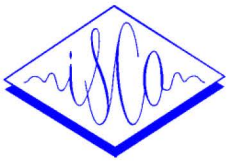

First European Conference on
Speech Communication and Technology
EUROSPEECH '89
Paris, France, September 27-29, 1989

VOICE BASED REMOTE DATA BASE ACCESS

A. Riccio, F. Carraro, E. Mumolo

ALCATEL FACE Research Centre
Via Nicaragua 10, 00040 POMEZIA
ITALY

ABSTRACT

The paper decribes an application of Speech Processing technologies in a real environment and the results of a preliminary field trial with real users. The aim was to integrate various Speech Processing capabilities, provided by a dedicated system, with a small data base and to make it possible an enquiry via a standard telephone; both public and private network have been used to test the performances.

Since the speech interface was the only way to access the data base, the experiment gave the opportunity to verify on field, both technical and ergonomics aspects.

1. INTRODUCTION

Thanks to the fast growing of Speech Processing capabilities, it become now possible to control even complex systems by using just voice; in this sense different applications have been proposed and developed [1]. In the next sections we describe the results of an experiment that involves the use of three different Speech Processing technologies at the same time, namely: speech recognition, speech coding and text-to-speech; all these technologies are made available on a PC based system capable to handle a telephone line and to manage a complete call progress and data retrieval.

The choice of the PC environment seems to be an interesting approach, because of the high flexibility offered for update and development, combined with the relatively low cost of the final system.

By means of such configuration, a voice driven telephone data access has been realized; users can dial the number of the extension connected to this system and, after listening to a presentation announcement, can ask for a specific menu (for istance : flight timetable, change of currency, etc.), just speaking the name of the menu. In case of doubt, an Help menu, always active and dinamically tailored for the current context, provides helpful information.

Information can be given to users by means of a speech coding tecnique and/or by a text-to-speech conversion, depending upon the menu.

2. System overview

The system is built around a personal computer equipped with two PC-compatible boards, entirely developed at our laboratory and able to perform various speech processing functions.

The boards have been designed to support both audio (for microphone and loudspeaker) and telephone line interfacing.
Actually the speech boards are able to handle two channels at the same time, but for this application only one channel was used.
A telephone line was connected to the boards allowing in such a way a remote access through our local PABX [2].
Finally a printer, as a data logging device, was used to report some statistics related to usage and performances of the equipment.
A schematic description is reported in fig. 1.

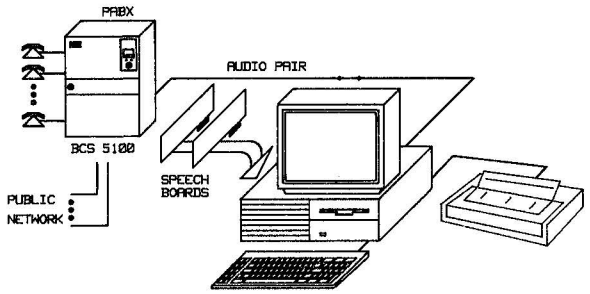

Fig. 1 SYSTEM OVERVIEW

With such an architecture, most of the the signal processing is performed by the DSP boards therefore no extra computational power and capabilities are demanded on PC, making this application feasible, flexible and easy up-gradable.
In fact, as new speech processing functions will be available or the existing ones will be improved, their integration in such a system won't raise particular problems.
Bearing this clear in mind, the sense of this application become a sort of test bed for a very advanced technology that seems to be at last ready to face the real environment.
In order to better illustrate the SW architecture, it could be useful to distinguish those SW modules implemented on the Host computer with respect to the specific speech processing SW modules implemented on the DSP

3. SW ON HOST COMPUTER

The following SW modules were implemented on the Host computer :

- *Call processing*
- *Voice commands interpretation*
- *Data retrieval*
- *Context sensitive HELP*

These modules will be briefly explained in the following.

3.1 Call processing

As a first step, the system waits for an incoming call, whose presence is detected by a dedicated industry standard I.C. which reveals the ring signal on the audio pair. The circuit shows a high immunity to line noise and extraneous spikes, featuring a highly reliable operation.

