Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 27 (2014) 389 – 397

Procedia Computer Science

5th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, DSAI 2013

Speech-Centric Multimodal Interaction for Easy-To-Access Online Services – A Personal Life Assistant for the Elderly

António Teixeira[a], Annika Hämäläinen[b,c], Jairo Avelar[b,c], Nuno Almeida[a], Géza Németh[d], Tibor Fegyó[d], Csaba Zainkó[d], Tamás Csapó[d], Bálint Tóth[d], André Oliveira[a], Miguel Sales Dias[b,c]

[a]*Department of Electronics Telecom. & Informatics/IEETA, University of Aveiro, Aveiro, Portugal*
[b]*Microsoft Language Development Center, Lisbon, Portugal*
[c]*ISCTE - University Institute of Lisbon/ADETTI-IUL, Portugall*
[d]*Department of Telecommunications & Media Informatics, Budapest University of Technology & Economics, Budapest, Hungary*

Abstract

The PaeLife project is a European industry-academia collaboration whose goal is to provide the elderly with easy access to online services that make their life easier and encourage their continued participation in the society. To reach this goal, the project partners are developing a multimodal virtual personal life assistant (PLA) offering a wide range of services from weather information to social networking. This paper presents the multimodal architecture of the PLA, the services provided by the PLA, and the work done in the area of speech input and output modalities, which play a key role in the application.

Selection and peer-review under responsibility of the Scientific Programme Committee of the 5th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2013).

Keywords: active aging, automatic speech recognition, elderly, multilingual, multimodal interaction, personalized synthetic voices, personal assistant, social interaction, spoken language modalities.

Selection and peer-review under responsibility of the Scientific Programme Committee of the 5th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2013).
doi:10.1016/j.procs.2014.02.043

1. Introduction

The population aged over 60 is rapidly growing worldwide [1]. This social change is so dramatic that the World Health Organization is actively promoting policies and programs to keep the older generations active, productive and independent as long as possible. From the point of view of information and communication technology (ICT), ensuring active ageing translates into developing applications that enhance the health, the social participation and the security of the elderly.

Previous research suggests that the elderly have more difficulties using ICT than younger adults do [2], [3]. The main reasons for this are the complexity of the existing user interfaces and the limited set of available interaction modalities caused by the technology mainly being designed with younger users in mind. One of the most promising ways of adapting the technology to better suit the needs of the elderly is to increase the choice of available interaction modalities. Adding speech to the available modalities is a particularly interesting alternative. The advantages of speech include it offering a natural and fast (about 150-250 words/minute) form of communication, and it requiring neither visual attention nor the use of hands [4]. In fact, speech is part of the three most popular input modality combinations mentioned in [5]: 1) speech and lip movements, 2) speech and gestures, and 3) speech, gestures and facial expressions. Several popular output modality combinations also include speech: 1) speech and graphics, 2) speech plus avatar, and 3) speech, text and graphics [5]. Usability evaluation studies (e.g. [6]) also suggest that speech is the easiest and most natural modality of human-computer interaction (HCI).

There is growing international interest, both in academia and in industry, in developing speech-driven applications aimed at improving the quality of life of the elderly. Past R&D projects in the area include, for instance, the Living Usability Lab (LUL) project [7], which developed a telerehabilitation service for the elderly [8], with a toolkit for multimodal interaction (including speech), and the possibility to adapt speech and graphical output to the context and to the user [9]. Speech input and output have recently also been used, for example, in a medication assistant for smartphones [10]. The PaeLife project, an ongoing industry-academia collaboration that is part of the Ambient Assisted Living (AAL) Joint Programme [11], is aimed at keeping the European elderly active and socially integrated. To this end, the project is developing a multimodal personal life assistant (PLA) offering the elderly a wide set of services from unified messaging (e.g. email, twitter, videoconferencing) through to relevant feeds (e.g. the latest news, weather information). The platform of the PLA comprises a personal computer connected to a TV-like big screen, as well as a portable device (a tablet) for mobility. One of the key modalities of the PLA is speech; speech input and output will be available in four European languages: French, Hungarian, Polish, and Portuguese.

