



MAIRIEVOX : A SPEECH-ACTIVATED VOICE INFORMATION SYSTEM

C. GAGNOULET, J. DAMAY

Centre National d'Etudes des Télécommunications
 B.P. 40
 F - 22301 LANNION

Abstract

In this paper, we describe an application of speaker independent speech recognition over telephone lines. The MAIRIEVOX system is a speech-activated interactive voice response system that provides informations of local interest via the ordinary telephone network. The system has been installed in the town hall of Lannion in France for more than one year. After presenting the main features of this system, we mainly point out the importance of human factors and related studies for the success of such an application. We also mention future improvements of the system and current industrial developments.

Description of the MAIRIEVOX service

MAIRIEVOX is a speech-activated interactive voice response system that provides information of local interest via the ordinary telephone network. The user can access a variety of informations on different subjects by making a simple call to the system.

MAIRIEVOX doesn't require the use of any specific customer equipment apart from an ordinary telephone set. Dialogue between the user and the system is entirely in speech form and the information is selected by spoken commands.

A tree structure is used to access informations (see Table I). The user makes successive selections within proposed menus to get the information he needs. The tree that describes the data base has three levels, each level offering a choice between two or three items. Limitation to three levels results from a need to avoid lengthening the dialogue. Thus the user needs only answering two or three questions in order to get the information he wants. Since only a small vocabulary can be recognized correctly at any time due to the telephone network degradations, each menu offers only a few choices.

Level 1	Level 2	Level 3	Messages level	
STANDBY DUTY	DOCTORS		Mess #1, Mess #2, ...	
	CHEMISTS		Mess #1, Mess #2, ...	
	VETS		Mess #1, Mess #2, ...	
LEISURE	ENTERTAINMENTS		Mess #1, Mess #2, ...	
	LOCAL EVENTS		Mess #1, Mess #2, ...	
	CINEMAS	LE CLUB		Mess #1, Mess #2, ...
		LE TAPAL'OEIL		Mess #1, Mess #2, ...
LES BALADINS			Mess #1, Mess #2, ...	
TOWN HALL	NEWS		Mess #1, Mess #2, ...	
	CITIZEN ADVICE		Mess #1, Mess #2, ...	
HELP	HELP	HELP	HELP NEXT PREVIOUS	
	CANCEL OTHER HEADING	CANCEL OTHER HEADING	CANCEL OTHER HEADING	

Table I. Tree structure of the service

10.21437/Eurospeech.1989-149

The final information headings proposed are the following:

DOCTORS : list of doctors on standby duty,
CHEMISTS : list of chemists on standby duty,
VETS : list of veterinary surgeons on standby duty,
ENTERTAINMENTS : list of local entertainments,
LOCAL EVENTS : list of various local events,
LE CLUB : "Club" cinema's program,
LE TAPAL'OEIL : "Tapal'oeil" cinema's program,
LES BALADINS : "Baladins" cinema's program,
NEWS : news concerning local life,
CITIZEN ADVICE : times when specific persons can be consulted at the town hall.

When he has come to the last level of a branch of the tree, the user can hear a series of vocal messages about the selected subject. There can be as many messages as necessary under a given heading. So as to move more easily from one message to another in a subject, user can interrupt at any time the vocal message to skip to the following or preceding one or to climb up one level or back to the main menu:

NEXT : go on immediately to the next information message,
PREVIOUS : go back immediately to previous information message,
CANCEL: immediately leave current heading going back to the top level,
OTHER HEADING : leave current branch allowing to choose another heading in the preceding menu.

Additionally, a **HELP** command makes it possible to have full instructions for using the system at any point in the tree. Moreover, during error recovery procedures, **YES** and **NO** are used.

