VOCAL INTERFACE FOR A MAN-MACHINE DIALOG

Dominique BEROULE

LIMSI (CNRS), B.P. 30, 91406 ORSAY CEDEX, FRANCE

ABSTRACT

We describe a dialogue-handling module used as an interface between a vocal terminal and a task-oriented device (for instance : a robot manipulating blocks). This module has been specially designed to be implanted on a single board using microprocessor, and inserted into the vocal terminal which already comprises a speech recognition board and a synthesis board. The entire vocal system is at present capable of conducting a real time spoken dialogue with its user.

I INTRODUCTION

A great deal of interest is actually being shown in providing computer interfaces through dialog processing systems using speech input and output (Levinson and Shipley, 1979). In the same time, the amelioration of the microprocessor technology has allowed the implantation of word recognition and text-to-speech synthesis systems on single boards (Liénard and Mariani, 1982 ; Gauvain, 1983 ; Asta and Liénard, 1979) ; in our laboratory, such modules have been integrated into a compact unit that forms an autonomous vocal processor which has applications in a number of varied domains : vocal command of cars, of planes, office automation and computer-aided learning (Néel et al., 1982).

Whereas most of the present language understanding systems require large computational resources, our goal has been to implement a dialog-handling board in the LIMSI's Vocal Terminal.

The use of micro-systems introduces memory size and real-time constraints which have incited us to limit ourselves in the use of presently available computational linguistic techniques. Therefore, we have taken inspiration from a simple model of semantic network ; for the same reasons, the initial parser based on an Augmented Transition Network (Woods, 1970) and implemented on an IBM 370 (Memmi and Mariani, 1982) was replaced by another less time- and memory-consuming one.

The work presented herein extends possible application fields by allowing an interactive vocal relation between the machine and its user for the execution of a specific task : the application that we have chosen is a man-machine communication with a robot manipulating blocks and using a Plan Generating System.

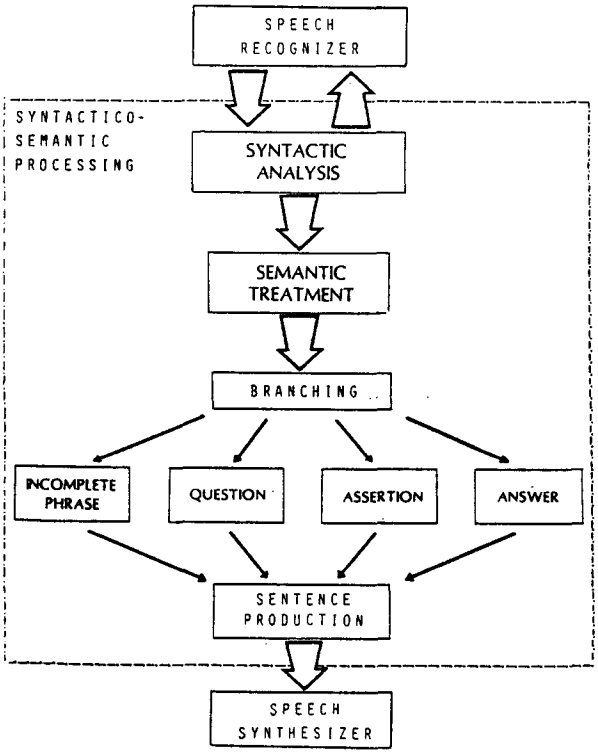

Figure 1. Block diagram of the system

II SYNTACTIC PROCESSING

A. Prediction Device

Once the acoustic processing of the speech signal is performed by the 250 word-based recognition board, syntactic analysis is carried out.

It may be noted that response time and word confusions increase with the vocabulary size of word recognition systems. To limit the degradation of performance, syntactic information is used : words that can possibly follow a given word may be predicted at each step of the recognition process with the intention of reducing vocabulary.

B. Parameters Transfer

In order to build a representation of the deep structure of an input sentence, parameters requested by the semantic procedures must be filled with the correct values. The parsing method that we developped considers the naturel language utterances as a set of noun phrases connected with function words (prepositions, verbs ...) which specify their relationships. At the present time, the set of noun phrases is obtained by segmenting the utterance at each function word.

