A SYSTEM OF PLANS FOR CONNECTED SPEECH RECOGNITION

Renato DE MORI, Yu F. MONG

Concordia University, Department of Computer Science,
1455, de Maisonneuve Blvd. Montreal, Quebec. H3G 1M8, Canada

Abstract

A planning system for recognising connected letters is described and some preliminary experimental results are reported.

1. Motivations and Relations with Previous Works

A number of researches on Automatic Speech Recognition (ASR) have been carried out using a recognition model based on feature extraction and classification. With such an approach, the same set of features are extracted at fixed time intervals (typically every 10 msecs.) and classification is based on distances between feature patterns and prototypes [LEVINSON 81] or likelihoods computed from a Markov model of a source of symbols generated by matching centisecond speech patterns and prototypes [BAHL 83].

These methods are usually speaker-dependent and are made speaker independent by clustering prototypes among many speakers. The classifier is not capable of making reliable decisions on phonemes or phonetic features, rather it may generate scored competing hypotheses that are combined together to form scored word and sentence candidates. If the protocol exhibits enough redundancy it is likely that the cumulative score of the right candidate is remarkably higher than the scores of competing candidates. If there is little redundancy in the protocols, like in the case of connected letters or digits or in the case of a large lexicon, then it is important that ambiguities at the phonetic level are solved before hypotheses are generated. Evidences of these difficulties are reported in recent literature [BAHL 84, RABINER 84]. For example, in the case of connected letters, in order to distinguish between /p/ and /t/ the place of articulation is the only distinctive feature and its detection may require the execution of special sensory procedures on a limited portion of the signal with a time resolution finer than 10 msec.

The need for a hierarchical application of recognition algorithms for plosive consonant recognition has been recently pointed out by many authors [DEMICHELIS 83,KOPEC 84]. This suggested that computer perception of speech can be modelled with a collection of operators for extracting and describing acoustic properties.Operator application is conditioned by the verification of some preconditions in the database that contains already generated descriptions of the signal under analysis. Sequences of operators belong to a system of plans where goals are the extraction and interpretation of various aspects of speech patterns. The input to the system is made of descriptions of acoustic properties obtained by hybrid (parametric and syntactic) pattern recognition algorithms and the outputs are hypotheses about phonetic features.

The recognition of unconstrained sequences of connected letters is a problem unsolved so far.

Using a redundant set of acoustic properties for characterising place and manner of articulation of some sounds makes it possible to have an accurate phoneme hypothesization even in difficult protocols. Nevertheless, the extraction of such properties requires the application of sequences of operators, some of them can be applied only if some preconditions are met. Useful sequences of operator applications is decided based on specific knowledge which establishes precedence relations depending on contextual constraints.

For example, burst properties are useful for hypothesising the place of articulation of plosive sounds [DEMICHELIS 83], but the operator that extracts them can be applied only after the successful application of another operator that detects and locates a plosive burst. Following a slightly different approach, Kopec [KOPEC 84] has shown that the point of consonant release has to be detected before applying efficient plosive recognition algorithms.

Preconditions for operator application are logical expressions of predicates. Predicates are defined over relations between acoustic properties. Both precondition expressions and predicate definitions are not known a priori and have to be learned. Learning is based on examples and is a search for plausible general descriptions of situations in which it is worth to apply an operator.

This paper describes the application of the planning concept and of AI learning methodologies to the conception of a system for the recognition of spoken connected letters belonging to the following set :

E1 := { P,T,K,B,D,G,C,V,E,3 }.

2. Overview of the System of Plans

The speech signal is first analysed on the basis of loudness, zero-crossing rates and broad-band energies using an expert system described in [DE MORI 82]. The result of this analysis is a string of symbols and attributes.

Symbols belong to an alphabet of Primary Acoustic Cues (PAC) whose definition is recalled in Table I.