Example of man-machine dialogue

As an example, here is a dialogue (translated in English) between a user and the system ("/..." indicates that the system is interrupted by user) :

S(erver) : Welcome to MAIRIEVOX. At any time, to know how to use MAIRIEVOX, use the word HELP. Now it is your turn :

S : Say "STANDBY DUTY", "LEISURE" or "TOWN HALL"

U(ser) : LEISURE

S : Say "ENTERTAINMENTS", "LOCAL EVENTS" or "CINEMAS"

U : CINEMAS

S : Say "LE CLUB", "LE TAPAL'OEIL" /...

U : LE CLUB

S : LE CLUB's program:

S : "Rain man", on June /...

U : NEXT

S : "Chinatown" /...

U : PREVIOUS

S : "Rain man" /...

U : OTHER HEADING

S : Say "LE CLUB", /...

U : HELP

S : Instructions for use : ...

System organization

From the hardware aspect, the system consists in a standard micro-computer (IBM-PC, XT compatible) and two specific boards : RDP50 (1), developed by CNET, for speech recognition and speech output, and COSETTE, marketed by the XCOM company, used as a telephone interface.

The software supplied with the system enables the information messages to be updated. The spoken messages need to be recorded periodically : some informations being valid for several months, other items for a few days only.

The total duration of the recorded messages shouldn't exceed the computer's disk capacity (20 minutes of speech requiring about 10 Mbytes of disk space, using 64 kbit/s PCM coding).

A table is used in which each message is associated with an expiration date (out-of-date messages are automatically eliminated by the system). The system's software then automatically bases the information retrieval on the this table content.

Speech recognition

The speech recognition system is speaker independent, and is restricted to isolated words recognition, although the RDP50 board also supports connected words recognition. For the MAIRIEVOX application, the vocabulary consists of the 21 words mentioned above. They are all eligible at any time by the speech recognition system, even if the dialogue filters those that are not valid in the context.

The algorithm used in the system is PHIL86 (2) developed by the CNET. The acoustical analysis consists of a Mel Frequency Cepstrum analysis. Every 16 ms, 6 cepstral coefficients, and 2 energy coefficients are computed. For the recognition, hidden Markov modeling is used for each word in the vocabulary (from 8 to 15 states according to the length of the word). To improve the speaker independence, three models are used for each word.

The evaluation of the recognition system was performed on a 600 speakers data base recorded over the telephone lines, and with the entire vocabulary (21 words). This database was divided in two parts of 300 speakers each (DB1 and DB2). Two tests were performed (TS1 with DB1 for learning and DB2 for test, and TS2 with DB2 for learning and DB1 for test). For each test, the results with 1 and 3 models per word are summarized in table II.

Models	Error rates		[95% confidence interval]
	TS1	TS2	
1 model	5.8	5.2	[0.54]
3 models	4.7	3.8	[0.48]

Table II. Speech recognition evaluation

In the application environment, very different recognition rates were observed : with inexperienced people, the recognition error rate is about 20% (3). Three different points must be emphasized about this deviation between simulation and actual field test results :

Probably the database used for learning the system is not sufficient to actually obtain a speaker independent system. Most of speakers recorded for this learning come from Brittany, and so, the different French dialects are not correctly represented. A larger database is now being recorded with 1000 speakers from different areas in France.

As a result of the dialogue and human factors studies, speech recognition must be always activated throughout the call, even when synthesis is on. Because of the coupling between synthesis and recognition, an electric echo cancelling algorithm is desirable. Such an algorithm is currently under evaluation in the MAIRIEVOX system.

An other problem occurs with the rejection of non allowed utterances. Different simple algorithms were evaluated in order to eliminate erroneous words, but today no satisfactory results were obtained. Actually, the best solution consists in comparing each utterance with the whole 21 words dictionary at any time, even if only 6 different words are allowed by the dialogue process. Thus, the non allowed utterances will be more probably recognized as forbidden words and therefore rejected.

Human factors studies

Speech recognition over the telephone network still has many limits that had to be taken into account when designing the vocal interaction between user and MAIRIEVOX so that speech convenience at the lexical level was not spoiled by too much artificialness in other aspects of dialogue.

