A Dialogue System for Telephone-based Services Integrating Spoken and Written Language

K. Georgila(1), A. Tsopanoglou(2), N. Fakotakis(1), G. Kokkinakis(1)

(1) Wire Communications Laboratory, Electrical and Computer Engineering Dept.,
University of Patras, 261 10 Rion, Patras, Greece
Tel:+30 61 991722, Fax:+30 61 991855, e-mail: {rgeorgil, fakotaki, gkokkin}@wcl.ee.upatras.gr

(2) KNOWLEDGE S. A., Human Machine Communication Dept.,
N. E. O. Patron-Athinon 37, 264 41 Patras, Greece
Tel:+30 61 452820, Fax:+30 61 453819, e-mail: atsopano@knowledge.gr

ABSTRACT

In this paper, we describe a Dialogue System for Telephone-based Services that integrates spoken (Dialogue Component) and written language (Optical Character Recognition-OCR). This system has been developed for a car insurance company in the framework of the LE-1 1802 project ACCeSS, but can be easily adapted to a different task. Handwritten application forms from prospective clients, concerning new contracts, are processed by the OCR component and the extracted information is stored on a local customers' database. The Dialogue Component reads the customers' database and collects the missing information of the application forms by calling the corresponding customer and asking about the blank or ambiguous fields. The system is currently tested at the company's site.

I. INTRODUCTION

During the last decade significant progress has been made in speech recognition and speech understanding as well as in the speech synthesis domain. This has encouraged the creation of Spoken Dialogue Systems that constitute the most effective and user friendly way for human-computer interaction by voice [1]. In particular, major efforts have been made in developing Spoken Dialogue Systems for real life applications. Although the current technology can support several applications, significant efforts are still necessary to develop systems capable to be widely introduced in the public and private sector.

In addition, significant achievements have been scored in the field of handwritten Optical Character Recognition (OCR) [2]. Thus, integration of speech and OCR technology in useful applications has been made possible and very promising.

In this paper we present a system developed in the framework of the LE-1 1802 project ACCeSS [3] for a car insurance company that combines spoken and written language. It consists of an OCR and a Dialogue Component. The OCR processes handwritten application forms filled out by potential clients and feeds a local database with the extracted information. The Dialogue Component aims at collecting the missing information of the application forms by initiating calls to the customers.

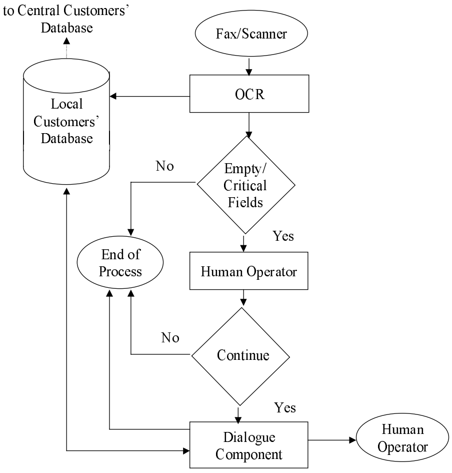

Figure 1 - System's function.

The steps of the procedure, shown in Figure 1, are: The system "reads" the handwritten application forms, which have been entered either by fax or scanning. The extracted optical information by the OCR component is stored on the local customers' database after confirmation by a human operator. In case that data are missing, the Dialogue Component initiates a call to collect them provided the human operator has given his/her permission. The system reads the stored information on the local database and prompts the customer to give the necessary answers to complete the blank fields. After the completion of this procedure the blank fields of the customers' records are filled. In case that the dialogue cannot be completed successfully, the human operator undertakes to resume the conversation with the user. Finally, a special module takes the responsibility to update the company's database with

