Voice Control in Smart Homes Using Distant Microphones: A VoiceXML-Based Approach

Gloria López, Victor Peláez, Roberto González, and Vanesa Lobato

Fundación CTIC, c/Ada Byron, 39, Edificio Centros Tecnológicos,
Parque Científico y Tecnológico Gijón, Asturias, 33203, Spain
{gloria.lopez,victor.pelaez,roberto.gonzalez,
vanesa.lobato}@fundacionctic.org

Abstract. This paper proposes the design of a voice control module for intelligent environments, primarily oriented to home environments. An intelligent environment is understood to be a ubiquitous space equipped with embedded devices. This solution is based on the main standards in the field of speech technologies (VoiceXML, MRCP, SRGS and SISR), dynamically adaptable to structural changes in the home automation system and scalable to the number of rooms and devices in the home. The final solution has been validated in a real home automation installation, using distant speech recognition and a keyword detection approach (keyword spotting, KWS). KWS works as an input filter for the dialogue system, making it more robust against noise. Test results have shown the technical feasibility of the solution and promising user acceptance.

Keywords: smart home, ambient intelligence (AmI), voice control, distant speech recognition, keyword spotting (KWS).

1 Introduction

In a world in which the home is a space composed of a growing number of increasingly complicated digital devices, the concept of ambient intelligence takes on new importance. In this context, it is particularly necessary to facilitate human interactions through natural and intuitive interfaces embedded in objects which form a part of the environment, and voice and multimodal interfaces play a key role [1]. The aim of these interaction types is to use the natural modes of human communication, such as language and behaviour, by applying different recognition technologies, such as those based on voice or gestures.

In the field of speech interfaces, there are many studies focused on natural interaction [2][3] with limited success. Therefore, limitations of natural speech interaction have led to usability being questioned even within the scientific community [4]. There are also less ambitious alternatives to natural interaction in terms of language restriction, but more suited to commercial applications in terms of effectiveness. These alternatives are based on the use of language models restricted to specific domains and speech recognition grammars [5].

Apart from the type of interaction, there are other two key factors when designing voice control solutions for home environments: the standardization of speech

D. Keyson et al. (Eds.): AmI 2011, LNCS 7040, pp. 172–181, 2011.

technologies and the evaluation of the voice control solution. VoiceXML is the most relevant standard for dialogue management with high commercial acceptance. However, although there is a wide range of VoiceXML products on the market, their use in the digital home is limited to remote voice control via telephone [6]. Other often overlooked aspect of voice control in ambient intelligence is the necessity of using pre-installed or embedded microphones, also known as distant microphones. These microphones are not intrusive but are more sensitive to the high degree of variability in the speech signal than close-talk microphones [7], resulting in an increase in the percentage of speech recognition errors. The improvement of robustness against noise in smart environments has been addressed from various perspectives, such as the use of microphone arrays [8] or keyword spotting approaches [9]. However, it is not usually addressed in evaluations of real installations [10][11].

The basis of this work is to develop a voice control solution for digital homes, without the restriction of a specific user group, based on the use of generic hardware and, as far as possible, based on the most widely used speech technology standards. It also includes the evaluation of the proposed system by real users in a home automation environment equipped with built-in microphones. As a continuously enabled speech recognition system is more sensitive to noise, a keyword spotting technique is used. In this way, the KWS approach can be evaluated using live audio, a type of test rarely documented in scientific literature although such techniques are widely accepted in smart home solutions.

2 System Overview

2.1 General Overview

Two aspects of the proposed solution are related to speech technology standardization. Firstly, the core of the software architecture is a VoiceXML platform with MRCP protocol support. The use of the MRCP protocol for communication with the speech synthesis (TTS) and speech recognition (ASR) technology offers several advantages. It guarantees the interoperability of these technologies with any other product with similar features on the market, and simplifies the distribution of the elements over the network. Secondly, taking into account that natural interaction is one of the remaining challenges of speech technology [4], restricted grammars are used to increase efficiency although the size of these grammars is large enough to ensure an adequate degree of naturalness in communication. These grammars fulfil the Speech Recognition Grammar Specification (SRGS) and also include semantic interpretation (SISR).

