LOGOS: A Multimodal Dialogue System for Controlling Smart Appliances

Theodoros Kostoulas, Iosif Mporas, Todor Ganchev, Nikos Katsaounos, Alexandros Lazaridis, Stavros Ntalampiras, Nikos Fakotakis

Artificial Intelligence Group, Wire Communications Laboratory
Dept. of Electrical and Computer Engineering,
University of Patras, 26500 Rion-Patras, Greece
{tkost, imporas, tganchev, nkatsaounos, alaza, dallas, fakotaki}@wcl.ee.upatras.gr

Abstract. The present work details a multimodal dialogue system, which offers user-friendly access to information, entertainment devices and white good appliances. We focus on the speech interface and the spoken dialogue management, with extensive description of the system's architecture and functionalities. The services supported are detailed, with comprehensive description of the scenarios implemented.

Keywords: Smart Home, Dialogue, Spoken Dialogue System, Multimodal Dialogue System, Speech Interaction.

1 Introduction

The increasing use of multimodal dialogue systems along with the progress of technology raises the need for user-friendly human-machine interaction. Intelligent interfaces of home appliances provide the means for facilitating the operation of these devices, within a dialogue system. Combining speech, which is the most natural communication mean between human beings, with other user inputs, like mouse or keyboard, would ensure successful interaction experiences.

To this end, a number of multimodal dialogue systems have been reported. In [1], Lu et al developed a dialogue system consisting of five main modules: sign-language recognition and synthesis, voice recognition and synthesis and dialogue control-management. A pair of colored gloves and a stereo camera had been used in order to track the movements of the signer. Sign-language synthesis is achieved by regenerating the motion data obtained by an optical motion capture system. In [2], MiPad is presented, a personal digital assistant, which gives to the users the ability to accomplish many common tasks, by using a multimodal interface and wireless technology. This prototype has been based on plan-based dialogue management [3]. Hemsen, [4], discussed design decisions for the combination of speech with other modalities, towards designing a multimodal dialogue system for mobile phones. Furui and Yamaguchi, [5], introduced a paradigm for designing multimodal dialogue systems, through designing and evaluating a variety of dialogue strategies. The system presented accepted speech and screen touching as input and presents retrieved

information on a screen display. In [6] an architecture that combines finite-state multimodal language processing, a speech-act based multimodal dialogue manager, dynamic multimodal output generation and user-tailored text planning, is described. The application provides access to restaurants and subway information.

In the present work, a multimodal dialogue system that offers user-friendly access to information and intelligent appliances is reported. This paper is organized as follows: Section 2 describes the system's general architecture. In section 3 the system's components and their function are detailed. Section 4 refers to the services supported by the LOGOS system, together with the scenarios implemented.

2 System Architecture

The LOGOS smart-home automation system offers user-friendly access to information, entertainment devices and provides the means for controlling intelligent appliances installed in a smart-home environment. Fig. 1 illustrates the overall architecture of the LOGOS system, and the interface to various appliances, such as high definition television (HDTV), DVD player, mobile phone, washing machine, refrigerator, etc. The multimodal user interface of that system allows home devices and appliances to be controlled via the usual infrared remote control device, PC keyboard, or spoken language. In the present study, we focus on the speech interface and the spoken dialogue interaction.

