USE OF PROCEDURAL KNOWLEDGE FOR AUTOMATIC SPEECH RECOGNITION

Renato DE MORI, Ettore MERLO, Mathew PALAKAL, and Jean ROUAT

McGill University
School of Computer Science
805 Sherbrooke St. W.
Montreal, Quebec, Canada
H3A 2K6

ABSTRACT

A paradigm for automatic speech recognition using networks of actions performing variable depth analysis is presented. The paradigm produces descriptions of speech properties that are related to speech units through Markov models representing system performance.

Preliminary results in the recognition of isolated letters and digits are presented.

1. INTRODUCTION

Recent results on Automatic Speech Recognition (ASR) and Speech Analysis suggest that progress in designing recognition devices and in advancing speech science knowledge may arise from an integration of the SO called cognitive and in formation-theoretic approaches [1].

The cognitive approach attempts to infer analytic knowledge about possible speech invariants and their relations. Work by Zue [2], Klatt [3], Stevens [4] and De Mori et al. [5,6] are along this line.

The information theoretic approach is based on a performance model containing states and transitions between any pair of states [7]. Probabilities can be learned that the system is in any of the model states or is changing state through any of the allowed transitions. Furthermore, the model generates in each state or in each transition, observable system parameters or descriptors according to some statistical distribution.

This paper proposes an attempt to integrate the two above mentioned approaches.

The idea is that of extracting speech properties using knowledge about acoustic correlates of linguistic units. For an abstract linguistic unit, the corresponding acoustic correlates have attributes which may differ from an instantiation to another due to the fact that different speakers produce different signals even if they intend to pronounce the same sound. Attribute statistics of sound properties collected on a large variety of pronounciations of the same sound from different speakers are probably the best knowledge we can gather today for characterizing different speaking styles. Furthermore, some expected acoustic properties can be missed in some cases and some unexpected properties can be detected in some other cases. These aspects can also be characterized by stochastic performance models.

An important problem arising when large vocabularies have to be recognized is that of identifying a possibly small set of Speech Units (SU) with which all the possible words and word concatenations can be obtained by com pi lotion. The learning problem is then reconducted to the conception of a performance model for each SU.

2. A MODEL FOR COMPUTER PERCEPTION OF SPEECH

The speech signal $x_1(t)$ is generated by a discrete and finite sequence of actions

$$A = a_1(t_1)a_2(t_2)a_3(t_3).....a_k(t_k).....a_K(t_K) \quad (1),$$

where $a_k(t_k)$ denotes an action ending at time t_k; $a_1(t_1)$ represents the silence preceding the beginning of a sentence.

When a person reads a sentence S, a relation

$$R_1(S, A) \quad (2)$$

is applied which produces A. The relation R_1 may depend on the speaker, his/her mood, state of health and history. As R_1 may produce several As for the same S, probability distributions for all the possible As can be derived using a generative model.

The speech signal $x_1(t)$ is generated by the sequence of actions A using another relation

$$R_2(A, x_1(t)) \quad (3).$$

R_2 depends on the anatomy of the speaker. Again, the same actions may produce different signals, because the speech production system is soft and its behavior is affected to some extent by the environment.

If the speaker does not read but generates a sentence from a set C of concepts, then a third relation is applied:

$$R_3(C, S) \quad (4).$$

R_3 may depend on the speaker and his/her culture. Statistical models can also be used for characterizing

3. PROCEDURAL NETWORKS

A Procedural Network *(PN)* can be described with a formalism similar to that used for an Augmented Transition Network Grammar *(ATNG)*. This formalism has been successfully used for Natural Language and Pattern Recognition [8]. A PN is a 5-tuple

$$PN = \{j, Q, A, q_0, q_f\} \qquad \text{(ii)}$$

where j is the network identifier, Q is a finite set of states, A is a finite set of directed arcs, $q_0 \in Q$ is the initial state and q_f is the final state. Without any loss of generality we consider only PNs with a single initial state and a single final state.