A Semantic-Syntax Directed Translation scheme operates on PAC descriptions and through the use of sensory procedures identifies the vocalic and the consonantal segments of syllable nuclei and for the vocalic segments hypothesizes the place of articulation of the vowel

Plans are then applied for interpreting the consonantal segment of every syllable. An overview of the plan for the recognition of the E1-set is shown in Fig.1.

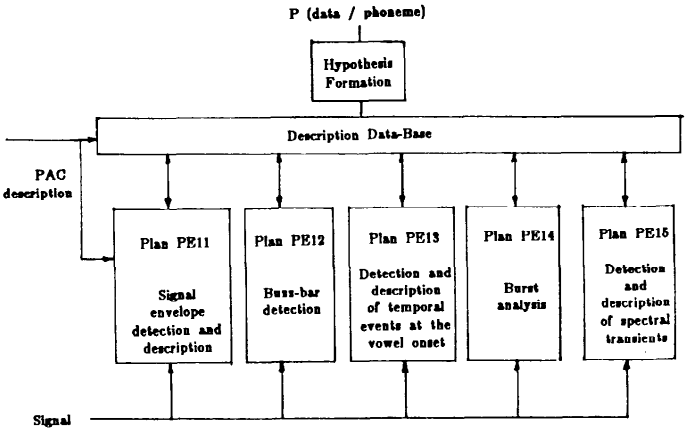

Fig. 1

The plan is subdivided into sub-plans (PE11,PE12,PE13,PE14,PE15).

PE11 produces an envelope description by analyzing the signal amplitude before and after preemphasis. Envelope samples are obtained every msec by taking the absolute value of the difference between the absolute maximum and the absolute minimum of the signal in a 3 msec interval. The envelope description is made by the following alphabet (~ represents negation) :

EDA = {SHORT—STEP(ST),
LONG—STEP(LST),
NO—STEP(NST),
STEP WITH HIGH LOW FREQUENCY ENERGY(BZ),
BURST—PEAK(BUR),
POSSIBLE—BURST(PBU),
NBZ=~BZ,
NBU =~BUR,
NPB=~PBU.}

PE12 detects a buzz-bar by analyzing the shape of FFT spectra before the voice onset. The alphabet of the descriptions it produces is :

BZA = {NOB,BU1,BU2,BU}

NOB means no buzz and the other three symbols describe degrees of buzz-bar evidence (BU1 : little evidence, BU : strong evidence)

PE13 analyzes temporal events at the voice onset. These events are related to voice onset time. They are :

D : the delay between the onset of low and high frequency energies,

ZQ : the duration of the largest zero-crossing interval of the signal at the onset,

ZR : the number of zero-crossing counts in the largest sequence of sucessive zero-crossing intervals with duration less than 0.5 msecs.

PE14 and PE15 perform respectively burst and formant transition analysis as described in [DEMICHELIS 83].

Preconditions for plan execution are learned with a general-purpose algorithm whose details are given in [DEMORI 84]. The highlights of this algorithm are summarized in the next section.

3. Learning Methodology

Learning rules from examples can be seen as the process of generalizing descriptions of positive and negative examples and previously learned rules to form new candidate rules. When applied incrementally this methodology can produce results which depend on the order in which examples are supplied and on the occurence of examples which are exceptions to the relevant rules. Incremental learning of rules has to come out with a set of rules that is the most consistent with the examples encountered so far.

In order to allow dynamic preservation of consistency among the set of rules, an algorithm has been conceived which uses the Truth Maintenance System formalism [DOYLE 79] and which is reminiscent of previous work by Whitehill [WHITEHILL 80]. The choice of a description language for examples and rules along with that of the generalizing algorithms is critical in a learning system in the sense that it may or may not allow the learning of relevant rules.