Apart from the challenges of developing speech technology for use in multimodal interaction in several languages, the project partners are faced with the challenges of customizing it for the elderly. It is well known that current speech recognizers do not work well with elderly speech. This is (in part) because many parameters of the speech signal (such as the fundamental frequency) change with age [12], and because most current speech recognizers have been optimized to recognize younger adult speech. To address this issue, we are collecting large databases of elderly speech for the four target languages, and training elderly-specific speech recognizers using those data. A successful speech interface should also be able to take into account the users' preferences. This is particularly important in the case of the elderly who might not be very familiar with technology. For increased user acceptance of speech output (speech synthesis), we are providing the elderly users with several synthesized voices to choose from based on their personal preferences [13].

In this paper, we present the multimodal architecture adopted for the PLA, the services planned and already available in the PLA, as well as the work aimed at tailoring automatic speech recognition and speech synthesis to the elderly and at making these technologies available to the four languages targeted by the project.

2. The PaeLife Personal Life Assistant

2.1. User-Centric Design

The PaeLife project adopted a user-centric approach for designing the PLA. We carried out extensive user studies amongst the elderly in France, Hungary and Poland to try to identify what kinds of online services the elderly are interested in, and what kinds of limitations and preferences they might have when it comes to HCI [14]. Based on the user studies, we made up personas (archetypal users of the application) and use scenarios to explore the set of tasks and interactions required for the application, and to help us evaluate the application in the future. The PaeLife personas are aged 60 or over and have some experience in using computers – although most of them are not expert users. While they do not have major health issues, they might suffer from typical age-related ailments (reduced dexterity, some degree of visual impairment etc.). In other words, they do not have any serious health-related conditions that would require physiological interfaces (e.g. electromyography), which are cumbersome, intrusive and difficult to use. However, as they might not be very proficient in using traditional user interfaces, we considered it very important to provide them with easy-to-use, natural interaction modalities – such as speech, touch and gestures. Table 1 presents part of an example use scenario featuring Mária Kovács, a relatively healthy 68-year-old Polish woman who still works part-time and, apart from her proofreading work, mainly uses the computer for social networking and for looking for new recipes and information about local cultural events.

Table 1: **Example use scenario with Mária Kovács, a 68-year-old Polish woman with basic computer skills**

Mária is watching TV in the living-room. She decides to use the tablet **to check the latest posts of her Facebook friends**. As one of the posts contains a video, Mária says, "Watch video". The video starts playing on the TV screen, and the TV channel that she was watching now appears as a miniature window in the corner of the TV screen. The tablet continues displaying the latest posts but the controls for the video are shown at the bottom of the screen, allowing the user to pause, stop, fast-forward and rewind the video. At the end of the video, three buttons appear on the TV: "Like", "Don't Like" and "Leave". Mária says "Like", and the video is tagged with her like. Mária then decides to use the tablet to **search for a friend on Facebook by her name**. She chooses the option "Search" and uses the tablet as a virtual keyboard; the results of her search are shown on the TV screen. Mária uses a swiping hand gesture to see the full list of results. When the photo of her friend shows up on the TV screen, she asks for more information by saying, "See the profile of the second photo". The profile of the friend now gets displayed on the TV screen. Mária says, "Send friend request".

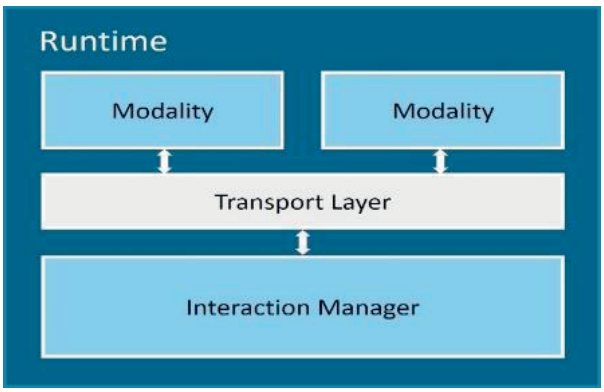

Figure 1: The W3C-recommended architecture adopted for supporting multimodal interaction in the personal life assistant (PLA).