After the occourrence of a programmable number of ring cycles, the SW module sets a flag that will be read by the supervisor module; this makes the following actions to be taken :

- The audio pair is connected to the telephone interface of the DSP boards
- The SW module for monitoring the telephone line signalling is activated in background mode
- The interactive session with the user is opened by sending the greeting voice message

Furthermore, if the user prematurely hangs up the handset, the call processing module is able to reveals this event and to inform the supervisor.

3.2 Voice commands interpretation

At the beginning of the man-machine dialog, a synthesized greeting message is sent to the user; this greeting message sounds like the following : " You are connected with the Alcatel Face Voice data bank : say your command or say HELP " Once the user has spoken the command, the recognized string and his score are forwarded by the recognition module running on the DSP boards to the supervisor module; provided that the score is above the current threshold value, the string is evaluated and the command executed. Otherwise a "Please repeat" message is generated .

3.3 Data retrieval

For the purpose of this experiment most of the data were entered and left unchanged (flight time table, touristic information etc.). Moreover, to better emphasize the capabilities of text-to-speech conversion, data in the "Currency Change" menu, were daily updated by replacing the file containing the relevant information. In this way it was simulated the real situation in which that file could be directly updated by means of a link between the PC and the local area network.

3.4 Context sensitive HELP

The user could ask for help in any situation and the system was able to give him/her the right information, pertinent to the current activated context. The Help messages gave the user information about the current menu and a list of the available commands.

4. SPEECH PROCESSING CAPABILITIES

As mentioned above, the speech processing capabilities are provided by a dedicated Hardware developed at ALCATEL FACE Research Centre by the Speech Processing Department; here a short description of this Hardware will be given, while major focus will be placed on the algorithms .

4.1 DSP Hardware

A block diagram of the Hardware is reported in fig. 2 ; as shown, two analog channels are available and both of them are able to be interfaced to either audio or telephone line.

The DSP functions are basically carried out by the TMS32010 processor that is configured to perform the requested algorithm by downloading the corresponding code in its program memory; the download is activated by the host and realized by the 68000 processor via the Shared Memory.
The same Shared Memory is used to transfer/ retrieve speech data to/ from the host mass memory (a Winchester hard disk); this is the case when the speech analysis/synthesis function is activated.
Finally, the DTW (Dynamic Time Warping) chip, an ALCATEL proprietary custom design, allows to perform most of the computations required for the speech recognition algorithm; the chip features a 450 ns cycle time and is capable to process over 800 templates in real time.

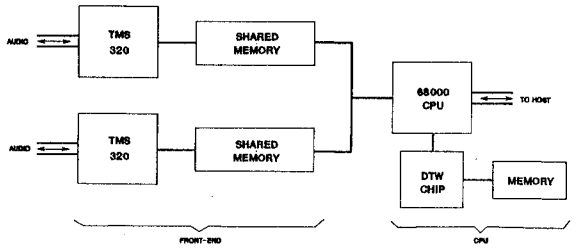

Fig. 2 HW-BLOCK DIAGRAM

4.2 DSP Software

DSP modules involved were, namely :

- Speech coding
- Isolated utterances recognition
- Text to speech conversion

In the following, a brief description of these modules is given.

- Speech coding

Two algorithms have been used :

- ADPCM featuring 32 Kbit/s
- RELP featuring 8.8 Kbit/s

The choice to use two different tecniques was taken on purpouse; it was decided in fact that messages and/or announcements were pre-coded with high quality while, for the voice mail box function, even a medium quality tecnique could be used.

As described in the following, a Voice Mail Box was available to each user to leave a message (for this experiment, generally a comment on the functionality of the system or some suggestion); in this case, the final user of this messages is supposed to be very cooperative and the quality of the recorded messages is not the main problem.

Neverthless, most of the users found the quality of RELP acceptable at least for male voices.

Both the algorithms have been implemented on the TMS 32010; the ADPCM program is completely contained on the DSP program memory and implements a version of ADPCM not CCITT compatible; it produces a 4 bit codeword for each 8 bit PCM sample, every 125 microseconds, resulting in a 32 Kbit/s data rate .

