A CONVERSATIONAL MODE SPEECH UNDERSTANDING SYSTEM

S.E. Levinson

Acoustics Research Department
Bell Laboratories, Murray Hill, N.J.

ABSTRACT

We describe a conversational mode speech understanding system which enables its user to make airline reservations and obtain timetable information through a spoken dialogue. The system is structured as a three level hierarchy consisting of an acoustic work recognizer, a syntax analyzer and a semantic processor. The semantic level controls an audio response system making two way speech communication possible. The system is highly robust and operates on-line in a few times real time on a laboratory minicomputer. The speech communication channel is a standard telephone set connected to the computer by an ordinary dialed-up line.

RÉSUMÉ

Nous décrivons un système de reconnaissance de la parole en mode conversationnel qui permet à l'utilisateur de réserver des places sur des avions ou d'obtenir des renseignements sur les horaires par voie de dialogue parlé. Ce système présente une structure hiérarchisée qui comporte trois niveaux, à savoir un organe de reconnaissance acoustique des mots, un analyseur syntaxique et une unité de traitement sémantique. Le niveau sémantique gère un système de réponse vocale qui rend le dialogue possible. Ce système est très résistant. En laboratoire, il est exploité en direct et parfois en temps réel sur un mini-ordinateur. La voie de transmission de la parole est un simple poste téléphonique relié à l'ordinateur par une ligne commutée ordinaire.

Recently there has been a substantial research effort in the area of speech understanding. Although the ultimate purpose of this work is clearly that of enabling natural spoken language human/machine communication, most of the work has actually been in the nature of building systems which transcribe speech. The system described in this paper is capable of conducting a complete spoken dialog with its user. The essential system architecture is hierarchical with three levels. These are an acoustic word recognizer, a syntax analyzer, and a semantic processor. The semantic level controls an audio response system which provides the speaking function. There is, of course, a significant interaction of the levels with one another.

The precursor to and incentive for this project was the flight information system built by Rosenberg and Itakura,[1] which consists of only a word recognizer and a voice response unit but which nonetheless is capable of conducting a limited dialog. The first two levels of the system and the voice response unit were taken in toto from previous investigations. The acoustic word recognizer was designed by Itakura[2] and is based on the computation of linear prediction coefficients (LPC), nonlinear time registration with stored reference patterns, and a maximum likelihood decision rule. The syntax analyzer is the maximum likelihood parser described by Levinson[3] and the voice response unit is based on an adaptive differential pulse code modulation (ADPCM) method of coding used by Rosenthal et al.[4] The unit uses the hardware coder/decoder of Bates.[5] These building blocks have been explained in detail by their original designers and will not be further described here. The interested reader is encouraged to consult the cited references for a complete discussion.

The goal of this project was to produce an on-line system which permits as nearly natural speech communication as possible. The system was to be robust, to understand the spoken input accurately over the standard telephone channel and respond quickly, and to require only moderate computational resources. To bring the goal within reach, certain constraints were deemed necessary. First, communication was restricted to pertain to a well-defined, limited subject. The flight information and reservation task is ideal for the purpose. It is a paradigm of the general data base information retrieval task for which natural language is appropriate. The tractability of the task domain allows us to impose some necessary restrictions on the flexibility of the language, limiting it to a small subset of natural English, which might otherwise be used for the purpose, generated by a formal grammar over a small vocabulary. Finally, we require that the input speech be disciplined in the sense that brief pauses between words are necessary. At the moment, the system must be trained by each of its users although this last restriction can be relaxed for a small additional increase in complexity.

Specifically, the flight information and reservation task includes 19 different semantic categories. Within each, several alternative and equivalent syntactic structures are permitted. The vocabulary of the speech recognizer is 127 words. The language is finite (regular), having 144 states and 450 transitions in its state diagram and contains over 6×10^9 sentences. A detailed specification of the language is given in Ref. 3. The voice response unit has a vocabulary of 191 words. Sentences are generated by a context-free grammar. The data base over which the system operates is the subset of the Official Airline Guide (OAG), pertaining to flights from New York to nine American cities.

A block diagram of the system is shown in Fig. 1. Speech input to the machine is in the form of a sentence, W. Brief pauses of approximately 100 ms between the words permit segmentation of the sentence. Each word is individually recognized by the minimum prediction residual principle of Itakura,[2] which provides an acoustic transcription of the input, $\tilde{W}$, and a distance matrix $[d_{ij}]$ whose ijth entry is a measure of the spectral distance between the ith word in the sentence and the reference

template for the jth vocabulary item. The parser takes this information and, using the technique described by Levinson,[3] produces the well-formed sentence, $\hat{W}$, having the minimum total distance. For efficiency, there is some communication between the acoustic and syntactic processing levels. Since the sentence is pre-segmented, the length of the sentence and the current word position can be given to the parser, which then returns a list of possible words to be matched to the input by the word recognizer. The parser also produces an explicit derivation of $\hat{W}$ in the form of a state sequence, Q. Since the grammar which generates the language is unambiguous, Q and $\hat{W}$ suffice (almost) to define the semantic meaning of the input.

