Manuscript - VDE Verlag GmbH (Hrsg): Tagungsband zu Ambient Assisted Living, 1. Deutscher Kongress mit Ausstellung, pp. 371-375, 2008

Developing Enabling Technologies for Ambient Assisted Living: Natural Language Interfaces, Automatic Focus Detection and User State Recognition

Florian Hönig, Institute of Pattern Recognition (LME), University of Erlangen-Nuremberg, Germany
Christian Hacker, LME, University of Erlangen-Nuremberg, Germany
Dr.-Ing. Volker Warnke, Sympalog Voice Solutions GmbH, Erlangen, Germany
PD Dr.-Ing. habil. Elmar Nöth, LME, University of Erlangen-Nuremberg, Germany
Prof. Dr.-Ing. Joachim Hornegger, LME, University of Erlangen-Nuremberg, Germany
Prof. Dr. med. Johannes Kornhuber, Department of Psychiatry and Psychotherapy, University Hospital Erlangen, Germany

Abstract

Parallel to the demographic change, research on technologies for assisted living to support the "silver generation" is growing. Indispensable for such technology are intuitive and practical user interfaces. We take an approach based on automatic speech recognition that satisfies these requirements. We present a prototypical system that provides e.g. control of household appliances or initiates and accepts telephone calls. It has a natural-language interface: spontaneous speech is allowed, and the user does not have to learn special commands. Also reminder functions for taking medicine and an emergency call are implemented. Wishing to avoid a "push-to-talk-button", we provide algorithms for automatic recognition of the intended addressee of the speech. Additionally, we classify the focus of attention from video recordings which can be useful if the system's interface uses an avatar. Another goal in assisted living is to automatically monitor the user's health which can be accomplished by measuring body functions. Sensors for acquiring these physiological signals can nowadays be integrated into the clothing; the signals can be transmitted wirelessly. In this paper, generic algorithms for automatic distinction between affective user states are presented. In the future, we will apply these solutions to health monitoring.

1 Introduction

Due to a low rate of birth and increasing expectation of life through improved health care, a dramatical demographic change is expected in most industrial nations in the near future. In Germany, the average age of the population is predicted to rise from about 40.9 years in 2003 to an estimated 50 years in 2050. A consequence is that more people will be in need of care while less people are available for the employment market. This will lead to a serious shortage of health care services.

The use of ambient assisting technologies at home [1, 2] could alleviate this imbalance by increasing the autonomy and quality of life of elderly people and supporting medical therapy by screening. Thus, the need to move to a health-care facility can be avoided or postponed.In this paper, we present a prototypical system for ambient assisted living: ISA-house („Intelligentes Seniorenangepasstes Haus“, intelligent house designed for elderly people), a system that provides e.g. control of household appliances or initiates and accepts telephone calls. Our approach implements the most natural interface: speech. It is based on a state-of-the-art automatic speech recognition system and a dialogue system that can process natural language: spontaneous speech is allowed, and the user does not have to learn special commands.

Further, we present new techniques that can be used to improve ambient assisting technology. Automatic user focus detection can help deciding whether the resident is addressing ISA-house or talking to another person. Our approach is able to utilize both audio and video input. Further, we present novel solutions for affective user state recognition with the help of physiological signals. The sensors can be integrated into the clothing, and the measured body functions can yield important information about the user, as they are often linked to affective processes. ISA-house could utilize the result of the user state classification for an adaptive dialogue, for example exhibit a special behavior if the user is stressed. These techniques can also be transferred to medical screening, allowing the system to support medical treatment or to alert the doctor when necessary.

2 ISA-House

ISA-House is a prototypical system for studying the possibilities of technology to support elderly people to stay in the comfort of their home rather than move to a health-care facility. It has a natural-language interface:

spontaneous speech is allowed, and the user does not have to learn special commands. It provides a central, unified interface to various technical equipment such as radio, television, DVD player, heating, lighting, and household appliances. Device-dependent details are hidden, and communication between the user and the devices is handled on the abstraction layer of a natural-language dialogue. ISA-House can autonomously handle in- and outbound telephone calls. The resident can e.g. ask "Please call my daughter!" which will cause ISA-House to query the personal address database and establish the desired telephone connection. Furthermore, reminder and monitoring functions support correct medical treatment and help regular food intake.