Each arc $a_i \in A$ is a 5 - tuple:

$$a_i = (q_{bi}, q_{ei}, P_i, condition_i, action_i) \qquad (12)$$

where $q_{bi} \in Q$ is the starting state of a_i, $q_{ei} \in Q$ is the terminal state of a_i, P_i is a measure associated to the arc (it can be a weight or a probability according to the scoring method used by the PN supervisor described later on), *condition,* is a condition and *action,* is an action; both of them are associated to the arc. The conditions can be categorized in two classes:

COND n

refers to a user defined condition n.

DEFAULT r

refers to a default condition (it is satisfied only if no other condition of any arc whose starting state is q_{bi} returns a measure greater than r).

The actions are executed by the PN supervisor and can be categorized in five classes

EXEn

executes a user defined action; such an action is usually a "matcher" which performs some computations on the input data and returns a result.

PUSH i

is defined as follows. Let's assume that PN_j has an arc that contains PUSH i. Let π_j be the process that executes PN_j. When the arc is reached whose associated action is PUSH i, the execution of π_j is suspended. The state of π_j is pushed on the top of the stack of the PN supervisor. A new process π_i that executes *PN,* is created and executed. When the final state of *PN,* is reached, the last arc of *PN,* is considered. It has associated either a POPABS f or a POPCOND f action. This action is executed. It returns scores computed by PN_i. These scores are passed to π_j whose execution is resumed while π_i terminates.

POPABS f

is associated to the final state of a PN. It stops the execution of the current network process and the result of the execution of the user defined function f is returned.

POPCOND f

This action is also associated to the final state of a PN. It has a synchronization capability that stops the execution of the current network if all the paths in the network leading to the final state have propagated their contribution to the computation of the scores the PN has to provide. If the condition is reached, then the result of the execution of the user defined function f is returned.

JMP

makes the score associated to q_{bi} propagate to q_{ei} without any change.

Each *PN* is associated a Working Memory (WM). Actions associated with the arcs of a subnetwork produce descriptions stored into the subnetwork WM. When a push to a subnetwork is made, the network supervisor may link the subnetwork WM with other WM, thus establishing the viewpoint within which conditions are tested. Most of the actions associated with arcs include plans, Hidden-Markov-Models (HMM), local parsers, rule-based inference units. All these tools are used for extracting an unambiguous description D of a speech pattern and for computing an a-priori probability for an hypothesis H:

$$P_r(D/H) \qquad (13)$$

The *PN* supervisor keeps up to date a search space where each node is represented by the following four-tuple:

$$(q, context, T, score) \qquad (14)$$

where:

-q is a state of a PN, with a buffer containing the information propagated by the actions executed before reaching it,

-"context" is the context (viewpoint) in which the conditions and actions of the arcs starting at q have to be executed,

~*T* is the starting time of the speech signal for the execution of sensory procedures invoked by the actions associated with the arcs starting at q,

-"score" is the score of the hypothesis contained in or implied by "context" up to T. Composite scores can be evaluated as likelihoods:

$$L(D,H) = Pr(D/H)Pr(H) \qquad (15)$$

where $Pr(H)$ is obtained by a language model.

The size of the search space can be kept small in spite of a large number of states in the *PN* if conditions and actions are properly chosen and placed in the network.

4. AN EXAMPLE OF APPLICATION

Multi-speaker recognition of isolated letters and digits was performed on four (two male and two female) speakers.

The vocabulary used for this experiment is defined in Table I. The speech utterance is segmented into Acoustic Segments (AS) with an algorithm described in [5].

this relation.

The generation of $x_1(t)$ can be seen as the application of the following composite relation:

$$G = R_3 o R_1 o R_2 \quad (5)$$

according to the scheme shown in Figure 1.

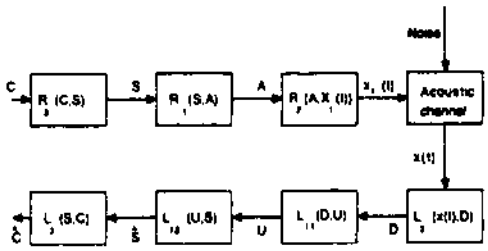

Fig 1 Speech Communication Channel

Recognition consists in applying the relations in the opposite direction. Unfortunately we have only a limited knowledge of these relations. We have used it for building speech synthesizers. We do not even know the alphabet $\sum_A \{a_k\}$ for the elements of A, although we know alphabets $\sum_C$ and $\sum_S$ for the elements of C and S respectively. Furthermore, signal $x_1(t)$ is affected by noise and is transformed into $x(t)$ through the acoustic channel.

