



## SPEAKER-DEPENDENT CONTINUOUS SPEECH RECOGNITION WITH KEAL

G. Mercier, D. Bigorgne, L. Le Guennec, L. Miclet, J. Monne, M. Querre,  
J. Vaissière, M. Cloatre\*

### ABSTRACT

In this paper, a Speaker-dependent continuous speech recognition system is described. An unknown utterance is recognized by means of the following procedures : acoustic analysis, phonetic segmentation and identification, word and sentence recognition.

In order to adjust some of the system parameters, a speaker adaptation module is able to match known utterances with their acoustical representation.

The task to be performed is given as a parameter of the system and is described by its vocabulary and its grammar. Recognition results of continuously spoken sentences extracted from a "pseudo-Logo" language are presented.

### INTRODUCTION

This paper describes the KEAL experimental continuous speech recognition system. It is designed to be the recognition component of a man-machine dialogue system able to provide voice access to a data base (Ref. 4).

Roughly speaking, Keal is a hierarchical bottom-up knowledge-based speech recognition system (Ref. 2), able to recognize discrete or continuous utterances.

Four experimental tasks are currently under investigation : Recognition of french numbers, recognition of italian numbers (Ref. 3), recognition of pseudo-Logo commands and recognition of spoken requests to a small subset of an electronic yellow pages data base.

In this paper, we describe recognition experiments of spoken sentences extracted from the "Logo" task domain. "Logo" is a well-known natural programming language used to teach computer science to children.

These experiments are described in the following sections. In section 2, the main components of Keal are recalled. In section 3, the vocabulary and the grammar of the application are described. In section 4, the experimental conditions and the training and test sets are presented. In section 5, recognition results and conclusions are discussed.

---

\*C.N.E.T. LAA/TSS/RCP, Route de Trégastel, 22301 LANNION (France).

## MAIN COMPONENTS OF KEAL

Keal integrates different knowledge sources :

- ACOUSTIC ANALYSIS is carried out by an n-channel, mel-scaled vocoder (n is a parameter of the system, usually 14, 16 or 29) ; this acoustic spectrum is computed every 13.3 ms or 10 ms.
- PHONETIC RECOGNITION is decomposed into a sequence of step activations: sentence onset detection, centisecond labelling, segmentation into pseudo-syllables and phones, primary phonetic feature recognition and phone recognition. A speaker-independent set of rules and speaker-dependent linear decision functions are used for this acoustic-phonetic decoding.  
A sequence of segments (called a phonetic spectrum or lattice) is provided by this module. Each of these segments has a finite number of attributes, for instance, a weight showing its probability of existence and a collection of possible phonemes ordered in decreasing likelihood.
- From the phonetic spectrum, LEXICAL ANALYSIS formulates hypotheses of words. It uses a dictionary which contains various phonetic transcriptions of each word of the application. Each of these transcriptions is matched against the phonetic spectrum by means of a dynamic programming algorithm. The output of this module is a lexical lattice including the set of the detected words in association with their limits and their similarity index.  
This dynamic programming algorithm can be optionally associated with a word verification module, using a set of phonetic and prosodic rules in order to modify the word matching procedure.
- SENTENCE RECOGNITION is performed by a left-to-right bottom-up parser using a context-free grammar. The words forming the sentence must satisfy different types of constraints such as grammaticality, adjacency- (two consecutive words must be not too far each from other and the overlapping, if any, must be short) and acoustic evidence (true detections show higher evaluations than false detections).  
In order to avoid looking through the whole lexical lattice, only the best partial solutions are kept, at each step, according to a beam-search strategy. The result of the parsing, if any, is the recognized sentence associated with a score and a parse-tree.
- The SPEAKER-ADAPTATION module is composed of two modules :
  - . a semi-automatic alignment program which maps an ideal phonemic transcription of words or sentences composing the reference data set into the acoustic-phonetic spectrum given by the system ;
  - . a stochastic approximation program which provides the optimal set of coefficients of the linear discriminant functions, obtained after computing the optimal hyperplanes dividing the reference data set of phones.

