Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 67 (2015) 283 – 292

Procedia Computer Science

6th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Infoexclusion (DSAI 2015)

Multilingual speech recognition for the elderly: The AALFred personal life assistant

Annika Hämäläinen[a,b], António Teixeira[c], Nuno Almeida[c], Hugo Meinedo[a], Tibor Fegyó[d], Miguel Sales Dias[a,b]

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

Abstract

The PaeLife project is a European industry-academia collaboration in the framework of the Ambient Assisted Living Joint Programme (AAL JP), with a goal of developing a multimodal, multilingual virtual personal life assistant to help senior citizens remain active and socially integrated. Speech is one of the key interaction modalities of *AALFred*, the Windows application developed in the project; the application can be controlled using speech input in four European languages: French, Hungarian, Polish and Portuguese. This paper briefly presents the personal life assistant and then focuses on the speech-related achievements of the project. These include the collection, transcription and annotation of large corpora of elderly speech, the development of automatic speech recognisers optimised for elderly speakers, a speech modality component that can easily be reused in other applications, and an automatic grammar translation service that allows for fast expansion of the automatic speech recognition functionality to new languages.

Peer-review under responsibility of organizing committee of the 6th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2015)

Keywords: ageing; automatic speech recognition; elderly; human-computer interaction; multilingual; multimodal; speech.

Peer-review under responsibility of organizing committee of the 6th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2015)
doi:10.1016/j.procs.2015.09.272

1. Introduction

Information and communication technology (ICT) has considerable potential when it comes to facilitating the lives of the elderly. However, due to the complexity of existing user interfaces and the limited set of available interaction modalities, combined with physical limitations such as poor eyesight, the elderly often have difficulties using ICT[1]. Therefore, it is very important to investigate the use of natural, easy-to-use interaction modalities in applications aimed at the elderly.

Speech would be a particularly interesting interaction modality in the case of the elderly, as it offers a natural form of human-computer interaction (HCI) that requires neither visual attention nor the use of hands[2]. However, our voices change as we age[3], and currently available automatic speech recognisers do not usually work well with elderly speech. This is because, to serve mainstream business requirements, they have been optimised for younger adults' speech. The degradation in automatic speech recognition (ASR) performance on elderly speech has been illustrated, for instance, by Wilpon & Jacobsen[4] and Vipperla et al.[5]. The same authors, however, also showed that performance improves considerably when automatic speech recognisers are specifically optimised for elderly speech. Such findings show that there is a real need for adapting ASR to elderly speech.

There is growing international interest, both in academia and in industry, in developing speech-enabled applications for improving the daily lives of the elderly. Examples of already developed products, applications and services include, for instance, a telerehabilitation service with multimodal interaction[6], Windows Phone applications tailored for the needs of the elderly[7], a humanoid robot to help the elderly in their daily activities at home[8], and a humanoid robot designed to be used for rehabilitation, fall detection and entertainment purposes at care institutions[9].

In this paper, we describe work done in the area of multilingual ASR in the AAL PaeLife project[10]. The goal of the project was to develop a multimodal virtual personal life assistant (PLA) that would help the elderly – in particular those who have retired recently and have some experience in using technology – to remain active, productive, independent, and socially integrated. The resulting application was named AALFred, and the ASR functionality was optimised for elderly speech in four European languages: French, Hungarian, Polish and Portuguese.

Despite its potential, the integration of multilingual ASR in applications aimed at the elderly poses various challenges. First, as mentioned before, automatic speech recognisers need to be optimised for elderly speech – a task that requires the availability of a sufficient amount of speech data collected from elderly speakers. In the context of the PaeLife project, we collected large corpora of French, Hungarian and Polish elderly speech – which, in itself, is a time-consuming, demanding effort – and then optimised the speech recognisers used in AALFred for elderly speech using these data. We also developed ASR for Portuguese elderly speech, using an elderly speech corpus collected in the context of the Living Usability Lab (LUL) project[6]. Second, the development of a multimodal application requires several components to seamlessly work together. To make this possible, we designed a speech modality component[11] that works decoupled from the services available in AALFred and, therefore, makes it easier to integrate ASR in any future services, as well as in any future applications or products. Third, the speech modality needs to work in multiple languages. In the absence of speech interaction designers with relevant language skills, we proposed a way of automatically deriving the first versions of ASR grammars, which define the allowed speech input in speech-enabled applications, for new languages.

