Giving Voices to Multimodal Applications

Nuno Almeida[1,2], António Teixeira[1,2(✉)], Ana Filipa Rosa[1],
Daniela Braga[3], João Freitas[4], Miguel Sales Dias[4,5], Samuel Silva[1,2],
Jairo Avelar[4], Cristiano Chesi[4], and Nuno Saldanha[4]

[1] Institute of Electronics and Telematics Engineering, University of Aveiro, Aveiro, Portugal

[2] Department of Electronics, Telecommunications and Informatics Engineering, University of Aveiro, Aveiro, Portugal
ajst@ua.pt

[3] Voicebox Technologies, Bellevue, WA, USA

[4] Microsoft Language Development Center, Lisbon, Portugal

[5] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisbon, Portugal

Abstract. The use of speech interaction is important and useful in a wide range of applications. It is a natural way of interaction and it is easy to use by people in general. The development of speech enabled applications is a big challenge that increases if several languages are required, a common scenario, for example, in Europe. Tackling this challenge requires the proposal of methods and tools that foster easier deployment of speech features, harnessing developers with versatile means to include speech interaction in their applications. Besides, only a reduced variety of voices are available (sometimes only one per language) which raises problems regarding the fulfillment of user preferences and hinders a deeper exploration regarding voices' adequacy to specific applications and users.

In this article, we present some of our contributions to these different issues: (a) our generic modality that encapsulates the technical details of using speech synthesis; (b) the process followed to create four new voices, including two young adult and two elderly voices; and (c) some initial results exploring user preferences regarding the created voices.

The preliminary studies carried out targeted groups including both young and older-adults and addressed: (a) evaluation of the intrinsic properties of each voice; (b) observation of users while using speech enabled interfaces and elicitation of qualitative impressions regarding the chosen voice and the impact of speech interaction on user satisfaction; and (c) ranking of voices according to preference.

The collected results, albeit preliminary, yield some evidence of the positive impact speech interaction has on users, at different levels. Additionally, results show interesting differences among the voice preferences expressed by both age groups and genders.

Keywords: Synthetic voices · Speech output · Multimodal interaction · Age effects

Daniela Braga — Work done while at Microsoft, now at Voicebox.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2015, LNCS 9170, pp. 273–283, 2015.
DOI: 10.1007/978-3-319-20916-6_26

1 Introduction

The use of speech in HCI is gaining more popularity [1]. Assistants such as Siri and Cortana are major examples of the success and user acceptance. Speech enabled applications should also offer the traditional ways of interaction. For example, in mobile phones the user should additionally be able to interact with touch and have visual feedback. So, developing such applications is more time consuming and implies some complex tasks. To add speech-based interaction to applications, developers use what the market offers in terms of speech engines. In this context, it is important to simplify the process of creating speech enabled applications, providing developers with easy methods and tools to embody speech into their application.

Another issue concerning the use of speech for interaction is the reduced variety of voices available. Most speech engines offer only one or, at most, two voices for each language, typically an adult male or female, and they do not support every language. This way, the user is forced to use that voice for all applications. For example, for European Portuguese, the language we adopt for this paper, there was only one female voice easily available, although it has been shown that the voice is an important factor of engagement [2], possibly leading to a stronger confidence in the application and improved comfort for the user. If multiple voices are made available, another problem that must be addressed is the selection of the more adequate voice for the user and application. In fact, choosing a particular voice might have different impacts depending on the target audience and on the context it is used. Sometimes, providing a gender match between the voice and the user might enable social-identification and enhance user engagement, but female voices that sound too masculine or vice versa, might have a negative impact [2]. The user might also be confused if a voice uses vocabulary that is inconsistent with its perceived age, e.g., a younger voice speaking as an older adult. Therefore, the available voices might influence what can be said and how it is said. There are issues that concern certain user groups. For example, if speech rate is too fast there is evidence that the elderly have problems in understanding it, which might constitute a strong barrier to its use [3–5]. Furthermore, when an application involves very different tasks or interface elements, using different voices for each might improve usability [6].

Considering all these aspects, the need for supporting multiple voices adapted to the HCI scenario, with different characteristics, is made clear. These voices will allow addressing, in a first instance, user preferences, and will serve as grounds for research to improve our understanding of the impact of individual voice characteristics on interaction and user experience, in a variety of applications and contexts.