A description language and rule generalization heuristics have been defined based on knowledge about rule-based Automatic Speech Recognition (ASR). A relevant aspect of the learning system developped for ASR is that generalization rules are not constrained by the Maximally Common Generalization property introduced in [WHITEHILL 80]. Positive and negative facts used for learning operators preconditions are described by their relevant concept and a conjunction of predicate expressions. Each predicate expression or selector [MICHALSKI 83] asserts that an acoustic property has been detected or that an acoustic parameter has been extracted with some specified value. A generalization rule derives from two conjunctions C1 and C2 a conjunction C3 that is more general than both C1 and C2 i.e. C1 $\Rightarrow$C3 and C2 $\Rightarrow$C3.

The generalized rules themselves are the nodes of a TMS [DOYLE 79]. Each node represents a rule of left-hand-side (LHS) CONJ and right-hand-side (RHS) CONC, having a support list SL whose IN and OUT parts are respectively the list of nodes with RHS CONC and LHS less general than CONJ and the list of nodes with RHS different of CONC and LHS less general than CONC. With each node are kept the lists of consistent examples (PE for positive evidence) and unconsistent examples (NE for negative evidence). Lastly each node has a STATUS property which is IN when the corresponding rule is believed to be true and OUT

otherwise. A node is IN i.e. its STATUS is IN if and only if all the nodes in the IN part and all the nodes in the OUT part of its SL are respectively IN and OUT and the numbers of examples in PE and NE satisfy a given predicate P (for example NE $\geq$ 2.PE). As the numbers PE and NE keep changing during learning, a generalization can be true at a certain moment and it can become false later. This justifies the use of TMS.

When a new example is learned a new node is created if necessary and this node is generalized with the existing ones to generate new nodes that are themselves generalized with other ones. Then the PE and NE of concerned nodes are updated and STATUS properties are modified when necessary and propagated through the network in order to maintain consistency. The stability of this process is guaranted together by the definitions of SL and the predicate P.

For each concept encountered so far a characteristic rule is derived from the network of nodes whose LHS is the disjunction of LHSs of all IN nodes with corresponding RHS. Using the just outlined learning algorithm, the following precondition expressions PR_j $(1 \leq j \leq 5)$ have been inferred.

PR1 = (LDD + SDD) * (LPK + SPK + MNS LPK)
PR2 = LDD * (LPK + SPK)
PR3 = (LDD + SDD) * (LPK + SPK)
PR4 = SPK + BUR + PBU
PR5 = LDD + LPK

Notice that '+' represents logical disjunction and A*B means that A preceeds B in time.

The two curves in Fig. 2a represent the time evolution of the signal energy (—) and the zero-crossings counts (---) in successive intervals of 10 msec of the first derivative of the signal. The phrase is the sequence of letters and digit E3G PCB. Fig. 2b shows the corresponding PAC description. Time unit is 0.01 sec.

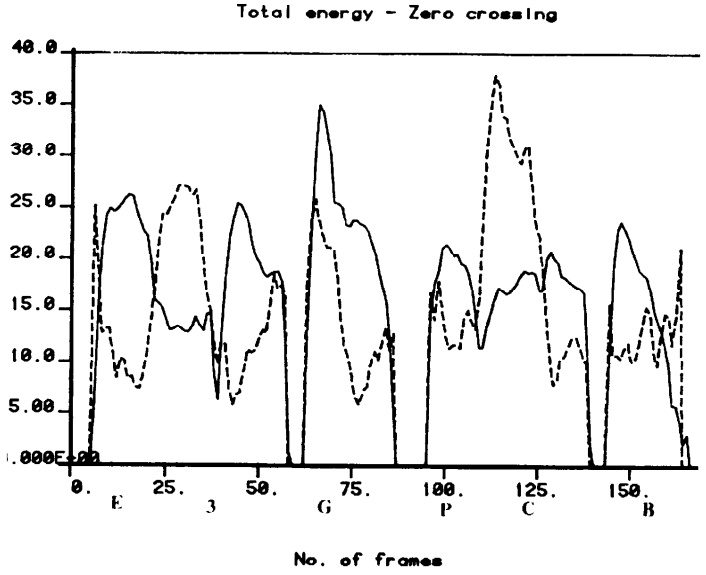

Fig. 2a

PAC	tb	te
LDD	1	7
LPK	8	28
LNS	29	38
SPK	39	41
LPK	42	60
SDD	61	65
MNS	66	72
LPK	73	89
LDD	90	96
LPK	99	111
SMD	112	113
LNS	114	126
LVI	127	141
SDD	142	145
LPK	146	168