* S1

Le petit chat gris attrape la souris

* S11 * S12

(the small grey cat is catching the mouse)

parameters :

* O11 ← chat   * O12 ← souris   * V1 ← attraper
* P11 ← (petit gris)
* P12 ← NIL

* S22

* S11 * S12 *S21 *S221 * S222

Prend la pyramide et pose la sur le gros cube

* S1 * S2

(grasp the pyramid and put it on the big cub)

parameters :

* O11 ← NIL	* O21 ← NIL	* O222 ← cube
* P11 ← NIL	* P21 ← NIL	* P222 ← (gros)
* V1 ← prendre	* V2 ← poser	
* O12 ← pyramide	* O221 ← pyramide	
* P12 ← (petite)	* P221 ← (petite)	

Figure 2. Parameters transfer

III SEMANTIC PROCESSING

A. System knowledge data

The computational semantic memory is inspired by the Collins and Quillian model, a hierarchical network in which each node represents a concept. Properties can be assigned to each node, which also inherits those of its ancestors. Our choice has been influenced by the desire to design a system which would be able to easily learn new concepts ; that is, to complete or to modify its knowledge according to information coming from a vocal input/output system.

Each noun of the vocabulary is represented by a node in such a tree structure. The meaning of any given verb is provided by rules that indicate the type of objects that can be related. As far as adjectives are concerned, they are arranged in exclusive property groups.

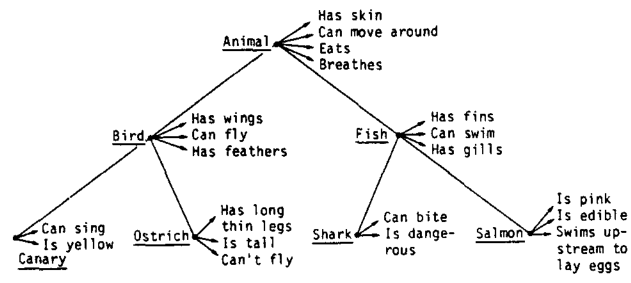

Figure 3. Semantic memory model

The knowledge-based data (which may be enlarged by information provided by the vocal channel) is complemented by temporary data which chronologically contain, in abbreviated form, events evoked during the dialogue.

B. Assertion processing

The small amount of data representing a given universe allows us to approach the computational treatment of these two complementary and contrary components of dialogue: learning and contestation.

Every time an assertion is proposed by the user a procedure parses its semantic validity by answering the question "Does this sentence fit with the current state of the knowledge data ?". If a contradiction is detected, it is pointed out to the user who must justify his proposal. If the user persists in his declaration, the machine may then modify its universe knowledge, otherwise the utterance is not taken into account.

When no contradiction is encountered, the program enters into a learning process adding to the temporary data or knowledge-based data.

User : Un poisson possède des plumes
(A fish has got feathers)
System : J'ai compris ... As-tu quelque chose à ajouter ?
(I have understood ... Would you like to say something else ?)
U : Une truite possède des écailles
(A trout has got scales)
S : Je croyais qu'un poisson possédait des plumes
(I thought that a fish had got feathers)
U : Non (No)
S : D'accord ... (All right)
U : Qu'est-ce qu'un poisson ? (What is a fish ?)
S : C'est un animal qui possède des écailles
(It is an animal which has got scales)

1. Teaching utterances

These assertions, characterized by the presence of a non-action verb, permit both the complete construction of the semantic network and of the concept relation rules specifying the possible entities that can serve as arguments for a predicate.

Although most of our knowledge results from long nurturing and frequent interactions with the outside world, it is possible to give an approximate meaning to concrete objects and verbs by using an elementary syntax. A new concept may be taught by filling in its position within the semantic network and possibly associating it with properties that will differenciate it from its brother nodes. Concept relation rules can be learned, too.