As we do not know $\sum_A$, nor we know R_2, we can characterize actions by descriptions of what they produce. According to this approach, the perception of $x(t)$ consists in extracting a sequence of descriptions:

$$D = d_1(\tau_1)\, d_2(\tau_2) \ldots d_i(\tau_i) \ldots d_f(\tau_f) \quad (6)$$

where $d_1(\tau_1)$ describes the silence preceeding the beginning of the speech signal and $d_i(\tau_i)$ describes the segment of $x(t)$ between the time instants τ_{i-1} and τ_i

Segments of D can be 10 msecs. frames or intervals of variable duration obtained by a segmentation algorithm like the one proposed in [5].

The descriptions D can be obtained by perceptual actions in analogy with the generative scheme. Perceptual actions, as well as generative actions have to be defined and used according to a criterion of economy. That is, there must be a limited number of actions (operators) based on which a variety of networks of actions, can be built.

Recognition can be seen as a combination of a relation

$$L_1(D, S) \quad (7)$$

that is the perceptual counterpart of relation R_1 used for speech generation, and a relation:

$$L_2(x(t), D) \quad (8),$$

that is the perceptual counterpart of $R_2(A, x_1(t))$.

The relation $L_2(x(t), D)$ is deterministic in the sense that it can produce only one description D for a signal $x(t)$. Description D can be of fixed duration, i.e. a descriptive phrase is generated at constant time intervals, or of variable duration, i.e. descriptions can be generated for intervals of different length. If we want to maintain the analogy with the production model just outlined, D should be of variable duration because the articulatory actions (gestures) are of variable duration.

Descriptions must refer to parameters, morphologies and properties that are characteristic for a sound and exhibit low variances when many speakers, different microphones and environments are considered.

In practice, fixed duration models have been developed and tested with a considerable degree of success mostly in speaker-dependent systems. In one of the most successful systems developed so far [7], D is a sequence of symbols obtained every 10 msecs. by vftptnr-qnunt.iration with a process that is speaker-dependent and context-independent.

Relation $L_1(D, S)$ has to capture two different types of knowledge. The first type of knowledge is a relation:

$$L_{11}(D, U) \quad (0)$$

between a sequence U of Speech Units (SU) and corresponding descriptions D. There are speech units like the plosive sound /b/ for which a large variety of different descriptions D are perceived as the same sound. Relation L_{11} is many-to-one and it could be interesting to collect statistics of the elements of the universe of acoustic descriptions that produce the perception of the same linguistic sound. These statistics may represent distributions of acoustic patterns produced by a single or many speakers having the intention of producing the same sound. Statistics may also take into account characteristics of background noise.

A second type of knowledge is a relation:

$$L_{12}(U, S) \quad (10)$$

where S is a linguistic entity like a sentence and U is a sequence of Speech Units. L_{12} can also contain statistics.

L_{12} may represent how different speakers may have different pronounciations of the same word. A stochastic model representing a word W in terms of SUs can be built.

An interesting possibility, we would like to explore in this paper, is that of designing L_2 and L_1 procedurally, through actions to be performed on $x(t)$ in order to obtain D, U and S.

It seems that Variable depth descriptions can be very useful in complex tasks where a preliminary selection of hypotheses has to be done based on robust but simple descriptions and then a more detailed analysis has to be performed involving levels of depth depending on the competing hypotheses, or just on acoustic preconditions.

The entire perception model can be represented by procedural networks which invoke subnetworks at several levels. At each level, different types of units can be defined and statistics of their components can be collected.

Table I The 36 word vocabulary

Zero	**One**	Two	Three
Four	**Five**	Six	Seven
Eight	**Nine**	A	B
C	**D**	E	F
G	**H**	I	J
K	**L**	M	N
O	**P**	Q	R
S	**T**	U	V
W	**X**	Y	Z

>rder to hypothesize the Speech Units contained in it. For this purpose, actions are introduced for describing the AS head, its vocalic part, and its tail according to the PN shown in Figure 2.