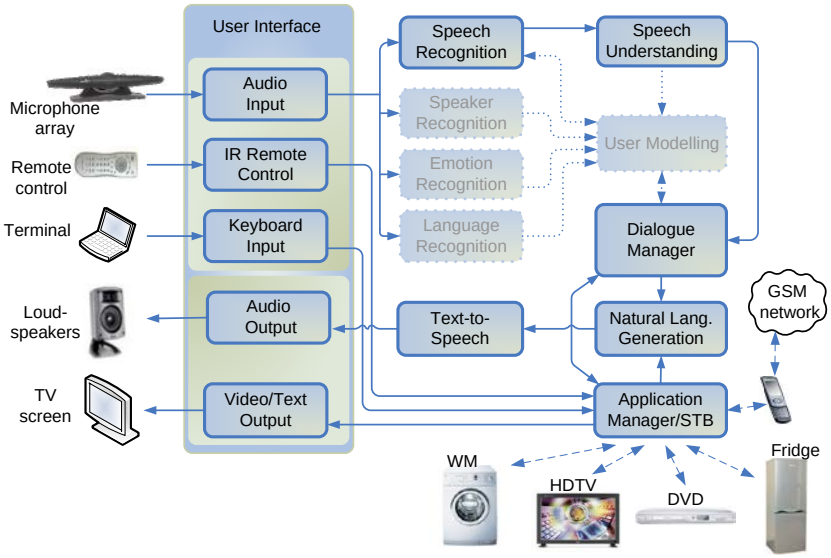

Fig. 1. Architecture of the LOGOS system.

In detail, the spoken dialogue interaction functions as follows: The speech is captured by a microphone array, which tracks the position of the user as s/he moves. Next, the preprocessed speech signal is fed to the speech recognition component, which utilizes domain-specific language models. The speech understanding component provides the interpretation of the output of the speech recognition component in terms of concept-command. The dialogue manager generates feedback to the user, according to the commands received from the speech understanding component, the device status and other information. This feedback is delivered via the text-to-speech component, which is fed by the natural language generation component. All components of the spoken dialogue system are deployed on a personal computer, and the communication with the intelligent appliances is performed through a set-top box (STB), which provides the interface to the intelligent appliances. The STB component is responsible for synchronization and efficient data flow among the dialogue manager and the smart-home appliances. Moreover, it provides additional visual feedback to the user, which can, optionally, accompany the audio output.

The smart-home system is designed with open architecture, which allows new components to be added, for extending the system's functionality. Such additions could be user modeling, speaker recognition, emotion recognition and language identification components (in Fig. 1 these are shown with dashed-line). These components could contribute to significant improvement of the user-friendliness of the system, as well as of its overall performance. For instance, by utilizing feedback from the emotion recognition component the system will be able to select an emotion-specific acoustic model, for improving the speech recognition performance, or to steer the dialogue flow accordingly, when negative emotions are detected. The speaker recognition component would provide the means for implementing access authorization for specific commands, but also for intelligent user profiling or using of user-specific acoustic models, etc. The language recognition component would enable multilingual access to the LOGOS system, i.e. using the acoustic and language models which correspond to the user's spoken language. The information carried out from the aforementioned components feed the user modeling component. User modeling is responsible for handling all user-related data. Thus, user's preferences, user specific settings and interaction schemes can be specified. Therefore, the system is able to adapt to different user requirements.

3 System Components

The present session describes in detail the aforementioned system components, which constitute the LOGOS system architecture. Each one of these components is independent to the rest, though interconnected, based on the system architecture.

3.1 User Interface

The user interface components are responsible for realizing the communication between the human and the machine. They consist of the acoustic input-output and the device input-output.

Acoustic Input. To facilitate acoustic input processing, speech enhancement methodologies are employed, so as to remove background additive noise. Most of the work has been focused on the improvement of synchronous communication devices, which operate in noisy environments [7]. The particularity of speech signal makes one able to estimate the noise part alone, during speech pauses, a property which is exploited by almost every algorithm. In brief, noise reduction methodologies are divided into the following classes: (i) Subspace algorithms, (ii) Spectral Subtractive algorithms, (iii) Wiener filter-based algorithms and (iv) Statistical model based algorithms.

Receiver's sensitivity can be reduced in the space of interference and noise, while concentrated on the desired signals (spatial filtering), by utilizing beamforming algorithms. Subsequently, noise suppressing methods can be used to achieve effective denoising. There are two categories of beamforming techniques, adaptive, which change their parameters based on received signals, and fixed, which keep them unchanged. A time delay in the time domain represents a negative phase shift in the frequency domain. This fact is utilized by a frequently used beamforming technique called delay-sum beamforming [8], where time delays are applied to the sensors for efficient alternation of the area of interest. This method is fixed, while generalized sidelobe canceller (GSC) [9], belongs to the adaptive methods. It encompasses two stages, in which one depends on the other. A fixed beamformer is employed, firstly, while an adaptive part is constantly filtering the signal, so as to assure noise elimination. In this particular task, noise is thought to be every sound which source is not placed in our area of interest.