Currently implemented functions of the demonstrator

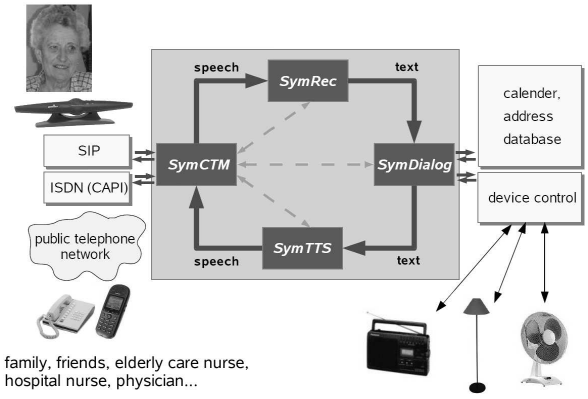

Fig. 1: System architecture of ISA-house

system are:

- Natural-language interface
- Autonomous handling of outbound telephone calls
- Control of the radio (volume and broadcast station)
- Control of simple devices like lamp or fan
- Query function for date and time
- Plug-in architecture for easy functionality upgrade

The architecture of the system is shown in Fig. 1. Speech input takes place either via a head-mounted microphone, a microphone array which can be placed anywhere in the room, or a telephone call. A microphone array can be used for echo cancellation which is otherwise a big problem for the accuracy of an automatic speech recognition system. Another approach is to use a mobile phone as an interface. An advantage is that the ISA-house interface can also be accessed from outside, e.g. for inquiring whether the oven is switched off. Inside the house, the telephone can be connected via bluetooth or WLAN with the system. Suitable mobile phones for the elderly with large keys etc. are already on the market[1].

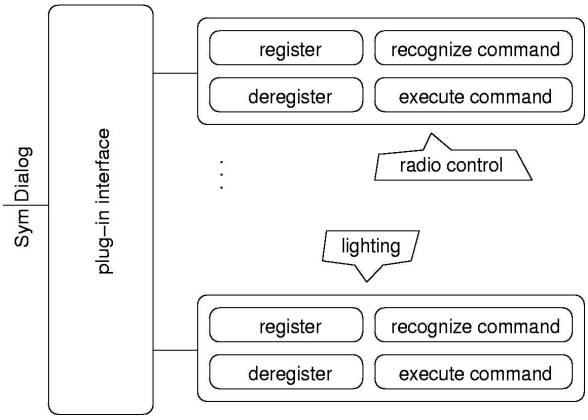

Fig. 2: Plug-in architecture of ISA-house

The core of the system is an ensemble of Sympalog tools (see Fig. 2). Sympalog[2] is a spin-off company of the Institute of Pattern Recognition; it operates several commercial telephone-based dialogue systems [3, 4]. SymCTM is a call and task manager which handles incoming requests and outgoing messages or puts through in- and outbound telephone calls. It has a robust barge-in module that activates the speech recognizer SymREC if speech is detected. SymREC decodes the utterance and passes the most likely spoken word sequence to SymDialog, an easily configurable dialogue manager. Based on the content of the recognized utterance this module analyses whether the system is being addressed and interprets the message. For this purpose, it can also access a calender or address database. Depending on the content, a desired command is executed or an answer is generated and passed to the speech synthesizer SymTTS. If the information from the user is incomplete, further enquiries to the user are made via SymTTS. The dialogue memory of SymDialog resolves back-references that occur in such conversations with the user.

For executing a command or acquiring external information, SymDialog communicates with several devices via a unified plug-in-architecture (see Fig. 2). Every plug-in has to implement four interfaces: "register", "deregister", "recognize command" and "execute command". SymDialog uses the "recognize command" interface to determine if any and which device is being addressed, and calls the "execute command" interface. Additionally to household appliances, various sensors can easily be integrated into this framework. For example, special sensors would allow un-intrusive control of the residents's behavior: if the water sensor indicates that the person is not using the toilet often enough, this can be a hint that the person is dehydrated and should be reminded to drink – a function that ISA-house is suited to fulfill.

One of the major advantages of the system is a free conversation – the user does not have to learn commands. This is exemplified by the following requests:

1 e.g. „Katharina das Große“, www.fitage.com

2 www.sympalog.com

- "I would like to talk to my daughter! – a telephone call is initiated.
- "I'd like to listen to Bayern 1!" – ISA-house starts the Internet radio tuned to desired broadcast station.
- "Switch the lights on, please!" – turns the light on.
- "What time is is now?" – the speech synthesis module informs the user about the current time.
- "Please call the nurse!" – ISA-house calls the health care center, autonomously conducts the conversation and then informs the resident when the nurse will come.