This paper is further organised as follows. Section 2 describes AALFred and the services it offers. Section 3 describes our efforts to develop multilingual ASR for the four PaeLife languages, and illustrates the performance improvements we achieved when optimising ASR for elderly speech. In Section 4, we present the speech modality component and the way new languages can be integrated into it. Finally, in Section 5, we formulate our conclusions.

2. The AALFred Personal Life Assistant

AALFred supports five human-computer interaction (HCI) modalities for easy, natural HCI: mouse, keyboard, speech, touch and gesture. Speech input is currently available and optimised for elderly speech in French, Hungarian, Polish and Portuguese. In those four languages, elderly users can use voice commands to access and operate services, and dictation to compose messages and to add descriptions of appointments (in the Messaging and Agenda services described below).

Physically, AALFred comprises a stationary main unit, a desktop computer connected to a large screen (e.g. an LCD TV), as well as a portable device, a tablet. In the main unit, the large screen supports graphical output, the internal microphone and speakers support speech input (ASR) and output (speech synthesis), 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 and portable units can work either together or separately as stand-alone devices, and can be connected to the internet and to the cloud for providing the user with online services (see Fig. 1).

AALFred offers the elderly a wide range of services in the areas of social communication, entertainment, information management and information search, accessible through different modalities:

- Agenda – managing appointments
- Contacts – managing contacts
- Messaging – receiving and sending messages (email, Twitter, Facebook, Skype)
- Audio or videoconference – establishing audiovisual communication (Skype)
- Media – viewing audiovisual information
- Find My – searching for local services (pharmacies, police, etc.)
- News Reader – having the latest news read out by a speech synthesiser
- WeatherForAll – checking the weather forecast

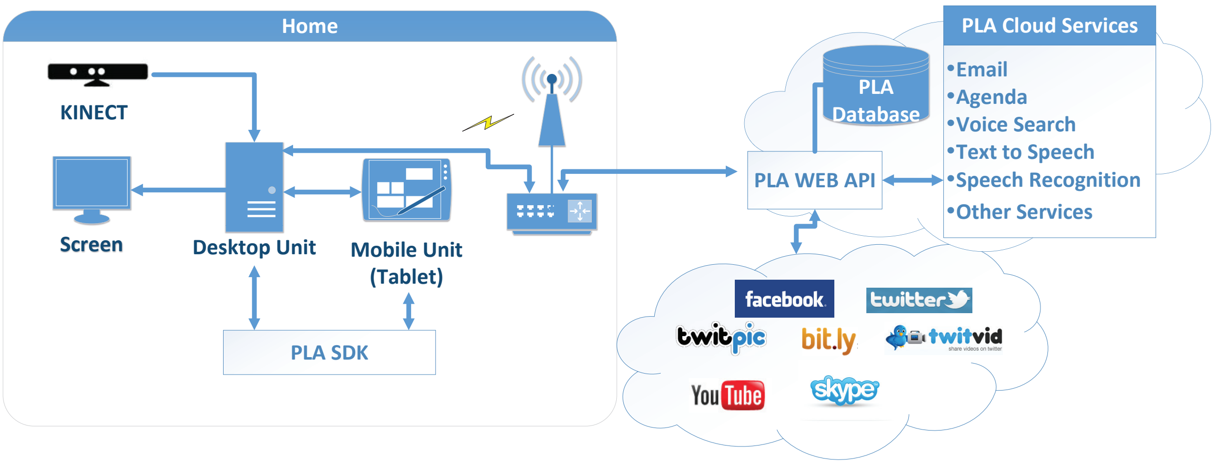

Fig. 1. The architecture of AALFred.