U : Le café ... (coffee ...)
S : Qu'est ce qu'un café (What is coffee)
U : Le café ressemble à du lait
(Coffee looks like milk)
S : ...
U : Le café est noir (coffee is black)
S : Je pense avoir compris ... à toi
(I think that I have understood ...)
U : Qu'est ce que du café ? (What is coffee ?)
S : Le café est un liquide noir
(Coffee is a black liquid)
U : Tu bois du café (You are drinking coffee)
S : C'est impossible (It is impossible)
U : Un ordinateur peut boire un liquide
(A computer is able to drink a liquid)
S : C'est bon ... j'ai compris
(That's right ... I have understood)
U : Tu bois du café (You are drinking coffee)
S : D'accord (Okay)
U : Qui bois du café ? (Who is drinking coffee ?)
S : Je le bois (I am drinking it)

2. Descriptive utterances

Sentences involving an action verb are translated into an unambiguous representation which condenses and organizes information into the very same form as that of the concept relation rules from knowledge data. Therefore, semantic validity can be easily tested by a pattern-matching process. A semantic event reduced to a nested-triplet structure and considered as valid is then inserted in the dynamic-events memory, and can be requested later on by the question-answering process.

Although the language is limited to a small subset of natural French, several equivalent syntactic structures are allowed to express a given event ; in order to avoid storing multiple representations of the same event, paraphrases of a given utterance are reduced to a single standard form.

One of the task effected by a language understanding system consists of recognizing the concepts that are evoked inside the input utterances. As soon as ambiguities are detected, they are resolved through interaction with the user.

U : Je prends le cube 1 (I am taking the cube 1)
S : S'agit-il du petit cube 1 ?
(Is the small cube 1 in question ?)
U : Oui (Yes it is)
S : O.K.

Relative clauses are not represented in the canonical form of the utterance in which they appear, but they are only used to determine which concept is in question.

article 1 - Nun 1 - Adjective 1 - Verb - article 2 - Adjec. 2 - Nun 2

abbreviated form : (V ((N1 A1)(N2 A2))) = semantic event E

relation rule n° i :

$$(R^i\ ((O^i_{11}\ P^i_{12})\ (O^i_{21}\ P^i_{21}))\quad ((O^i_{12}\ P^i_{12})\ (O^i_{22}\ P^i_{22}))\quad ((O^i_{1j}\ P^i_{1j})\ (O^i_{2j}\ P^i_{2j}))$$

Relation
Object
Property

$$E \text{ allowable} \iff \exists\,(i,j)\ /\ \forall\, k = 1, 2$$

$$\begin{cases} V \equiv R^i \\ N_k \in \mathcal{D}(O^i_{kj}) & (\mathcal{D} : \text{Descendants}) \\ P_{kj} \in \mathcal{P}(N_k) & (\mathcal{P} : \text{Property group}) \\ P_{kj}\ \mathcal{C}\ A_k & (\mathcal{C} : \text{Compatible with}) \end{cases}$$

Figure 4. Pattern-matching of a simple sentence

saisis les cubes 2 et 5 (grasp cubes 2 and 5)

prends le cube 2 et le 5
(take hold of the cube 2 and the 5 one)

attrape le cube 2 et saisis le cube 5
(lay hold of the cube 2 and grasp the cube 5)

⇩

canonical form :

(ASK((YOU)((TAKE((I)(CUBE(2))&(CUBE(5)))))))

Figure 5. Paraphrases translation

3. Orders

Input utterances beginning with an action verb specify an order that the machine connected to the vocal interface is supposed to execute ; in addition to the deep structure of this natural language message, a formal command language message is built and then sent to the machine. The task universe memory is modified in order to reflect the execution of a user's command.

User : Prends la pyramide qui est sur la table et pose la sur le gros cube (grasp the pyramid which is on the table and put it on the big cube)
Machine : S'agit-il du gros cube 3 ?
(are you talking of the big cube 3 ?)
User : Oui (yes)
Machine : D'accord (all right)

Deep structure

(ASK((YOU)((TAKE((I)(PYRAMID(SMALL)))))
&((PUT((I)((ON((PYRAMID(SMALL))(CUBE(GROS 3))

formal command : ((HOLD PYR2)(ON PYR2 CUB3))

User : Qu'ai je fait (What did I do)

Deep structure : (DO ((YOU) ?)

Machine : Tu as demandé que je prenne la petite pyramide et que je pose la petite pyramide sur le gros cube 3 (You have asked me to grasp the small pyramid and put the small pyramid on the big cube 3)
User : Qui a saisi un objet (Who has taken an object)

Deep structure : (TAKE (? (OBJ)))

Machine : J'ai pris la petite pyramide (I have grasped the small pyramid)
User : Qu'as-tu fait après ? (What have you done afterwards ?)