2.2. Multimodal Interaction

Based on the user studies, we identified the following key requirements for HCI in the PLA: 1) support for several input and output modalities, 2) support for the distribution of modalities across different devices (PCs, tablets etc.), 3) adhering to international standards and avoiding closed solutions, and 4) possibility to change or add modules (supporting services or interaction) without the rest of the system being affected. These requirements were met by developing an integrated framework that supports multimodal interaction specifically tailored to the elderly, with the available interaction modalities including speech, touch, gestures, keyboard and mouse. The framework is based on the recent recommendations of the World Wide Web Consortium (W3C) regarding multimodal interaction [15] – a choice motivated by the open-standard nature of the recommended architecture and the ease of integrating new modules and already existing tools into the system.

The architecture (see Figure 1) has three major components:

- The **Interaction Manager** (IM), which manages the different interaction modalities
- The **Modality Components**, which represent the input and output modalities
- The **Runtime Framework**, which acts as a container for all other components and provides communication between the different modalities and the IM

From the point of view of the PLA, the Modality Components are the most important components of the architecture because they provide a simple way to integrate the chosen input and output technologies – including speech input and output, which are discussed in more detail in Section 3 – into the system.

2.3. The PLA and the Available Online Services

The PLA itself comprises a stationary main unit that runs on a desktop computer, as well as a portable unit that runs on a tablet (see Figure 2). In the main unit, a big screen (e.g. an LCD TV) supports graphical output, the internal microphone and speakers support speech input and output, and a Kinect sensor supports gesture input. In the portable unit, on the other hand, the display supports graphical output, the internal microphone and speakers enable speech input and output, and the multi-touch support of the operating system makes touch input possible. The main unit and the portable units work together and can be connected to the internet and to the cloud for providing the user with online services. The two units can also work as stand-alone devices. As illustrated in Figure 2, two different types of tablets can be used as the portable unit: premium tablets, with all the services available for the PLA, and low-cost tablets, with a subset of those services. Apart from a more limited set of services, the low-cost tablets might have to use remote or cloud-based services, for example, for handling speech input and output.