In this paper we present contributions regarding: (a) our method to deliver speech interaction to applications, simplifying developers work and making available a simple way of choosing the voice (in Sect. 2); (b) the process that we have used to create new voices for Portuguese with different genders and ages (in Sect. 3); and (c) the results of three preliminary evaluations assessing voice preferences by adult and elderly users of both genders (Sect. 4). Conclusions are presented in Sect. 5.

2 Speech Modality – Generic Support to Multiple Languages and Voices

Taking into consideration the challenges identified in the previous section, it is important to foster ways to easily support speech output in different languages, with a rich set of voices, and enabling their selection both during development and runtime.

Our consideration of speech interfaces is inserted in the wider scope that encompasses the design and development of applications supporting multimodal interaction. In this context, speech is addressed as a modality, aligned with the concepts of the W3C recommendations for a multimodal architecture [7]. The speech modality communicates with the interaction manager (IM), responsible for managing all the modalities and their communication with the application, through standard markup messages. One of the major features provided by this architecture is that its different components can be developed as standalone modules [8] that communicate with the remaining architecture elements in a standard way. Therefore, developers do not need to master all the constantly evolving technologies considered in multimodal interaction scenario. Experts may be given specific tasks on that subject proposing a standalone module that is easily integrated. This also has a positive impact on how easily a modality can be improved, without need of changing the application core, and on how existing modalities can be reused in new applications adopting the same architecture.

Following this rationale, a generic modality was proposed [9] supporting speech input and output. Considering speech output, which has the most relevance to the subject matter of this paper, this modality uses the installed engines and language dependent language packs and voices, and provides a small set of simple methods for application developers to change/select voices and send texts for being synthesized and transmitted to sound output devices, such as speakers or headphones (Fig. 1).

It is in the context of this generic modality that the outcomes of the work described in the following section were considered and made available in different applications, in the scope of project AAL4ALL and PaeLife, resulting in a modality supporting a rich set of voices created for multiple languages in a unified way.

3 Development of New Voices

Until recently, for a given language, only one or two voices were available, as the methods to create voices for Speech Synthesis were complex, time consuming and very expensive. Recent technological developments and the work done in collaboration with Microsoft Language Development Center in Portugal, in projects such as Smartphones for Seniors[1] [10–12] and AAL Paelife[2] [13], allowed the creation of a set of new voices. More specifically, the application of robust stochastic training methods (e.g. Statistical Parameter Synthesis, SPS, or Voice Adaptation, VA), which are extensions of simpler methods based on Hidden-Markov-Models (HMM) applied to TTS [14, 15],

[1] http://www.smartphones4seniors.org/en-us/home.aspx.

[2] http://www.microsoft.com/portugal/mldc/paelife/.

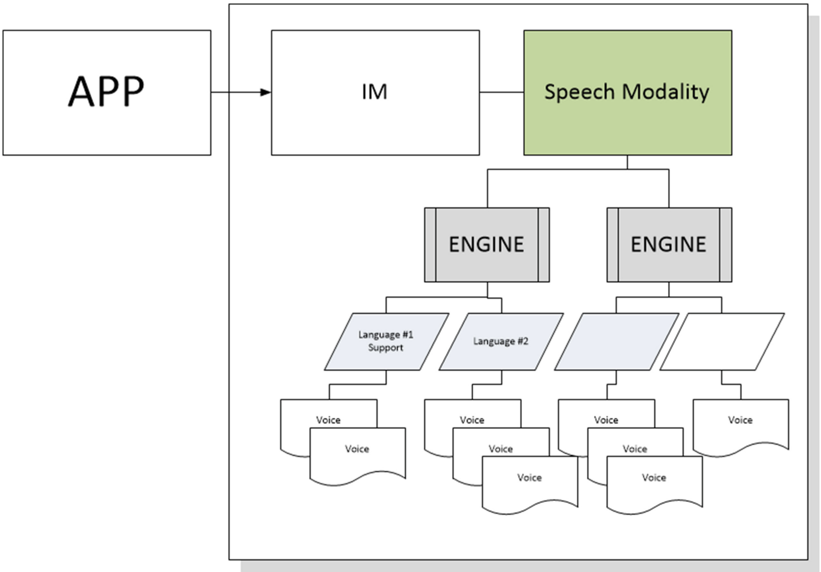

Fig. 1. The generic speech modality offers the means to add multilanguage support to an application, offering easy access to multiple languages and voices without changes to the core of the application.

required fewer recordings with respect to the ones required using classic approaches to speech synthesis (e.g. Units Selection based methods [16]) and permitted faster realizations of high-quality, flexible, personalized voices.