Currently, many extensions of the system are being implemented. Two issues receive special attention: A robust detection of the user focus that utilizes paralinguistic information, and a user state recognition module using physiological signals that may be used to adapt the dialogue to the current affective state of the user. These aspects are discussed in the two following chapters.

3 Classification of User Focus

The most natural interface between the user and ISA-house is speech. The speech recognition engine should assist the resident whenever it is necessary. However, it is also important that the system does not disturb in everyday life, unless it is required. It has to work unobtrusively in the background. Besides the challenges of robust speech recognition and natural dialogue, it is also important that the system recognizes whether it is addressed by the user or not. This means that the system has to perform the following analysis steps:

1. Is speech present in the moment, or is there silence or background noise that could possibly be produced by household appliances or by the television or the radio?
2. If speech, does this speech address the system, or does the resident talk to s.o. else? Maybe he or she even talks to herself or is reading aloud?

Whereas in state-of-the-art speech recognition systems usually the first step is implemented using voice-activity-detection and background acoustic models, the classification of the addressee of speech is subject of current research. In the German SmartWeb project[3] we have analyzed the user's focus of attention, which can be seen as baseline research for follow-up projects on ambient assisted living. If the user talks to the system, *On-Talk* is classified, whereas *Off-Talk* summarizes all kinds of speech that is not directed to the system [5]. In SmartWeb, additionally the images of a video camera are analyzed (the dialogue system was running on a smartphone; here, the integrated camera was used to analyse, whether the user looks onto the display or not). Looking onto the display and talking to the system is denoted as *On-Focus; Off-Focus* means that the system is not addressed. In the ISA-house scenario this video information is only available, if the user interface additionally provides e.g. an embodied conversational agent or an avatar that attracts attention. However, in some cases the system must prioritize speech and react to an utterance even if the resident does not show visual attention, e.g. when the user is crying for help.

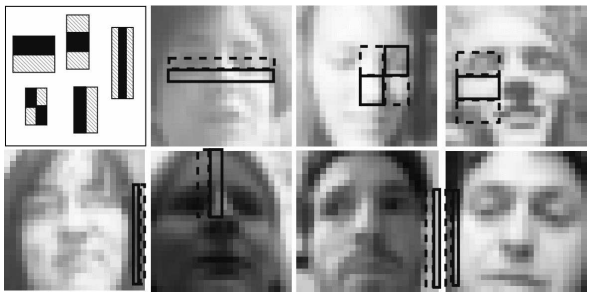

Fig. 3: The 7 best features of the SmartWeb face detector. Top left: Different shapes of Haar-wavelets

In SmartWeb, the focus of attention was analyzed using prosodic, video, and linguistic information. Prosodic information is extracted from the speech signal by analyzing the energy of the speech and its variation, the fundamental frequency (pitch) and its variation, jitter, shimmer, and the rate-of-speech, as well as the duration of words, syllables, and pauses. The video signal is analyzed employing face-detection algorithms and counting the portion of detected faces for different parts of the utterance. On the SmartWeb data, a Viola-Jones face detector [6] that is based on Haar-like wavelet features was trained. The 7 best features are shown in Fig. 3. Finally, for the linguistic analysis a very generic, scenario independent approach has been chosen: each word of the spoken word sequence is replaced by its part-of-speech (POS) category (e.g. noun, verb, particle, etc.). Then, the variation of categories is analyzed, as well as the portion of function words and content words. Algorithmic details are described in [7].

The system was evaluated using the SmartWeb Video Corpus which contains 3.2 hours of speech from 99 speakers. The test candidates had to interact with an automatic dialogue system and were disturbed by a second person in order to evoke Off-Talk. The candidates did not know anything about On-Talk/Off-Talk.

For the automatic detection of On-Talk/Off-Talk it could be shown that energy is important, but also jitter and shimmer. Using 100 prosodic features for automatic classification, on average each of the two classes On-Talk and Off-Talk are correctly classified with 77 % recognition rate. When combining prosody with linguistic information and video information, On-Focus vs. Off-Focus is detected with 85 % recognition rate.