Deep structure : (DO ((I) ?))

Machine : J'ai posé la petite pyramide sur le gros cube 3 (I put the small pyramid on the big cube 3)

C. Question processing

In everyday language, intonation often contitutes the marker that discriminates between questions and assertions. Since prosody information is not presently taken into account by the word recognition system, the presence of an interrogative pronoun switches on the information research processing in permanent knowledge-data or in dynamic-events memory.

1. Research in permanent knowledge-data

The program is allowed to express its knowledge at the user's request, for instance, on concept meanings, or the systems abilities.

U : Qui es-tu ? (Who are you ?)
S : Je suis un ordinateur bavard (I am a talkative computer)
U : Puis-je parler avec toi ? (Can I speak with you ?)
S : Oui, c'est possible (Yes, that is possible)
U : Tu es une souris (You are a mouse)
S : Je pensais que j'étais un ordinateur (I thought that I was a computer)
U : C'est faux (That's false)
S : O.K.
U : Qui es-tu ? (Who are you ?)
S : Je suis une souris bavarde (I am a talkative mouse)
U : Est-ce qu'un chat peut parler avec toi ? (Is a cat able to speak with you ?)
S : Bien sûr que non (Of course not)

2. Research in dynamic-events data

The abbreviated semantic events list is closely examined, from recent to older data, until the question-pattern approximately matches one of the memorized events. Possible analogy between a memorized event and one evoked by the question is then analysed. Coincidences rarely happen, so the system must be able to ask for full specifications about the event that interests the user ; at that time there is a vocal discussion aimed at leading the system to that event in a step-wise manner.

U : Une femme blonde cherche une brosse (A blond woman is looking for a brush)
S : D'accord, je comprends (All right, I understand)
U : Je cherche un peigne (I am looking for a comb)
S : Très intéressant ... (Very interesting ...)
U : Quel être humain cherche quelque chose ? (What human being is looking for something ?)
S : S'agit-il de toi ? (Are you talking about yourself ?)
U : Non (No)
S : S'agit-il d'une femme blonde ? (Are you talking about a blond woman ?)
U : Oui (Yes)
S : Elle cherche une brosse (She is looking for a brush)

D. Processing a user's incomplete utterance

An important specific quality of the semantic process is that it is able to accomodate bad acoustical recognition through intelligent interactive feedback.

So, when one part of a given sentence has not been recognized, because of mispronunciation or background noise, the system produces a suitable question bringing the user to repeat the unrecognized word within his answer.

Two cases can occur :
if the word is again unrecognized, the system assumes that the entity is not in the prescribed vocabulary (containing the acoustic features of the words). An explanatory message is then produced through the synthesis module.
if the lexical entity is well recognized this time, it is added to the previous utterance and computed in the same manner as the others.

U : Je (?) un livre (I am (?) a book)
S : Que fais-tu avec le livre ? (What are you doing with the book)
U : Je le mange (I am eating it)
S : C'est impossible ... je ne te crois pas (It is impossible ... I do not believe you)
...
U : Une (?) femme boit du thé (A (?) woman is drinking tea)
S : Comment est la femme ? (What is the woman like ?)
U : Elle est grande (She is tall)
S : O.K.
U : Est-ce qu'une femme boit du thé ? (Is a woman drinking tea ?)
S : Oui, la grande femme (Yes, a fat woman is)
...
U : Un homme lit un gros (?) (A man is reading a thick (?))
S : Que lit-il ? (What is he reading ?)
U : Un gros livre (A thick book)
S : J'ai compris (I have understood)

U : Qui lit un livre ? (Who is reading a book ?)
S : Un homme lit un gros livre
(A man is reading a thick book)

When a certain amount of acoustical components in a sentence have not been recognized, the system asks for the user to repeat his assertion.

U : Le (?) (?) un petit (?)
S : Peux-tu répéter s'il te plait ?