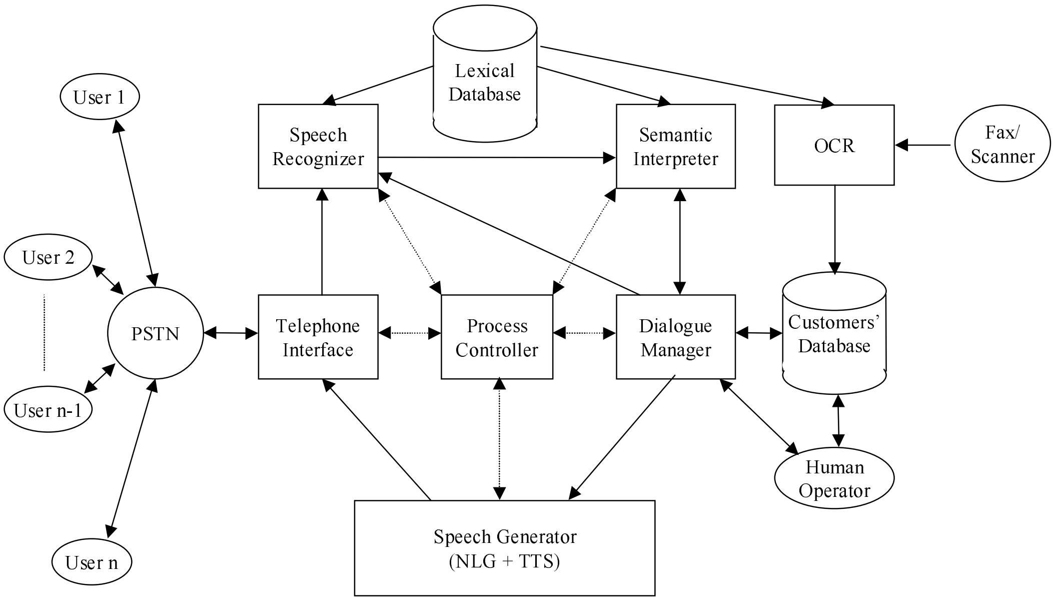

Figure 2 – System's structure.

the collected information. In this way, the company's central database is protected from being contaminated by erroneous entries.

The paper is organized as follows: The system architecture is given within Section II. Section III describes the hardware and software specifications of the system. Evaluation experiments are given in section IV. Finally, a short summary and some conclusions are given in section V.

II. SYSTEM ARCHITECTURE

The system consists of two functional components: The OCR and the Dialogue. The Dialogue Component includes the following modules: The Speech Recognizer, the Semantic Interpreter, the Dialogue Manager, the Speech Generator, the Telephone Interface and the Process Controller. There are also a lexical database accessed by the Speech Recognizer, the Semantic Interpreter, the Speech Generator and the OCR component and a customers' database, which connects the OCR component with the Dialogue. The structure of the system is shown in Figure 2.

A. Speech Recognizer

The speech recognition module is based on continuous Hidden Markov Models (HMMs). Each basic recognition unit (phoneme or syllable) is described by a 5-state left to right HMM the parameters of which, have been defined during the off-line segmental k-means training procedure. Syllables i.e. two-phone, three-phone, four-phone and five-phone combinations, are formed by two or more consonants always ending in a vowel. The entire set of the used units consists of 808 units that are:

- 35 phonemes
- 108 two-phone combinations (diphones)
- 419 three-phone combinations (triphones)
- 238 four-phone combinations and
- 8 five-phone combinations

Experimental tests have proven that these combinations give the best recognition results for Greek, which is a syllable-based language [4].

During the recognition phase two lexicons are activated at each dialogue stage, one common for all the nodes and another depending on the type of input expected. Transition probabilities between the models and bigram probabilities between the words have been estimated by using large speech and text corpora. During the on-line recognition procedure, the Frame Synchronous Viterbi Beam Search algorithm is used in order to produce the most probable sequence of basic recognition units taking into consideration the transition probabilities between them [5]. Consequently word sequences are formed using the current lexicons and the estimated bigram probabilities. In this way, the recognizer finds the N-best hypotheses and passes them to the Semantic Interpreter with their corresponding scores.

B. Semantic Interpreter

The linguistic analysis is based on Island Parsing, Pattern Matching and Frame-based Representation techniques. The main knowledge sources are a Semantic Network and Frame-slot structures connected with each other. Simple bigram grammar rules are also used to assist the parsing process as well as to evaluate the recognition output [6].

Expectations that emerge from the current state of the dialogue model are used. These Dialogue Expectations are applied very early in the linguistic analysis so that pragmatic-irrelevant solutions are not considered in the first place. The results are structured appropriately and sent to the Dialogue Manager.