The methodology used for creating these new voices, one of the main contributions presented in this article, consists of a set of stages that can be summarized as follows:

1. The voice donors, providing the recordings to be used for training a personalized voice, need to be selected. Since recording voices is time consuming, a first selection, based on small samples of audio recorded from different individuals, is implemented using an online interface.
2. The collected recordings go through a two-stage evaluation procedure:

   (a) First, a group of speech experts rate the voices based on criteria such as pleasantness, intelligibility, expressiveness, attitude, accent and perceived age, as well as the adequacy of the voice profile with respect to the application scenario.
   (b) Second, a group of end-users listen to the shortlisted recordings and score them on a scale similar to the one used by experts (Mean Opinion Score, MOS [17]).
3. The higher ranked individuals record less than 2000 phonetically rich prompts. The recordings should be performed in a sound treated room to suppress noise.
4. These recordings get automatically chucked according to various acoustic features and aligned with a phonetically annotated version of the script used to elicit them.

5. The processed audio materials are used to train the artificial voice model using either a Statistical Parameter Synthesis approach or the Voice Adaptation approach (this is possible only if a rich acoustic language model already exists based on previous recordings).

With the increasing number of elderly and their involvement in the development of applications for seniors and/or AAL, it became relevant to explore if the availability of seniors' voices as an option would facilitate voice enabled applications' adoption by elderly users. Therefore, following this methodology, we have added two senior voices to the existing two adult voices.

4 Users' Preferences

To gather information regarding preferences – needed for guidelines for developers, to accompany the multimodal framework – it was decided to have information from users on their voice preferences. A first objective, reflected in the choice of users, was to understand whether the senior users prefer to have TTS systems interacting with them using young or senior voices. The second objective was to perform the evaluation in the context of applications we have developed using the generic speech modality and the new voices (the personal assistant AALFred [13] and the Medication Assistant for Windows Phone [10–12]). Due to practical constraints, our initial plan for evaluating the two applications with the same users was replaced by separate evaluations.

4.1 Evaluation of the Voices by Different Age Groups in AALFred

An online MOS evaluation was conducted with a paragraph of news and a sentence from the AALFred application produced with the 5 voices (4 new ones plus the existent female voice, Hélia).

Test Conditions. The evaluation started with an introductory text providing general information about the test. Before beginning, data pertaining the user profile and context was collected (gender, year of birth, loudspeakers/headphones, silent/noisy environment, and whether the speaker was native or not). During the test, the listener played a sound sample and, after that, answered 8 questions, all of which were on a 5 point scale (1–5). The questions of the test are listed in the following Table 1:

Listeners. Thirty-two **Portuguese** speaking listeners participated in the test (11 of them were elderly - above 60 years old - and 21 were young adults). The average age of the elderly listeners was 65 years old, whereas for the young listeners it was 29 years old. All of them except three were native speakers. Twenty-six of them were males and fourteen of them used loudspeakers, while 18 used headphones. Thirty of them listened to the test in a silent environment, and two users were in noisy conditions.

Results. The average values for each parameter were considered for analysis. As average performance of the voices is different, as a first step, scores were preprocessed by subtracting the average value for the parameter.

Table 1. Information on the parameters, questions and scales used in the evaluation of the voices.

ID	Parameter	Question	Scale
Q1	Pleasantness	How do you like this voice?	Not at all! (1) … Yes, a lot! (5)
Q2	Intelligibility	How understandable is this voice?	I missed many words (1) … Everything was crystal clear (5)
Q3	Expressiveness	How dynamic is this voice?	Do you feel the person speaking bored? (1) … Is the person speaking exited and motivated while speaking? (5)
Q4	Attitude	Do you find the speaker charismatic?	Yes, a lot! (1) … Not at all! (5)
Q5	Rhythm/speech rate	How do you like the pacing rate?	Not at all! (1) … Yes, a lot! (5)
Q6	Accent	Do you hear an accent in the voice?	Yes, I perceive an accent and this is unacceptable (1) … No, this is perfect Portuguese (5)
Q7	Perceived age	Is the person speaking matching the age range?	Not at all! (1) … Yes, a lot! (5)
Q8	Artificiality	Do you feel the voice has an annoying artificial flavor?	Yes, this is extremely annoying (1) … No, at all! (5)

The new scores were used in a MANOVA with listener age and voice age as factors. MANOVA results indicated as significant, for example, the effect of listeners age for Q3 and Q8. MANOVA results were also confirmed, in general, with an ANOVA made with the sum of the several evaluation parameters.