To give a simple illustration of the interaction modalities, let us consider the view of available services in AALFred in Fig. 2. To see the services hidden on the right side of the screen, the user can, for example, swipe right with their hand (gesture), drag the screen right with their finger (touch), or say, "Move right" (speech). To access a service, the user can touch the relevant icon, or use a voice command such as "Show my Agenda" or "Open my Agenda". In fact, several different voice commands are able to perform the same action. The user can also use various interaction modalities to operate a service. In the case of the Agenda service, (s)he might, for instance, want to add a new appointment. As illustrated in Fig.3, the user can touch the desired day or speak out the day of the week (e.g. "Thursday" or "Open Thursday"). When the desired day is displayed, (s)he can add a new appointment by tapping the +-button on the screen or say, for example, "Add a new appointment". (S)he can then add the details of the appointment using the keyboard (either an on-screen, inbuilt or external keyboard, depending on the device used) or speech input (dictation). The details of the appointment will be saved when the user, for instance, says, "Save."

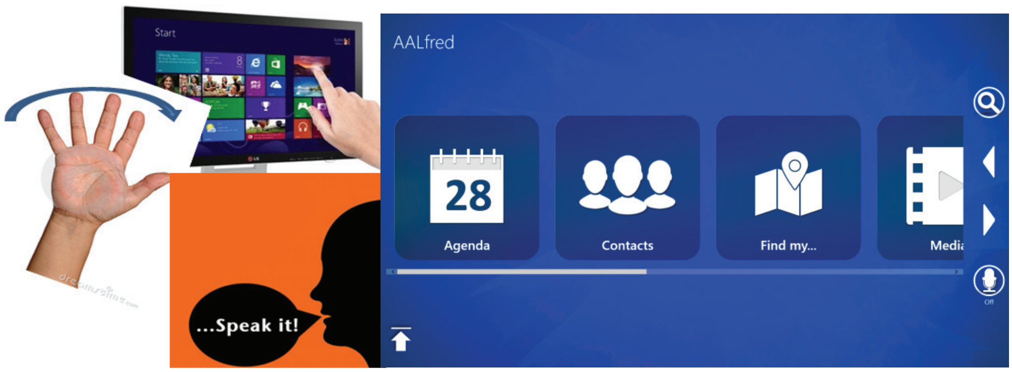

Fig. 2 Interaction modalities available in AALFred include gesture, touch and speech.

Fig. 3 Operating the Agenda service using different interaction modalities.

3. Providing State-of-the-Art Multilingual ASR for the Elderly

In this section, we summarise how ASR works and describe how it was optimised for the elderly and for multiple input languages in the PaeLife project. ASR is technology that translates acoustic speech signals into the corresponding sequences of (written) words. To be able to do that, automatic speech recognisers typically use three types of language-specific knowledge bases: 1) a language model, which contains information about the possible words and sequences of words in the input language, together with their probabilities of occurrence (e.g. 'I use Bing' is a probable sequence of words while 'I chair Bing' is not), and/or grammars, which contain information about the allowed speech input in different situations (e.g. numbers from 1 to 31 are allowed in the case of grammars meant for recognising dates, while e.g. 32 is not), 2) a pronunciation lexicon, which represents each word in the ASR vocabulary (the finite set of words the system can recognise) in terms of individual speech sounds (phones) (e.g. 'Bing' word contains three phones: /b ɪ ŋ/), and 3) acoustic models, which represent the stochastic time-based relationship between the input speech signal and the phones occurring in the speech. Together, the language model and/or the grammars, the lexicon and the acoustic models are used by the speech recogniser to find the most probable sequence of words for the input speech signal. State-of-the-art ASR systems employ statistical modelling techniques and are developed using large quantities of speech for training the acoustic models, and large quantities of text data for training the language model. Acoustic models are typically trained using speech collected from young to middle-aged adults. Because the acoustic properties of speech produced by elderly speakers differ from those produced by younger adults[3], acoustic models expected to successfully recognise elderly speech have to be (re)trained using a sufficient amount of elderly speech.