Dynamic adaptability to changes in the home automation installation is mainly related to the way in which the initialization of the context information is addressed. In this work, this information consists of an XML configuration file for each type of device in the installation. This file defines the actions supported by each type of device and any other relevant information regarding these actions, such as configuration parameters and correspondence with actions in the automation middleware.

Focusing on the specific logic of the dialogue manager, a frame-based approach [12] is used. The slots of these frames are generated dynamically based on the state of

the dialogue and the context information handled by the dialogue system. In a home automation installation, each device belongs to a class shared with other devices of similar functionality. Each device also has a unique name which distinguishes it from other devices of the same type. With this device identification and given that actions can have a variable number of parameters, each frame is composed of four elements: action, type of device, device name and list of action parameters (optional).

As the design and evaluation of the system uses distant microphones, the proposed architecture includes the use of a keyword spotting technique, which activates and deactivates the dialogue system.

The proposed hardware solution is based on common computers, one for each voice-enabled room, and other generic audio devices such as speakers, microphones and optional microphone preamplifiers. This feature facilitates the scalability of the solution in terms of the number of rooms that can be controlled. In addition, the solution is independent of the specific hardware in the home automation installation.

2.2 Software Architecture

Fig. 1 shows the proposed software architecture based on a distributed model with two blocks: a specific module for the speech, called the voice control module; and a second block, called the hardware controller, which handles the communication with the hardware devices to be controlled. The integration of these two independent blocks is based on two aspects. Firstly, the inclusion of the information related to the translation of each action into tasks on the home automation installation in the XML configuration files. Secondly, the integration of the dialogue manager and the alert generator as two new services of the service architecture used by the hardware controller module.

The dialogue manager has been implemented as a web application, and is responsible for controlling dialogue flow and turning dialogue actions into hardware controller tasks. The flow of the dialogue depends on the previous dialogue turns and other contextual information obtained from the automation hardware controller. The output of the dialogue manager is based on a set of predefined templates which will be completed dynamically, depending on the state of the dialogue.

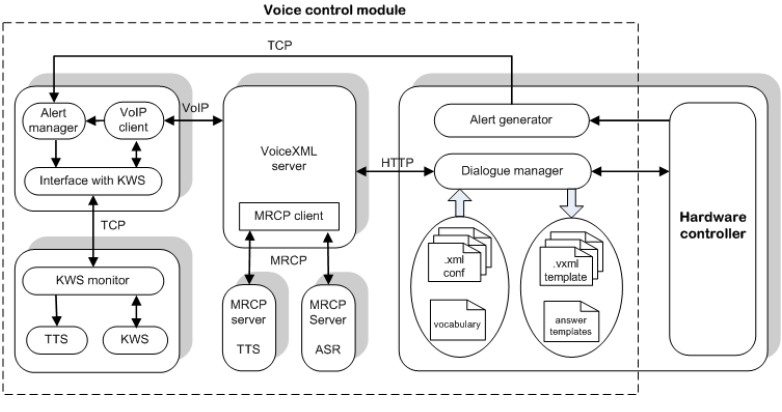

Fig. 1. System software architecture

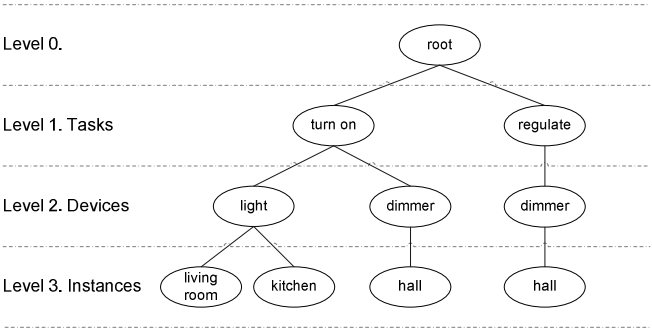

Fig. 2. Intermediate tree-shaped representation

For the internal structure of the dialogue manager, we propose the use of an intermediate tree-shaped representation which allows the storage of environment information. This tree-shaped representation facilitates the access and processing of environment information during the dialogue as in [13]. This intermediate representation is generated during the start up of the dialogue manager, using a sequential scan over all the home automation devices managed by the automation hardware controller. The evaluation of each device instance for which the Visitor design pattern is essential involves the processing of the XML file for that particular type of device, as well as the processing of the linguistic information associated with that file. In addition, if there is some specific information about the device itself (e.g. its name), that information will also be loaded at this point. As a result, the representation shown in Fig. 2 is obtained. Each level contains the following information:

- Level 1 (tasks): all actions that can be performed on the devices via voice
- Level 2 (devices): the types of devices that are related to each action
- Level 3 (instances): the instances or specific devices identified by their name

With this representation, each dialogue turn is translated into an in-depth search over the described tree to determine the type and content of the response according to the current state of the dialogue. Each type of response will be associated with a VoiceXML template (e.g. question, answer, confirmation, etc.) and also with other templates that contain the guidelines for generating the answer sentences.