The identified significant differences were further investigated. As representative examples, the results for listeners' age and voice age interaction are presented in Fig. 2 and for listeners' age and voice's gender in Fig. 3.

It is noticeable, in Fig. 2, that the two groups of listeners (elderly and young adults) have very different preferences and perceptions regarding speech rate and artificiality. Elderly consider the elderly voices as having the best speech rate, while young listeners have an opposite opinion. Regarding artificiality, the elderly assigned a score above average to the voices for both age groups, contrasting to the below average scores attributed by young listeners

When looking at the results for each of the two tasks (AALFred and News), as presented in Fig. 3, it is again noticeable that elderly and young listeners have different preferences regarding attitude (Q4) and speech rate (Q5). Elderly listeners score male voices with higher values in the two tasks. The effect of the utterance type is also visible.

4.2 Evaluation of TTS in AALFred

As part of a broader evaluation of the PaeLife assistant, information was gathered on the TTS quality of the new voices.

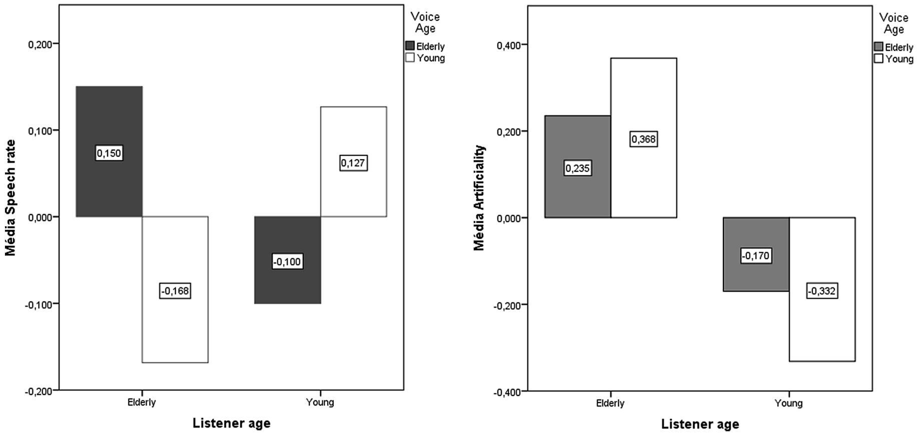

Fig. 2. - Results regarding speech rate (Q5) and artificiality (Q8) as function of listener and voice ages.

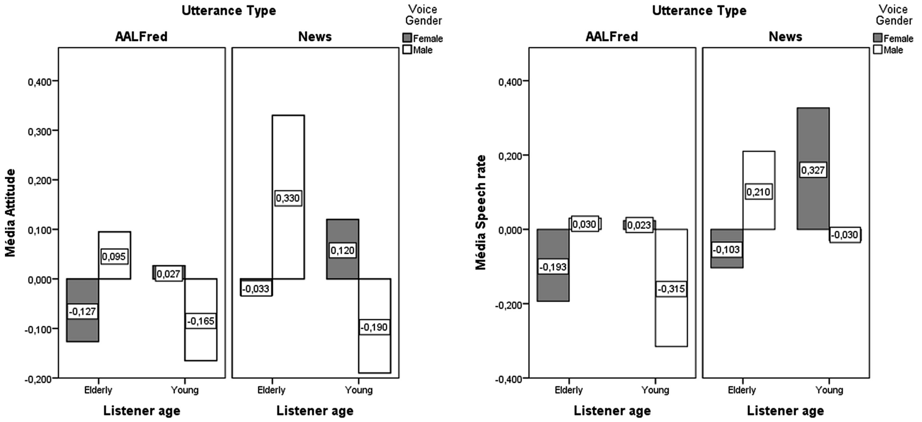

Fig. 3. - Results regarding attitude (Q4) and speech rate (Q5) as function of listener age, voice gender and utterance type.

Method and Users. The evaluation adopted a direct observational method, which involved a researcher observing users behavior and actions while taking notes.

Five users were recruited and had the opportunity to experience and use AALFred for a period of time. The field trials started in the beginning of November 2014 and are planned to be finished by the end of February 2015. The emphasis was placed on speech and touch usability evaluation. Regarding TTS voices, 5 were made available for selection by the users, 4 of them specifically developed for the Project, as described earlier.

