An Expert System for Interpreting Speech Patterns

Renato De Mori°, Attilio Giordana°, Pietro Laface§, and Lorenza Saitta°

° Istituto di Scienze dell'Informazione
Università di Torino
Corso Massimo d'Azeglio 42
10142 -TORINO Italy

§Istituto di Elettrotecnica Generale _ CENS
Politecnico di Torino
Corso Duca degli Abruzzi 24
10129 Torino Italy

ABSTRACT

Efficient syllabic hypothesization in continuous speech has been so far an unsolved problem. A novel solution based on the extraction of acoustic cues is proposed in this paper. This extraction is performed by parallel processes implementing an expert system represented by a grammar of frames.

1. INTRODUCTION

Central to the organization of a Speech Understanding System (SUS) are the representation of knowledge structured on several levels of abstraction, and the control strategy that has to use the knowledge efficiently. This paper introduces a general framework for interpreting speech patterns and describes a set of rules which have been succesfully applied in a task-independent multi-speaker system for speech decoding.

In principle, the system should be capable of accepting any sentence of any speaker in any language. For every analyzed sentence it produces a lattice of structured phonetic hypotheses. These hypotheses are obtained using relations between phonetic features and acoustic cues.

In the present implementation the system contains a set of rules which have been tested extensively giving good results for the Italian Language. Surprisingly, good results were also obtained with limited tests on the English and the Japanese Languages. The interesting aspect of the system is that the present set of rules can be considered as a kernel which can be enriched as new knowledge is acquired. Knowledge updating is presently performed by the designers, but it is hoped that some automatic learning will be introduced in thefuture.

Particular care has been taken in selecting rules which use robust, easily detectable and possibly speaker-invariant acoustic cues.

A frame language is proposed which describes a planning system for controlling rule application.

2. RELATION BETWEEN PHONETIC FEATURES AND ACOUSTIC CUES.

A phoneme is represented by a set of phonetic features. For example, the phoneme /g/ is represented by the following set:

/g/ = "nonsonorant-interrupted-consonant(NI)-
, lax(L), palatal(P)"

The phonetic features NI and L are related to acoustic cues by context-independent rules while P is involved in a relation in which also the context is taken into account. For the sake of brevity, the rules will be introduced with an example.

The phonetic feature "palatal" is involved in a relation with the acoustic cues "compact--burst-spectrum", "pseudo-loci" and "slopes of the second formant transition".The algebraic relation between the phonetic feature and the acoustic cues depends on wheter it is associated with the feature "tense" or "lax". In both cases the relation has the following general form:

"palatal" = p1."pal-pseudo-loci"."pal-slopes" +
p2."compact-burst" +
p3."compact-burst" . "pal-pseudo-loci".
"pal-slopes" ;

p1, p2, p3 are measures of the importance of the logical conjunction (indicated by a dot) following them in the relation and + indicates logical disjunction.

The acoustic cues "pal-pseudo-loci" and "pal-slopes" are defined by other relations involving judgements expressed on parameters contained in the detailed descriptions of the acoustic cues. Let F2B and F3B be the pseudo-loci of the second and third formant before the plosive and let F2A, F3A be the pseudo-loci just after the consonant burst, "pal-pseudo-loci" is defined as follows, in conjunction with the feature "lax" in a single intervocalic nonsonorant consonant:

"pal-pseudo-loci" = p4."high-pseudo-loci before" +
p5."high-pseudo-loci after" +
p6."high-pseudo-loci before" .
"high-pseudo-loci after" ;

where "high-pseudo-loci before" is a fuzzy set defined in the plane of the coordinates F2B, F3B and "high-pseudo-loci after" is another fuzzy set defined in the plane of the coordinates F2A, F3A.

Analogously, "pal-slopes" is defined as follows:

"pal-slopes" = p7."rising SB" + p8."falling SA" +
p9 . "rising SB" . "falling SA" ;

SB is the slope of the second formant transition before the consonant and SA is the slope of the second formant transition after the burst.

Context dependencies are limited to pseudo-syllabic segments. The detection of bounds of pseudo-syllabic segments is a side effect of a Semantic Syntax Directed Translation (see Tai and Fu[1]) which generates primary phonetic hypotheses from acoustic cues. These hypotheses can be ambiguous. Primary phonetic hypotheses are:

VF	: front vowel
VB	: back vowel
VC	: central vowel
SON	: sonorant consonant
SNCL	: cluster of sonorant consonants
NI	: nonsonorant interrupted consonant
NA	: nonsonorant affricate consonant
NC	: nonsonorant continuant consonant .

Primary phonetic hypotheses are used as a preliminary constraint for lexical access with some similarity with a recent work by Shipman and Zue [2].