Fig. 2b

4. Hypothesis Generation Rules

Expressions made of symbols extracted by subplans PE11 and PE12 and representing positive and negative information have been inferred for each PAC description and for each phoneme using the outlined learning algorithm.

An example of such rules is given in the following :

E := NOB NBZ NST NBU NPB
B := BU BZ NST NBU PBU

There are 96 of such rules in the system.

A PAC description is used for indexing a set of rules that is matched against the input description produced by the plan. As rules and descriptions contain the same number of symbols, a similarity index between a rule and a description is computed in closed form.

The parameters extracted by PE13, PE14 and PE15 are used in fuzzy relations. There is a fuzzy relation for each phoneme and the invocation of a fuzzy relation is conditioned by PAC, PE11 and PE12 descriptions. Fuzzy relations are conjunctions of disjunctions of fuzzy sets. A fuzzy relation computes another similarity index between phonemes and data. A fuzzy relation can be seen as a conjunction of clauses. Each clause contains a disjunction of fuzzy sets defined over a parameter extracted by the planning system. Fuzzy sets have been derived from a-priori probability distributions of parameters.

A similarity index is computed by using the max operator for disjunctions and by summing the contributions of each clause and then dividing the sum by the number of clauses.

An example of fuzzy relations is the following :

E := short D short ZQ low ZR
K := long D short ZQ high ZR

where "short, long, high, low"are fuzzy sets. There are 43 of such relations.

A-priori probabilities of the two similarity indices are inferred from experiments for every phoneme. These probabilities can be supplied to the language model for further preprocessing.

A simple recognition strategy based on similarity indices has been used for the experiment described in the next section. Its details are omitted for the sake of brevity.

5. Results and Conclusion

Experiments on 500 samples of sequences of symbols in the E1 set pronounced by two male and one female speakers have given an error rate of 0.5% in segmentation without requiring any speaker adaptation.

The proposed approach has been tested on a protocol of 400 connected pronounciations of symbols of the E1 set in strings of five symbols each. The strings were pronounced by one male speaker, the voice of which was used for deriving the rules. As the recognition algorithm is syllable based, it is not constrained by the number of syllables.

Performances are shown in Fig. 3 (speaker #1). The curve shows the error rates obtained in the following cases :

top 1 : there is an error when the right candidate is not ranked in the first position;

top 2 : there is an error when the right candidate is not ranked in the first two positions;

top 3 : there is an error when the right candidate is not ranked in the first three positions.

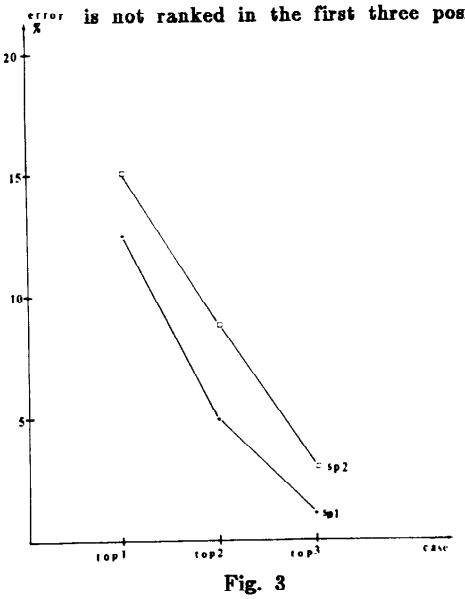

Fig. 3

Performances on the voice of a new male speaker are also shown by the curve labelled speaker #2. An analysis of the most frequent errors is summarized in Fig. 4.