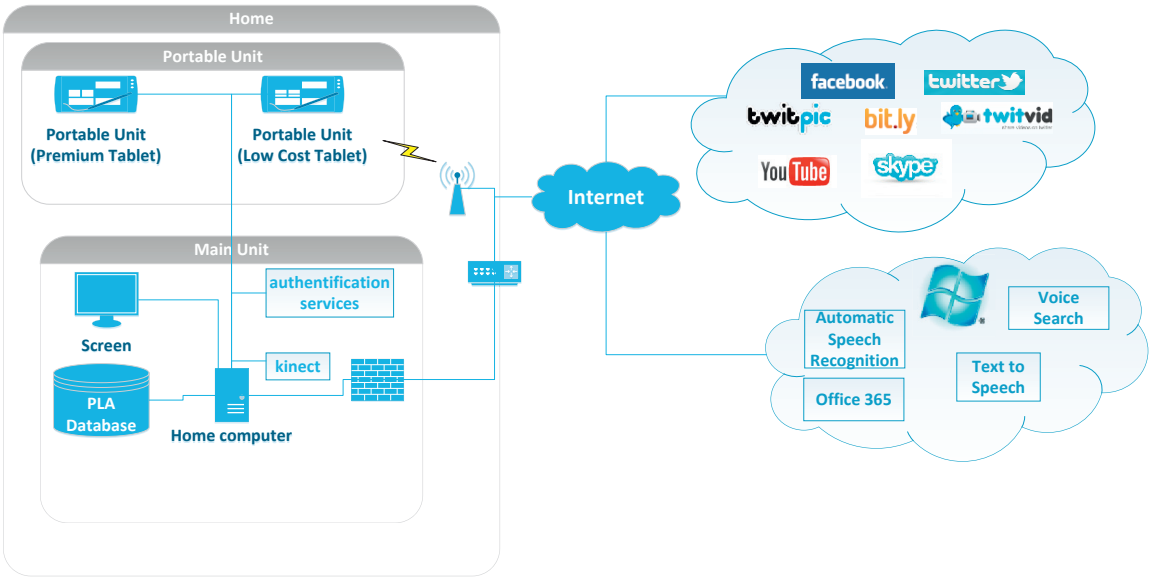

Figure 2: The architecture of the PLA.

In practice, the PLA functions as the multimodal assistant for the elderly and supports a wide range of services, for instance, in the areas of social interaction and entertainment.The PLA is divided into modules that are responsible for delivering different types of services. The modules and the corresponding services are presented in Table 2.

Table 2: The modules and services of the PLA.

Module	**Services**
Unified messaging	1. Quasi-instant messaging 2. Voice call 3. Videoconferencing on Skype 4. Messaging using email, Facebook and/or Twitter 5. Voice mail
Calendar	1. To-do list 2. Birthday reminders 3. Name day reminders 4. Cultural events
Social networking	1. Geographical location of contacts 2. Graphical representation of contacts' availability 3. Email, Facebook and/or Twitter activity with contacts 4. Visualisation of the time that has passed from previous contact (e.g. photos of contacts getting larger), aimed at encouraging regular communication 5. Information about contacts (e.g. status)
TV schedule	
Weather information	

Media manager	1. Photos 2. Videos 3. Music 4. Text documents etc.
Relevant feeds	Examples: latest news, accessibility of places for mobility-impaired users, local health care services and pharmacies
Facebook groups of interest	Examples: recipes, organised travel, and clubs

3. Towards Better Speech Recognition and Synthesis for the Elderly

In the PLA, speech input (automatic speech recognition; ASR) and output (speech synthesis; TTS) are handled using two different speech platforms: one provided by Microsoft [16] and already supporting French, Polish and Portuguese, and another provided by the Budapest University of Technology & Economics (BME) and already supporting Hungarian. In this section, we describe how these speech technologies are optimized for the elderly users in the PaeLife project.

ASR is technology that translates acoustic speech signal into a sequence of words. To be able to do that, automatic speech recognisers typically use three knowledge bases: 1) a language-specific language model (or grammar), which contains information on the possible sequences of words and their probability in the language in question (e.g. 'I use Bing' is a probable sequence of words, while 'I chair Bing' is not), 2) a language-specific pronunciation lexicon, which represents words in terms of individual speech sounds, or phonemes (e.g. Bing contains three phonemes: /b ɪ ŋ/), and 3) language-specific acoustic models, which model the acoustic properties of the phonemes used in the pronunciation lexicon.

Before a recognition task can be performed, the language model and the acoustic models must be trained using large corpora of language-specific texts and speech, respectively. Together, the language model, the lexicon and the acoustic models then 'model' the acoustic realisations of all possible sentences in the language in question, and are used to find the most probable sequence of words to represent the incoming acoustic speech signal.