E. Sentence production

1. Translation of a deep structure into an output sentence

This process consists of inserting semantic entities into the suitable syntactic diagram which depends on the computational procedure that is activated (question answering, contradiction, learning, asking for specifications ...). Since each syntactic variation of a word corresponds to a single semantic representation, sentence generation makes use of verb conjugation procedures and concordance procedures.

In order to improve the natural quality of speech, different types of sentences expressing one same idea may be generated in a pseudo-random manner. The same question asked to the system several times can thus induce different formulated responses.

2. Text-to-speech transcription ambiguities

A module of the synthesis process takes any French text and determines the elements necessary for the diphone synthesis, with the help of a dictionnary containing pronunciation rules and their exceptions (Prouts, 1979). However, some ambiguities concerning text-to-speech transcription can still remain and cannot be resolved without syntactico-semantic information ; for instance : "Les poules du couvent couvent" (the convent hens are sitting on their eggs) is pronounced by the synthesizer : / l ɛ p u l d y k u v ã k u v ã / (the convent hens convent).

To deal with that problem, we may send the synthesizer the phonetic form of the words.

IV CONCLUSION

The dialog experiment is presently running on a PDP 11/23 MINC and on an INTEL development system with a VLISP interpreter in real-time and using a series interface with the vocal terminal.

The isolated word recognition board we are using for the moment makes the user pause for approximately half a second between each word he pronounces. In the near future we plan to replace this module by a connected word system which will make the dialog more natural. It may be noted that the compactness of the understanding program allows its implantation on a microprocessor board which is to be inserted in the vocal terminal.

At present we apply ourselves to make the dialog-handling module easily adaptable to various domains of application.

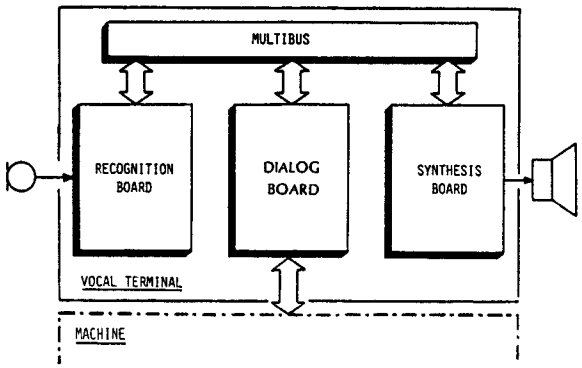

Figure 6. Multibus configuration of the Vocal Terminal

Acknowledgements

We are particulary grateful to Daniel MEMMI, Jean-Luc GAUVAIN and Joseph MARIANI for their precious help during the course of this work. Special thanks to Maxine ESKENAZI, Françoise NEEL and Michèle CHASTAGNER.

REFERENCES

V. ASTA, J.S. LIENARD - L'icophone logiciel : un synthétiseur par formes d'ondes - 10e JEP, Grenoble, 1979.

E. CHARNIAK, Y. WILKS (editors) - Computational Semantics - North-Holland, 1976.

A.M. COLLINS, M.R. QUILLIAN - Retrieval time from semantic memory - Journal of Verbal Learning and Verbal Behavior, 1969.

J.L. GAUVAIN - Reconnaissance de mots enchaînés et détection de mots dans la parole continue - Thèse 3e cycle, Orsay, 1982.

S.E. LEVINSON, K.L. SHIPLEY - A conversational system using speech input end output - The Bell System Technical Journal, vol. 59, n° 1, january 1980.

J.S. LIENARD, J.J. MARIANI - Système de reconnaissance de mots isolés : MOISE - Registred Technical Report ANVAR n° 50312, juin 1980.

D. MEMMI, J.J. MARIANI - ARBUS : A tool for developing application grammars - Coling, Prague, 1982.

F. NEEL, J.S. LIENARD, J.J. MARIANI - An experiment of vocal communication applied to computer-aided learning - IFIP WCCE81, juillet 1981.

B. PROUTS - Traduction phonétique de textes écrits en français - 10e JEP, Grenoble, 1979.

R. SCHANK - Conceptual information processing - North Holland, 1975.

T. WINOGRAD - Understanding natural language - Academic Press, 1972.

W.A. WOODS - Transition network grammar for natural language analysis - Communication of the ACM, vol. 13, n° 10, 1970.