.

In Section 2 a user study was presented evaluating the acceptance of a voice controlled calendar application. Furthermore, the user's tolerance limit for erroneous recognitions caused by a mock-up simulation of an ASR system was evaluated for the targeted user group of age 63-75 years in this study. The result was that speech input is accepted by the older users even when the system produces recognition errors.

Section 3 presented an exemplary approach of a personal activity and household assistant. Based on this exemplary application, possible challenges for the use of ASR systems have been discussed which typically will occur when ASR is used for hands-free acoustic user interaction with assistive technologies at home. In this context, different concepts for improving robustness of ambient voice controlled devices have been presented. Also a brief insight in acoustic feature selection has been introduced as an example for a current field of research in ASR.

Finally, in Section 4 results of an ASR experiment with a hands-free system were shown. This was done to demonstrate the importance of considering the specific environment in which the ASR system will be used. It could be shown that recognition rates will considerably fall if known distortions are not taken into account. The considered distortions in this experiment are room reverberations and the microphone characteristic that has an effect in the use of different microphone types.

As an overall conclusion it can be summarized that the development of ASR as a technology for user interaction with assistive technologies has reached sub-goals, so far. The performance of state-of-the-art systems sill does not reach the recognition

performance of humans. This holds particularly when ASR is used with hands-free equipment (e.g., ambient microphones) in arbitrary acoustical environments.

Since the use of ASR in combination with hands-free equipment is a very natural way to interact with assistive technologies and because hands-free ASR has a large potential to make assistive technologies unobstrusive and much easier to use at the same time, the concepts for enhancing the robustness in ambient ASR (as presented in this contribution) are promising from a technology as well as from an AAL application oriented perspective.

Acknowledgments. This work was supported in parts by the Lower Saxony Ministry of Science and Culture, Germany, through the "Niedersächsisches Vorab" grant programme within the Lower Saxony Research Network "Design of Environments for Ageing (GAL)".

References

1. European Commission Staff: Working Document. Europes Demografic Future: Facts and Figures. Commission of the European Communities (2007)
2. Statistical Federal Office of Germany: Demographic Changes in Germany: Population Development in Germany (In German original language: Demografischer Wandel in Deutschland - Heft 1 - Bevölkerungs- und Haushaltsentwicklung im Bund und in den Ländern) (2007)
3. Statistical Federal Office of Germany: Demographic Changes in Germany: Impacts on Hospital Treatments and People in Need of Care (In German original language: Demografischer Wandel in Deutschland - Heft 2 - Auswirkungen auf Krankenhausbehandlungen und Pflegebedürftige im Bund und in den Ländern) (2008)
4. European Ambient Assisted Living Innovation Alliance: Ambient Assisted Living Roadmap. VDI/VDE-IT AALIANCE Office (2009)
5. Alexandersson, J., Zimmermann, G., Bund, J.: User interfaces for AAL: How can I satisfy all users? In: Proc. Ambient Assited Living Congress, Berlin, Germany (2009)
6. Rennies, J., Goetze, S., Appell, J.-E.: Personalized acoustic interfaces for human-computer interaction. In: Ziefle, M., Röcker, C. (eds.) Human-Centered Design of E-Health Technologies: Concepts, Methods and Applications. IGI Global (2010)
7. Meyer, E.M., Heuten, W., Meis, M., Boll, S.: Multimodal Presentation of Ambient Reminders for Older Adults. In: Proc. 3. Deutscher AAL Kongress, Berlin, Germany (2010)
8. Wölfel, M., McDonough, J.: Distant Speech Recognition. Wiley, Chichester (2009)
9. Mildner, V., Goetze, S., Kammeyer, K.-D.: Multi-Channel Noise-Reduction-Systems for Speaker Identification in an Automotive Acoustic Environment. In: Proc. Audio Engineering Society (AES), 120th Convention, Paris, France, May 20 -23 (2006)
10. Lippmann, R.: Speech recognition by machines and humans. J. Speech Communication 22, 1–15 (1997)
11. Goetze, S., Moritz, N., Appell, J.-E., Meis, M., Bartsch, C., Bitzer, J.: Acoustic User Interfaces for Ambient Assisted Living Technologies. In: Proc. Informatics for Health and Social Care, vol. 35(4), pp. 161–179 (2010)
12. Meis, M., Fleuren, T., Meyer, E.M., Heuten, W.: User centred design process of the personal activity and household assistant: Methodology and first results. In: Proc. 3. Deutscher AAL Kongress, Berlin, Germany (2009)