The RELP program occupies a larger memory size, therefore the complete coding is performed by two separate modules, namely Relp Analysis and RELP synthesis that were downloaded into the program memory depending on the context. As known RELP belongs to the LPC class, where the excitation signal is replaced with the LPC residual. In order to limit the data-rate, the residual, during the Analysis phase, is low pass filtered near to 800 Hz; during the synthesis, the high frequencies are reconstructed by spectral folding. The transmission of residual requires 6.4 Kbit/s; LPC coefficients and energy need 2.4 Kbit/s yielding a total data rate of 8.8 Kbit/s.

- Speech recognition

The speech recognition algorithm is based on single stage dynamic time warping for connected word recognition [3], [4]; the same tecnique can be used for isolated utterance recognition. Generally, using the algorithm for connected words recognition, a finite state syntax has to be defined, with a starting and an ending node; when the system is used for isolated utterances recognition, a single node is active.

In order to allow a multi-user operation, the system has been trained by a relatively large number of speakers so that, by producing multiple reference patterns, a speaker independent operation has been made possible. This approach is feasible by means of clustering techniques that allow to reduce the total amount of templates to a reasonable dimension.

Moreover, the training procedure has been performed using the same environment used for the real operation; the selected speakers in fact used a standard telephone handset of one of the subsets connected via the PABX to the speech recognizer. In this way, the templates were created taking into account both telephone noise and bandwidth limitations.

- Text to speech conversion

For text to speech synthesis, a diphone concatenation approach has been used. This choice has been already chosen for many languages and even for italian has shown good results.

Actually text-to-speech synthesis consists of two main parts : a linguistic analysis (necessary to segment the written text into diphones and to generate the proper intonation) and the speech synthesis itself that allows to produce the sound [5],[6].

Diphones have been extracted from natural speech by using an interactive tool for segmentation; the segments have been then LPC coded and stored.

The quality is quite intelligible even though not completely natural; this is the main problem for this technology and it is common also to other implementations even in other languages, so that it can be considered as the current status of art.

5. SYSTEM OPERATION

5.1 Human factors considerations

Two basic consideration have driven the design and development of the system; on the one hand, several field trials have shown that, besides the technical performances (in terms of recognition accuracy , vocabulary size, kind of syntax, quality of pre-coded messages etc.), what makes a speech processing based application successful, is the easy of use and the kind of friendliness the user perceives at his first approach. In this sense, the human-machine interaction plays a basic role; therefore, for such application, the dialog must take care of the consideration that the user will probably access the system using just his voice, without looking at any written menus; but, even in this case, that will be the one where this kind of system will be useful, the user must be able to get the service.

On the other hand, the adoption of precoded messages in place of written menus need special cares; the messages indeed shouldn't be too verbose or long in order to avoid the user get bored or confused; further, if any speech compression tecnique is used, it should be guaranteed a pleasant quality; finally, the combination of male and female voice should be possible to make the messages more incisive and clear.

5.2 Access to menus

The operation of the system was extremely easy; the user could access the Voice Data Bank by using a standard telephone either inside our laboratory or outside; after the connection was established, the user received the greeting message and was prompted to speak a command or to say HELP.

Depending on the user's command and on the recognition result (successful or not), the operating context could either remain the same or be changed according to the spoken command.

Therefore the user could move through menus, get the desired information, receive help messages in case of doubt or just to listen to the summary of entries available in the current menu.

To quit the enquiry session, the user could simply hang up the handset or say the corresponding command, "END OF ENQUIRY"; in the latter case he received a goodbye message.

5.3 Available Menus

Since the aim of the experiment was the global evaluation of the performances, the choice of the type of menus was not a critical point. We only tried to encourage the usage of the Data Bank by including data of general usefulness; actually the following menus, fig.3, were available:

- Flight time table
- Currency change
- Touristic information
- Voice Mail Box
- System description

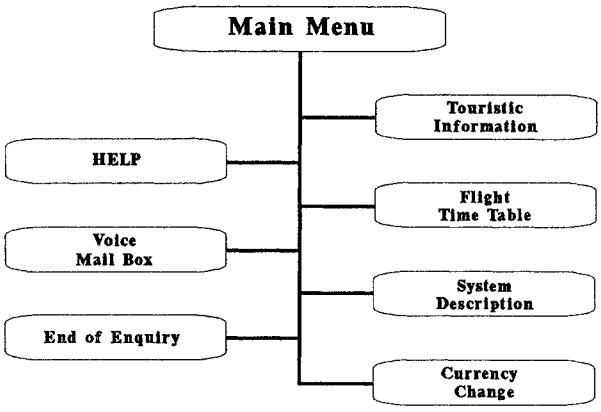

Fig. 3 MAIN MENU

This last menu contained a description of the system itself, explaining the technology, the capabilities, the basic problems and other related topics.

