Speech Processing in the Networked Home Environment A View on the Amigo Project

[1]*Reinhold Haeb-Umbach,* [2]*Basilis Kladis,* [1]*Joerg Schmalenstroeer*

[1]Department of Communications Engineering
University of Paderborn, Germany
{haeb,schmalen}@nt.uni-paderborn.de

[2]Knowledge S.A. - LogicDIS Group
Patras, Greece
bkladis@knowledge-speech.gr

Abstract

Full interoperability of networked devices in the home has been kind of an elusive concept for quite some years. Amigo, an Integrated Project within the EU 6-th framework program, tries to make home networking a reality by addressing two key issues: First, it brings together many major players in the domestic appliances, communications, consumer electronics and computer industry to develop a common open source middleware platform. Second, emphasis is placed on the development of intelligent user services that make the benefit of a networked home environment tangible for the end user. This paper shows how speech processing can contribute to this second goal of user-friendly, personalized, context-aware services.

1. Introduction

Traditionally home automation, mobile communication, consumer electronics and personal computing were strictly separate domains all having their own industrial players and their own standardization efforts. The vision of a networked home, in which several pieces of equipment are connected using an infrastructure, has been put forward for a couple of years. However, except for a few "converged" products, such as smartphones or mobile digital assistants, that combine aspects of the mobile and PC domains, or media centers that combine aspects of the PC and CE domains, the reality is far behind this vision.

The European 6-th framework Integrated Project Amigo – Ambient Intelligence for the Networked Home Environment (contract number IST 004182) [1] aims at making the vision become a reality by focussing on two key issues.

First, it brings together for the first time major players of all four aforementioned domains in order to develop open, standardized, interoperable middleware such that devices and services built by different manufacturers can interoperate at different levels. This includes automatic discovery of devices and services as well as service composability, upgradeability and self-administration, which are a necessity for easy installation and use by the end consumer.

Although focused on a more restricted application domain, the ITEA Ambience [2] and IST Ozone projects [3] had similar objectives as Amigo. Indeed, both projects have promoted system architectures based on service-orientation for ambient intelligence systems. Previous work of the scientific community that also is being used as a base line for the work in the Amigo project is coming from the EU-FET proactive initiatives "The Disappearing Computer", in particular the project "Ambient Agoras" [4].

The second objective of Amigo is the development of attractive services based on intelligent user interfaces which are simply not possible in a traditional non-connected home and which therefore make it worthwile to buy and install a network infrastructure at home.

This paper is concerned with the question how speech processing can help realize intelligent user services. To make a user interface being perceived as intelligent requires automatic context aggregation and personalization. This asks for speech recognition and dialog systems which personalize to the user and his preferences, both, user and preferences, being identified automatically. We discuss issues of personalized dialog modeling in this paper.

As we all know, speech conveys more than just the linguistic contents of what has been spoken. Therefore, the speech subsystem not only consists of an automatic speech recognition server and dialogue modules. The captured speech signals are also analyzed to glean information about who is speaking (speaker recognition), how many people are in a room (speaker change detection), where they are (speaker localization), and, possibly, in which social setting or context they are (emotion recognition). This topic of "acoustic scene analysis" is also discussed in the sequel.

A further objective of this paper is to point to the new challenges that a networked home environment places on the speech processing subsystem, coming from privacy and security concerns.

The paper is organized as follows. In the next section we introduce the architecture of the speech processing subsystem of Amigo, Section 3 presents the notion of "acoustic scene analysis". Next, the problem of dialog personalization is addressed in Section 4, some remarks on security and privacy are presented in Section 5, and we finish with some conclusions.

2. Speech Processing System Architecture

The speech processing subsystem of Amigo consists of four main units (see Fig. 1): the Speech Signal Preprocessing Server, the Speaker Recognition Server, the Speech Recognition and Dialog Server, and the Speech

Output Server.

In the Speech Signal Preprocessing Server the signals captured by microphone arrays and wireless microphones are processed to deliver two output streams: one stream contains an enhanced speech signal, which is obtained by microphone array beamforming and single-channel speech enhancement techniques, and the second stream is a stream of accompanying metadata about speaker identity, number of speakers, speaker position etc. obtained from acoustic scene analysis (see Section 3).