In order to overcome a number of different limitations deriving from a realistic room environment (such as generic speaking scenario and room layout, multiple moving speakers and low computational cost) we utilized a commercial microphone array with eight sensors. The device, called Voice Tracker^{TM}, achieves high signal-to-noise ratio (SNR) values by concentrating on a specific talker and filtering out background noise and reverberations in the acoustic environment [10].

Acoustic Output. The actions taken by the system are given to the user as the output of two conventional loudspeakers which are set up outside the range of the microphone array to avoid any interference. They are placed in a symmetrical way, according to the room boundaries, and we tried to keep acoustic echo level as low as possible.

Device Input-Output. Additional input from the user is retrieved from the IR remote control and the keyboard. These data are feeding the application manager / STB component, which can either directly execute the retrieved command, or pass this command to the dialogue manager. In the later case, the dialogue manager will handle

the request by modifying the dialogue flow. Device output consists of the TV screen where notifications are displayed.

3.2 Speech Recognition Component

The Speech Recognition Component, as well as the Speech Understanding Component are based on the ©ScanSoft SpeechPearl [11] speech recognizer. The speech recognition process can be divided into two phases, namely preprocessing and decoding.

In the preprocessing phase, the speech input is recorded and digitized. Afterwards, an acoustic analysis is performed. Specifically, using the fast Fourier transform (FFT), the front end processor breaks down the recorded audio sequence into small units of frames, where each frame covers some milliseconds of speech. For each frame, a feature vector is created. Each feature vector contains frequency and energy information. The frames overlap each other, by using a sliding window. The feature vectors are the input of the decoding process.

In the decoding phase, a word graph is built on the basis of an acoustic model, an application lexicon and a language model. The acoustic model is based on the widely used hidden Markov models (HMMs) and consists of one HMM model for each uniphone and triphone of the language. Furthermore, HMM models for whole words such as digits, yes/no and natural numbers exist. Using the Viterbi algorithm, the feature vector sequence is compared against the HMM models. The Viterbi algorithm searches for the most probable of the word observations in the application lexicon. The lexicon is organized in a tree and contains only the words that are part of the related task grammar. In addition to the acoustic score, a language model score is counted in. Language model consists of word bigrams, concept bigrams and rule probabilities. The Speech Recognition Component combines the acoustic and language scores, to construct a word graph. The word graph contains nodes connected by edges, which are labeled with the recognized words. This structure represents a set of the most probable spoken word sequences. Finally, the *n* most probable paths through the word graph are sorted to an n-best list in order to be further used from the Speech Understanding Component.

3.3 Speech Understanding Component

When the speech recognition process is finished, the resulting word graph is passed to the speech understanding component, for further processing. The speech understanding component searches for meaningful word sequences, by parsing the incoming word graph and using the related grammar to find matching concepts. A concept is a non-terminal that defines a certain word or word sequence, forming a sense unit within a sentence. Concepts are computed according to the rules defined in the corresponding grammar. The defined grammars can have multiple concepts, which may occur in arbitrary order. Any parts of the sentence that are not covered by a concept are taken as so-called fillers. Fillers are not taken into account, when processing the meaning of the sentence.

From the found concepts and the remaining fillers the speech understanding component creates an acyclic directed graph called concept graph. The concept graph represents the most probable spoken sentences, which were found by the speech understanding module. Fillers are taken as meaningless gaps in the concept graph. The optimal path through the concept graph is determined by internal scores of the speech understanding module. The nodes of the concept graph are identical to those of the word graph. Its edges are not labeled with words, but with concepts. Concepts will not always be adjacent to one another. In any case, finding an instance of a particular concept does not necessarily mean that the corresponding words were actually said, i.e. their inclusion in the word graph may be a result of recognition errors. The sentence alternatives according to the concepts from the concept graph are presented in a n-best list. This list is sorted by internal scores and starts with the best matching sentence alternative.