In the following subsections, we describe the work we did to collect, transcribe and annotate large corpora of elderly speech, to train acoustic models optimised for this kind of speech, and to test the performance of the models.

3.1. Collecting, Transcribing and Annotating Corpora of Elderly Speech

To support the development of acoustic models optimised for the elderly, we collected a large corpus of elderly speech for each of the languages supported by AALFred. We selected speakers from 3rd age universities, care institutions, and social clubs and associations for seniors in different parts of France, Hungary, Poland and Portugal, to ensure a variety of regional accents in the data. All speakers were 60 years of age or older. The French and Polish corpora consist of read newspaper sentences, while the Portuguese corpus also contains a large amount of read command & control prompts. The Hungarian corpus comprises both read newspaper sentences (about 80% of the recordings) and spontaneous command & control utterances (about 20%). The main statistics of the corpora are detailed in Table 1.

Table 1. Main statistics of the EASR corpora.

	#Speakers	Total Audio (hh:mm:ss)
Portuguese	986	185:10:25
French	328	76:09:07
Hungarian	1229	183:42:15
Polish	781	203:17:23

Once the data collection was finished, we used the prompts presented to the speakers as the starting point for orthographically transcribing the recordings with read speech. We verified and corrected those initial transcriptions to ensure that they matched what the speakers said in the recordings. In the case of the Hungarian spontaneous speech, we transcribed the recordings from scratch. In addition, using an annotation scheme, we marked the presence of noises and other audio events in the recordings. These audio events included filled pauses (e.g. *ah*, *hmm*), non-human (e.g. door banging) or human (e.g. coughing) noises, damaged words (e.g. false starts, mispronounced, unintelligible or truncated words), and speech from non-primary speakers (e.g. the recording supervisor). The corpora are collectively called the EASR Corpora of European Portuguese, Hungarian, French and Polish Elderly Speech, and are described in detail by Hämäläinen et al.[12]. In the case of French, Hungarian and Polish, the data-related work was done in the PaeLife project, whereas the Portuguese data was collected, transcribed and annotated in the scope of the Living Usability Lab[6] and Smartphones for Seniors[7] projects.

3.2. Development of Elderly-Specific Acoustic Models

In AALFred, speech input is handled using two different ASR systems: Microsoft Public Speech Platform Runtime version 11[13], which supports French, Polish and Portuguese, and VOXerver[14], which is a SAPI/Microsoft Speech Server-compatible system that supports Hungarian. The goals of the ASR-related work were to create French, Hungarian, Polish and Portuguese acoustic models optimised for the elderly users of AALFred, and to obtain the best possible recognition performance using existing techniques and tools compatible with the requirements of the Microsoft Public Speech Platform.

We divided all elderly speech corpora into three datasets (training set: 85% of the speakers; development test set used for optimisation purposes: 5% of the speakers; evaluation test set used for measuring the final performance of the acoustic models: 10% of the speakers). In the case of French, there was a total of 60.3 hours of audio in the training set, whereas the same figures were 107 hours, 165.8 hours and 147.5 hours for Hungarian, Polish and Portuguese, respectively.

We adapted the French, Polish and Portuguese acoustic models to elderly speech by taking the standard acoustic models that come with Microsoft Public Speech Platform Runtime and retraining them with the elderly speech from the training sets. The standard acoustic models comprise a mix of Gaussian Mixture Model (GMM) -based gender-dependent whole-word models and cross-word triphones that have been trained using several hundred hours of read and spontaneous speech collected from young to middle-aged adult speakers. They also include a silence model, a hesitation model for modelling filled pauses, and a noise model for modelling human and non-human noises. The hesitation and noise models were retrained using the stretches of audio signal that correspond to the hesitation and noise tags inserted into the transcriptions during the transcription and annotation phase.

