

LARGE VOCABULARY ISOLATED WORD RECOGNITION: A REAL-TIME IMPLEMENTATION

C.Vicenzi, C.Favareto, D.Sciarra, A.Carossino, A.M.Colla, C.Scagliola,
P.Pedrazzi. (*)

ABSTRACT

In this paper a Large-Vocabulary, Real-Time, Isolated Word Recognition system is presented. While the final goal of the project is that of recognizing words from a vocabulary whose size is of the order of 10-20 thousands, the present system is intended to perform real-time recognition on vocabulary subsets (cohorts) of up to 2 thousand words. The system is implemented on an EMMA2 multiprocessor for a fast response. Basic features of the system are the use of sub-word units (diphones) for the acoustic measurements and the derivation of synthetic symbolic word templates directly from the required lexicon.

INTRODUCTION

In view of the realization of a word recognizer able to identify one word out of a 10-20 thousand words vocabulary, the present work has the goal of achieving both real-time operation and an optimal accuracy over quite large subsets of the complete vocabulary. Actually, run-time detection of these subsets (or cohorts) should be done by a following system version performing a sort of pre-classification of the input speech into some acoustic gross features. In this paper we are going to describe a real-time implementation of isolated word recognition applied to a cohort (the "vocabulary" hereinafter) whose size could range up to 2 thousand words. To obtain fast comparison of the unknown word with each entry in the vocabulary, the system is implemented on an EMMA2 multiprocessor (ref 1); the hardware/software realization is extremely modular and thus computing resources can be dynamically allocated to the various recognition processes involved. Parallel architecture and efficient implementation of the analysis and matching algorithms ensure fast response times (about 1 sec in worst cases). Multiplexing of a certain number of speech channels on the same hardware is possible too.

The language is represented in terms of diphones (ref 2): these are sub-word units describing both transient and steady-state sounds. They are economical (about 300 units for the Italian language), almost invariant with the context and require only few minutes of speech material for training a new speaker to any application. Each word is then represented by one or more prototypes, each one being a sequence of unit labels with associated duration bounds (ref 3).

Results obtained from preliminary tests carried out with vocabularies of 686 and 1392 words respectively are reported in a following section.

SYSTEM DESCRIPTION

Our system can be represented (Fig.1) as a sequence of six modules: Acoustic Analysis, End-point Detection, Diphone Spotting, Classification or Labelling, Word Decoding and, finally, Reliability Evaluation. The first module attends to the real-time speech signal analysis to feed an on-line automatic word end-point detector; the next two modules apply the acoustic knowledge about the current speaker in order to compute the dissimilarity pattern of the speech units with respect to input stream. The decoding module then performs the recognition phase by matching such pattern against the current vocabulary description and sorting the results. The last module evaluates the confidence of each recognition and it was inserted in the system in order to provide a robust output interface to any applicative program using the recognizer.

In the following sections each module will be described in more detail.

Language Model - The use of sub-word units is one of the practical solutions to the problem of recognizing large vocabularies, allowing an easier and faster training phase at least in a speaker-dependent mode (ref 3). We chose to represent the vocabulary words through sequences of labels with associated duration hypotheses (minimum and maximum) which depend on the context; such compact descriptions are derived by rule from the lexicon.

(*) ELSAG S.p.A. - Via G.Puccini,2 - 16154 GENOVA (ITALY)

Each label then is acoustically described by one or more diphones and its distance is assumed to be the smallest among those computed on the relevant diphone templates. Such diphones in our definition include transient portions between adjacent sounds (sometimes comprising three-sounds events) and stationary sounds represented by single spectral state templates. The training session for a new speaker is fast and simple, consisting of the utterance of a few connected sentences (about 5' of speech); diphone occurrences are extracted by means of a "forced" strategy in a connected speech diphone-based recognizer. A template set of about 540 speaker-dependent elements, with an average length of 3.1 frames and representing about 300 Italian diphones, is automatically extracted from this material; such a set is representative of all the Italian diphones, thus being suitable for any application. Besides, cepstral descriptions of background noise are continuously extracted during the recognition sessions, in order to improve the matching accuracy. One or more templates have been provided for each word, in order to allow for phonetic or pronunciation variations.

Acoustic Analysis - The input speech is collected by a close-talking microphone, sampled at a 10 KHz rate, quantized on 12 bits and then pre-emphasized. Every 12.8 ms a Hamming frame 25.6 ms long is taken; from each frame 16 Mel Based Cepstral parameters are computed, plus an extra one representing the total energy in the frame. Each parameter is finally scaled to an 8 bit value.

End-point Detection - This module performs detection of the speech segments from the input stream of acoustic parameters and it has been designed to work also in a high level-varying environmental noise. Each new incoming energy value is used to update the current best segmentation hypotheses; primary choice is made on the basis of the energy contour in a time window of approximately 100 msec. Word endpointing is accomplished by means of time separation thresholds among the candidate speech segments found, and then refined by investigating energy local behaviour before and after the first and last segment respectively. Energy thresholds are made to be self-adaptive with some range constraints.