The voice control module has two additional software components (see the left-hand section of Fig. 1). The first one includes a VoIP client, which is responsible for contacting the dialogue system through the VoiceXML platform. The second component implements the keyword spotting module and also has a local TTS system, whose function within the architecture is essentially limited to the voice conversion of the alerts that come from the home automation hardware. The keyword spotting technique detects a specific keyword and triggers an alert addressed to the VoIP client indicating the need to start a new connection with the dialogue system.

Focusing on keyword spotting approaches, and taking into account that there are several alternatives, some of the most common HMM-based approaches (Hidden

Markov Models) have been developed and evaluated. Three keywords (habitación -room-, apartamento -apartment- and casa -house-) were selected because of their relevance in the context of this work and their high frequency in the training corpus. The KWS approaches evaluated were the filler models based on triphones, filler models based on phonemes, a large vocabulary continuous speech recognition / LVCSR system, and a hybrid system in which the approach based on LVCSR and the approach based on triphones run in parallel.

3 Evaluation and Results

A completely functional prototype of the system was developed using a commercial VoiceXML platform and some specific tools (ATK and Flite) for the KWS module in order to evaluate the proposed design. The prototype was deployed in a laboratory equipped with a home automation network.

3.1 Evaluation Methodology

Testing was carried out in two phases. Firstly, various KWS approaches were evaluated using the Figure of Merit metric [14]. The performance of these techniques was compared and some of their configuration parameters were adjusted.

The second part of the test was the validation of the complete voice solution. Although the PARADISE methodology [15] is the most popular initiative in the evaluation of dialogue systems, there is still no single standard criterion for evaluating, comparing and predicting the performance and usability of such systems. Therefore, in the absence of a standard evaluation methodology and considering the possible limitations of PARADISE, such as excessive coupling between usability and user satisfaction [16] or a very limited predictive capacity [17], qualitative and quantitative measurements are addressed separately in this work. Later, the relations between both types of measurements are statistically analyzed.

Some of the most common quantitative measurements, also known as interaction parameters, are presented in [18][19][20]. After studying these parameters and the characteristics of the proposed system, the following parameters were considered:

1. Number of dialogues (ND).
2. Dialogue duration (DD). Average duration of a dialogue in seconds.
3. System-turn duration (STD) and user turn duration (UTD). Average duration of a system/user turn in seconds.
4. System response delay (SRD). Average delay of a system response in seconds.
5. User response delay (URD). Average delay of a user response in seconds.
6. Number of turns (NT). Average number of turns uttered in a dialogue.
7. Number of system turns (NST). Average number of system turns in a dialogue.
8. Number of user turns (NUT). Average number of user turns in a dialogue.
9. Number of helps (NH). Average number of user-help requests in a dialogue.
10. Number of ASR rejections (NAR). Average number of ASR rejections in a dialogue (the system was unable to "hear" or to "understand" the user).
11. Number of ASR errors (NAE). Average number of ASR errors (the system understood incorrectly the user prompt).

12. Task success (TS). Average number of tasks completed successfully.
13. Incomplete dialogues (ID). Average number of incomplete dialogues.
14. Number of KWS rejections (NKR). Average number of times the user has to repeat a keyword until the system detects it.
15. Voicexml response delay (VRD). Average time delay between the KWS system calling the VoiceXML platform and the user receiving the prompt response.