In addition to the above, in the case of French, we tested a new acoustic modelling paradigm. This paradigm, based on Deep Belief Neural Networks (DNNs), is currently the state-of-the-art acoustic modelling approach, with ample evidence in the literature showing significant performance gains when compared with classic approaches, such as GMM-based acoustic models[15]. For our work, we used existing French DNN-based acoustic models as a starting point, and adapted them to elderly speech using the data in the French training set. Similar DNN-based acoustic models for Polish and Portuguese are currently under development.

The Hungarian acoustic models are trained using a Gaussian Mixture Model (GMM) -based, gender-independent, position-dependent cross-word triphone approach. The training methodology included speaker normalisation and discriminative training. Similar to the other PaeLife languages, the Hungarian acoustic models also include silence, hesitation and noise models. In the case of Hungarian, however, we merged the noise and silence models into an extended silence model. Unlike the acoustic models for the other languages, we trained the Hungarian models from scratch (rather than using younger adult speech models as a starting point). As the transcription and annotation work was still ongoing when we trained the models, the training set included both read speech with manually verified transcriptions and annotations, spontaneous speech with manual transcriptions and annotations, as well as read speech without annotations. We used a lightly supervised selection method to eliminate mispronounced sentences from the unannotated read speech. In the future, we also intend to train DNN-based acoustic models for Hungarian.

3.3. Evaluation Results

To illustrate the improvements in ASR performance that can be achieved by using acoustic models optimised for elderly speech, we trained comparable bigram language models (LMs) for all four PaeLife languages. The French, Hungarian and Polish corpora are the most comparable with each other in terms of contents; they mainly contain read out newspaper sentences (the Hungarian corpus also contains some spontaneous commands; cf. Section 3.1). In addition to newspaper sentences, the Portuguese corpus contains a considerable amount of command & control material; only about half of the recorded utterances are read out newspaper sentences. To keep the ASR results as comparable as possible across languages, we used the sentences in the Hungarian and Polish training sets – excluding the sentences that also appear in the test sets – to train the LMs but, in the case of Portuguese, we also excluded the command & control material from the training material. To compensate for this loss of training sentences in the case of Portuguese, we appended the training material with unused sentences from the original pool of newspaper texts available for collecting the EASR corpus. In the case of French, we had to use all the sentences in the training set for training the LM. Otherwise, the number of words in the test set that are not included in the ASR vocabulary, the so-called out-of-vocabulary (OOV) words, would have been very high; this would have masked the performance improvement arising from the elderly-specific acoustic models. In the case of Hungarian, the LM perplexity (an information theory -derived measure of how well a probability model (the LM) is able to predict a sample (the test set utterances); the lower a perplexity values is, the more accurately the LM is able to predict the word sequences in the test set utterances) and OOV rate are significantly higher than in the case of the other languages. This is due to the agglutinative nature of the language, resulting in thousands of possible word forms for a single stem. Table 2 summarises the key details of the LMs, as well as the results and improvements gained with the specialised acoustic models, as compared with standard acoustic models trained with young to middle-aged adults' speech (baseline).

Table 2. ASR results with acoustic models optimised for elderly speech. The 6th and 7th columns present the word error rates (WERs) obtained on the evaluation test sets using the baseline and the elderly-specific acoustic models. The last column indicates the relative reduction in the word error rates.

Language	ASR vocabulary words	Word tokens (test set)	LM perplexity (test set)	OOV words (test set)	WER (%) Baseline AMs	WER (%) Elderly Speech AMs	WER (%) relative reduction
French	11287	36423	31.9	68	24.2	20.9	13.6
French - DNN	11287	36423	31.9	68	20.2	13.7	32.2
Hungarian	39151	40865	147.1	2372	27.0	19.4	28.1
Polish	19781	83135	54.3	283	16.0	13.6	15.0
Portuguese	8934	52970	60.0	219	18.3	16.4	10.4

As we can see in Table 2, the elderly-specific acoustic models provide considerable improvements in ASR performance over the baseline models. As expected, DNN-based models result in better ASR performance and a higher relative WER reduction than comparable GMM-based models. Once we are able to test ASR performance using LMs specifically developed for the dictation scenarios in AALFred (composing messages and agenda appointments), which – from the language point of view – are much harder ASR tasks than "predicting" newspaper sentences, the gains obtained from the acoustic model optimisation will be even higher. Conversely, the gains are expected to be lower in the case of commands, which are usually easy to recognise using relatively simple grammars. For now, we do not have suitable or enough data for the PaeLife languages to run experiments representing such scenarios.