In addition, semantic information can be included in a grammar. In this way, it is possible to interpret the recognition results on a high level of performance. Concepts can define attributes, which contain the semantic information. Then, the grammar contains information on how to compute attribute-value pairs. The understanding result provides the computed attribute-value pairs. This enormously improves the interpretation of a recognition result.

3.4 Natural Language Generation Component

Natural Language Generation (NLG) is the research area investigating the ability of the machine to produce high-quality natural language text from a machine representation system, such as a knowledge base or a logical form. There are four major categories in NLG concerning the degree of the complexity and the flexibility of each method [12]: canned text method (simply prints a sequence of words from existing list), template-based method (use of pre-defined templates or scenarios), phrase-based method (using generalized templates) and feature-based method (each expression is represented by a feature – combining features to create a sentence) [13]. In the present work we inherit the template-based method, once it suits to the needs of the provided services.

3.5 Text To Speech Component

Text to Speech (TtS) conversion is the process of converting a string into spoken language. The TtS of the LOGOS system is a diphone based residual excited linear predictive coding (LPC) synthesis technique based on the Festival Speech Synthesis system [14]. A diphone is a unit of speech starting in the middle of one phone till the middle of the next one.

The first phase of the realization of the Greek TtS component is to parameterize all the diphones. Each diphone is divided into pitch synchronous frames and, for each frame, the LPC coefficients and the residual signal are extracted in order to be used during the synthesis stage.

The second stage is interrelated with the prosody of the synthesized speech. In order to produce as natural synthetic speech as possible, emphasis must be given to the duration model of the system. The duration model is coping with the task of determining the length of phones in speech, considering various levels of signal representation [15]. The classification and regression tree (CART) [16] machine learning method is used for the duration modeling. The feature set for training the model is composed of features which can be extracted only from text such as phonological, morphological, linguistic and syntactical features. For some features, a window around the investigated phone was applied, exploiting the knowledge concerning the characteristics of the neighboring phones [16].

3.6 Application Manager / STB

The application manager / STB component is utilizing the efficient synchronization of the data among the dialogue manager and the smart-home appliances. This component keeps track of the current state of each device, providing this information to the dialogue manager, when such a request arises. Moreover, the component provides visual feedback to the user, realizing the graphical user interface. Additionally, the component is fed from the dialogue manager, for parsing the suitable commands to the connected intelligent appliances.

3.7 Dialogue Manager

The dialogue manager component determines the tasks that should be activated, taking into consideration the dialogue history and the current user input. At each stage of the dialogue flow, the dialogue manager provides to the speech recognition the necessary information for loading the corresponding language models, in order to perceive high recognition accuracy and reliability. This language model is chosen with respect to the expected action-request of the user. The model relies on a text corpus built from all the possible ways the user can use to express her/his request. The dialogue manager retrieves information concerning the current status of all the devices from the application manager/STB. Moreover, it feeds the application manager/STB with the appropriate commands, so as to control the connected devices, correspondingly to the user's request. The feedback to the user is adapted to the orientation of the current task, i.e. when needed, audio and video/text outputs coexist.

As the human-machine interaction progresses, within a session, the dialogue manager builds a tree, whose nodes constitute the dialogue history. When needed, based on the scenario implemented, the user can abort the procedure, or change the dialogue flow. The implemented dialogue relies on the mixed-initiative model. According to the mixed-initiative dialogue model, the dialogue steps / actions can be designed in a more flexible and dynamic way. Specifically, the system orders or wait the user to make her/his request. The system decides its next step dynamically, according to the semantic information extracted from the user's utterance, by the speech understanding module. The mixed-initiative dialogue model offers the ability

of indirect confirmation, i.e. the dialogue corrects itself by using user's requests, without prompting the user, explicitly, to clarify the last request.

