World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:2, No:2, 2008

A System of Automatic Speech Recognition based on the Technique of Temporal Retiming

Samir Abdelhamid, and Noureddine Bouguechal

***Abstract*—**We report in this paper the procedure of a system of automatic speech recognition based on techniques of the dynamic programming. The technique of temporal retiming is a technique used to synchronize between two forms to compare. We will see how this technique is adapted to the field of the automatic speech recognition. We will expose, in a first place, the theory of the function of retiming which is used to compare and to adjust an unknown form with a whole of forms of reference constituting the vocabulary of the application. Then we will give, in the second place, the various algorithms necessary to their implementation on machine. The algorithms which we will present were tested on part of the corpus of words in Arab language Arabdic-10 [4] and gave whole satisfaction. These algorithms are effective insofar as we apply them to the small ones or average vocabularies.

***Keywords*—**Continuous speech recognition, temporal retiming, phonetic decoding, algorithms, vocal signal, dynamic programming.

I. Introduction

THE algorithms of dynamic programming are a true success in the field of the speech recognition especially for the small ones and average vocabularies.

These algorithms were introduced in order to ensure temporal retiming between the vocal signal and the forms of reference. This temporal retiming is made necessary by the variations in the speed and the rhythm of elocution between the sentences of training and recognition.

We will see that these techniques are adaptable to the automatic speech recognition. In this case, the objective is to compare a graph of reference, including the various phonological variations of a word, with the phonetic representation extracted from the vocal signal.

In this approach, temporal retiming aims to synchronize the form of reference and the representation of the signal when an error of segmentation or labelling appears at the time of phonetic decoding.

We will expose initially the theoretical aspect of the function of retiming then we will give the various algorithms necessary to its implementation on machine.

Samir Abdelhamid is with Institute of computer science, University of BATNA, 05000 Algeria (e-mail: samir_abdel@hotmail.com).

Noureddine Bouguechal is with Institute of Electronic, University of BATNA, 05000 Algeria (e-mail : Bouguechal@univ-batna.dz).

II. Structure of the Studied System

The systems guided by syntax allow only the recognition of sentences corresponding rigorously to the grammar of the definite language. However, in speech recognition, it is necessary to take account of errors which come primarily from:

- addition of parasitic words or noises
- errors made by the phonetic decoder
- local syntactic alternatives used by the speaker

A) Structure Adopted

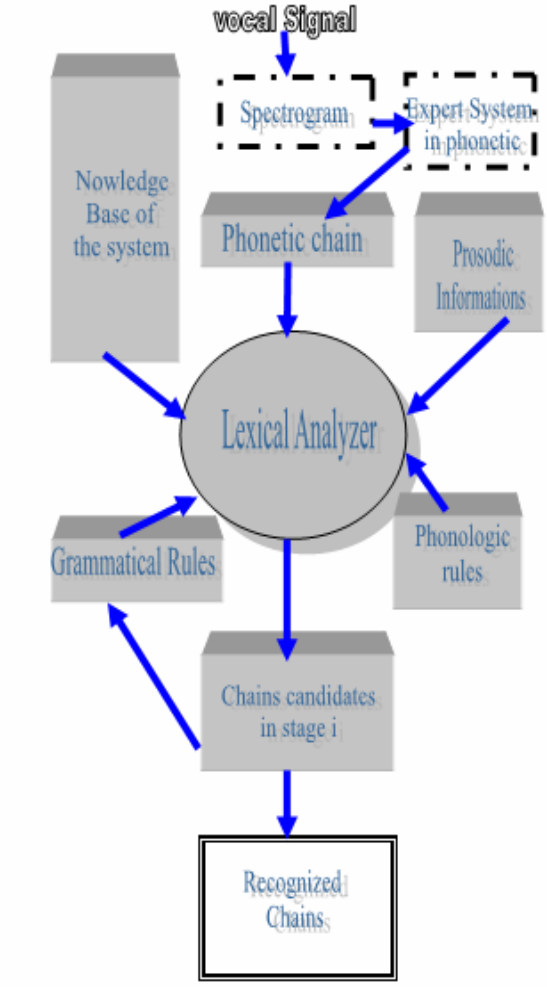

Fig. 1 Structure of the system

It is thus noted that if a traditional syntactic analysis is applied, it is irremediably led to a fatal error which leads to false results of recognition. To solve this problem, various heuristic were proposed. We quote, for example, the techniques calling upon other sources of information that syntax like the grammatical rules and the emission of assumptions on the nature of the next word to be recognized.