4. Multilingual Speech-Enabled Interaction in AALFred

In this section, we briefly describe the implementation of the speech modality in AALFred: a generic speech modality component that works decoupled from AALFred services, the interpretation of speech input using Spoken Language Understanding (SLU), and support for the fast integration of new languages by automatically deriving the first versions of semantic and ASR grammars.

4.1. Generic Speech Modality Component

AALFred is based on a MultiModal Interaction (MMI) framework[16], which follows the W3C recommendation of a multimodal architecture[17]. The major components of the architecture that are relevant for this paper include the modalities and the interaction manager (IM), which controls the HCI-related information. Communication between the modalities and the IM is based on events (life-cycle events[17]), and information is encoded for transmission using a mark-up language (Extensible MultiModal Annotation (EMMA)[17]). One important benefit of this architecture is its decoupled nature; the components are developed independently of the application and, as such, can easily be integrated in other applications.

In the PaeLife project, we developed a generic speech modality component[18] to support speech-based interaction with AALFred[19]. One major benefit of the speech modality component is that it is decoupled from the services available in AALFred and can, therefore, easily be integrated in new services (see Fig. 4). Furthermore, it can handle multiple languages and is very scalable when it comes to adding new languages (see Section 4.3).

Similarly to the touch and gesture modality components, the speech modality component communicates with the IM. Whenever an event occurs in a modality component, the component in question uses an EMMA-encoded message wrapped inside an MMI life-cycle event to send the event information to the IM for processing and, if needed, the IM creates a new MMI life-cycle event with the same EMMA-encoded message to be forwarded to the application (AALFred). For instance, if a user has opened the Agenda service and uses a voice command (e.g. "Open Thursday") to request the application to display their schedule for Thursday, the speech modality component will send the corresponding event to the IM, with the semantic output resulting from the SLU processing of the ASR output (see Section 4.2). The IM processes the event and creates a new event to be sent to the application, which then displays the user's schedule for Thursday (see Fig. 3).

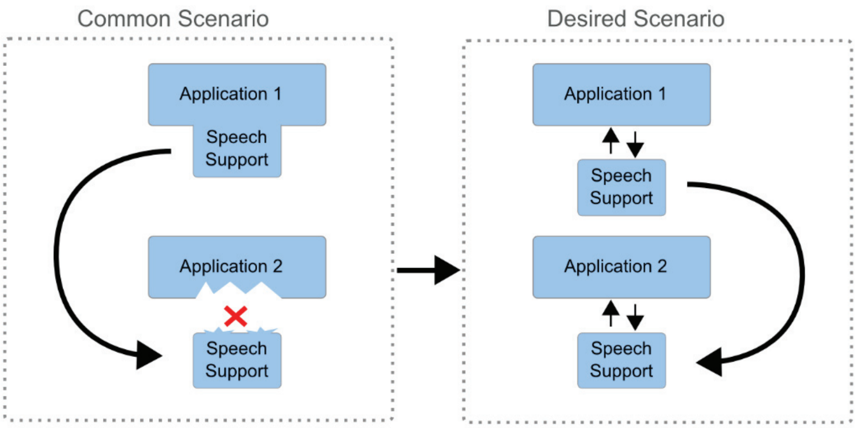

Fig. 4 AALFred uses an architecture in which the speech modality is decoupled from the available services.

```
[Main]
        ([ACTION])
        ([HELP])
;
[ACTION]
        ([AGENDA])
        ([APPOINTMENTS])
[...]
;
[AGENDA]
        (agenda)
        (show my agenda)
        (go to my agenda)
        ([CHANGEDATE])
        (*open [WEEKDAYS])
[...]
;
[WEEKDAYS]
        ([MONDAY])
[...]
;
[THURSDAY]
        ([Thursday])
;
[...]
```

Fig. 5. Example of a semantic grammar used in AALFred.

```
ASR output: open Thursday
Semantic Result: [ACTION].[AGENDA].[WEEKDAYS].[THURSDAY]
```

Fig. 6. Example of the semantic tags resulting from the semantic parsing of the ASR output "open Thursday".