To be able to train acoustic models that model the acoustic properties of phonemes as accurately as possible, the speech corpus used for the training must contain orthographic (word-level) transcriptions of the spoken material. The lexicon will then be used to identify the underlying phonemes, whose acoustic properties in the speech signal will be used to train the acoustic models. Standard speech recognisers are usually trained using speech corpora collected from younger adult speakers. Because the acoustic properties of speech produced by elderly speakers differ from those produced by younger adult speakers [12], they are not able to recognise elderly speech as well as they are able to recognise younger adult speech. To successfully recognise elderly speech, it is important to collect a sufficient amount of elderly speech for training elderly-specific acoustic models [17–19]. In the PaeLife project, we have already finished collecting large corpora of domain-specific speech from subjects aged 60 or over for two of the target languages: Portuguese (180 hours of read speech) [20], [21], and Polish (170 hours of read speech). At the time of writing this paper, we have also already collected about 46 and 35 hours of read speech for French and Hungarian, respectively. The goal is to collect 200 hours of read speech for both of those languages, as well as an additional 60 hours of spontaneous speech for Hungarian. One of the benefits of collecting read – rather than spontaneous – speech is that, apart from some small corrections, the sentences presented to the speakers can also be used as the orthographic transcriptions of the spoken material. On the other hand, spontaneous speech must be transcribed from scratch.

We have already trained elderly-specific acoustic models for Portuguese, and integrated them into the ASR-based services that are already available in the PLA. We will do the same for the remaining three languages as soon as the data collection and/or the transcription work has ended; for now, the other languages are using standard acoustic models that have been trained using younger adult speech.

TTS is technology that converts text into artificially produced speech. Users are more likely to identify with and accept synthesized voices that match their preferences and/or their own age group, gender etc. [13]. In terms of TTS, the main goal of the PaeLife project is to increase the user acceptance of synthesized voices by offering the elderly users of the PLA a wide variety of voices to choose from. In practice, they are currently provided with the

younger adult voices (female and male) that the two project partners developing speech technologies for the PLA, Microsoft and BME, already have available in their speech platforms. In addition, one elderly voice will be developed for each target language, and the users of the PLA will also be offered the possibility to create personalised voices. To generate new synthesised voices, speech is usually recorded from people with the desired kind of voice, and the recorded speech is then manipulated to form new spoken words and sentences. As few as 200 sentences are enough to generate intelligible voices using the Microsoft methodology [22]. Therefore, personalised voices can even be the voices of the relatives of the users of the PLA – an alternative that might be particularly attractive to the elderly.

4. Currently Available Speech-Driven Services in the PLA

The speech-driven services that are already available and in integration in the PLA system include the following:

- Weather information service. This service uses Hungarian TTS voices to read out weather information from a Hungarian weather information website.
- Unified messaging. This service employs speech input and output for using email, Twitter, YouTube, Skype and Facebook via a single, simplified, easy-to-use interface. The service is currently available in French and Portuguese, with the Portuguese ASR already optimized for elderly speech.
- News feeds. This service offers ASR- and gesture-driven interaction with news feeds. It is already possible to navigate news items in French, Polish, English and Portuguese using simple voice commands (e.g. “to the right”) and by starting to read the contents (e.g. the first 3-4 words) of a news item select that specific item. Hungarian news feeds are also already available in the system but can only be accessed using ASR when acoustic models have been trained for Hungarian. Figure 3 illustrates the easy-to-use multimodal interaction with news feeds.