For the sake of brevity, the problem of generating lexical hypotheses won't be discussed in this paper. Nevertheless it is worth mentioning that lexical hypotheses may constrain the application of context-dependent relations at the syllabic level involving places of articulation. The extraction of acoustic cues used by these relations can be based on top-down predictions.

3. PARALLEL ALGORITHMS FOR GENERATING SYLLABIC HYPOTHESES

The extraction of acoustic cues and the application of relations is performed under the control of an expert system. The procedural knowledge as well as the structural one are integrated in a collection of frames described by a frame language.

The frame language and the the structure of the expert system allow a certain degree of parallelism to be achieved in speech decoding.The system has been conceived in a framework of distributed problem solving and has been simulated on a DEC VAX 11/780 computer.

Parallel algorithms are conceived by decomposing the task of hypothesis generation into a number of subtasks. Subtasks are accomplished by reasoning programs called experts. The main motivation for using a distributed model is that parallel execution of tasks can make the system work close to real-time facing ambiguous data and a very large variety of possible solutions.

Each expert uses some knowledge written in a Long Term Memory (LTM) and may write hypotheses or results of intermediate computations into a Short Term Memory (STM).

The extraction of reliable descriptions of the speech signal is a difficult task and is accomplished by a society of auditory experts.

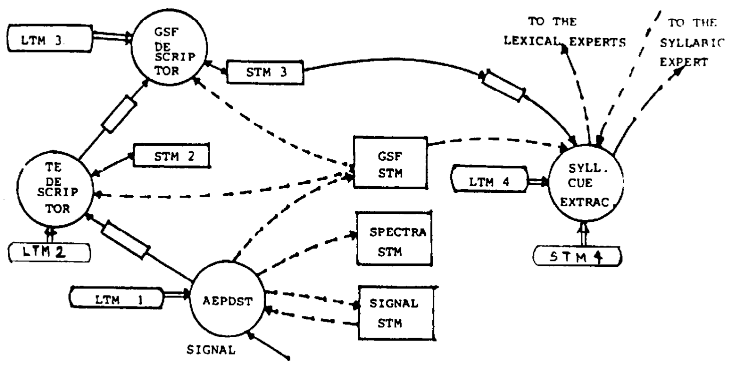

Fig. 1 - Experts of the auditory society

Fig. 1 shows the auditory experts.
Actions of writing into and reading from Short-Term Memories are represented by dashed arrows. Message passing between experts is represented by arrows. When a message contains pointers to a STM, a link is established between the arrow representing a message passing and the arrow representing the action of writing into the STM.

The speech signal is sampled, quantized, stored into a "SIGNAL-STM" and transformed by an expert called "Auditory Expert for End-Points Detection and Signal Transformation" (AEPDST). AEDPST looks for the starting point of a sentence by using a set of rules for end-points detection When this point has been detected, AEDPST starts transforming the signal in order to obtain a frequency-domain representation of it which is stored into the "SPECTRA-STM".
Some gross spectral features (GSF) are computed from the spectra and stored into the "GSF-STM".

The LTM of AEDPST, denoted LTM1 contains rules for end-point detection and spectral transformation.After a long enough part of the signal has been transformed, a synchronization signal is sent to the Expert for the description of the time evolution of the total signal energy (TE-DESCRIPTOR). TE-DESCRIPTOR has the task of describing the time evolution of the total energy of the signal (TE) in terms of peaks and valleys. At the same time, AEDPST goes on, transforms another portion of the signal and sends a message to TE-DESCRIPTOR. This operation is repeated until a sentence end-point is detected.

The LTM of TE-DESCRIPTOR, denoted LTM2, contains a grammar GTEDES that controls a coding of TE in terms of peak and valleys. This grammar and its use were described in [3].

Descriptions of the signal energy (TE) are sent to another expert called "GSF-DESCRIPTOR", which provides the acoustic cues for segmenting the sentence. These acoustic cues are sent to the "SYLLABIC-CUE-EXTRACTOR" which determines pseudo-syllabic bounds and extracts, sometimes upon re-