Diphone Spotting and Classification - The Diphone Spotting module continuously measures the distance between each input frame (of the end-pointed sequence) and all the diphone templates. The distance measure between two spectral states is computed as a Chebyshev distance plus a penalty that is function of the energetic difference between the two frames. A simple frame-to-frame distance can be adopted when templates are one frame long, while for a multiple-frame template the distance from an input frame is the minimum normalized distance between the template itself and each of the same length portions of the input sequence comprising that frame. A "frame vs. diphone distance" array, which is vocabulary independent, is produced.

Subsequently, the Classification module operates on the above array, by assigning to each of the predefined labels the minimum among the distances measured on the associated diphones. In this implementation the labels almost correspond to the diphones themselves, except for labels describing the possibility of alternative pronunciations (i.e. different /S/ or /Z/ sounds) or the recovery of some diphone spotting failures (i.e. weak voice bar versus silence). The result of this computational phase thus becomes a "frame vs. label distance" array.

Word Matching - The algorithm adopted here consists of an exhaustive matching between the input frame/label array versus each word template; the matching strategy adopted, in spite of the sub-word environment, is actually a word matching one; each template is in practice expanded into a true word template before the application of a specific Dynamic Programming technique. Such an expansion makes it easier the adoption of a regular and fast algorithm for the propagation of the best path through the template (ref 3). The local DP constraints adopted allow neither duplication nor skip of any state (no time warping) while the path is in the "minimum duration" portion of each label contained in the word template; states in the following portion can instead be skipped to directly access the first state of the next label. The label is used as a pointer to the frame/label distance array, so each best path calculation reduces to a sequence of table accesses and sums.

The choice of a word matching strategy leads to fast recognitions with constant work memory requirements; moreover, it generates an ordered list of word candidates instead of a single "best" candidate.

Reliability Evaluation - The output from the decoding phase consists of an ordered list of the words most similar to the input speech. In order to evaluate the minimum list depth with almost 100% probability of including the right word, a reliability algorithm examines the ordered sequence of the ratios $R(k) = D(1)/D(k)$ for $k=2,\dots,K$, where $D(k)$'s are the accumulated scores and K is the length of available list. Comparison of $R(k)$ sequence with a statistically generated reference one allows automatic shortening of the list under predictable error conditions.

MULTIPROCESSOR IMPLEMENTATION

The previously described recognition system has been implemented as shown in Fig.2. A host monitor, which runs on a mainframe computer, controls both an acoustic Front-End processor (FE) and a multiprocessor unit EMMA2(*) whose task is the LV-IWR one. The FE processor, built around a Zilog Z80 uP and a couple of TI TMS320 DSP's (ref 4), sends cepstral coefficients of speech to the EMMA2 receiver along a 56 kbit/sec serial line. EMMA2 is a company manufactured powerful multiprocessor that can be configured with a high degree of modularity. In the current implementation of the system a substantially single bus architecture has been chosen; on this bus (a "family" bus) the following boards co-work: (A) a "P1" board, which contains a single Processing Element (PE) and acts as the Master of the family; (B) an adequate quantity of "PN" boards, each one equipped with 3 independent PE's, which constitute the computational resources of the system; (C) an "HCSM" high capacity memory board which is used as a mass storage unit for the large static and dynamic data structures involved in the system. Each PE has a "private" environment (memories, I/O) and also a "local" one. Local memories are shared on the family bus and are used as communication areas among the processors. Interrupts and broadcasting facilities are also available in EMMA2 architecture to provide synchronization and faster data movements. Speech parameters are collected by the Master process which performs real-time automatic end-point detection as a background task; it also controls synchronization with host activity by receiving its commands and sending back recognition results and status reports. The most important task it performs is anyway the management of the "slave" PE's of the family: the Master organizes them into two functional groups, the Spotting and the Decoding ones, and then, during the recognition phase, controls their proper sequential executions and checks the flow of their own I/O data. The Spotting group is composed of a suitable number of PE's and is activated by the reception of the current end-pointed frame sequence which is broadcasted by the Master process. Each PE then matches all the diphone acoustic templates against a proper portion of the sequence; spectral distance measure is fast accomplished by means of a co-processing gate array chip which can perform, besides other functions, the Chebyshev distance computation. The complete frame vs. label array previously mentioned becomes then automatically reconstructed in the family shared memory, becoming available to the next Decoding process. Signalling of the completion of such a phase causes the Master process to awake the Decoding (word matching) group, whose population can be easily tailored as needed. Each PE in this group executes the comparisons on a proper subset of the current vocabulary (retrieving it from the mass storage board) and then releases in the family memory a final (partial) list of its own best N candidates. Finally, the Master process checks the completion of such activities, gathers partial results, sorts them and then sends final top-N list back to the host for further interpretation. A schematic diagram of the EMMA2 implementation is given in Fig.3; data buffering between consecutive processes is realized in the shared family memory while almost all of the large static knowledge is kept in the HCSM board and retrieved only for initialization purposes.