The subjective evaluation of spoken dialogue systems is usually done with user questionnaires. Some of the most common questionnaires in the evaluation of spoken dialogue systems are not specifically designed for voice, such as AttrackDiff [21] or SUMI [22]. However, users often have preconceived ideas of what the interaction should be, closely linked to the characteristics of conversation between humans. The naturalness, intuitiveness and ability are particularly important in the evaluation of spoken dialogue systems but are not covered in sufficient detail in the general questionnaires [23]. Some of the most common voice-specific questionnaires are: SASSI [23], SERVQUAL [24], the questionnaire for telephone services based in dialogue systems proposed in [17] or the questionnaire proposed in [18]. For the reason described above, the SASSI questionnaire was used in this test. In addition, questions relating to user antecedents and to general opinions about the overall system proposed in [17] were given to each user at the beginning and end of the test session.

3.2 Test Set-Up

To perform the KWS test, three men and three women reported a set of utterances, some of which contain keywords and other not. These utterances were recorded using a headset and a distant microphone placed 1.80 meters and 3 meters from the speaker.

Eleven men and six women between 23 and 40 years of age (mean=28) participated in the second part of the test. Users were cited individually in the testing room where they received a sheet containing brief instructions as well as a map of the testing room. The map showed the locations and the names of the voice-enabled devices and a summary of the supported actions.

During the tests users were required to control the status of the lights in the room. There were several lights with adjustable intensity and several lights with only on/off function. The option of requesting contextualized help or of cancelling the ongoing dialogue was also continuously available. The users filled out a SASSI questionnaire after completing each scenario. The following scenarios were used:

1. Turn on / off two lights, with or without adjustable intensity.
2. Set the intensity of two lights at a level between 0 and 100%.
3. Ask for help and exit the system.

The interaction parameters will be calculated for each dialogue and averaged later for the three scenarios described.

3.3 Results

In the first part of the test, the KWS technique that best suited the test conditions was the filler models method, based on triphones.

In the second part of the test, 169 dialogues (Turn on/off: 94, Regulate: 48, Help/Exit: 27) were obtained. Thus, the results commented in this section are based on the analysis of 169 samples of the interaction parameters, 54 SASSI questionnaires (3 per user) and 18 ITU-T questionnaires (1 per user).

Table 1 shows the results obtained for the questionnaires which in both cases have a 7-point Likert scale. Before analyzing their results, a value of 7 was assigned to the most positive category of the scale and a value of 1 to the most negative category. A value of 4 represents a neutral judgment.

SASSI results for each task are logical given the estimated level of complexity of the three tasks. The SASSI overall satisfaction has been calculated as the average of the values obtained for the three individual tasks, resulting in a low standard deviation. In general, per task or per system, the SASSI results show that users valued positively the questions about system response accuracy, likeability, habitability and speed categories. In addition, a neutral-low score in the most negative categories (cognitive demand and annoyance) can be interpreted as a positive user opinion about the system. The first question of the ITU-T questionnaire refers specifically to the overall impression or satisfaction, classifying it as extremely bad (value=1), bad, poor, fair, good, excellent or ideal (value=7). Most of the users responded that their overall impression was good (mean=4,78 and mode=5, with 80% of responses equal to 5 and 6), giving the worst score to the question related to the help expected from the system. The help provided by the system is an aspect to be improved in the future along with a study of the users' suggestions collected at the end of the ITU-T questionnaire.

Table 1. Summary of SASSI and ITU-T questionnaires results

		Tasks			**Overall Satisfaction**
		Turn on/off	Regulate intensity	Help / Exit	
SASSI Categories	System Response Accuracy	4,94 (σ =1,50)	4,60 (σ =1,77)	5,53 (σ =1,91)	5,02 (σ =0,47)
	Likeability	5,59 (σ =1,18)	5,36 (σ =1,29)	5,79 (σ =1,46)	5,58 (σ =0,22)
	Cognitive Demand	3,20 (σ =1,36)	3,30 (σ =1,37)	2,60 (σ =1.44)	3,03 (σ =0,38)
	Annoyance	3,70 (σ =1,76)	4,24 (σ =1,60)	3,02 (σ =1,76)	3,66 (σ =0,61)
	Habitability	4,31 (σ =1,57)	4,28 (σ =1,55)	4,58 (σ =1,87)	4,39 (σ =0,17)
	Speed	4,58 (σ =1,50)	5,06 (σ =1,24)	5,44 (σ =1,42)	5,03 (σ =0,43)
ITU-T (question 1)		----	----	----	4,78 (σ =1,48)

In the same way as the PARADISE philosophy, multiple linear regression (MLR) was used to determine the relationship between user satisfaction (dependent variable) and the values obtained for the interaction parameters (independent variables). Before doing the regression, the correlation between the parameters was studied. Thus, the parameters highly correlated (coefficient greater than 0.7) and less statistically significant (with greater p values) to the dependent variable were removed. In addition, the standard Z-score normalization function was applied to the dataset.