Table I
Rewriting rules of the frame-structure grammer

```
FRAME          := (‹ NAME› ‹ SLOT-LIST› )
SLOT-LIST      := ((‹SLOT› ((‹SLOT›) ) )
SLOT           := ((‹NAME› [( DESCRIPTION )]) )
DESCRIPTION    := ( described-as ‹ NAME ›)
                                          k)1
               := (( CONNECTIVE› ‹ DESCRIPTION›    )
               := (not ‹ DESCRIPTION› )
               := (filled-by ‹ NAME›)
               := ‹ CONDITIONAL›
               := (result-of ‹ PROC›)
CONDITIONAL    := (when ‹ NAME ›
                  ‹ DESCRIPTION› ‹ DESCRIPTION›
                  [ (else‹ DESCRIPTION› ) ])
               := (when ‹ PREDICATE EXPRESSION›
                     DESCRIPTION
                     (else  DESCRIPTION ) )
               := (unless  DESCRIPTION  DESCRIPTION )
CONNECTIVE     := or
               := and
               := xor
PREDICATE
EXPRESSION     :=  PREDICATE
               := ( not  PREDICATE )
                                       k)1
               := ( CONNECTIVE  PREDICATE      )
PROC           := F- function
               := P- procedure
```

quest from a syllabic expert, detailed acoustic cues to be used for pseudo-syllable hypothesization. Syllabic hypothesization is performed by a SYLLABIC EXPERT (SE) which receives lexical expectations and an unambiguous description of acoustic cues and sends syllabic hypotheses to the lexical level. These hypotheses are affected by degrees of plausibility.

The organization of knowledge stored into the LTM of the GSF-DESCRIPTOR is introduced in the next Section.

4. INTEGRATION OF STRUCTURAL AND PROCEDURAL KNOWLEDGE IN THE LONG TERM MEMORIES OF THE AUDITORY EXPERTS.

The LTM of GSF-DESCRIPTOR, denoted LTM3, contains an integration of the structural and procedural knowledge for obtaining a description of the gross spectral features represented by the time evolution of the following parameters:

TE : the total energy of the signal,
E12 : the energy in the 3 - 5 KHz frequency band,
R12 : the ratio of the energies in the frequency bands B1 = 0.2 - 0.9 KHz, B2 = 5 - 10 KHz.

The knowledge in LTM3 is a hierarchical network of plans represented by a grammar of frames. The network represents a control strategy according to which knowledge is applied for extracting acoustic cues from spectral information.

A frame is an information structure made of a frame-name and a number of slots. A slot is the holder of information concerning a particular item called "slot-filler" (Minsky [4]). Slot-fillers may be descriptions of events, relations or results of procedures. Attempts to fill the slots are made during a frame instantiation. A frame instantiation can be started by a simple reasoning program of an expert after having received a message. After a frame is instantiated , a copy of its LTM structure is created into the STM. At the beginning all the slots in the STM are empty and the expert which created the instantiation attempts to fill the slots sequentially. Frame structures are precisely defined by the rules of a grammar defining all the acceptable composition of the attibute relations. Table 1 shows the rules of this frame-structure grammar. The exponent K >1 of an expression means that the expression can be rewritten any number of times greater than 1. The asterisk means that the expression can be absent, present, or repeated any number of times. Brackets in Table 1 contain optional items which can be repeated any number of times.

The frame-structure grammar defines a language for representing LTM knowledge.

Table II contains a part of the description of the frames stored in the LTM of GSF-DESCRIPTOR. Predicates are indicated in capital letters by words ending with -P and are defined by semantic attachments which will be described informally. Functions are indicated by names starting with F-. Procedures are indicated by names starting with P-.Whenever the frame GSFDFR is instantiated a process for filling the frame slots is created along with a node in the output queue QUOUT.

Whenever a description of a total energy peak is received by the GSF-DESCRIPTOR, an instantiation of the frame PKTE is created into the STM of GSF-DESCRIPTOR by the attempt of filling the slot FRSTR of GSFDFR. The execution of the corresponding plan is then initiated. This process attempts to fill sequentially the slots of PKTE. Receiving dip descriptions causes the instantiation of a frame DPTE. PKTE and DPTE are complex

Table II
The LTM of GSF-DESCRIPTOR

```
(GSFDFR
         (INPUT result-of P-READ(PARAMETERS))
     (FRSTR (or (when PK-P(INPUT)
                         (filled-by PKTE))
                    (when DP-P(INPUT)
                         (filled-by DPTE))))
     (TERM(result-of P-APPEND(QUOUT))))
(PKTE      ; peak of total energy
     (INTPTE (result-of F-INT(INPUT)))
     (PEAKE12 (result-of F-DESCRPEAK(Fa,Fb,INTPTE)))
     (HR (result-of F-FHGR12(INTPTE)))
     (VINT (result-of F-CVINT(PEAKE12,HR)))
     (PCONT    (unless (filled-by (or
         (VOCPEAK)(SONPEAK)(NSPEAK)(BRSTPEAK)))
         (described-as UPK(INTPTE)))))

(VOCPEAK  ; Vocalic peak
         (WCONT (when (and (HDURPKTE-P)(HPR12P))
         (filled-by (or(VOCCUESET)
         ((LEFTVOW))(CONSVOW))))))
(VOCCUESET
     (LOWR (result-of F-FLOWR(INTPTE)))
     (TRNINT (result-of F-TRNF(INTPTE)))
     (VWINT (result-of F-INT(VCINT))
     (HGR (result-of F-CONSHR(INTPTE,VWINT)))
     (VCONT (filled-by (or(VOW)
         (CONSVOW)(VOWCONS)))))
```

structures. Attempting to fill their slots causes the extraction of acoustic cues.