13. Schröder, J., Wabnik, S., van Hengel, P.W.J., Goetze, S.: Detection and Classification of Acoustic Events for In-Home Care. In: Proc. 4. German AAL-Congress, Berlin, Germany (2011)
14. Breining, C., Dreiseitel, P., Hänsler, E., Mader, A., Nitsch, B., Puder, H., Schertler, T., Schmidt, G., Tilp, J.: Acoustic Echo Control – An Application of Very-High-Order Adaptive Filters. IEEE Signal Processing Magazine, 42–69 (1999)
15. Hänsler, E., Schmidt, G.: Speech and Audio Processing in Adverse Environments. Springer, Heidelberg (2008)
16. Goetze, S., Xiong, F., Rennies, J., Rohdenburg, T., Appell, J.-E.: Hands-Free Telecommunication for Elderly Persons Suffering from Hearing Deficiencies. In: Proc. 12th IEEE International Conference on E-Health Networking, Application and Services (Healthcom 2010), Lyon, France, July 1-3 (2010)
17. Benesty, J., Morgan, D.R., Sondhi, M.M.: A Better Understanding and an Improved Solution to the Specific Problems of Stereophonic Acoustic Echo Cancellation. IEEE Trans. on Speech and Audio Processing 6(2), 156–165 (1998)
18. Goetze, S., Kallinger, M., Kammeyer, K.-D., Mertins, A.: Enhanced Partitioned Residual Echo Estimation. In: Proc. Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA (2006)
19. Bitzer, J., Simmer, K.U.: Superdirective microphone arrays. In: Brandstein, M.S., Ward, D. (eds.) Microphone Arrays: Signal Processing Techniques and Applications, pp. 19–38. Springer, Berlin (2001)
20. Doblinger, G.: Localization and Tracking of Acoustical Sources. In: Topics in Acoustic Echo and Noise Control, pp. 91–122. Springer, Berlin (2006)
21. Wolf, M., Nadeu, C.: On the potential of channel selection for recognition of reverberated speech with multiple microphones, Interspeech, Japan (2010)
22. Benesty, J., Sondh, M.M., Huang, Y.: Springer Handbook of Speech Recognition. Springer, New York (2008)
23. Rabiner, L., Juang, B.-H.: Fundamentals of speech recognition. Prentice-Hall, Englewood Cliffs (1993)
24. Tchorz, J., Kollmeier, B.: Automatic classification of acoustical situation using amplitude modulation spectrograms. J. Acoust. Soc. Am. 105(2), 1157 (1999)
25. Bishop, C.M.: Pattern recognition and machine learning. Springer, Heidelberg (2006)
26. Kollmeier, B., Koch, R.: Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. J. Acoust. Soc. Am. 95(3), 1593–1602 (1994)
27. Moritz, N., Meyer, B.T., Anemüller, J., Kollmeier, B.: Robustness of automatic speech recognition with amplitude modulation spectrograms (In German original language: Robustheit automatischer Spracherkennung mit Amplitudenmodulationsspektrogrammen). In: German Annual Conference on Acoustics (DAGA), Berlin, Germany (2010)
28. Kanedera, N., Arai, K., Hermansky, H., Pavel, M.: On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Communication 28, 43–55 (1999)
29. Kanedera, N., Hermansky, H., Arai, T.: On properties of modulation spectrum for robust automatic speech recognition. In: Proc. ICASSP 1998, pp. 613–616 (1998)
30. Hermansky, H.: RASTA processing of speech. IEEE Trans. Speech and Audio Processing 2(4), 578–589 (1994)