When designing MAIRIEVOX dialogue, we thus had to answer to the following questions related to the vocal input mode :

- how should we indicate to the user when and how he should speak (prompt messages) ?
- how could we manage the non allowed words problem ?
- what kind of correction procedure could be offered ?
- how could we help some badly recognized user to get some information anyhow ?

While answering these questions, we also had to progressively improve vocal feedbacks, structure of interaction, wording of services and the service itself, as for any interactive voice response system. These improvements require to answer to such questions as:

- how should we organize the access to information ?
- how could we offer a more rapid access to expert users ?
- how could we help the users get acquainted with the system ?
- can we shorten the vocal prompts ?
- is the vocabulary sufficient and adequate ?
- etc...

The following paragraphs describe the laboratory and field experiments conducted from March to June 1988 to get answers to these questions. It must be stressed that making human factors and service studies early in such a project is essential since late modifications of vocabulary and dialogue structure imply expensive changes in speakers data base and service software.

"When and how should the user and MAIRIEVOX speak ?" - Experiments in CNET

After a first prototype had shown the feasibility of MAIRIEVOX, it was decided to conduct an experiment regarding dialogue ergonomics. It had three objectives :

- make a definite choice for speech turn-taking between user and system.
- analyse user behaviour and opinions regarding : MAIRIEVOX welcome message, help, speech recognition and dialogue appreciation, information given.
- test convenience of the error recovery strategy. An appropriate error message is given to the user in case his utterance is not detected (low sound level or even no response) or is recognized as an invalid word; then the menu is repeated. If a second error occurs, the error message is repeated and the user is asked to answer YES/NO questions corresponding to each choice of the menu. If too many errors occur, MAIRIEVOX interrupts the communication.

Two versions of MAIRIEVOX were thus compared : in the first one ("beep" version), each vocal prompt was followed by a "beep" after which a silence was reserved for the user to give his choice, in the second one ("any time" version), vocal choice corresponding to a dialogue state can be expressed at any time, during or after the prompt. By example:

- S : "Say STANDBY DUTY, LEI..."
- U : "LEISURE".

In both versions, users could use the control commands (NEXT, PREVIOUS, HELP, OTHER HEADING) during information retrieval.

Two groups of 10 CNET employees executed 5 scenarios of information research, each group using one of the two versions, then completed an evaluation questionnaire. During the scenarios and next to the subjects, the experimenter checked which words were pronounced, when (during a silence or during a prompt) and whether they were correctly recognized by the system.

As we had chosen to have two different groups to avoid learning and interference effect between the two versions, it was checked that the two populations could be compared as regards recognition during silences.

From the results given by this experiment we got the following answers to our questions :

- For both versions, the speech recognition rate during post-prompt silences was near 74 percent.
- The overall recognition rate from the user point of view is worse in the "beep" version than in the "any time" version. Actually, users tend to speak before silence in both cases, thus being unsuccessful in the "beep" case (= subjective misunderstanding) and being recognized, although with a smaller rate than during silences, in the "any time" case.
- The "any time" version allows a quicker access to information nodes.
- Users don't appreciate to wait for a signal before speaking.
- For the "any time" version, subjects succeeded in more scenarios and there were less error recovery sequences.

Although these first results already lead to prefer the "any time" version, a control experiment was conducted to check some remaining questions:

- with the "beep" version, would some learning effect help more experienced user improve their performance (i.e. by speaking only when it is allowed to) ?
- with the "any time" version, would a reverse learning effect lead users to deteriorate their overall recognition rate (i.e. by speaking too frequently outside silences, thus when the system has worse performance) ?

For this second experiment, the same groups of subjects executed the same scenarios 15 days later. We obtained the following results:

- speech recognition performances of both groups improved, coming from two different learning effects: for the "beep" version, a better respect of prompt before speaking; for the "any time" version, more interruptions were compensated by a louder elocution than at the first trial.
- it is not possible to defend one version against the other on the only basis of speech recognition rates during this experiment.