The speaker recognition Server provides the identity of the speaker contained in a segment of the speech input signal. Homogeneous segments are obtained by a speaker change detection algorithm. Clearly, speaker recognition is used to identify individual family members in order to allow for, e.g. dialog personalization (see Section 4). Further, non-family members should be granted restricted access to the system.

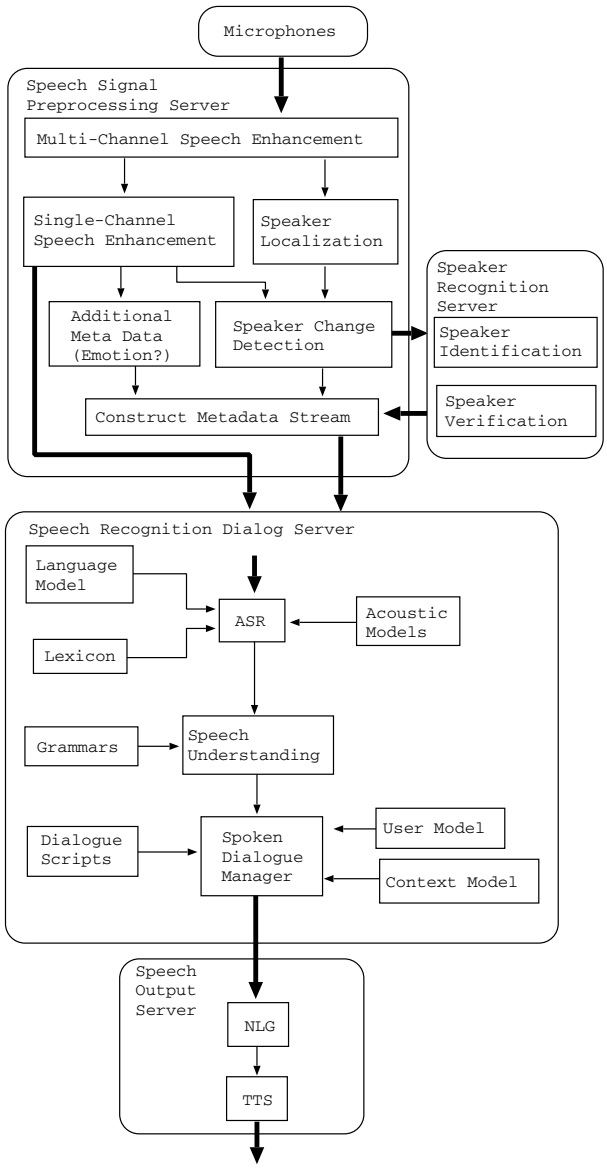

Figure 1: *System architecture*

The Speech Recognition and Dialog Server consists of three main modules: The automatic speech recognizer (ASR) transcribes the acoustic signal into written text. The speech understanding module (SU) interprets the output of the speech recognition module on the basis of language rules (grammar), specifically compiled for the application in question, and generates the semantic output. Finally, the spoken dialog manager (SDM) generates the response of the system based on the output of SU by taking into account what the system knows about the current context (context model) and what it knows about the user (user model).

The Speech Output Server produces a spoken response. Depending on the current status of the interaction the response may be a question, a confirmation or the information requested by the user. A spoken response is generated in two steps: The natural language generator (NLG) prepares the text to be spoken by taking into account the relevant context and assures compliance with grammatical rules. The text-to-speech synthesis (TTS) converts the textual output of the NLG to a speech signal to be played back to the user.

Speech is a powerful input modality, but not equally well suited for any kind of user input. Gestures and pointing devices can serve as a good complement [5]. The system architecture must then be extended to gesture recognition and an overall input modalities fusion to base the system reaction on all available sensor informations. Likewise the system output has to be presented in whatever modality is suited best for the current situation, be it an acoustic output through the loudspeakers or visual output on a screen. Moving and adapting content and functionality between output devices depending on their degree of compatibility can be perceived as an "intelligent" functionality, which underlines the usefulness of a networked home environment.

An important issue to be addressed in the home environment is the multi-user situation since several family members and guests may be in a room and want to address the system. A "multiple-user shared experience" is envisaged, where distributed applications for sharing information allow for remote presence and ambient sharing, allthough the location of the parties may be physically far apart. Then the speech processing services will also be distributed.