INTPTE is filled by the result of the application of the function F-INT on the argument INPUT. This function gives the time of beginning, the time of ending and the duration ot the peak described in INPUT. INTPTE is written into the STM after PEAKTE in the instantiation of PKTE. The next slot of PKTE is filled by the result of the function F-DESCRPEAK which describes the peak of energy in the frequency band Fa-Fb and in the time interval written in INTPTE.Successively, the function F-FHGR12 fills the slot HR.It gives the description of the time intervals inside INTPTE in which the ratio R12 is high (greater than a threshold TH1).The function F-CVINT computes the time intervals inside the peaks in PEAKE12 in which R12 is high.The last plan of the sequence attempts to fill the slot PCONT. This slot can be filled by a disjunction of frame instantiations called VOCPEAK, SONPEAK, BRSTPEAK. Each invoked frame corresponds to a hierarchy of more detailed plans which are executed for attempting to fill the frame slots.If no frame instantiations can be completed, a default condition is assumed consisting in filling PCONT with the description UPK(INTPTE). UPK is the description of an uncertain peak detected in the time interval INTPTE.

A similar network of plans is used for attempting to fill the slots of DPTE.

The execution of more detailed plans for filling the slots of VOCPEAK is conditionned by the verification of the truth of the two predicates HDURPKTE-P and HPR12-P. HDURPKTE-P is true if the duration of the signal energy peak is high, HPR12-P is true if there is at least one peak of R12 in INTPTE whose maximum value is higher than a threshold TH2.VOCCUESET has slots which are filled by the extraction of acoustic cues which usually appear in a total energy peak containing at least one vowel. F-TRNF(INTPTE) extracts an interval at the beginning of a peak where cues of a consonantal transient, typical for example of plosive sounds, has been found. F-INT(VCINT) looks for the description of vocalic cues in VINT. The time interval in which these cues have been found fills the slot VWINT. F-FLOWR extracts the intervals in which R12 is low, F-CONSHR extracts the consonantal interval in which R12 is high. The default value for both the functions is zero. The predicate HGVINT-P is true when the maximum energy in the band Fa-Fb is high in the time interval VINT. AEQ-P(VWINT,INTPTE) is true when the two time intervals VWINT and INTPTE are almost coincident. If the above predicates are both true, the peak is described as a vocalic one and the description is VOC(VWINT).

Comments after the colons in Table II help in understanding frame instantiations. Frame instantiations in a conjunction or in a disjunction can be performed in parallel. Detailed acoustic cues are related to phonetic features which are hypothesized under model-driven constraints. A detailed description of the complete planning system and its frame based representation can be found in [5].

5. CONCLUSIONS.

A new model for representing and using the syllabic knowledge of a Speech Understanding System in terms of acoustic cues has been introduced.

Experiments on several hundreds of syllable uttered in continuous speech by four male and one female speakers gave the right interpretation with the highest evidence value in more than 90% of the cases. The results refer to syllables extracted from spoken sentences of the every day language. More than hundred syllables were analyzed for each talker.

The system has been simulated with a general purpose program for parallel problem solving. Preliminary results show that with an average degree of parallelism of 12, syllabic hypothesization, excluding signal processing, can be done in real-time using standard multi-microprocessors architectures.

6. REFERENCES.

1 - Tai,J.W. and Fu,K.S.,"Semantic Syntax-Directed Translation for Pictorial Pattern Recognition",Purdue University Report TR-EE 81-38, 1981.

2 - Shipman,D.W. and Zue,V.W.,"Properties of Large Lexicons: Implication for Advanced Isolated-Word Recognition Systems",Proc. ICASSP-82,Paris 1982,pp.546-549.

3 - De Mori, R.,"Computer Models of Speech Using Fuzzy Algorithms." New York: Plenum Press 1982.

4 - Minsky,M.,"A Framework for Representing Knowledge" In The Psychology of Computer Vision, Winston,P. Ed.,McGraw Hill, 1975.

5 - De Mori,R.,Giordana,A.,Laface,P. and Saitta L.,"Parallel Algorithms for Syllable Recognition in Continuous Speech", To appear on IEEE Transactions on Pattern Analysis and Machine Intelligence.