The results seem interesting even if a large population of speakers has to be analyzed before deriving robust fuzzy sets capable of giving the same performances on different speakers.

Error Analysis in the Recognition of the E1 Set

Type of Error	Contribution to the Overall Error (%)
Confusion among p, t, k	13
Confusion among b, d	6
Confusion between cognate consonants	4
Confusion between b, v	13
Confusion between e and p or t and vice-versa	41
Confusion among g, c, 3	10
Other	3

Fig. 4

Nevertheless, the idea of using a number of phonetically significant properties in a recognition system based on the planning paradigm appears very promising. The analysis of the behavior of each plan and of the errors generated by their application suggests the actions that have to be taken in order to improve recognition accuracy.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada With grant no. A2439.

B. Delgutte, (CNET, Lannion A) suggested the introduction of temporal cues for characterizing plosive phonemes against vowels. A. Fron (ENST, Paris) wrote the program for extracting these cues. M.Gilloux (CNET Lannion A) wrote the learning program. The Authors wish to thank all of them.

References

[BAHL 83] Bahl L. R., Jelinek F., Mercer R. L. : A Maximum Likelihood Approach to Continuous Speech Recognition; IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-5, Num. 2, March 1983.

[BAHL 84] Bahl L. R., Das S. K., De Souza P. V., Jelinek F., Katz S., Mercer R. L., Pichenx M. A. : Some Experiments with Large-Vocabulary Isolated Word Sentence Recognition; Proc. IEEE Conference on Acoustic Speech and Signal Processing, San Diego, Col. 2651 - 2653.

[DEMICHELIS 83] Demichelis P., De Mori R., Laface P. and O'Kane M. : Computer Recognition of Plosive Sounds Using Contextual Information; IEEE Transactions on Acoustic Speech and Signal Processing, Vol. ASSP-31, Num. 2, p 359-377, April 1983.

[DE MORI 82] De Mori R., Giordana A., Laface P., Saitta L. : An Expert System for Interpreting Speech Patterns; Proc. of AAAI-82, p. 107-110, 1982.

[DE MORI 83] De Mori R. : Computer Models of Speech Using Fuzzy Algorithms; Plenum Press N.Y. 1983.

[DE MORI 84] De Mori R. and Gilloux M. : Inductive Learning of Phonetic Rules for Automatic Speech Recognition; Proc. CSCSI-84, London, Ontario, p. 103-106, 1984.

[DOYLE 79] Doyle J. : A Truth Maintenance System; Artificial Intelligence, Vol. 12, Num. 3, p 231-272, 1979.

[KOPEC 84] Kopec G. E. : Voiceless Stop Consonant Identification Using LPC Spectra; Proc. IEEE Conference on Acoustic Speech and Signal Processing, San Diego, Cal. 4211 - 4214.

[LEVINSON 81] Levinson S., Rabiner L. R. : Isolated and Connected Word Recognition Theory and Selected Applications; IEEE Trans. on Communications, Vol. COM-29, Num. 5, p. 621-659, May 1981.

[MICHALSKI 83] Michalski R. S. : A Theory and Methodology of Inductive Learning; in Machine Learning : an Artificial Intelligence Approach, Tioga, p. 83-134, 1983.

[RABINER 84] Rabiner L. R., Wilpon J. G., Terrace S. G. : A Directory Listing Retrieval System Based on Connected Letter Recognition; Proc. IEEE Conference on Acoustic Speech and Signal Processing, San Diego, Cal. 3541 - 3544.

[WHITEHILL 80] Whitehill S. B. : Self Correcting Generalization; Proc. of AAAI-80, p. 240-242, 1980.