4 Dialogue Flow Scenarios

The LOGOS system supports two voice-controlled services: (a) Home devices control and (b) SMS Messaging. In the home devices control service, the user is able to control entertainment devices and white good appliances. The first scenario implemented provides access to the TV Tuner via voice. The home user is interacting with the dialogue system, issuing commands, for checking which one of the available TV channels has a movie starting within the next 1 hour. The LOGOS system asks the user if s/he prefers a foreign or Greek movie. The user makes a decision and the assistant informs her/him about the available movies, so that the user can choose the one s/he wishes to watch. Then, the system informs the user for the exact time the movie starts. The home user may decide to switch to the specific channel and can increase/decrease the volume of the TV.

The second scenario provides stored video control. The home user subsequently issues commands to check the available stored movies. The system asks the user if s/he prefers a Greek or non-Greek movie. The user makes a selection and the system informs her/him about the available movies. The user can decide to watch one movie, so he issues the relevant command to start watching the movie. Afterwards, s/he can control the volume of the projection.

In the third scenario, the user has the ability to control and monitor the white good appliances connected to the application manager / STB component. The user can issue commands to retrieve a list of all the connected devices. The system informs the user, which are the available devices. Then, the user can issue commands to get informed about the current status of each device. Moreover, the user can control a selected device. An example dialogue of the aforementioned service is shown in Table 1.

Table 1. Example dialogue of the white good appliances control.

Turn	Dialogue
1. [User]	Assistant, please provide me the list of the available devices
2. [Assistant]	The available devices are "Fridge1" and "WashingMachine1".
3. [User]	Assistant, which is the current status of "WashingMachine1"?
4. [Assistant]	The current status of "WashingMachine1" is "Wash" or "Ready".
5. [User]	Assistant, please "Start" or "Stop" the "WashingMachine1".
The new status of "WashingMachine1" is displayed on the TV screen for verification.	
6. [User]	Assistant, which is the current temperature of "Fridge1"?
7. [Assistant]	"Fridge1" current temperature is "x.x" Celsius.
8. [User]	Assistant, please decrease "Fridge1" temperature.
The application is programmed to decrease the temperature by 1 degree of Celsius (default) at every call.	
9. [User]	Assistant, please start monitoring "Fridge1" temperature.
No reply from Assistant is needed since the outcome will be displayed on the TV screen. When a change on the temperature of the fridge is taking place a notification on the screen will be displayed.	

In the SMS messaging service the home user may read, compile and send an SMS. In the scenario implemented the user decides to send an SMS to a target user in order to inform her/him for the movie that s/he prefer to watch and ask the target user if s/he wishes to join. During the SMS compilation the user is prompted to utter the SMS. After the SMS compilation, the home assistant reads to the user the content of the SMS. Afterwards, the home user is prompted to provide the name or the phone number of the target user (if the name of the target user is not listed in the list of the known SMS recipients). After successful compilation, the home user prompts the home assistant to send the SMS to the target user.

On successful reception, the target user replies back to the SMS with a confirmation that s/he will join the home user in order to watch the movie. The LOGOS multimodal dialogue system informs the home user via audio that a new SMS is received from the target user and, also, by displaying a text notification on the TV display. When the home user observes this notification s/he commands the home assistant to read/display and then delete the received SMS from the LOGOS system inbox. Table 2 illustrates an example dialogue for compiling SMS.

Table 2. Example dialogue for compiling SMS.

Turn	Dialogue
1. [User]	Assistant, please compile an SMS.
2. [Assistant]	SMS Body?
3. [User]	"yyyy" will start at <hh:mm>. Do you wish to join?
4. [Assistant]	Name?
5. [User] *	<Target User>.
6. [Assistant]	Unknown recipient. Please, provide phone number.
7. [User] **	697xxxxxxx.
8. [Assistant]	<Target User>, 697xxxxxxx. Do you confirm?
9. [User] *,**	Yes.
10. [User] *	Send SMS.
11. [Assistant]	SMS has been successfully sent.
Alternative Dialogues:	
*[User] *: Cancel SMS.*	
Then, the user needs to issue the compile SMS command in order to start again.	
*[User] **: No.*	
Then, the dialogue goes to turn 4.	