## TASK DESCRIPTION

Each word of the vocabulary is described by means of a sequence of basic units. At the lexical level, the basic units of the Keal system are the phonemes. Each word has usually a standard phonemic form and eventually one, two or more phonemic variants. Some sequences of phonemes, like the

diphone /w a/, being difficult to separate into two different sounds, are considered as basic units. In this case, this particular diphone is named /oi/ like in the word /s ois an t/ ("soixante" or "60").

The lexicon is composed of 122 basic words, from which each sentence can be generated by means of a context-free grammar. This dictionary is speaker-independent.

At the phonetic level, the set of phonemes is divided into two main classes : vowels and consonants (some phonemes like /r/ can belong to both classes).

Within each class, some allophonic units are introduced in order to have homogeneous classes of phonemes ; for instance, the following phonemes /r/, /l/, /a/ and /o/ have two allophones. Each stop consonant is itself divided into two allophones corresponding respectively to the occlusion part and to the burst part ; these allophones correspond to the following set : /p/, /p1/ ; /t/, /t1/ ; /k/, /k1/ ; /b/, /b1/ ; /d/, /d1/ ; /g/, /g1/.

The context-free grammar of Logo sentences is composed of 32 non-terminal or syntactic categories, 76 terminal or lexical categories and 100 rules. The language has a branching factor of 25.

## EXPERIMENTS

### 4.1. Speech data

Speech collected from six speakers (3 female and 3 male speakers) was recorded, digitized at a sampling rate of 12.8 khz and F.F.T. was carried out, each 13.3 ms. The log-energy in each of 14 critical bands was computed. This speech data was divided into two sets :

- The TRAINING set was composed of 73 short task-independent sentences uttered once by each speaker.
- The TEST set is composed of 39 logo sentences per speaker.

### 4.2. Speaker training experiment

For each speaker, a reference set of phones is automatically extracted from the TRAINING set by means of the automatic alignment program which maps each phonetic lattice given by the phonetic analyser on the ideal phonetic transcription of each training sentence.

This phonetic reference set is then checked or improved by hand, in case of mapping errors. From this set, two new sets of coefficients (corresponding respectively to vowels and consonants), representing the optimal hyperplanes separating the phones in each of these two classes are computed. Accordingly for each speaker, at the end of this training time, the system is provided with these two sets of speaker-dependent coefficients used to refine the first phonetic recognition given by the rule-based part of the acoustic-phonetic decoding module.

## RESULTS

The percentage of correct sentence recognition obtained on the test set is 80 % (187 sentences out of 234). At the word level the recognition percentage is 89 % (806 out of 903). These results are shown in table 1. At the phonetic level, 80 % of the phonemes are identified within the first four candidates, without using phonotactic or lexical or syntactic constraints.

Speakers	Female			Male		
	AB	MG	DD	CG	JM	RV
Word Recognition Percentage	90	92	93	92	81	90
Sentence Recognition Percentage	79	82	87	79	77	74

Table 1 Speaker-dependent correct word and sentence recognition for 6 speakers

#### CONCLUSIONS

These results show that reasonable continuous speech recognition performance can be obtained, even with a partially rule-based, bottom-up approach. Of course, different ways are possible to improve these results :

The phonetic lattice will be refined by adding new contextual rules and by improving the set of basic allophones. We hope also to improve the word recognition level by including a statistical error model for phonetic errors (confusions, deletions and omissions) which are automatically estimated by the alignment module.

The current context-free grammar of the "Logo" language is not optimal and a new active chart parser using A.T.N. grammars is under test (Ref. 1).

#### REFERENCES

1. COZANNET A., 1986 : "ALOEMDA, analyseur linguistique pour l'oral et l'écrit par la méthode des diagrammes actifs", P1015 ESPRIT project Palabre Rome meeting, report IM22010.
2. LEA W.A., Ed, 1980 : "Trends in Automatic Speech Recognition", Englewoods Cliffs, Prentice-Hall, New York.
3. MERCIER G., LE GUENNEC L., LAFACE P., 1987 : "Recognition of italian numbers and connected digits", (deliverable), P 1015 ESPRIT project (Palabre).
4. SIROUX J., GILLET D., 1985 : "A system for man-machine communication using speech", Speech communication, Vol 4, N° 4 pp 289-315.