3 http://www.smartweb-projekt.de/

In the future, this recognition results can be improved using scenario dependent linguistic information (e.g. if the resident talks about calling s.o., this utterance is more likely to be addressed to the system) and by applying speaker adaptation techniques (the number of users is restricted by the number of residents). Decision thresholds can be adjusted depending on whether it is more problematical to reject a request that is directed to the system or to response to speech that is not directed to the system. Additionally, a clever call back initiated by the system can clarify ambiguous cases and avoid confusion.

4 Affective User States

Recent research on Human-Machine-Interfaces increasingly tries to enhance user interfaces by incorporating the affective state of the user, aiming at increased pleasantness, effectiveness and safety. The user interface of ISA-house for example, could show a specific dialogue behavior if the user is stressed. An intriguing possibility of determining the affective user state is the use of physiological signals such as skin conductivity or heart rate, as the measured body functions are often linked to affective processes [8]. Wireless and integrated sensors are currently being developed by a number of research institutes; unlike visual or auditory clues, most physiological signals are not under voluntary control and cannot be masked up to the same extent.

In this work, we present methods for automatically recognizing the affective user state from physiological signals in real-time. These methods have been developed within the SmartWeb[3] and HUMAINE[4] research projects. Our approach [9] to this pattern recognition task is very generic and relies on data-driven methods. It concentrates on the problems of intra- and interpersonal variability of the signals, the liability to artefacts and problems of real-time classification both in terms of computational effort and signal input.

For the present work, six physiological signals were used: electrocardiogram (ECG), electromyogram measured at the neck, skin conductivity between index and middle finger, skin temperature at the little finger, blood volume pulse at the ring finger (BVP) and abdominal respiration (Resp). Four derived signals are computed: The heart rate acquired from the ECG and BVP channel, the lag between ECG and BVP, which can be regarded as a surrogate parameter of the systolic blood pressure, and the respiration rate from the Resp channel. This approach has the advantage that no signal-specific algorithms have to be included into the feature extraction module.

Artefacts can render a signal useless for whole passages. Therefore, an artefact detection module marks signals that are unplugged or represent physically unlikely values. Corrupted signals are excluded from further processing for the current point in time. In order to deal with the variable number of remaining input channels, each valid signal is analyzed separately; the single results are later combined for the final classification. For signal analysis, a multi-resolution approach is taken, using analysis windows covering the past 1, 5, 20 and 60 seconds of a signal. This aims at combining the stability of large analysis windows and the capability of small windows to reflect quick changes which is needed for real-time classification. From each analysis window, a large number of generic features like mean, standard deviation or slope is extracted. A labeled dataset is then utilized to create features specialized to the set of affective states to be recognized and signal at hand by means of a data-driven transform, the Fisher linear discriminant analysis (LDA): the generic features from all analysis windows of a signal are stacked into a single feature vector which is then projected into a lower-dimensional space.

Two different feature sets are provided: 50 *moving features* are computed recursively for each new sample and thus have a constant computational complexity with respect to the length and step size of the analysis windows. This is important for real-time computation of the larger analysis windows and makes the approach well-suited for a possible implementation on limited hardware. The *sliding features* go further and have a memory requirement independent of the analysis window length. This is favorable for a possible implementation on hardware with small memory. The 44 sliding features approximate most of the 50 moving features.

The final feature vectors of each signal resulting from the LDA transformation of the generic features are scored with a Gaussian Mixture Model consisting of 10 mixture components. The resulting scores from all valid signals are, assuming statistical independence between the different physiological signals, combined probabilistically by multiplication, yielding a final score for each class.

Our approach was evaluated on the multi-modal Drivawork (Driving under Varying Workload) database [10] which contains 15 hours of physiological recordings from 24 participants during different stress levels in a simulated driving context. All evaluations are done using person-independent 10-fold cross validation, i.e. each pair of train and test set is disjoint with respect to the participants. The class-wise averaged recognition rates are reported. For binary classification of stress/non-stress, recognition rates between 88.8% and 96.1% (depending on the evaluated subset) were achieved using both feature sets. Using either the moving or sliding feature set alone yielded similar results. Results can be further improved by a user adaption as a simulated user adaption by means of a mean-variance-normalization per participant showed.