Table 2. Multivariate Linear Regression Models

Dependent Variable	Significant Predictors			Coefficient of determination R2
User Satisfaction (SASSI Likeability Score)	+ 0.798 * TS (sig 0,010)			0.401 (sig 0.024)
TS	- 0.657 * NAR (sig 1,949E-08)	- 0.601 * NAE (sig 2,0629E-06)	- 0.384 * ID (sig 0.0008)	0.847 (sig 1,0428E-11)

Table 2 shows the regression results. In the first model, user satisfaction was used as a dependent variable and was calculated as the sum of the responses corresponding to the Likeability category of the SASSI questionnaire. Only the percentage of tasks completed successfully (TS) proved to be a significant predictor, explaining 40% of the variance of the dependent variable.

User satisfaction was removed in the second regression model and task success was used as the dependent variable. The second regression model shows that the number of ASR rejections (NAR), the number of ASR errors (NAE) and the number of incomplete dialogues (ID) are, in this order, the parameters that have the most influence on the task success. In the second model, the significance and the coefficient of determination are higher than in the previous approximation.

Among the parameters that have the most direct influence on task success, and therefore indirectly on user satisfaction, only the number of incomplete dialogues (ID) is related to the combination of a VoiceXML platform with a KWS technique, the main peculiarity of the proposed solution. The number of incomplete dialogues alludes to communication errors between the VoIP client and the VoiceXML server and could probably be minimized by using a different VoiceXML platform.

The other parameters closely related to the type of solution proposed in this work were the VoiceXML response delay (VRD) and the number of KWS rejections (NKR). Although none of these parameters appeared as a predictor factor in the regression models, it is worth commenting the results obtained in both cases. One of the doubts cleared during the tests is whether the VRD remains constant during tests (mean=1,74, standard deviation=0,82, 90th percentile=1,76). The NKR mean was 1,73 and the NKR value was less than 2 in the 78,11% of every cases. Although this seems a positive result, it is true that the standard deviation in this case was 2,87 which justifies that many users complain about difficulty in accessing the system due to the high number of rejections.

4 Conclusions and Future Work

This work proposes a voice control solution for the devices installed in houses using main voice standards. It uses technology available in the market and facilitates the integration of new voice products such as speech recognition engines or VoiceXML platforms. The solution is independent of the devices to be controlled and it is also able to adapt dynamically to changes in these devices. The proposal has been implemented and then validated with user tests characterized by the control of real devices and the use of distant speech recognition.

Test results have shown the technical feasibility of the solution and promising user acceptance. The most important factor to maximize user satisfaction was the task success, which is mainly related with the number of ASR rejections, ASR errors and incomplete dialogues due to unexpected system errors. In addition, user answers to the last questionnaire report that the number of KWS rejections was high.

As future work, it is important to minimize the ASR rejections, ASR errors, unexpected system errors and the number of KWS rejections to improve user acceptance. In addition, it would be interesting to perform a second evaluation study by installing the solution in a real house and collecting data over an extended period of time. Finally it would be advisable to design new mechanisms to improve the dialogue and facilitate the correction of errors by users.