The semantic processor takes Q and $\hat{W}$ and interprets them in the context of the conversation stored in the u-model to generate "actions" which involve searching the data base, altering the context of the conversation, and generating a response. Although it is not conceptually important, for the sake of completeness we point out that the model of the conversation is encoded in an internal representation which is different from the one in the data base. Therefore, a translation process takes place between the semantic analysis and storage in the model. The external form of an item is denoted c_i and its internal code u_i.

The data base search routine can take a set of items $\{c_i\}$ and either match it to a complete flight description, C, or determine in what way the set is insufficient. A complete flight description can be used to answer a question and/or update the u-model.

The response generator takes the current flight description, C, the recognized input sentence, $\hat{W}$, and a semantic code, K, and generates a reply to $\hat{W}$ from the grammar G_S. The reply is a string of symbols, $\mathcal{B}$, representing a sentence, one symbol per word.

A subroutine which controls the ADPCM hardware uses $\mathcal{B}$ to access a file of pre-coded isoalted words and concatenate them into the speech waveform x(t) of the desired reply. Details of the voice response hardware and software are given in Refs. 5 and 4, respectively.

The system requires four distinct pieces of hardware, a laboratory minicomputer which in the present implementation is a Data General S-230, a CSPI MAP array processor which performs some of the computation for word recognition, the ADPCM coder, and a data set which provides the interface between the computer and the telephone network.

All other functions are implemented in software. Since the address space of the S-230 is limited to 32K, the software configuration is an overlay structure with some communication via disk files. The individual overlays, the word reference templates, $R(\tau)$, the data base, and the speech files reside on disk files totaling a few hundred thousand bytes of storage. The input and output grammars and the semantic table are core-resident.

The performance of the system is highly encouraging. The accuracy of the speech recognition portion of the system was reported previously by Levinson et al.[6] to be over 96 percent on sentences. With the addition of the semantic processor, 6 of the 21 sentence errors encountered by one set of test sentences were corrected without intervention by the user. In the remaining 15 instances, the system recognized the error and it was corrected by the user on his next input sentence. In no case was communication seriously disrupted. This phenomenon has a profound effect on a user of the system. His attention is drawn away from speech recognition accuracy and sharply focused on the exchange of information between himself and the machine. This points very strongly to the conclusion that progress in speech recognition can be made by studying it in the context of communication rather than in a vacuum or as part of a one-way channel.

The response time is currently about five times real time but can easily be reduced. The naturalness of the system is low due to the discipline

required in speech input, and we are working to improve it. Overall, we are confident that our continuing efforts will result in increasing accuracy, flexibility, efficiency, and habitability of the system.

REFERENCES

1. A. E. Rosenberg and F. Itakura, "Evaluation of an Automatic Word Recognition System over Dialed-up Telephone Lines." J. Acoust. Soc. Amer., 60 supp. 1, S12, 1976.
2. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Trans. Acoust. Speech Sig. Proc., ASSP-23 (1975), pp. 67-72.
3. S. E. Levinson, "The Effects of Syntactic Analysis on Word Recognition Accuracy," B.S.T.J., 57, No. 5 (May-June 1978), pp. 1627-1644.
4. L. H. Rosenthal, L. R. Rabiner, R. W. Schafer, P. Cummiskey, and J. L. Flanagan, "A Multiline Computer Voice Response System Using ADPCM Coded Speech," IEEE Trans. Acoust. Speech Sig. Proc., ASSP-22 (1974), pp. 339-352.
5. S. L. Bates, "A Hardware Realization of a PCM-ADPCM Code Converter", Massachusette Institute of Technology, M.S. unpublished thesis, Cambridge, Massachusetts, 1976.
6. S. E. Levinson, A. E. Rosenberg, and J. L. Flanagan, "Evaluation of a Word Recognition System Using Syntax Analysis," B.S.T.J., 57, No. 6 (May-June 1978), pp. 1619-1626.

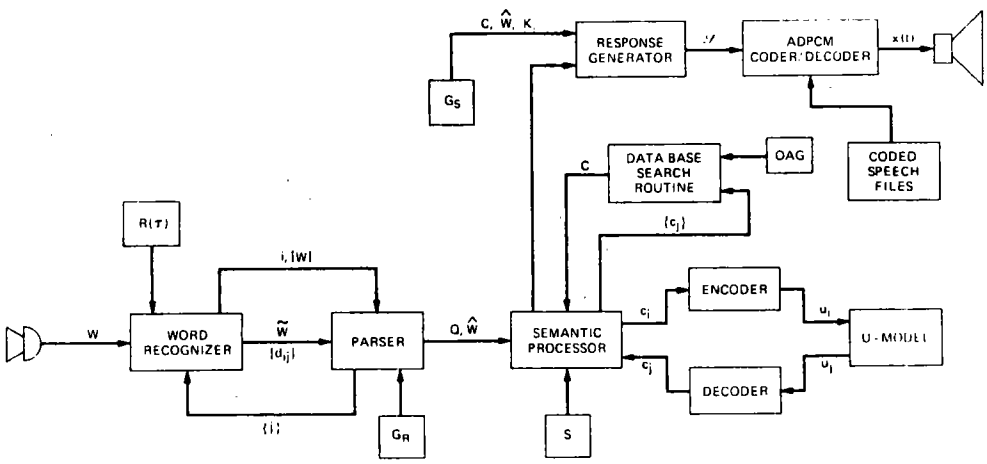

Fig. 1—Block diagram of the system.