The module of recognition is built around a lexical analyzer as shown in Fig. 1. In the general strategy adopted of prediction and checking, the objective of the analyzer is the evaluation of a rate of dissimilarity between a portion of the phonetic lattice and the whole of the words emitted in assumption by the syntactic level.

The process of recognition adapted to the level of the sentence is inspired of the algorithm of level building with a multiple research in beams.

We point out that a research in beam is a form of the strategy of "some the best initially". Where we preserve the most probable assumptions (simultaneously) with search in parallel and for synchronized way.

The choice of the algorithm of the level building was selected because of prevalence of the short words in continuous speech recognition. Indeed, from a probabilistic point of view, it is easier to decode without errors a short word than a long word. What implies that, among the whole of possible interpretations of the lattice, the recognition of the short words is always prevalent compared to that of the long words.

B) General Diagram of the Algorithm Used

The recognition is carried out by an iterative process where we proceed in the following way:

With each stage i of the recognition, the grammatical model provides to the module of recognition some assumptions on the class of the next word to be recognized. It provides also some information on the syntactic and semantic features of the same word.

Inspired of these assumptions, the lexical analyzer proceeds to a comparison between the partial chain of the phonetic lattice and the whole of the words of reference of the lexicon pertaining to the grammatical class emitted like assumption.

This comparison consists in making a certain rate of dissimilarity between the two forms quoted before. This rate of dissimilarity is calculated by the analyzer only after consultation of prosodic information and the rules of phonology. Thus, we can build a base of partial assumptions where we find, at stage i of the recognition, the whole of the chains of N words giving a partial interpretation of the phonetic lattice. With each chain three types of information are associated:

- The score of recognition
- The number of the last interpreted segment of the lattice
- The part of speech and the list of the syntactic features that each word of the chain in the context where it were recognized.

The final phase of the recognition process consists in removing a chain of the base as soon as this one provides a complete interpretation of the lattice to arrange it in the base of the recognized chains. The process of recognition continues until the base of the partial chains is empty. This process is schematized by the following algorithm:

INITIALIZE the base of the candidate chains

WHILE nonempty Base **DO**
  FOR each class of the lexicon **DO**
    TREATMENT
  ENDFOR

ARRANGE in the base of the recognized chains the various chains interpreting the lattice
ENDWHILE

TREATMENT : stages of the process of recognition

SELECT among the chains recognized at the preceding level those which can be followed of a word pertaining to the current class

/* Hypothèses on the next word to be recognized * /

CALCULATE, by interpreting the grammatical rules for each selected chain, the features which the word must have obligatorily which will be identified in the lattice and will be then concatenate with the treated chain

/* Evaluation of the syntactic and semantic features of the next word to be recognized */

LAUNCH the lexical analysis on the current class

/* Comparaison with emission of a rate of dissimilarity*/

CONCATENATE words recognized with the chains selected previously and calculate the cumulated scores

/* Construction of the chains recognized with scores of recognition */

ELIMINATE the chains according to the score obtained

/* Procedure of elimination of the chains having a weak rate of recognition */

III. Structure of the Lexicon

In a system of automatic speech recognition, the dictionary must comprise syntactic, phonetic and phonological information.

Indeed, this information is useful for the checking of assumptions (words) based on the calculation of similarity between the phonetic representation of a word and a portion of the phonetic transcription of the treated sentence.

The lexical level must thus comprise a syntactic component and a phonetic and phonological component.

