A new statistical approach for the automatic segmentation of continuous speech signals

Régine André-Obrecht

To cite this version:

Régine André-Obrecht. A new statistical approach for the automatic segmentation of continuous speech signals. [Research Report] RR-0511, INRIA. 1986. inria-00076043

INRIA

CENTRE DE RENNES
IRISA

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Rapports de Recherche

N° 511

A NEW STATISTICAL APPROACH FOR THE AUTOMATIC SEGMENTATION OF CONTINUOUS SPEECH SIGNALS

Régine ANDRÉ - OBRECHT

Mars 1986

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
35042 - RENNES CÉDEX
FRANCE
Téléphone : 99 36 20 00
Télex: UNIRISA 950 473 F
Télécopie : 99 38 38 32

Publication Interne n° 287
38 pages
Février 1986

A NEW STATISTICAL APPROACH FOR THE AUTOMATIC SEGMENTATION OF CONTINUOUS SPEECH SIGNALS

Régine ANDRE-OBRECHT

Résumé:

Le but de la segmentation proposée est la détection d'évènements acoustiques significatifs de changements articulatoires.

L'approche repose sur une modélisation statistique du signal et la détection séquentielle sur le comportement de statistiques de test. Quatre méthodes sont étudiées et comparées.

Les résultats indépendant du locuteur ont permis de caractériser les frontières et de mettre en évidence trois types de segments infraphonémiques:
- les zones stationnaires des phonèmes
- les zones transitoires mais homogènes
- les segments courts correspondants à des évènements courts (explosion de plosives,...).

Abstract:

We discuss a statistical approach for the automatic segmentation of the continuous speech signal. The purpose is to detect acoustic events which reveal articulatory changes, such as voice or frications onset and termination, closure, formantic variations, etc...

The main idea is to model the signal by a statistical model and to use test statistics to detect sequentially abrupt changes in the parameters of this model. Four segmentation algorithms are presented here and compared.

The results, obtained by each procedure, are similar: a detection correspond actually to an acoustic event. The so defined elementary units are infra-phonemic, they correspond to stationary parts of the phonemes, to homogeneous transitions during which the formantic variations are monotonous and to short segments, such as the bursts of unvoiced plosives.

Travail soutenu par un contrat CNET.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L. A. 227)
UNIVERSITÉ DE RENNES 1 I. N. S. A. DE RENNES

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE
(LABORATOIRE DE RENNES)

A NEW STATISTICAL APPROACH
FOR THE AUTOMATIC SEGMENTATION OF CONTINUOUS SPEECH SIGNALS

Régine ANDRE-OBRECHT
I.R.I.S.A.
Campus de Beaulieu
F-35042 RENNES Cédex
tel:99 36 20 00, telex:UNIRISA 950473 F

EDIC numbers: 2.3.2. and 2.5.3.

ABSTRACT.- We discuss a statistical approach for the automatic segmentation of the continuous speech signal. The purpose is to detect acoustic events which reveal articulatory changes, such as voice or frication onset and termination, closure, release, formantic variations, etc...

The main idea is to model the signal by a statistical model and to use test statistics to detect sequentially abrupt changes in the parameters of this model. The four segmentation algorithms presented here differ in the nature of the model of the acoustic channel (AR, ARMA), in the nature of the excitation of the model (the glottal impulses can be taken into account or not in the model) and in the nature of the test statistics (generalized likelihood, statistics of cumulative sum type).

After a first experiment to tune each procedure to the speech signal, the performance of each one is evaluated using a record of ten phonetically balanced sentences and sequences of numbers pronounced by ten speakers (four male, six female). Except for the pulse method which takes into account the pitch, the learning of the parameters (threshold, window-size...) is speaker-independent.

The results, obtained by each procedure, are similar : a detection correspond actually to an acoustic event. The so defined elementary units are infra-phonemic, they correspond to stationary parts of the phonemes,

to homogeneous transitions during which the formantic variations are monotonous and to short segments, such as the bursts of unvoiced plosives.