On average, each menu contained about 15 entries; including the key-words "HELP", "CHANGE MENU" and "END OF ENQUIRY".

5.4 Voice Mail Box

This function was added to give the user the opportunity to leave comments or suggestions about the Voice Data Bank just using the system itself; by selecting the "MAIL BOX" menu, it was possible to record, in real time, a message of any duration. A silence detector, implemented on the DSP was used to identify the end of the recording; at this point the same message was immediately played back to the user during the same session for verification purpose. Of course, this was an extrapolation of the true Mail Box operation, but the service was considered so nice that almost all users wanted to use it.

It was chosen to use for this function a RELP coding at 8.8 Kbit/s to limit the mass memory occupation; the quality of the coded voice was poorer compared to ADPCM, but this aspect was not pointed out in the final questionnaire.

6. USERS' SIDE ACCEPTANCE

In order to make this part of the experiment effective and meaningful, a questionnaire form was prepared and distributed among users to solicit them to answer some critic questions. These forms were used together with the session reports produced via a printer connected to the system, to have an overview of the kind of problems actually encountered using the system [7].

Questions were basically addressed to three different aspects :

- Performances of the speech functions

- User friendliness

- Usefulness of the system

Concerning the performances, the basic problem is the recognition accuracy that, particularly when the system was accessed from outside our laboratory, via the the public network, gave unsatisfactory results. Text-to-speech was generally accepted even though the quality was judged slightly boring.

User friendliness was considered good but in some cases the messages were considered too long; the possibility to interrupt in some way the synthesized message was expressely requested.

Finally, the general feeling was positive and most of users found the system very attractive with respect to the traditional access performed via a keyboard.

7. CONCLUSIONS

Speech processing technology seems to be at last mature to be integrated in real applications; so far, however, many technical problems are still pending, especially concerning the speech recognition aspect; with respect to traditional experiments dealing with systems used in favourable operating environment, the application described shows that constraints like bandwidth limitations and/or distorsions, telephone grade microphones, telephone network quality, noise and so on, heavily affects the system performances.

Apart from that, users found the idea of a voice access to a remote data base very exciting, suggesting several additional services and improvements they would have appreciated.

This encourages our efforts to make such a system more reliable and effective.

REFERENCES

[1] M. Immendoerfer, E. Mumolo "Recent Achievements in Speech Processing Technologies and Their Applications", Electrical Communication, Technical Journal of ALCATEL NV, Vol. 62, No. 3/4 1988, pp. 288-293

[2] M. Bazzani, F. Carraro, G. Colangeli, E. Mumolo, P. Pierucci, A. Riccio, "Application of Speech Processing to a New Generation PABX", International Workshop on Recent Advances and Applications of Speech Recognition, Rome, May 86 , Proceedings Supplement, pp. 79 - 108

[3] J.S. Bridle, M. D. Brown, "Connected Word Recognition Using Whole Word Templates", Proceedings of the Institute of Acoustics Autumn Conference, Nov. 79, pp. 25-28

[4] J.S. Bridle, M.D. Brown, R.M. Chamberlain, "A One Pass Algorithm for Connected Word Recognition", ICASSP 82, Paris, pp. 899, 902

[5] P. Pierucci, E. Mumolo, C. Labonia, "Multichannel Text-to-Speech System for Electronic Mail Applications", European Conference on Speech Technology, Edinburgh, Sept. 87, pp. 264-267

[6] M. Bazzani, E. Mumolo, "PC-Based Telephone Communication System for Deaf-Blind People", Globecom 88, Hollywood, Florida, Nov. 88, pp. 43-47

[7] J. Vaughan, "Human Factors Field Trial", ITT Internal Report, Nov. 84