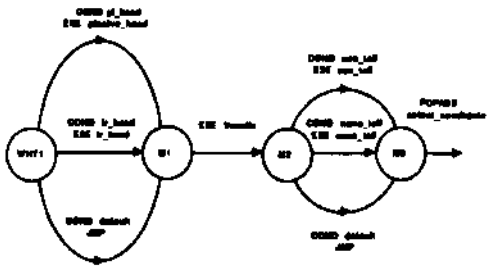

Fig. 2 Example of a Procedural Network

Speech Units correspond in this example to phonemes and are characterized by "place and manner of articulation".

The AS head is analyzed by attached *procedures* (actions) performing an Elaboration-Decision (ED) paradigm. There are two possible ED-actions for the head of an AS, namely:

-plosive head
-fricative (including affricate) head

The choice of the ED action is made by disjoint conditions associated to arcs. These conditions are regular expressions of Primary Acoustic Cues introduced in [5], State MI in Fig. 2. will be reached by only one arc. Using techniques partially described elsewhere [6], hypotheses about the place and manner of articulation for the speech unit in the head subsegment are generated and scored by the following a-priori probability:

$$\mathbf{Prh = Pr\ (data\ (h)\ /\ place \cap manner)} \qquad \mathbf{(16);}$$

After state M1, the ED-action "vocalic" is executed. It segments the vocalic part of AS into stationary and transient units.

Let

$$v_1 v_2 \ldots v_x \ldots v_X$$

be such subsegments. For each segment v_x spectral lines are considered as data (see [0] for details) and a-priori probabilities about place and manner of articulation are obtained by HMMs of spectral lines in the segment v_t. For each subsegment and for each consistent "place-manner" pair, the following probability is computed.

$$Pr\ (v_x) = Pr\ (data\ (v_x)/place \cap manner\ ) \qquad (17).$$

From state M2 to state M3, ED-actions for the tail of AS are executed similar to those used for the head. A probability

$$\mathbf{Prt = Pr\ (data\ (t)\ /place \cap manner)} \qquad (18)$$

scores the hypotheses of the tail subsegment. The data extracted in the head, the subsegments of the vocalic part and the tail -can be assumed to be independent. The "select" action associated to the POPABS arc computes for each candidate hypothesis the probability Pr (data /hyp) by multiplying the probabilities of the phonemes which appear in "hyp" :

$$P_{h\,1} P_{h\,2}, \ldots, P_{hi}, \ldots, P_{hI}$$

Preliminary results have been obtained on a test set of 400 utterances of words belonging to the alphabet defined in Table I and pronounced by four speakers. The voices of these speakers were used for learning only head and tail statistics. An overall recognition rate superior to 90% was achieved.

ACKNOWLEDGEMENTS

This work was supported by the National Science and Engineering Research Council of Canada under Grant A2439. We would like to thank Lorraine Harper for her help in the final preparation and layout of the paper.

REFERENCES

[1] S.E Levinson, "Structural methods in automatic speech recognition", IEEE Proceedings, pp 1625-1650, November 1985

[2] V. W Zue, "The use of speech knowledge in automatic speech recognition", IEEE Proceedings, pp. 1602-1615, November 1985.

[3] D H Klatt, "Review of the ARPA Speech Understanding Project", J Acoust Soc Amer., vol 62, pp 1345-1366, 1977

[4] K. N Stevens, "Acoustic correlates of some phonetic categories", J Acoust Soc. Amer , vol. 68, pp 836-842, 1980

[5] R De Mori, P Laface, and Y Mong, "Parallel algorithms for syllable recognition in continuous speech", IEEE Trans Pattern Anal Machine Intell., vol 39, pp 1-88, 1985

[6] R De Mori,L Lam, M Gilloux, "Learning and plan refinement in a knowledge-based system for automatic speech recognition" IEEE Trans Pattern Anal Machine Intell., vol 41, pp , 1987

[7] L R Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition", IEEE Trans Pattern Anal. Machine Intell, vol. PAMI-5, pp 179-190, 1983

[8] K S. Fu, "Syntactic pattern recognition and applications", Prentice Hall, 1982

[9] E Merlo, R. De Mori, M. Palakal and G. Mercier, "A continuous parameter and frequency domain based Markov model", in Proc. Inter. Conf. on Acoust., Speech, Signal Processing, pp 1597-1600, Tokyo, Japan, 1986