4 www.emotion-research.net

5 Discussion

In the future, new forms of living will be necessary in order to face the demographic change. With technologies like ISA-house, normal homes can be converted into a highly supportive environments without too much effort – an alternative to expensive integrated living projects. However, those systems are required to adapt themselves to the user. It would never be accepted that the users have to adapt themselves to the system. Therefore, a natural language interface as provided by ISA-house seems a very suited approach, as no keys, menus or commands have to be memorized. On top of that, speech is a valuable alternative to a state-of-the-art emergency button: given a sufficient configuration of microphones, the system can be addressed reliably from anywhere in the house without the need to remember to carry the portable emergency device. Also, a push-to-talk button to activate the speech recognizer would not be acceptable. Therefore, an automatic, robust detection of the user focus is important. A continuous medical screening, as it could be developed based on our methods for user state recognition with physiological signals, would improve medical care and reassure the resident and affiliated persons. Summing up, the ISA-house prototype and the methods for user focus and user state classification open up promising possibilities for ambient assisted living.

Future work will focus on user adaptation which makes speech recognition and the other classification modules more accurate. Also the collection of speech and other data from the target group will improve the reliability of the recognition. A cooperation between physicians and engineers will allow to design new therapy and screening paradigms. The above presented approaches will be integrated into the ISA-house prototype, rendering it a complete demonstration system that can be tested and evaluated.

6 Summary

With ISA-house, we have presented a promising research platform that serves as a basis for exploring the possibilities of ambient assisted living for the silver generation. Several unique features like natural language interface or autonomous conduct of outbound telephone calls are already provided; new services can be readily integrated due to the unified plug-in architecture of the system. Novel approaches for advancement of the user interface have been given: methods for automatic recognition of the user focus both in audio and video, and techniques for the classification of the affective user state using physiological signals. The latter techniques are an ideal starting basis for developing algorithms for continuous medical screening with physiological signals, another promising field of application in ambient assisted living.

7 Acknowledgment

Part of this work was funded by the EC within HUMAINE (Grant IST-2002-507422), by the German Federal Ministry of Education and Research (BMBF) within SmartWeb (Grant 01IMD01 F) and by the German Research Foundation (DFG) under grant SFB 603/TP B2. The responsibility for the content lies with the authors.

8 Literature

[1] Grinewitschus, V.: Vernetzung optimieren – inHaus II. In: Electronic Home Jahrbuch 2007, Nr. 2, pp. 159-161, 2006

[2] Borodulkin, L.; Ruser, H.; Tränkler, H.-R.: 3D Virtual Smart Home User Interface. In: Proc. IEEE International Symposium on Virtual and Intelligent Measurement Systems, pp. 111-115, 2002

[3] Nöth, E.; Horndasch, A.; Gallwitz, F.; Haas, J.: Experiences with Commercial Telephone-based Dialogue Systems. In: it – Information Technology 46, Nr. 6, pp. 315-321, 2004

[4] Horndasch, A.; Gallwitz, F.; Haas, J.; Nöth, E.: Der mixed-initiative Ansatz als Basis für benutzerfreundliche Sprachdialogsysteme. KI, 2005, 3/05: pp. 38-41.

[5] Batliner, A.; Zeissler, V.; Nöth, E.; Niemann, H.: Prosodic Classification of Offtalk: First Experiments. In: Proc. 5th Int. Conf. on Text, Speech, Dialogue (TSD), Lecture Notes in Artificial Intelligence, pp. 357-364, 2002

[6] Viola, P.; Jones M.J.: Robust Real-Time Face Detection, Int. J. Comput. Vision, 57(2), pp. 137-154, 2004

[7] Batliner, A.; Hacker, C.; Kaiser, M.; Mögele, H.; Nöth, E: Taking into Account the User's Focus of Attention with the Help of Audio-Visual Information: Towards less Artificial Human-Machine-Communication. In: Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP), pp. 51-56, 2007

[8] Picard, R.W.; Vyzas, E.; Healey, J.: Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(10), pp. 1175-1191, 2001

[9] Hönig, F.; Batliner, A.; Nöth, E.: Fast Recursive Data-driven Multi-resolution Feature Extraction for Physiological Signal Classification. In: 3rd Russian-Bavarian Conference on Biomedical Engineering, pp. 47-52, 2007

[10] Hönig, F.; Batliner, A.; Nöth, E.: Real-time recognition of the affective user state with physiological signals. In: 2nd Int. Conf. on Affective Computing and Intelligent Interaction, Proceedings of the Doctoral Consortium, pp. 1-8, 2007