The AALFred configuration process was performed individually for each one of the users and email and Facebook accounts where created for those who did not use these services before but were interested in trying them out in AALFred.

Users were then introduced to AALFred and a short demonstration was done for each one of the participants. A small instruction leaflet was also delivered to the users for them to consult if they had difficulties interacting with the tablet and/or with the AALFred application itself. The media diary (where participants were asked to register all interactions with AALFred) was then explained and a few examples were discussed and registered.

Results. Regarding TTS and the voices, the main outcomes of the evaluation were:

1. Output using synthetic speech was felt as important, necessary and very useful in some particular cases like messages and news;
2. Although there was a preference for the speech modality, most of the users found that speech and touch were noteworthy and complementary;
3. All the users enjoyed the fact that they could choose between 5 available voices;
4. The Microsoft standard voice "Hélia" was not chosen and used by any of the users;
5. Some of the users also stated that they had chosen a voice that somewhat resembled the researchers' own voice – in this study the researcher was also the one responsible for solving the existing problems and helping the elderly user whenever needed. Therefore, the researcher and the help topic seemed to be closely linked;
6. The importance of feedback by voice and the recognition that this somewhat filled the void that is present when users interact with technology. The possibility of having AALFred's reply seemed to be well appreciated by all seniors and was considered one of the most useful features of the application. Two users stated that this kind of interaction was an important help when they were feeling lonely as voice simulates presence, enabling the application to communicate with the user in the same way the user communicates with their family and friends: by using speech. This brings another meaning to "assistant applications for the elderly".

4.3 Evaluation in a Medication Assistant for Smartphones

The Medication Assistant application [10, 12] allows the user to manage his/her medication intakes by showing reminders, but it also aims to go beyond that by providing information and advice in case he/she forgets to take the medicine or feels side effects.

Method and Users. A set of 6 elderly with ages from 57 to 90, two male and four female, participated in this third evaluation. They were instructed to select one of the voices, and after this, they were asked to perform a task in the application consisting in navigating in the application through the different menus and obtain the information of any drug. While browsing, the application gives spoken feedback and at the end reads information about the drug. They repeated the process for all the voices available (only the 4 new ones were considered, leaving 'Hélia' out). After using the application, users were asked to rank the voices.

Results. The results obtained (ranks from 1 to 4, with 1 meaning first choice) were analyzed using repeated measures ANOVA, that showed as significant ($p < 0.05$) the effect of voice gender, the gender of the listener and the interaction between gender of the voice and gender of the listener. The interaction effect, the most relevant for this

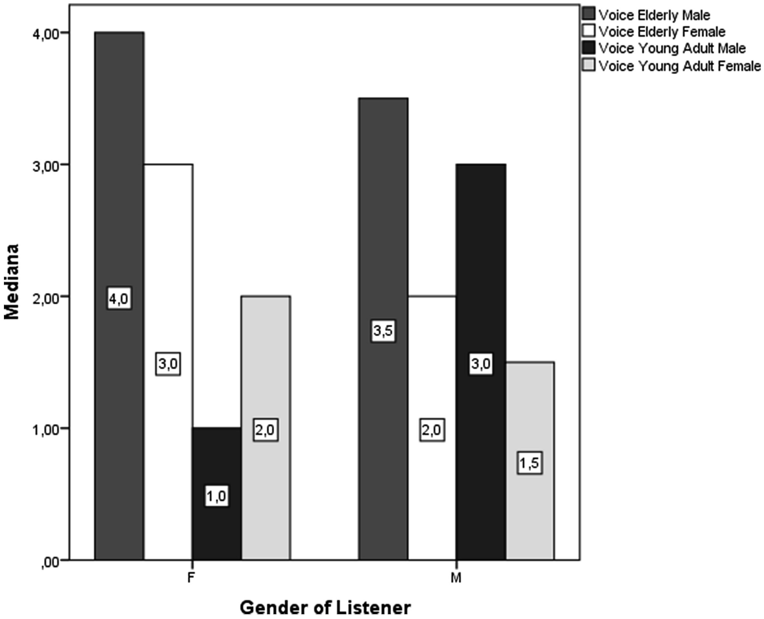

Fig. 4. Median voice ranks as function of the listener's gender. As bars represent ranks, smaller bars represent better positions, contrary to the most common interpretation.

paper objective, is depicted in Fig. 4, using the median of the ranks (considered more adequate than average values).

The figure shows that while female listeners prefer the voice of the young adult male (with a median rank of 1), male listeners prefer the two female voices, with some preference for the young one.