C. Dialogue Manager

The Dialogue Manager constitutes the heart of the Spoken Dialogue System and is responsible for controlling the human computer interaction by following predefined dialogue scenarios. These scenarios have been modeled off-line in a graphical notation by using Dialogue Flow Graphs (DFGs) that have been translated into a node structure by using the Dialogue Structure Editor (DSE), a generic, graphical and user friendly tool. The output of the DSE tool is an intermediate format that is read directly and executed by the Dialogue Manager.

The Dialogue Manager takes the results of the Semantic Interpreter as input and forms a node-dependent structure that contains all the useful information supplied by the customer (i.e. the function name that must be executed at the current node, the variable names that hold the parameters given by the Semantic Interpreter etc). Following, it queries/updates the local database, or prompts the user to give extra information or confirm the validity of the data provided at the previous node. It notifies the Speech Recognizer and the Semantic Interpreter on the dialogue expectations and the Speech Generator about the system messages.

D. Speech Generator

The speech generator combines a Natural Language Generator (NLG) and a Text-To-Speech Synthesizer (TTS). The NLG constructs the system's response according to the dialogue design. It generates a natural language message, which is then converted into speech.

The speech output of the system is a combination of prerecorded messages and synthesized speech. Prerecorded phrases are used for standard responses, speech synthesis for names, numbers etc., that cannot be prerecorded. In this way, the system responds fast and is quite intelligible.

In case that prerecorded phrases do not exist for a particular part of the sentence, the module can replace the missing part with synthesized speech.

E. Telephone Interface

This module controls the special hardware, which is responsible for handling the various PSTN calls. It directs the Initiator (sub-module of the Process Controller) to activate the various modules for the service of the incoming call and the Speech Recognizer and Speech Generator modules, to get input and send output respectively to the specific network line.

F. Process Controller

The initiation of different instances of the system's modules in order to service multiple calls as well as to handle other system resources is undertaken by the Process Controller. This module acts as a server, which initiates instances of the Dialogue Component modules, gives an id number to each connection and notify the functional modules about the id number of the call they have been assigned.

The use of this identifier is necessary for the establishment of a single communication channel between the initiated modules, when more than one calls must be handled simultaneously.

In cases where the input supplied by the user cannot be understood by the system even after clarification and/or confirmation questions, the Process Controller after notifying the Dialogue Manager directs the call to a Human Operator who undertakes to resume the conversation with the user and recover possible errors, taking into consideration the history of the dialogue transaction and the already collected information.

G. Optical Character Recognition (OCR)

This module converts a hand-written application form into a set of information, which is understandable by the system. If the application form comes from an existing customer, it updates only the necessary information in the customer's record, and in case some critical information is missing, it updates an appropriate table. This is checked by the Dialogue Controller, which initiates a telephone call to the customer.

A pre-processor submodule has the task to refit the input form and to extract the content of the slots of the inserted page. In order to do that, it compares the inserted document with the prototype form and refits the input page. Afterwards the module undertakes to extract the text of each slot of the refitted page and to store it at the slots database, in binary matrices (1s and 0s).

From the database the matrices are read and handled as strings of characters. After removing the noise, the recognizer divides the strings into individual characters. Each character is recognized separately and the output of the sub-module is a sequence of characters.

The recognized characters and the information about the corresponding field are used by the OCR component to activate the appropriate lexicon(s) (e.g. if the field corresponds to the license number, it activates the number lexicon). Then, by using this lexicon as input, and implementing the level building algorithm (dynamic programming), it finds the most probable word(s) corresponding to the recognized characters and updates the appropriate field in the customers' database [7].

H. Databases

The integrated system uses three databases: One containing the lexicons activated at each dialogue node by the system's modules, one local database containing personal information about the customers and the company's central database.

The lexical database consists of two correlated tables, one containing special fields necessary for the correct function of each module e.g., orthographic and phonetic transcriptions of words for the Speech Recognizer, syntactic and semantic information for the Semantic Interpreter, filenames that point to prerecorded messages for the Speech Generator and a second table filled with node specific information and constraints. The fields of the lexical database are independent of the application domain. By having one database for all the modules it is ensured that all the subparts of the system share the same lexicons.

The customers' database contains application dependent fields. In our case, for the insurance company's task these fields are the customer's telephone number, date of birth, driving license number, identity card number, city, postal code, etc. The customers' database is updated by the Dialogue Manager according to the queries created by the information taken from the semantic interpreter. The Dialogue Manager provides a generic interface for any kind of SQL server commands embedded in C.