In our system, the lexicon contains three types of information:

- Phonetic and phonological descriptions allowing interpretation of the phonetic representation resulting from decoding acoustico-phonetics [7]
- Syntactic information establishing the link between the lexical level and the syntactic model
- The orthographical form of each word.

Here an extract of the lexicon where we define a grammatical class.

We chose the class of the verbs, of the French language, where we define the three types of information quoted above.

VERB [13]
TYPE
NR = [1...6] /* person and number of the verb*/
A, I, Q, N, = BOOLEAN; /* Syntactic elements */

ENTER

AIMERAIS	[1NR, Q, N, I]	:	**ɛ/m/e/r/ɛ**
APPARAIT	[3NR, Q]	:	**a/p/a/r/ɛ**
CONSTITUENT	[6NR, N]	:	**c/õ/s/t/i/t/u**
REMERCIE	[1NR, N]	:	**r/e/m/ɛ/s, , r/i**

.....
.....
ENDCLASS

Each class of the lexicon comprises three zones of descriptions of data:

1. The name of the grammatical class treated (VERB),

2. A zone of declaration in which we define the nature of the features syntactic related to the words of the lexicon,

3. A part named, enter lexicon, in which is defined each word of the grammatical category.

The end of each class is indicated by the terminal "ENDCLASS".

IV. Representation of the Lexicon

The lexicon comprises two lexical structures gathering the unit of linguistic information. It must contain the phonetic and phonological transcription of each word and its orthographical form.

The phonetic transcription is used at the time of the comparison of the lattice to the various words candidates of a definite class.

Once the word is recognized, we must then consult the grammatical lexicon in order to recognize the orthographical form of the found word.

It is thus necessary to have for each word of the lexicon, a phonetic form describing it. This phonetic description integrates, of course, various phonological alternatives [14]. Thus, the suggested representation would be a graph in which the standard phonetic transcription of each word is supplemented by the various elisions, substitutions and insertions foreseeable (see Fig. 2).

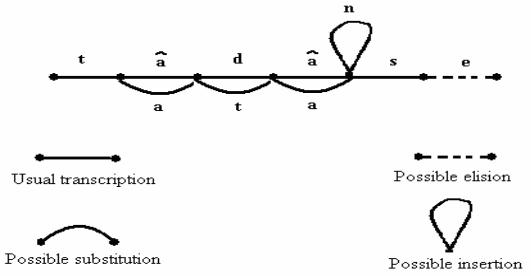

Fig. 2 Phonetic representation of the word « tendance »
In the system 'Myrtille I'

The organization of these phonetic lexicons is an important aspect which affirms the correct operation of the system.

Indeed, of optimizations can be implemented. They consist in factorizing in a phonetic lexicon the beginnings of identical words. Thus, two words begin with the same phonetic graph amalgamate their common part as indicated on Fig. 3.

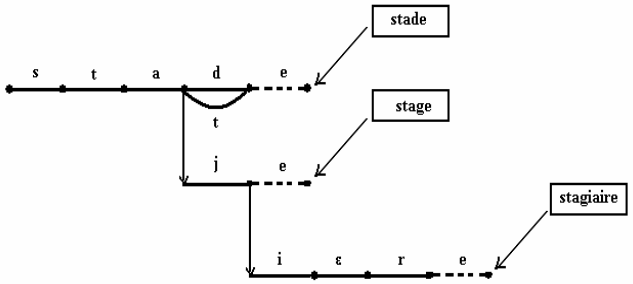

Fig. 3 Arborescent representation of the words of the lexicon

It is noticed that each sheet of the built tree corresponds to the result of a phonetic description.

V. The Function of Retiming

It is a question of comparing a graph of reference comprising various phonological variations of a word with representation of the phonetic type extracted from the vocal signal.

In this approach, temporal retiming aims to synchronize the form of reference and the representation of the signal when an error of segmentation appears at the time of phonetic decoding. [11]

The dynamic comparison between the two phonetic forms consists in looking for in the graph GRAPH X VEC an optimum way (see Fig. 4.).

