



PUBLIVOX : A VOICE CONTROLLED CARD PAY PHONE

C. Gagnoulet (*), F. Zurcher (*), J. Tirbois (*), T. Serradura (**)

ABSTRACT

This paper concerns an application project for the use of a speech recognition system in public phones. The project is jointly run by the CNET and the French company CROUZET. An evaluation should be forthcoming in early 1988 with ten prototypes of public phones with voice access, located in different areas around France.

This paper describes the main features of these phones : the vocal dialogue, the speaker-independent recognition system, the hand-free dialer. The preliminary results obtained in simulation and real-life prototypes will also be given.

INTRODUCTION

The number of industrial recognition systems continues to increase, though there are as yet very few applications to actually make use of them.

In France, a demand for a telecommunication application has appeared in the area of public call-boxes, which are often made unusable by vandalism. The vulnerable elements here are firstly the money container, and to a lesser extent the telephone handset. Various methods have been investigated for reducing this vandalism. One step in this direction was the introduction of I.C. card phones; this removes one of the causes of vandalism (the money box), though other elements remain vulnerable. It was with these considerations in mind that the CNET and the French company CROUZET decided to undertake this project in 1985. It is intended to install prototypes of card-operated public phones having neither handset nor dial. These are to go under the name of PUBLIVOX. The role of the handset and the dial are both replaced by voice-operated functions : a "hand-free" telephone will replace the handset and a voice recognition system will replace the dial. A trial batch of ten or so PUBLIVOX call-boxes should be available for use by the public in early 1988.

This paper will give details on certain aspects of the project: the overall architecture of PUBLIVOX; the basis for the establishment of dialogue; technical characteristics for the voice functions. The results from preliminary trials will also be set out.

SYSTEM ARCHITECTURE

The PUBLIVOX system consists of :

- a public phone, operated by I.C. cards, handling all the telephone functions,
- an I.C. card reader, with locking and unlocking flap,
- a "hand-free" telephone board,
- a speech recognition board,
- a board emitting coded signals in voice forms,

(*) CNET : LAA/TSS/DAP, Lannion, FRANCE.

(**) CROUZET, Terminaux et Systèmes, Valence, FRANCE.

- a CPU board managing the interface between the phone and the voice elements via links of the RS232 type.

From the user's point of view, the PUBLIVOX looks very much like a conventional call-box; from the outside, the booth resembles those found in Paris. However, the use of a speech recognition system will require an increased acoustic insulation; with thicker glass and less reverberant panels, insulation figures of better than 6 dBA have been obtained. Another difference is that the technical column has been enlarged to ensure adequate ventilation and to take the additional electronic boards.

There is no change to the card reader, nor to the liquid crystal display (two lines of twenty characters). The microphone is hidden behind a protective grid in the place of existing dial. A second grid conceals a loudspeaker, this taking the place of the existing handset. Finally, a single pushbutton is used for both "pick-up" and "hang-up" operations.

In the first models, the speech-processing boards will be located in the plinth.

VOCAL DIALOGUE WITH THE USER

It is obvious that the dialogue with the user will be a very important factor in any system based on speech recognition; an unsatisfactory dialogue could well lead to the system's being rejected out of hand, regardless of what might be the quality of the speech recognition as such.

The user dialogue for this system has been designed with two basic principles in mind. Firstly, the novelty of the new system should not require too great an effort of adaptation from the user. Secondly, the system should seek to dampen any negative effects caused by imperfections in the speech recognition function.

It was decided to adopt the principle of expressing telephone numbers in groups of two digits. The decision was based on the limitations of the recognition system used, and the existing habits of French telephone users. Services other than the request of a normal number (emergency services, abbreviated dialing) are requested by pronouncing a single word corresponding to the name of the service.

User guidance will chiefly take place via the LCD display, as is the case in existing call-boxes. Vocal guidance will be reserved for assisting the user if difficulties arise; it will come into operation after pre-established time delays and on user errors or failures to recognize words.

Functions have been included in the dialogue for rectifying recognition errors. A "Confusion matrix" models the most frequently found recognition errors and rectifies repetitive errors (repeated incorrect recognition for numbers requested by the same speaker).