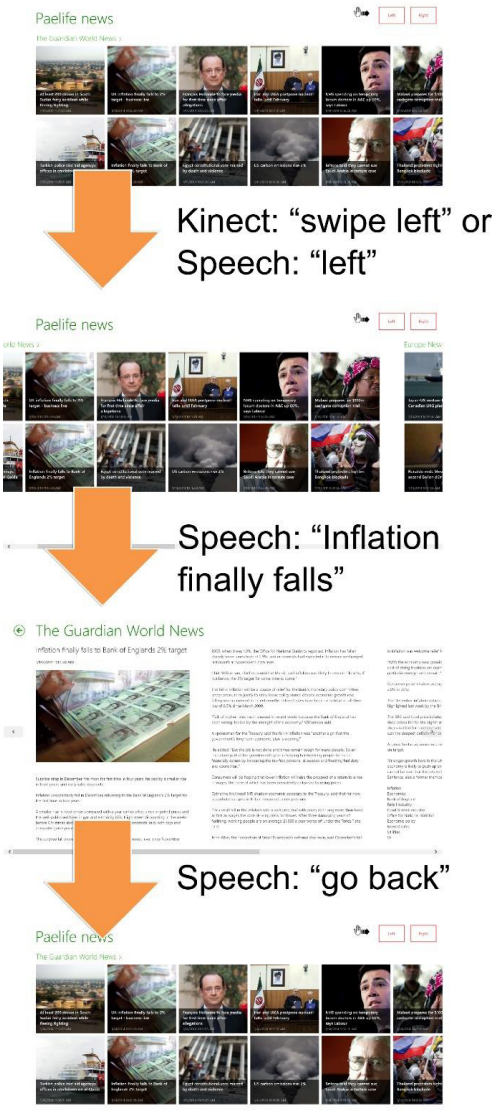

Figure 3: Interaction with the news module of the PLA.

5. Conclusions

Due to the rapid, worldwide growth of the elderly population, it is of paramount importance to devise applications for enhancing the health, the social participation and the security of the elderly. Because of the special needs and limitations of the target population, such applications need to pay special attention to easy and natural human-computer interaction. In this paper, we presented a multimodal personal life assistant aimed at providing the elderly with easy access to a wide range of online services – in particular, information services and services related to social interaction – and, thus, making their lives easier and encouraging their continued participation in the society. As speech is one of the easiest and most natural modalities of human-computer interaction, speech input (automatic speech recognition) and output (speech synthesis) pay a key role in the application, and will be available

in four European languages (besides English): French, Hungarian, Polish and Portuguese. In this paper, we discussed the framework that we have adopted for handling the interaction modalities available in the application (speech, touch, gestures, keyboard and mouse), the services planned and already available in the personal life assistant, as well as our approach to adapting automatic speech recognition and speech synthesis to the elderly and making these technologies available in the four target languages.

Acknowledgments

Authors acknowledge the funding from AAL JP and national agencies: MLDC was funded by Portuguese Government through the Ministry of Science, Technology and Higher Education (MCES); University of Aveiro was funded by FEDER, COMPETE and FCT in the context of AAL/0015/2009 and IEETA Research Unit funding FCOMP-01-0124-FEDER-022682 (FCT-PEstC/EEI/UI0127/2011). BME acknowledge the support of the FuturICT project (TÁMOP-4.2.2.C-11/1/KONV-2012-0013) and the PAELIFE project (AAL-08-1-2011-0001).