References

1. López-Cózar, R., Calleja, Z.: Multimodal dialogue for ambient intelligence and smart environments. In: Nakashima, H., Aghajan, H., Augusto, J.C. (eds.) Handbook of Ambient Intelligence and Smart Environments, pp. 559–579. Springer US, Boston (2010)
2. Florencio, E., Amores, G., Manchón, P., Pérez, G.: Aggregation in the In–Home domain. Procesamiento del Lenguaje Natural (40), 17–26 (2008) ISSN 1135-5948
3. Yates, A., Etzioni, O., Weld, D.: A reliable natural language interface to household appliances. In: Proc. 8th International Conference on Intelligent User Interfaces, pp. 189–196. ACM Press (2003)
4. Edlund, J., Gustafson, J., Heldner, M., Hjalmarsson, A.: Towards human-like spoken dialogue systems. Speech Communication 50(8-9), 630–645 (2008)
5. Montoro, G., Alamán, X., Haya, P.A.: Spoken interaction in intelligent environments: a working system. In: Ferscha, A., Hoertner, H., Kotsis, G. (eds.) Advances in Pervasive Computing, pp. 217–222. Austrian Computer Society, OCG (2004)
6. Dimopulos, T., Albayrak, S., Engelbrecht, K., Lehmann, G., Moller, S.: Enhancing the Flexibility of a Multimodal Smart Home Environment. In: Proc. Int. Conf. on Acoustics (DAGA), Stuttgart, Germany, pp. 639–640 (2007)
7. McTear, M.: Spoken dialogue technology: Enabling the conversational user interface. ACM Computing Surveys 34(1), 90–169 (2002)
8. Coelho, G.E., Serralheiro, A.J., Neto, J.: Microphone array front-end interface for home automation. In: Proc. Hands-free Speech Communication and Microphone Arrays (HSCMA), Trento, Italy, pp. 184–187 (2008)
9. Potamitis, I., Georgila, K., Fakotakis, N., Kokkinakis, G.: An integrated system for smart-home control of appliances based on remote speech interaction. In: Proc. 8th European Conference on Speech Communication and Technology, Geneva, Switzerland, pp. 2197–2200 (2003)
10. Möller, S., Krebber, J., Raake, A., Smeele, P., Rajman, M., Melichar, M., Pallotta, V., Tsakou, G., Kladis, B., Vovos, A., Hoonhout, J., Schuchardt, D., Fakotakis, N., Ganchev, T., Potamitis, I.: INSPIRE: Evaluation of a smart-home system for infotainment management and device control. In: Proc. 4th International Conference on Language Resources and Evaluation (LREC), Lisbon, pp. 1603–1606 (2004)
11. Gárate, A., Herrasti, N., López, A.: GENIO: an ambient intelligence application in home automation and entertainment environment. In: Proc. Joint Conference on Smart Objects and Ambient Intelligence, pp. 241–256 (2005)

12. Neto, J.P., Mamede, N.J., Cassaca, R., Oliveira, L.C.: The Development of a Multi-purpose Spoken Dialogue System. In: Proc. Eurospeech 2003, Genéve, Switzerland (2003)
13. Haya, P.A., Montoro, G., Alamán, X.: A Prototype of a Context-Based Architecture for Intelligent Home Environments. In: Meersman, R. (ed.) OTM 2004. LNCS, vol. 3290, pp. 477–491. Springer, Heidelberg (2004)
14. Rohlicek, J.R., Russell, W., Roukos, S., Gish, H.: Continuous hidden Markov modeling for speaker-independent word spotting. In: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 627–630 (1989)
15. Walker, M.A., Litman, D., Kamm, C., Abella, A.: PARADISE: a general framework for evaluating spoken dialogue agents. In: Proc. of the 35th Annual General Meeting of the Association for Computational Linguistics, ACL/EACL, pp. 271–280. ACL, Madrid (1997)
16. Dybkjaer, L., Bernsen, N., Minker, W.: Evaluation and usability of multimodal spoken language dialogue systems. Speech Communication 43, 33–54 (2004)
17. ITU-T Rec. P.851, Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems. International Telecommunication Union, Geneva (2003)
18. Walker, M.A., Kamm, C.A., Litman, D.J.: Towards developing general models of usability with PARADISE. Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems (2000)
19. Möller, S., Smeele, P., Boland, H., Krebber, J.: Evaluating Spoken Dialog Systems According to De-facto Standards: A Case Study. Computer Speech and Language 21(1), 26–53 (2007)
20. ITU P series Rec, Parameters Describing the Interaction with Spoken Dialogue Systems, ITU, Geneva (2005)
21. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Ziegler, J., Szwillus, G. (eds.) Mensch & Computer 2003. Interaktion in Bewegung, pp. 187–196 (2003)
22. Kirakowski, J., Corbett, M.: SUMI: The software usability measurement inventory. British Journal of Educational Technology 24(3), 210–212 (1993)
23. Hone, K., Graham, R.: Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). Natural Language Engineering 6(3-4), 287–303 (2000)
24. Hartikainen, M., Salonen, E., Turunen, M.: Subjective evaluation of spoken dialogue systems using SERQUAL method. In: Proc. of International Conference on Spoken Language Processing (ICSLP) (2004)