3. Acoustic Scene Analysis

Distant microphones are preferred to close-talking microphones from an ease-of-use perspective, since they are unobtrusive, pre-installed , and the number of microphones is independent of the number of speakers.

Multi-channel (microphone array) signal processing has to be employed to make up, at least in part, for the detrimental effect of reverberation and distance from source to sensor on the speech quality.

Recently, we have proposed the concept of an adaptive Filter-and-Sum (FSB) beamformer, which blindly, i.e. without explicit source localization, directs a beam of increased sensitivity towards the desired source [6]. This is done by adaptively estimating the dominant eigenvector of the cross power spectral density of the microphone signals. The experiments revealed fast adaptation and high robustness of the derived algorithms. While a

conventional Delay-and-Sum (DSB) beamformer realizes pure delays in the microphone paths, an FSB employs FIR filters which can realize any transfer function.

The adaptive FSB was able to attain almost completely the theoretically maximum signal-to-noise ratio (SNR) gain in spatially uncorrelated and correlated noise fields [7]. For spatially uncorrelated noise, this limit is $10 \log_{10} M$ dB, where M is the number of microphones.

However, if the array should achieve the same output SNR as a close-talking microphone, this may result in an unjustifiably large number of microphones. In [8] it was calculated that a microphone array of at least $M = 10$ omnidirectional microphones would be required if the array at a distance of 1 m should maintain the same SNR as a close-talking microphone at 0.1 m distance from the source. For 2 m the number would have to be increased to more than $M = 20$. This simple argument alone, not to mention the detrimental effect of reverberation on speech recognition accuracy [9] shows that a replacement of close-talking microphones by a distant microphone array has to be taken with great caution, and viability depends on the particular circumstances.

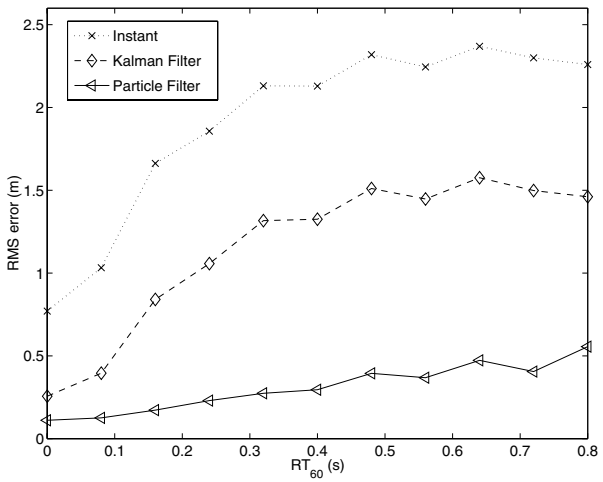

Figure 2: *RMS postion error as a function of room reverberation time with additive noise at microphones (SNR=20dB)*

While for some situations a close-talking wireless microphone is not avoidable, microphone arrays may still be used to deliver additional information about the acoustic scene. In [10] we could show that adaptive beamforming combined with particle filtering can deliver very accurate estimates of the position of a moving speaker. As an example, Fig. 2 shows the rms position error as a function of room reverberation time RT_{60} for a signal-to-noise ratio at the microphones of 20 dB. The instantaneous estimate can be greatly improved by a postfilter, where a particle filter clearly outperforms a Kalman filter. An rms position error of less than 0.5 m even for highly reverberant environments is sufficient for most applications.

Algorithms for speaker change detection have been developed in the course of the DARPA Hub-4 Broadcast News Transcription effort. The application of the Bayesian Information Criterion (BIC) to this problem was one of the most successful approaches [11]. Together with speaker recognition algorithms [12] this would allow for a continuous tracking of who is currently speaking, which is an important information for acoustic model adaptation or dialog personalization/customization.

The usage scenarios discussed within the Amigo project ask for automatic gathering of information about user emotion and user mode of communication (casual dialogue among participants in the same room, attentive communication with a physically remote partner, voice commands to the system). Speech processing can help here to some extent. However, emotion recognition by speech is still at its infancy [13], and reliable recognition of other than strong emotions seems to be beyond the current state-of-the-art.

4. Dialog Personalization

In general, the concept of user modeling addresses issues of understanding users in order to make a system useful and make user-system interaction user friendly and universal. User modeling and system personalization will be addressed in three different layers

Firstly, in design-time adaptation to users, the information gathered during the extensive user requirements studies will be used to anticipate users' needs, goals, interactive behavior, etc., in order to proactively fit the system to its future users. The system will be programmed at design time to dynamically respond in different ways depending on the user's behavior.