5 Conclusion

In this paper we report the three lines of work we are exploring to give more and better voices to HCI applications. First, we report on the generic/unified way we developed to ease the inclusion of voice output in applications, even by developers with a lack of knowledge in the area. We also describe how new voices are being created, profiting from recent advances in corpus based speech synthesis is also described. We regarded the inclusion of these additional voices as a very important addition to the proposed speech modality, particularly addressing both genders and young adult and senior voices.

Last, but not least, results on differences in preferences regarding two age groups (young adult and older-adult) and their correlation with applications are presented, providing very useful first insights and contribution for guidelines on how to select voices for certain applications and interaction models.

The results highlight the importance of giving the users the possibility of choice and also illustrates that sometimes the preferred TTS voice is not the voice with more quality (from the technical point of view) but a more familiar sounding voice.

Acknowledgments. Research partially funded by IEETA Research Unit funding (PEst-OE/EEI/UI0127/2014), project Cloud Thinking (funded by the QREN Mais Centro program, ref. CENTRO-07-ST24-FEDER-002031), Marie Curie Actions IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP), project Smart Phones for Seniors (S4S), a QREN project (QREN 21541), co-funded by COMPETE and FEDER, project PaeLife (AAL-08-1-2001-0001) and project AAL4ALL (AAL/0015/2009).

References

1. Bijani, C., White, B.-K., Vilrokx, M.: Giving voice to enterprise mobile applications. In: Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI 2013, p. 428. ACM Press, New York, USA (2013)
2. Nass, C., Brave, S.: Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship. MIT Press, Cambridge, MA, USA (2007)
3. McCoy, S.L., Tun, P., Cox, L., Wingate, A.: Aging in a fast-paced world: rapid speech and its effect on understanding. ASHA Lead. **12**, 30–31 (2005)
4. Gordon-Salant, S., et al.: Sources of age-related recognition difficulty for time-compressed speech. J. Speech Lang. Hear Res. **44**, 709–719 (2001)
5. Vipperla, R., Wolters, M., Renals, S.: Spoken dialogue interfaces for older people. Adv. Home Care Technol. **1**, 118–137 (2012)
6. Hale, K., Reeves, L., Stanney, K.: Design of systems for Improved Human Interaction (2006)
7. Bodell, M., Dahl, D., Kliche, I., Larson, J., Porter, B., Raggett, D., Raman, T., Rodriguez, B.H., Selvaraj, M., Tumuluri, R., Wahbe, A., Wiechno, P., Yudkowsky, M.: Multimodal architecture and interfaces: W3C recommendation. http://www.w3.org/TR/mmi-arch/
8. Dahl, D.A.: The W3C multimodal architecture and interfaces standard. J. Multimodal User Interfaces **7**, 171–182 (2013)
9. Almeida, N., Silva, S., Teixeira, A.: Design and Development of Speech Interaction: A Methodology. In: Kurosu, M. (ed.) HCI 2014, Part II. LNCS, vol. 8511, pp. 370–381. Springer, Heidelberg (2014)
10. Teixeira, A., Ferreira, F., Almeida, N., Rosa, A., Casimiro, J., Silva, S., Queirós, A., Oliveira, A.: Multimodality and adaptation for an enhanced mobile medication assistant for the elderly. In: Proceedings of the Third Mobile Accessibility Workshop (MOBACC), CHI 2013, France (2013)
11. Ferreira, F., Almeida, N., Rosa, A.F., Oliveira, A., Teixeira, A., Pereira, J.C.: Multimodal and adaptable medication assistant for the elderly: A prototype for interaction and usability in smartphones. In: 2013 8th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–6. IEEE, Lisboa (2013)
12. Ferreira, F., Almeida, N., Rosa, A.F., Oliveira, A., Casimiro, J., Silva, S., Teixeira, A.: Elderly centered design for interaction – the case of the S4S medication assistant. In: 5th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, DSAI (2013)

13. 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: Proceedings of the DSAI 2013, Procedia Computer Science, pp. 389–397 (2013)
14. Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Commun. **51**, 1039–1064 (2009)
15. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K.: The HMM-based speech synthesis system version 2.0. In: Speech Synthesis Workshop, Bonn, Germany, pp. 294–299 (2007)
16. Hunt, A.J., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing Conference, pp. 373–376. IEEE (1996)
17. Viswanathan, M., Viswanathan, M.: Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. Comput. Speech Lang. **19**, 55–83 (2005)