4.2. Automatic Speech Recognition with Spoken Language Understanding

To be able to use speech input to access and operate the AALFred services, we needed a way of robustly extracting the intended meaning of the speech input from the ASR output (i.e. the string of words that the automatic speech recogniser "thinks" it "heard"). This can be done using an SLU parser, which makes use of semantic grammars. Semantic grammars map possible ASR output to semantic tags that capture the intended meaning of speech input, and the SLU parser parses ASR output into a sequence of semantic tags defined in the grammars. The tags are then used by the application to produce the desired actions (e.g. opening the Agenda service). We chose to use the Phoenix[20] parser and grammar specification format. This is because the Phoenix parser has been designed to be robust to errors in ASR output and disfluencies in speech.

Let us consider SLU in AALFred. Whenever a user speaks and the ASR engine recognises the speech input, the speech modality component requests the service in question to extract the semantic tags for the ASR output, and sends the tags for the IM to process. Fig. 5 presents an example of a semantic grammar used in AALFred, and Fig. 6 is an example of the semantic tags for the ASR output "open Thursday".

4.3. Multilingual Support

The speech modality needs semantic grammars for all supported languages. To produce these semantic grammars, we developed a service[21] that enables the automatic translation of an English grammar into other languages, for instance, using Microsoft Translator API[22]. In other words, the developer only needs to create a semantic grammar for English, and this grammar can then automatically be translated into other languages. The semantic tags are the same for all languages, i.e., the tags in the English grammar are copied over to the automatically translated grammars. The ASR grammars, which are used by the automatic speech recognisers during recognition time (cf. Section 3), are generated by extracting all possible commands from the semantic grammars. Due to the limitations of machine translation, the grammar translation service also supports the manual revision and correction of the automatically translated grammars. Furthermore, AALFred is capable of dynamically updating grammars based on user input (e.g. when a user adds new contacts in the Contacts service). Of course, the semantic and ASR grammars could also be created manually. However, the grammar translation service that we developed allows for a fast integration of new languages into AALFred.

5. Conclusions

In this paper, we described the work done in the area of multilingual automatic speech recognition in the scope of the PaeLife project[10], which had the goal of developing a multimodal virtual personal life assistant to help the elderly remain active and socially integrated. The personal life assistant, AALFred, supports four European languages: French, Hungarian, Polish and Portuguese. We implemented the speech modality of AALFred as a generic component that is decoupled from the available AALFred services and can, therefore, be easily integrated in any future services. To be able to quickly increase the number of supported languages, we developed a grammar generation service that allows the automatic translation of English-language grammars into other languages, and supports the manual verification and correction of these automatically translated grammars. To provide the best possible user experience for the target audience, the automatic speech recognisers used in AALFred were optimised for elderly speech. For this purpose, we collected large corpora of elderly speech for all the supported languages. The results of our experiments show that the optimised speech recognisers can provide a considerable improvement in automatic speech recognition performance on elderly speech as compared with standard speech recognisers tuned for younger adult speech. Furthermore, the collected corpora are valuable resources for the international speech community. They will soon be available for research and development purposes through the Linguistic Data Consortium (LDC)[23].

Acknowledgements

Authors acknowledge the funding from AAL JP and national agencies: MLDC was funded by the 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 acknowledges the support of the FuturICT project (TÁMOP-4.2.2.C-11/1/KONV-2012-0013) and the PaeLife project (AAL-08-1-2001-0001).