Because of user satisfaction, easiness and speed of dialogue, the "any time" version was chosen. Due to the observations made during these experiments, other improvements were brought to the dialogue such as:

- Many prompts and help messages were modified or added to better fit the context in which they occur and be expressed in a clearer and simpler manner. Thus, the first message was modified (see dialogue example). This first sentence gives all necessary information to a beginner, prevent him from feeling obliged to say HELP and help them distinguish between introduction and first menu.

The number of yes/no recovery attempts allowed before ending an unsuccessful session was raised in order to let the user decide himself if he wants to give up; the yes/no procedure was improved by defining better reaction in case neither NO nor YES is understood by using confusion matrix data.

- First reactions to the information messages were gathered: suggestions about new headings, updating rhythm, way to speak telephone numbers and messages, ...

"How do users really use and appreciate MAIRIEVOX ?" - Field study

The preceding experiment had to be completed in a more realistic context: more spontaneous use, variety of phone line conditions,... so as to get a better knowledge of service acceptability. A limited advertising campaign was made by sending information and an evaluation questionnaire to 400 people (60 of which returned it completed). The MAIRIEVOX system was installed at the Lannion town hall. An experimenter was listening to every call made to MAIRIEVOX during day time and checking user behaviour as in the preceding experiments (80 calls to MAIRIEVOX could be analysed). A fictitious command, **REMARK**, was added so as to let users come in communication with the experimenter at any moment in a session. This experiment gave the following results:

- the overall recognition rate reached an acceptable level of 79% and 80% of the subjects expressed their satisfaction with the speech recognition quality.
- the yes/no strategy proved its efficiency: there were 50 YES/NO recovery dialogues, 38 of them leading to the desired end.
- most people felt the access was easy, some of them nevertheless finding it too slow. They used the help feature 15 times.
- 70% subjects expressed their satisfaction towards MAIRIEVOX.
- suggestions were made to improve content and access to information: giving information in date order announcing the date first, accessing to movie programs by words like TODAY, MONDAY, etc..., having a fix order in announcing information fields (date, time, program, content, ...). These suggestions would lead to offer a real data base structure with various types of interrogations.

Since April 1988, statistics are collected on the MAIRIEVOX system. As the speech signal cannot be recorded during the call, only state changes in the dialogue automaton are used. So these results are very hard to interpret.

For a one year period, the average number of calls per week was 289, for a town of about 20000 inhabitants. Among these calls, 62% concerned the leisure informations, 16% the standby duty informations and 22% the town hall informations. 77% of the calls obtained one or several informations. Today, we obtain a mean value of 150 calls each week, and most of the users are now experienced users.

Industrial developments

The CNET know-how has been transferred to more than 10 different companies, mainly in France. Today, similar prototype systems are available or under evaluation in France. These interactive speech-activated vocal systems are developed by ALCATEL-TITN (Rennes), FERMA (Paris), ACSYS (Paris), SEFER (Poitiers) ... Specific developments were undertaken by these companies to offer multiple ports systems and ergonomic software to build various applications. Many other applications are thus expected before the end of 1989.

Conclusion

Giving speech recognition capability to a telephone interactive response system presents many potential advantages. From the user point of view, it is hoped that the interaction with the machine will seem more natural and pleasant than using a DTMF keyboard and entering codes to get information. From the information provider point of view, it is an opportunity to open their market to every telephone set owner. User satisfaction and industrial efforts that followed MAIRIEVOX first implementation has shown that these potential advantages can turn into reality, provided the present capabilities of speech recognition are well identified and necessary human factors studies undertaken.

REFERENCES

- (1). J.P. TUBACH, C. GAGNOULET, J.L. GAUVAIN, "Advances in speech recognition products from France", Speech Tech'89, pp.266-269; New York, 1989.
- (2). D. JOUVET, J. MONNE, D. DUBOIS, "A new network-based speaker independent connected word recognition system", IEEE ICASSP 1986, pp. 1109-1112; Tokyo, 1986.
- (3). B. COLIN : "Evaluation du micro-serveur vocal MAIRIEVOX", CNET, rapport de stage; Lannion, 1988.