More formally, it is a question of finding among the whole of the possible functions of retiming that which is optimal within the meaning of certain metric.

Various functions of retiming W are in the following way defined:

$$\{1,...,I+J+1\} \rightarrow \{1,...,I\} \times \{1,...,J\} \times \{I+1, J+1\}$$

$$k \quad \rightarrow \quad W(k)=\{I(k),J(k)\}$$

Example of Retiming Function:

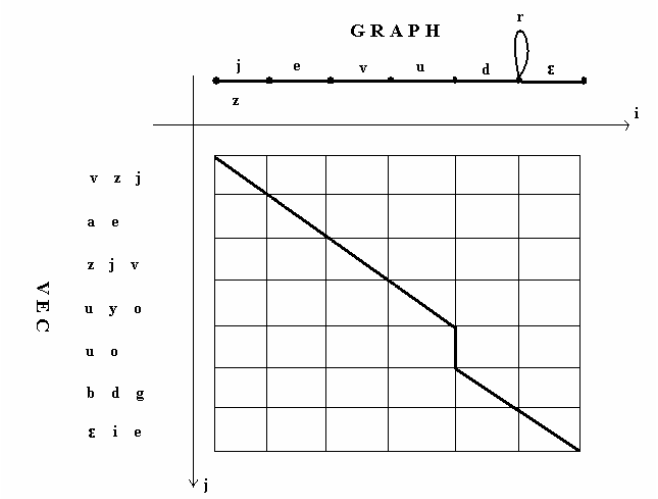

Fig. 4 The search of an optimum way in the graph GRAPH X VEC

By applying the principle of optimality of BELLMAN, it is possible to write the definition of the optimal function of retiming in the following way:

$$W = \underset{I(k),J(k)}{ARG} D\left[I(k),J(k)\right]$$

With D recursively defined by:

$$\begin{cases} D[I(k),J(k)] = \underset{I(k),J(k)}{Min}\left(D[I(k-1),J(k-1)]+d([I(k-1),J(k-1)],[I(k),J(k)])\right) \\ D[I(1),J(1)]=0 \end{cases}$$

Where $d([I(k-1),J(k-1)],[I(k),J(k)])$ represent a weighting associated to each elementary arc $[W(k),W(k+1)]$ of displacement in the graph and k the length of the way.

By integrating the constraints of displacements, we obtain:

$$\begin{cases} D[I(k),J(k)] = Min \begin{cases} D(i-1,j)+d[(i-1,j),(i,j)] \\ D(i,j-1)+d[(i,j-1),(i,j)] \\ D(i-1,j-1)+d[(i-1,j-1),(i,j)] \end{cases} \\ D(1,1)=0 \end{cases}$$

VI. Evaluations of the Penalties Associated to the Arcs

A. Diagonal Arc

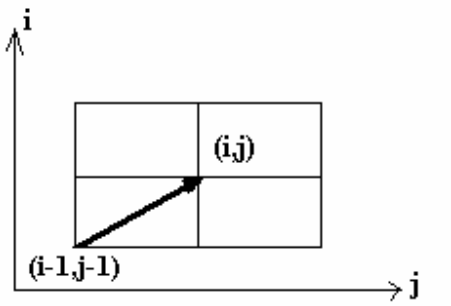

Fig. 5 Diagonal displacement

It is the ideal case where the optimal way passes by the point $(i-1,j-1)$. The continuation at the point (i,j) will not be made that if there is correspondence between the current elements of the two forms to compare (see Figure 5.). The penalization associated to the arc will be:

$$D([i-1,j-1],[i,j])=Min\,Ph(GRAPH[i-1,m],VEC[j-1,n])$$

Where Ph : represent the distance inter-phoneme.