The company's database includes the final and correct data of the clients. A special module has the responsibility to update this database with the collected information.

III. HARDWARE AND SOFTWARE SPECIFICATIONS

The Dialogue Component runs on a Pentium PC using an industry-standard telephone line interface card (Dialogic D41ESC). All the modules of the Dialogue Component are developed in C/C++ apart from the Semantic Interpreter, which is implemented in Prolog. The OCR component is implemented in C/C++ and runs on a separate Pentium PC. The two PCs access the same local customers' database.

IV. EVALUATION EXPERIMENTS

The speech recognizer has been trained with speech from the telephone recordings of the LE-2 4001 project SPEECHDAT-II [8], using 1000 male and female speakers, which cover a wide range of speech variation. The OCR component has been trained and tested using a large amount of completed handwritten forms (more that 2000 forms) supplied by the insurance company.

The integrated system has been recently installed at the premises of a car insurance company and is tested by personnel of this company. The results are regularly communicated to the developers and serve as basis for improvements. After this stage which will last for six months, the system will be given to customers' service. A formal evaluation lasting for three months will then be carried out.

One of the main problems is that the system initiates the call to the customer. In this case it is very difficult for the system to find out whether it is contacting the appropriate person or not. Another problem is that users tend to give more information than needed at a specific node, which causes additional problems to the speech recognition and semantic interpretation modules. For example when the system asks for the postal code many users give the city too, which complicates the task of the speech recognizer and the semantic interpreter that have activated lexicons containing numbers but not cities. The above phenomenon has led to the expansion of the activated lexicons at each node, in order to handle unexpected user utterances.

V. SUMMARY AND CONCLUSIONS

An integrated Dialogue System combining spoken and written language has been presented. It consists of an OCR and a Dialogue Component. Information extracted from application forms by the OCR component updates a local database. The Dialogue Component takes the responsibility to call the customer in order to gather data for the blank or ambiguous fields of the local customers' database. When all the fields have been completed and checked by a human operator, the company's database is updated. The system is currently running at a car insurance company and tested by its personnel. The results are used for further improvements. A formal evaluation of the system will be carried out using real customer calls.

REFERENCES

[1] P. Dalsgaard and A. Baekgaard, "*Spoken Language Dialogue Systems*", Proceedings in Artificial Intelligence, Volume: "Prospects and Perspectives in Speech Technology", Muenchen, Germany, September 1994, pp. 178-191.

[2] A. Dengel, R. Bleisinger, R. Hock, F. Fein and F. Hones, "*From Paper to Office Document Standard Representation*", Computer, Vol. 25, No. 7, July 1992, pp. 63-67.

[3] ACCeSS, "*Automated Call Centre Through Speech Understanding System*", LE-1 1802.

[4] A. Tsopanoglou and N. Fakotakis, "*Selection of the most effective set of subword units for an HMM-based speech recognition system*", Eurospeech '97, 5th European Conference on Speech Communication and Technology, Rodos, Greece, Sept. 22-25, 1997, Vol. 3, pp. 1231-1234.

[5] Chin-Hui Lee and L. R. Rabiner, "*A Frame-Synchronous Network Search Algorithm for Connected Word Recognition*", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 37, No II, November 1989, pp. 1649-1658.

[6] A. Thanopoulos, N. Fakotakis and G. Kokkinakis, "*Linguistic Processor for a Spoken Dialogue System based on Island Parsing Techniques*", Eurospeech '97, 5th European Conference on Speech Communication and Technology, Rodos, Greece, Sept. 22-25, 1997, Vol. 4, pp. 2259-2262.

[7] N. Liolios, N. Fakotakis and G. Kokkinakis, "*Form Identification and Skew Detection from Projections*", EUSIPCO '98, 9th European Signal Processing Conference, Rodos, Greece, Sept. 8-11, 1998 (to be presented).

[8] I. Chatzi, N. Fakotakis and G. Kokkinakis, "*Greek speech database for creation of voice driven teleservices*", Eurospeech '97, 5th European Conference on Speech Communication and Technology, Rodos Greece, Sept. 22-25, 1997, Vol. 4, pp. 1755-1758.