References

1. Teixeira, V., Pires, C., Pinto., F., Freitas, J., Dias, M.S., Mendes Rodrigues, E. Towards elderly social integration using a multimodal human-computer interface. In *Proc. Living Usability Lab Workshop on AAL Latest Solutions, Trends and Applications*, Vilamoura, Portugal; 2012.
2. Bernsen, N.O. Towards a tool for predicting speech functionality. *Speech Communication* 1997; 23(3):181-210.
3. Xue, S.A., Hao, G.J. 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* 2003; 46:689-701.
4. Wilpon, J.G., Jacobsen, C.N. A study of speech recognition for children and the elderly. In *Proc. ICASSP*, Atlanta, GA, USA; 1996.
5. Vipperla, R., Renals, S., Frankel, J. Longitudinal study of ASR performance on ageing voices. In *Proc. Interspeech*, Brisbane, Australia; 2008.
6. Living Usability Lab. [Online]. Available: http://www.livinglab.pt/. [Accessed: 5-Feb-2015].
7. Smartphones for Seniors. [Online]. Available: http://www.smartphones4seniors.org/. [Accessed: 5-Feb-2015]
8. Project ROMEO. [Online]. Available: http://www.projetromeo.com/. [Accessed: 5-Feb-2015]
9. Zora. [Online]. Available: http://www.zorarobot.be/. [Accessed: 5-Feb-2015]
10. PaeLife: Personal Assistant to Enhance the Social Life of Seniors. [Online]. Available: http://www.paelife.eu/. [Accessed: 5-Feb-2015]
11. Francisco, P., Almeida, N., Pereira, C., Silva, S. Services to support use and development of speech input for multilingual multimodal applications for mobile scenarios. In *Proc. ICIW*, Paris, France; 2014.
12. Hämäläinen, A., Avelar, J., Rodrigues, S., Dias, M.S., Kolesiński, A., Fegyó, T., Németh, G., Csobánka, P., Lan, K., Hewson, D. The EASR Corpora of European Portuguese, French, Hungarian and Polish Elderly Speech. In *Proc. LREC*, Reykjavik, Iceland; 2014.
13. Microsoft Speech Platform 11.0. [Online]. Available: http://www.microsoft.com/en-us/download/details.aspx?id=27224. [Accessed: 12-Feb-2015].
14. Tarján, B., Sárosi, G., Fegyó, T., Mihajlik, P. Improved recognition of Hungarian call center conversations. In *Proc. SpeD*, Cluj-Napoca, Romania; 2013.
15. Dahl, G.E., Yu, D., Deng, L., Acero, A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing - Special Issue on Deep Learning for Speech and Language Processing* 2012; 20(1):33-42.
16. Almeida, N., Teixeira, A. Enhanced interaction for the elderly supported by the W3C multimodal architecture. In *Proc. Interacção*, Vila Real, Portugal; 2013.
17. Bodell, M., Dahl, D., Kliche, I., Larson, J., Porter, B. *Multimodal Architecture and Interfaces*. http://www.w3.org/TR/mmi-arch/; 2012.
18. Almeida, N., Silva, S., Teixeira, A. Design and development of speech interaction: A methodology. In *Proc. HCI International*, Crete, Greece; 2014.
19. Teixeira, A., Hämäläinen, A., Avelar, J., Almeida, N., Németh, G., Fegyó, T., Zainkó, C., Csapó, T., Tóth, B., Oliveira, A., Dias, M.S. Speech-centric multimodal interaction for easy-to-access online services: A personal life assistant for the elderly. In *Proc. DSAI*, Vigo, Spain; 2013.
20. Ward, W. Understanding spontaneous speech: The Phoenix system. In *Proc. ICASSP*, Toronto, Canada; 1991.
21. Teixeira, A., Francisco, P., Almeida, N., Pereira, C., Silva, S. Services to support use and development of speech input for multilingual multimodal applications for mobile scenarios. In *Proc. ICIW,* Paris, France; 2014.
22. Microsoft Translator API. [Online]. Available: http://www.microsoft.com/translator/translator-api.aspx. [Accessed: 12-Feb-2015].
23. Linguistic Data Consortium. [Online]. Available: https://www.ldc.upenn.edu/. [Accessed: 12-Feb-2015].