$$m:1\rightarrow 2 \quad and \quad n:1\rightarrow 3$$

B. Vertical Arc

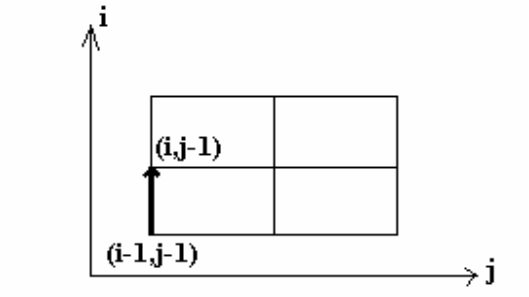

Fig. 6 Vertical displacement

The continuation of the way by displacement $[(i-1,j),(i,j)]$ corresponds in this case to the resolution of two possible cases (see Fig. 6.):

- There is an error of a less segmentation during phonetic decoding
- Segment i of the phonetic graph is optional. It is thus about an elision and the distance is expressed by :

$$d([i-1,j],[i,j])=ELIS$$

C. Horizontal Arc

We can obtain a horizontal arc only in the two following cases (see Fig. 7.):

- The phonetic decoder made an error of over segmentation
- $GRAPH\ [i,3]\ et\ VEC\ [J-1]$ coincide. An insertion is thus envisaged in the phonetic graph.

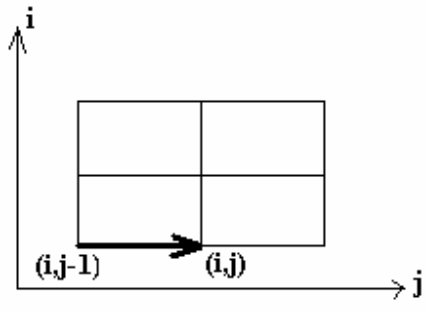

Fig. 7 Horizontal displacement

VII. Algorithms of Evaluation

From the practical point of view, we present here the algorithms of evaluation of the function of retiming, of elementary displacements like and of the cumulated distances.

STAGE 1 : Initialization

For i = 1 to I **Do**
  For j = 1 to J **Do**
    $D(i, j) \leftarrow \infty$
  EndFor
EndFor
$D(1,1) = 0$
STAGE 2

STAGE 2 : Evaluation of the retiming function

For j = 1 to J **Do**
  For i = 1 to I **Do**
    STAGE 3
    STAGE 4
  EndFor
EndFor

STAGE 5

On the level of the stage two, we refer to the stages three, four and five. It is in fact of the calls of procedures that we could have indicated by **'Call stage i'**.

STAGE 3 : Evaluation of elementary displacements

d[(i , j) , (i+1 , j+1)] ← ∞
d[(i , j) , (i+1 , j)] ← **Rules_interpretation (i , j)**
d[(i , j) , (i , j+1)] ← **Rules_interpretation (i , j)**

For n = 1 to 3 **Do**
  For m = 1 to 2 **Do**
    /* Diagonal arc
    Distance ← **Ph(GRAPH[i , m] , VEC[j , n])**
    If Distance < d[(i , j) , (i+1 , j+1)]
    Then d[(i , j) , (i+1 , j+1)] ← **Distance**
    EndIf
  EndFor

  /* Horizontal arc - Insertion -
  Distance ← **Ph(GRAPH[i , 3] , VEC[j , n])**
  If Distance < d[(i , j) , (i , j+1)]
  Then d[(i , j) , (i , j+1)] ← **Distance**
  EndIf

 /* Vertical arc – Elision –
 If GRAPH[i , 1] < 0
 Then d[(i , j) , (i+1 , j)] = Elis
 EndIf
EndFor

STAGE 4 : Calculation of the cumulated distances

$$D(i+1, j) = Min \begin{cases} D(i+1, j) \\ D(i, j) + d[(i, j), (i+1, j)] \end{cases}$$

$$D(i, j+1) = Min \begin{cases} D(i, j+1) \\ D(i, j) + d[(i, j), (i, j+1)] \end{cases}$$

$$D(i+1, j+1) = Min \begin{cases} D(i+1, j+1) \\ D(i, j) + d[(i, j), (i+1, j+1)] \end{cases}$$

STAGE 5 : Results of the evaluation

$$D(W) = D(I+1, J+1)$$

Explanations:

- The procedure Rules_interpretation activates the interpreter of rules of phonology and of strategy and it allows the calculation of certain local distances according to the context of analysis given in parameter.
- The recognition of a word is obtained by iteration of the algorithm above on the whole of the vocabulary. The recognized word is then that which has the best function of retiming.
- The elementary distances are calculated at the stage three and are cumulated in the field of definition of W.