The experiments show the robustness of the segmentation algorithms which are used, though the speech signal is not stationary. This study is completed by a comparison between the experimental results and a hand made segmentation, it shows the accuracy of the change times estimates.

I. INTRODUCTION

These last years significant advances have been achieved in the field of word recognition, but the continuous speech recognition still raises many problems.

A too great computation cost and storage make the isolated and connected word recognition strategies based on a word reference dictionary unfeasible. A solution lies in the phonetic recognition on the speech input. Such a system may be divided into several components (Figure 1).

The objective of the first step, called the acoustico-phonetic processor, is to accept the speech signal as input and to produce a string of discrete units. This transformation from continuous to discrete, is the only one of this type inside the whole system and requires a segmentation of the continuous speech; but an accurate segmentation and labeling of the signal are an extremely complicated and difficult task. The difficulties are to state which type of units is desired, by which methods the signal will be segmented to find these units and how the identification will be made. These problems are of course interconnected and their good solution depends on our knowledge about acoustic and articulatory phonetics.

Because of the size of the vocabulary, the basic recognition units are smaller than words; they are usually phonetic in nature, such as phones, diphones, syllables... Several approaches have been proposed, and they can be classified in the two following categories :

* Segmentation with recognition : the signal is processed by overlapping blocks. A first method consists in defining a pattern as a finite sequence of blocks and in identifying it as the realization of a sequence or a network of known models (template matching [13]). An other technique is based on the extraction of acoustic cues and on hierarchical identification of each block [10], and the segmentation is obtained through this labelling.

*Sequential segmentation : after a block preprocessing and an extraction of acoustic cues, each variation of the parameters involves an increment of a function. The parameters of the current block can be also compared with those of the first block of the current segment by some distance. When the value of the segmentation function [13] or of the distance [15] crosses a threshold, a new segment is detected.

Our approach differs from the above techniques : no block preprocessing is made, and the segmentation we propose is strictly defined with statistical signal processing methods, and consists in detecting non stationarities. The main idea is to consider the signal as a string of stationary units and to detect by a statistical test changes in the parameters of the unit model. The signal preprocessing and the statistical segmentation are jointly performed on line, a decision is taken after every sample.

Such a statistical segmentation without defining a priori the phonetic nature of the units, is unusual in speech recognition systems : as we will see, it results in a more reliable representation of each unit and in a good segmental information.

Four processing procedures are presented in Section II. In section III, experimental results are reported and the comparison of the performance of the methods leads us to discuss the nature of the units. The conclusion is devoted to some remarks about the advantages that such a signal preprocessing may provide in an analytic speech recognition system.

II. PRESENTATION OF STATISTICAL METHODS OF SEGMENTATION

For the four methods of segmentation which will be presented, the signal is assumed to be described by a string of homogeneous units, each of which is characterized by a statistical model. In order to detect the jumps in the model parameters and thus to detect a boundary between two units, two (three or four) models are estimated at different time locations in the signal, and their similarity is evaluated by a test statistics.

The segmentation problem is threefold :

. the selection of the model structure : autoregressive model, pulse or noise excitation...

. the choice of a test statistics : likelihood ratio, Kullback's divergence...

. the practical implementation : choice of the identification method, learning of the parameters...

The presentation of the four methods follows this organization and each method will be illustrated on a small example

II.1 - BRANDT'S GENERALIZED LIKELIHOOD RATIO TEST [6,7]

a. The model

Every homogeneous segment of the signal (y_n) is assumed to be described by an autoregressive model of order p, denoted M(A,σ), i.e.

$$\begin{cases} y_n = \sum_{i=1}^{p} a_i \, y_{n-i} + e_n \\ \\ \mathrm{var}\ (e_n) = \sigma^2 \end{cases} \qquad (1)$$

where $(e_n)_n$ is a zero mean white noise with variance σ^2.