By way of example, a simple communication is described by the figure 1. Payment is by prepaid card and the user is "experienced", i.e. there are no vocal assistance messages.

SPEECH RECOGNIZER

The recognition system used is speaker-independent and capable of recognizing isolated and connected words from a vocabulary of about one hundred words (ref 1). For PUBLIVOX, the recognition of two-digit figures will require a vocabulary of 28 words. The remaining vocabulary space will be taken up by isolated words used for auxiliary commands.

The technique used is a MFCC acoustic analysis (8 coefficients computed every 20 ms), and statistical modeling of the application (hidden Markov Models). Various basic units were used to define the whole model,

Mechanical action	Display	Vocal input	Dialogue state
Press button	Please pick-up		Inactive
Card inserted, flap closed	Insert card		Card validation
	Pronounce your number in groups of two digits		Vocal dialing
	--	96	
	96 --	05	
	96 07 --	Correction	
	96 --	05	
	96 05 --	11	
	96 05 11 --	11	
	96 05 11 11	Send	Dialing, communication established
		Communication ...	
Push button			Hang-up
Recover card			Inactive

Figure 1

the best results being obtained using allophonic models initialised after manual segmentation by a phonetician expert (ref 2).

The most difficult problem still to be overcome is that of environmental noise. Various techniques have been tested for eliminating noise by spectral subtraction, though as yet it has not been possible to obtain a significant improvement in this area. On the first prototypes, it will not be possible to include additional noise-elimination processing; the sites for this first call-boxes will therefore be chosen so as to ensure that the signal/noise ratios are above the threshold required for correct operation of the speech recognition system. The system is to be trained in conditions as close as possible to those encountered in the real-life situation.

More sophisticated noise-elimination techniques are under study for use in later versions. One such technique involves dividing the spectrum into two sub-spectrums; low frequencies are processed by Widrow filtering, and high frequencies by spectral subtraction. Such a system would require the use of two microphones.

HAND-FREE TELEPHONE

With the hand-free telephone, the user has a relative freedom of movement and position in the booth. Two peak level equalizers are used :

- one is used for transmission. This is placed after the microphone amplifier and corrects differences in sound levels caused by changes in voice intensity and changes in the speaker's position in the booth.
- one is used for reception. This corrects differences in sound levels resulting from the long distance communications and from the different attenuations introduced by different call-box locations.

An additional variable gain unit maintains strictly constant the gain of the loop as a whole (reception plus transmission), and guarantees unconditional stability (no feedback). At the same time, it ensures comfortable listening conditions at both ends. This gain control is accurate and rapid enough to avoid the word clipping effects, and the circuit's operation is not sensitive to environmental noise. In then absence of speech, the gain is equally divided between transmission and reception, thus minimizing the amount of noise heard at either end of the line.

FIRST QUALITY EVALUATION

Trials of the system under real-life conditions of public use are scheduled for 1988, though some preliminary results have already been obtained by the CNET. Laboratory tests using two-digits figures were carried out on the speech recognition system. In these ideal conditions (low noise, experimented speakers), a recognition rate of more than 95 % was obtained with 26 speakers. In the very first tests run on the actual prototype, a recognition rate of 93 % was obtained with 13 speakers.

PUBLIVOX includes a semi-automatic self-evaluation system; each time the call-box is used, a recording is made of the words recognized, times elapsed and dialogue automaton states. This data can be periodically read using a portable computer. An analysis program will give real-life conditions rates of recognition and indications on the dialogue effectiveness.

As a conclusion, at the present time, PUBLIVOX can be considered as a highly ambitious project from the technical point of view; the speech recognition system will be called upon to operate in difficult conditions. Whatever the outcome of the final evaluation, the spin-offs from the project can be expected to be of considerable importance, both as regards investigation of real limits of current speech recognition systems, as regards vocal dialogue.

ACKNOWLEDGEMENT

The authors will to thank D. Jovet and J. Monne for many useful discussions.

REFERENCES :

1. D. Jovet, J. Monne, D. Dubois, "A new network-based speaker-independent connected-word recognition system", ICASSP_86, p.1109, Tokyo 1986.
2. K. Bartkova, D. Jovet, "Speaker-independent speech recognition using allophones", ICPhs_87, Tallin 1987.