VIII. Conclusion and Perspectives

Temporal retiming is an effective technique for synchronization between the form of reference and the vocal signal.

Into our team of research, we introduced these algorithms in order to stage with the errors of segmentations or labelling which appear at the time of phonetic decoding. However, these algorithms can present failures as soon as we are confronted with the problem of the localization of borders of the words. This problem emerges when we make for example the connection between two successive words. In this case, the algorithms of dynamic programming present an indeterminism at the localization of the words.

This problem can be solved while taking into account, for example, the whole of the possible combinations of the words of a vocabulary. This could be the subject of another article.

ACKNOWLEDGMENT

The authors would like to thank Dr. Chaouki Abdelhamid and PhD student Azzeddine Abdesselam for their help and their fruitful discussions about the problem of the affectation of the rates of dissimilarity to the recognized chains.

REFERENCES

[1] J. P. Haton, D. Fohr et M. Djoudi: Un système expert pour le décodage acoustico-phonétique pour l'Arabe standard. Conférence Maghrébine, Septembre 1989.
[2] Y. Belkaid: Les voyelles de l'Arabe littéraire moderne. Analyse spectrographique Rapport N° 16, travaux de l'institut de phonétique de Strasbourg, 1984.
[3] O. Deroo, C. Ris: Hybrid HMM/ANN Systems speaker independent continuous speech recognition in French Travaux de l'école Polytechnique de MONS Belgique, 2000.
[4] S. Abdelhamid: Contributions à l'étude et à la réalisation d'une machine à dicter en Français. Thèse de Magister de l'institut d'informatique de l'université de Batna, Algérie, 1994.
[5] M. Guerti: Contribution à la synthèse de la parole en Arabe standard. Actes des 16ème journées d'études sur la parole. Hammamet, Tunisie 1987.
[6] Benhamouda: Morphologie et syntaxe de la langue Arabe. Nationale Edition, 1983.
[7] N. Carbonell, J. P. Haton, D. Fohr Aphodex, design and implementation of an acoustic-phonetic decoding expert system. IEEE International conference on Acoustics, speech and signal processing, 1986.
[8] V. Barreaud : Reconnaissance automatique de la parole continue: compensation des bruits par transformation de la parole. Thèse de l'université de Nancy1, 2004.
[9] S. Stuker :Automatic Generation of Pronunciation Dictionaries For New, Unseen Languages by Voting Among Phoneme Recognizers in Nine Different Languages, Master thesis, Carnegie Mellon University, Pittsburgh, PA, USA, April, 2002.
[10] D. Vaufreydaz, M. Akbar, J. Caelen : Environnement Multimédia pour l'Acquisition et la gestion de corpus Parole, JEP'98, pp. 175-178, Martigny, Switzerland, June 1998.
[11] H-F. Silverman, D-P. Morgan: The application of dynamic programming to connected speech recognition, IEEE ASSP magazine, vol.7, pp.6-25, 1990.
[12] L. R. Rabiner : A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, L.R. Rabiner, Proceedings of the IEEE, vol 77, No 2, 1989.
[13] N. Carbonell, J.P Haton, F. Lonchamp, JM. Pierrel : Élaboration expérimentale d'indices prosodiques pour la reconnaissance; application à l'analyse syntaxico-sémantique dans le système MYRTILLE II", *Séminaire Prosodie et Reconnaissance*, Aix-en-Provence, 1982.
[14] J. M. Pierrel : Utilisation des contraintes linguistiques en compréhension de parole continue dans le système Myrtille II. TSI, Vol 1, N° 5, 1982, pp. 403-421.