EXPERIMENTAL RESULTS

The presented system is now being tested in a relatively small hardware configuration with 6 PE's for the Spotting group and 12 for the Decoding one, yet obtaining almost real-time responses. Two sets of on-line tests have been carried out in a high noise environment, one with a 686 words vocabulary (818 templates) and another with a 1392 words one (1683 templates). A first speaker obtained 85.4% and 85% in word recognition rates respectively, while a second one obtained 91.7% and 88.7% rates.

(*) Patent pending.

CONCLUSIONS

In the present work, an IWR system able to handle quite large cohorts of word templates in real-time was developed. This goal is achieved by means of a modular EMMA2 multiprocessor on which different specialized tasks are implemented. A general purpose language modelling and a flexible system design ensure easy adaptation to any Large Vocabulary isolated word application.

On-line preliminary tests, performed in high background noise on moderately large cohorts, gave promising results in terms of time response and also in recognition accuracy, although no linguistic information was used.

REFERENCES

- 1- E.Appiani, G.Barbagelata, F.Cavagnaro, B.Conterno, R.Manara, "EMMA2 : An Industry-Developed Hierarchical Multiprocessor for Very High Performance Signal Processing Applications", 1st Int. Conf. on Supercomputing Systems, St.Petersburg, Florida, pp.310-319, Dec.1985.
- 2- A.M.Colla, C.Scagliola, D.Sciarra, "A Connected Speech Recognition System Using a Diphone-Based Language Model", Proc. ICASSP 1985 (31.9), Tampa, Florida, 1985.
- 3- C.Vicenzi, D.Sciarra, "Using Diphones in Large Vocabulary Word Recognition", Montreal Symposium on Speech Recognition, Montreal, July 1987.
- 4- M.Cavazza, A.Ciaramella, R.Pacifici, "Implementation of an Acoustic Front-End for Speech Recognition", CSELT R.D. 84.274, 1984.

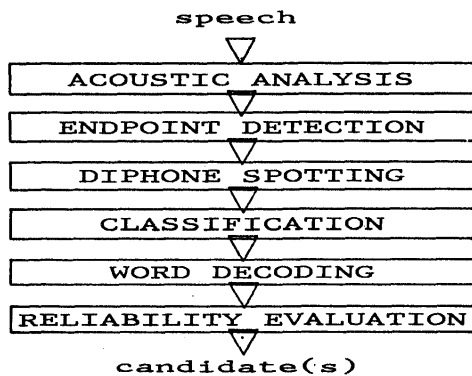


Fig.1 - The system modules.

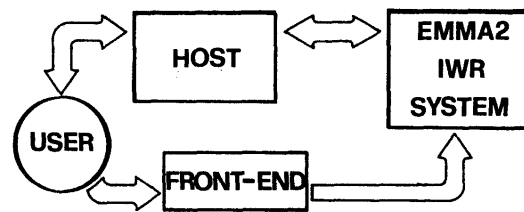
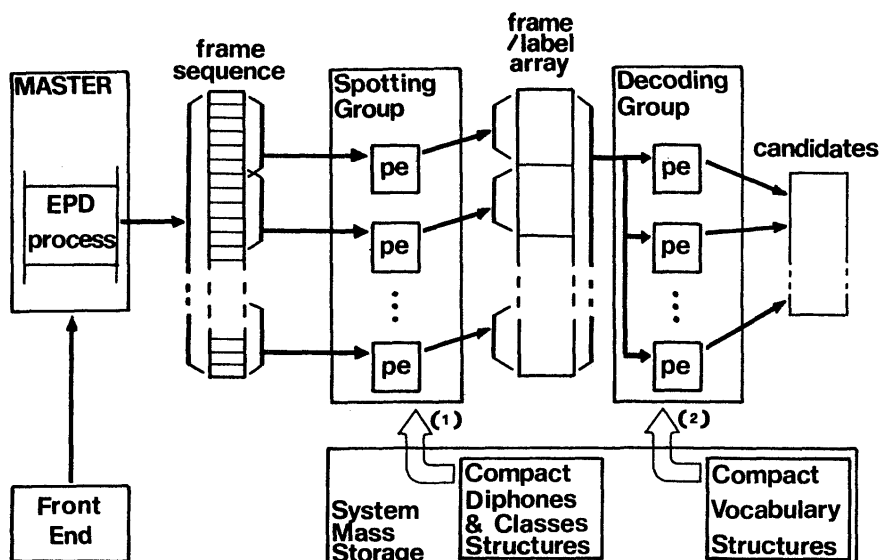


Fig.2 - Basic architecture of the system.



(1) only at start-up or speaker change.
 (2) only at start-up or vocabulary change.

Fig.3 - Data flow and processes in the system.