5 Conclusion

In the present work a multimodal dialogue system that provides the means for user-friendly access to information and smart appliances is presented. The smart-home system is designed with open architecture, thus allows the integration of additional components (e.g. speaker, language, emotion recognition components), for extending the system's functionalities. We deem that these components would contribute towards enhancing the system's performance and lead to more successful interaction experiences.

Acknowledgments. This work was partially supported by the LOGOS project (*A general architecture for speech recognition and user friendly dialogue interaction for advanced commercial applications*), which is funded by the General Secretariat for Research and Technology of the Greek Ministry of Development.

References

1. Lu, S., Igi, S., Matsuo, H., Nagashima, Y.: Towards a dialogue system based on recognition and synthesis of Japanese sign language. Gesture and Sign Language in Human-Computer Interaction, 259--271 (2006)
2. Huang, X., Acero, A., Chelba, C., Deng, L., Duchene, D., Goodman, J., Hon, H.W., Jacoby, D., Jiang, L., Loynd, R., Mahajan, M., Mau, P., Meredith, S., Mughal, S., Neto, S., Plumpe, M., Wand, K., Wang, Y.: MIPAD: A next generation PDA prototype. Proc. ICSLP 2000, Beijing, China, Oct., vol. 3, 33--36 (2000)
3. Wang, K.: A Plan-Based Dialog System with Probabilistic Inferences. Proc. ICSLP 2000. Beijing, China, vol. 2, 644--647 (2000)
4. Hemsen, H.: Designing a multimodal dialogue system for mobile phones. Proc. of the first Nordic Symposium on Multimodal Communications, Copenhagen, Denmark, 25--26 (2003)
5. Furui, S., Yamaguchi, K.: Designing a multimodal dialogue system for information retrieval. In Proc. ICSLP 1998, vol. 4, 1191--1194
6. Johnston, M., Bangalore, S., Vasireddy, G., Stent, A., Ehlen, P., Walker, M., Whittaker, S., Maloor, P.: MATCH: An architecture for multimodal dialogue systems. In Proc. Annu. Meeting of the Association for Computational Linguistics (2002)
7. Ntalampiras, S.: MoveOn Deliverable D5.10.1: Overview on Speech Enhancement Algorithms. (2007)
8. McCowan. I.A.: Robust Speech Recognition using Microphone Arrays. PhD thesis, Queensland University of Technology, Australia (2001)
9. Griffiths L., Jim, C.: An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. On Antennas and Propagation, vol. 30(1), 27--34 (1982)
10. Acoustic Magic Microphone Array, http://www.acousticmagic.com/
11. ScanSoft Speech Recognizer, http://www.scansoft.com
12. Konrad, K.: Model generation for natural language interpretation and analysis. Berlin London, Springer (2004)
13. Cao, W., Zong, C., Xu, B.: Approach to interchange-format based Chinese generation. In Proc. Interspeech 2004, 1893--1896 (2004)
14. Black, A., Taylor P.: The Festival Speech Synthesis System: System Documentation. Technical Report HCRC/TR-83, Human Communication Research Centre, University of Edinburgh, Scotland, UK, http://www.cstr.ed.ac.uk/projects/festival.html
15. Lazaridis, A., Zervas, P., Kokkinakis, G.: Segmental duration modeling for Greek Speech Synthesis. In Proc. of the 19th IEEE International Conference on Tools with Artificial Intelligence, Patras, Greece, 518--521 (2007)
16. Chung, H., Huckvale, M.A.: Linguistic factors affecting timing in Korean with application to speech synthesis": In Eurospeech 2001, Denmark, vol. 2, 815--818 (2001)