References

[1] W. H. Organization, "Active aging: A policy framework," in *Second United Nations World Assembly on Ageing*, 2002.
[2] D. A. C. Stephanidis, "Universal accessibility in HCI: Process-oriented design guidelines and tool requirements," in *ERCIM Workshop on User Interfaces for All*, 1998.
[3] V. Teixeira, C. Pires, F. Pinto, J. Freitas, M. S. Dias, and E. M. Rodrigues, "Towards elderly social integration using a multimodal human-computer interface," in *Proc. International Living Usability Lab Workshop on AAL Latest Solutions, Trends and Applications, AAL*, 2012.
[4] N. O. Bernsen, "Towards a tool for predicting speech functionality," *Speech Communication*, vol. 23, no. 3, pp. 181–210, Nov. 1997.
[5] T. H. Bui, "Multimodal Dialogue Management - State of the art," 2006, no. TR-CTIT-06–01.
[6] A. Teixeira, D. Braga, L. Coelho, A. Fonseca, J. Alvarelhão, I. Martín, A. Queirós, N. Rocha, A. Calado, and M. Dias, "Speech as the Basic Interface for Assistive Technology," in *DSAI*, 2009.
[7] "Living Usability Lab." [Online]. Available: http://www.livinglab.pt/. [Accessed: 18-Mar-2013].
[8] A. J. S. Teixeira, C. Pereira, M. Oliveira e Silva, J. Alvarelhão, A. Silva, M. Cerqueira, A. I. Martins, O. Pacheco, N. Almeida, C. Oliveira, R. Costa, and A. J. . Neves, "New Telerehabilitation Services for the Elderly," in *I.M. Miranda and M.M. Cruz-Cunha [Eds], Handbook of research on ICTs for healthcare and social services: Developments and applications*, IGI Global, 2013.
[9] A. Teixeira, C. Pereira, M. Silva, O. Pacheco, A. Neves, and J. Casimiro, "AdaptO - Adaptive Multimodal Output," in *Proc. PECCS*, 2011.
[10] A. Teixeira, F. Ferreira, N. Almeida, A. Rosa, J. Casimiro, S. Silva, A. Queirós, and A. Oliveira, "Multimodality and Adaptation for an Enhanced Mobile Medication Assistant for the Elderly," in *Proc. Third Mobile Accessibility Workshop (MOBACC), CHI 2013*, 2013.
[11] "Ambient Assisted Living Joint Programme." [Online]. Available: http://www.aal-europe.eu/. [Accessed: 11-Jul-2013].
[12] S. A. Xue and G. J. Hao, "Changes in the human vocal tract due to aging and the acoustic correlates of speech production: A pilot study.," *Journal of Speech, Language and Hearing Research*, vol. 46, no. 3, pp. 689–701, 2003.
[13] C. Nass and S. Brave, "Wired for speech: How voice activates and advances the human-computer relationship," in *MIT Press*, 2007.
[14] N. Saldanha, J. Avelar, M. Dias, A. Teixeira, D. Gonçalves, E. Bonnet, K. Lan, N. Géza, P. Csobanka, and A. Kolesinski, "A Personal Life Assistant for 'natural' interaction: the PaeLife project," in *AAL Forum 2013 Forum*, 2013.
[15] M. Bodell, D. Dahl, I. Kliche, J. Larson, B. Porter, D. Raggett, T. Raman, B. H. Rodriguez, M. Selvaraj, R. Tumuluri, A. Wahbe, P. Wiechno, and M. Yudkowsky, "Multimodal architecture and interfaces: W3C Recommendation," 2012. [Online]. Available: http://www.w3.org/TR/mmi-arch/. [Accessed: 18-Mar-2013].
[16] "Microsoft Speech Platform 11.0." [Online]. Available: http://www.microsoft.com/en-us/download/details.aspx?id=27224. [Accessed: 18-Mar-2013].
[17] R. Vipperla, S. Renals, and J. Frankel, "Longitudinal study of ASR performance on ageing voices," in *Proc. Interspeech*, 2008.
[18] A. Baba, S. Yoshizawa, M. Yamada, A. Lee, and K. Shikano, "Acoustic models of the elderly for large-vocabulary continuous speech recognition," *Electronics and Communications in Japan*, vol. 87, no. 7, pp. 49–57, 2004.
[19] T. Pellegrini, I. Trancoso, A. Hämäläinen, A. Calado, M. Dias, and D. Braga, "Impact of age in ASR for the elderly: Preliminary experiments in European Portuguese," in *Proc. IberSPEECH*, 2012.
[20] A. Hämäläinen, F. Pinto, M. Dias, A. Júdice, J. Freitas, C. Pires, V. Teixeira, A. Calado, and D. Braga, "The first European Portuguese elderly speech corpus," in *Proc. IberSPEECH*, 2012.
[21] A. Júdice, J. Freitas, D. Braga, A. Calado, M. Sales Dias, A. J. S. Teixeira, and C. Oliveira, "Elderly speech collection for speech recognition based on crowd sourcing.," in *Proc. DSAI*, 2010.
[22] D. Braga, P. Silva, M. Ribeiro, M. Henriques, and M. Dias, "HMM-based Brazilian Portuguese TTS," in *Propor 2008 Special Session: Applications of Portuguese Speech and Language Technologies*, 2008.