Secondly, the system will be designed to be customizable by users who can modify the system's behavior in various ways, both concerning functionality and with respect to their preferred style of interaction, by selecting among customization options. An example of such a system customization is the option for users to define custom actions. Custom actions are sequences of requests to the system which are identified by a single descriptor; therefore, uttering a custom action identifier prevents the user from performing sequences of elementary requests which he perceives as a whole complex request. The purpose of this feature can be easily seen in the "Watch TV environment" scenario where the user speaking a simple utterance can trigger a sequence of actions for setting up the lighting, blinds, sound systems, TV appliance, etc., for watching his favorite TV channel/show.

Thirdly, in user model-based adaptation, the system may itself observe the user's behavior and adaptively modify its interactive behavior on the fly as a function of the data gathered. Thus, user model-based adaptation will be carried out automatically by the system, using collected and stored user information. Two forms may be distinguished:

- In history adaptation the system uses implemented user models for each user; these models are built from information on the individual user's behavior collected and stored during user-system interaction.
- In instant adaptation it uses short-term buffered user information that is discarded after the end of each session. Thus, in instant adaptation the system may treat the user anonymously, in contrast to history adaptation, where a reliable user recognition process is required to support user model-

ing and personalization for recognizing the current user and applying his individual settings.

5. Privacy and Security

Privacy and security issues are very important, as personal data and preferences must be protected not only against intruders, but also against other users. The user field studies have shown that people fear the possibility of being observed through the system by others ("Big Brother feeling") or the feeling of loosing control. Collecting information about the user to automatically build up a user modell supports the idea of an "intelligent" user service that personalizes to the user, but it also raises concerns on security. It is important to keep the user in control of which data will be collected and how they will be used.

Personal data and usage information must be reliably protected against unauthorized access. Open source software can be a part of the solution for this problem, as the applications and system components can be tested for security holes by a large community of developers.

6. Conclusions

Speech, the most important communication modality among humans, is also an important building block of intelligent user services. Using acoustic scene analysis, automatic aggregation of context and user models, and dialog personalization, user services can be realized, that are perceived as "intelligent". This notion is further enhanced by fusing different input modalities and selecting the most appropriate output device and presentation mode, depending on the current situation. If this can be realized without raising privacy and security concerns by the user, the goal of the Amigo project, that a networked home enjoys wide customer acceptance, can become a reality.

7. Acknowledgments

This work has been supported by the European union project "Amigo - Ambient intelligence for the networked home environment". More informations on "www.amigo-project.org"

8. References

[1] www.amigo-project.org

[2] www.hitech-projects.com/euprojects/ambience

[3] www.extra.research.philips.com/euprojects/ozone

[4] www.ambient-agoras.org

[5] www.smartkom.org

[6] E. Warsitz and R. Haeb-Umbach, "Acoustic Filter-and-Sum Beamforming by Adaptive Principal Component Analysis", in *Proc. ICASSP*, Philadelphia, March 2005.

[7] R. Haeb-Umbach and E. Warsitz, "Performance Evaluation of Filter-and-Sum Beamforming in Spatially Uncorrelated and Correlated Noise", submitted to *IWAENC: Int'l Workshop on Acoustic Echo and Noise Control*, Eindhoven, Sept. 2005.

[8] G.W. Elko, "Microphone Arrays", in *Proc. Workshop on Hands-Free Speech Communication*, Kyoto, April 2001.

[9] Jinachitra P. and Prieto R., "Towards Speech Recognition Oriented Dereverberation", in *Proc. ICASSP*, USA, 2005.

[10] E. Warsitz, R. Haeb-Umbach, and S. Peschke, "Adaptive Beamforming Combined with Particle Filtering for Acoustic Source Localization", in *Proc. ICSLP*, Jeju, Korea, Oct. 2004.

[11] S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, Feb. 1998.

[12] J.P. Campbell, "Speaker Recognition: A Tutorial", in *Proceedings of the IEEE*, Vol. 85, No. 9, pp 1437-1462, Sep. 1997.

[13] *Speech Communication*, Special Issue on Speech and Emotion, Vol. 40, Nos 1-2, Apr. 2003.