$A = (a_1,\dots,a_p)$ and σ are the parameters of the model.

b. The test

To detect a jump in the parameters (A,σ), two hypotheses are tested against each other (Figure 2).

H_0 : the signal $(y_0,\dots,y_N)$ is described by the model $M_0(A_0,\sigma_0)$.

H_r : there exists a jump time r, such that the signal $(y_0,\dots,y_r)$ is described by the model M_1 (A_1,σ_1) and the signal $(y_{r+1},\dots,y_N)$ by the model $M_2(A_2,\sigma_2)$.

The test statistic is based upon a generalized likelihood ratio (G.L.R) :

. If, for k = 0, 1, 2

$$A_k = \arg \min_{A} \sum_{n \in W_k} (y_n - \sum_{i=1}^{p} a_i y_{n-i})^2$$

$$A^t = (a_1,\dots,a_p)$$

$$\sigma_k^2 = \min_{A} \frac{1}{|W_k|} \sum_{n \in W_k} (y_n - \sum_{1=1}^{p} a_i y_{n-i})^2$$

where W_k denotes one of the three windows shown in the figure 2, and $|W_k|$ its cardinal, the logarithm of the maximum likelihood of each hypothesis is :

$$L\ L\ (H_0) = -\ N \operatorname{Log} \sigma_0 - \frac{N}{2}$$

$$L\ L\ (H_1) = -\ r \operatorname{Log} \sigma_1 - (N-r) \operatorname{Log} \sigma_2 - \frac{N}{2}$$

Then, the ratio of maximum likelihood gives the statistics

$$D_N(r) = -\ (N-r) \operatorname{Log} \sigma_2 - r \operatorname{Log} \sigma_1 + N \operatorname{Log} \sigma_0$$

A jump is detected if

$$\max_{r} D_N(r) > D_0 \qquad (3)$$

where D_0 is a fixed threshold, and the instant r of change is estimated as the argument of the maximum in (3).

Unfortunately, the simultaneous detection-estimation is of high computational cost. Hence, in practice a simplified version is used by separating the change detection from the estimation of the instant r.

c. The practical implementation

It is made in two steps :

Step 1 : change detection

The length of the model M_2 is forced to be a fixed length L, and the statistics

$$D_n(n-L)$$

is computed on-line.

A change detection occurs when

$$D_n(n-L) > D_0 \qquad (4)$$

Step 2 : change time estimation

When a detection occured, say at time n_D, it is reasonable to expect that the true change time r* satisfies :

$$n_D - L < r_* < n_D$$

A new G.L.R. decision variable ΔD_n is monitored for $n > n_D$ to compare the following hypotheses (Figure 3) :

H'_0 : there exists a jump at time r $(r \leqslant n)$

H'_1 : there exists a jump at time n-L (recall L is a fixed lag)

The initial conditions are given by :

$$r = n_D - L$$

$$n = n_D$$

If $\Delta D_n < 0$, r is unchanged; otherwise, r is updated to n-L, and the comparison goes on until $n = n_D + L$. r* is given by the last value reached by r.

The different models M_0, M'_0, M_1 are identified by a growing memory ladder method and the model M_2 by a sliding block covariance ladder method.

For more details on the implementation, see [2], [6,7].

d. A first experiment

One sentence, pronounced by a male speaker, is used to tune the method (the same signal will be process by each method). The sampling rate of these signals is equal to 12.800 Hz.

Three parameters must be fixed during this first experimental stage :

- the length of the sliding window L
- the order of the model p
- the threshold D_0(4).

The length of the sliding window must be chosen long enough to have a good identification, but not too long to obtain small units as the plosive burst : this compromise leads us to set L equal to 100 samples, i.e. approximately 8ms

Several trials are made with an order equal to 10, 12 and 16; but as the other methods will show, an under-estimation of the model order is a cause of some omissions, we prefer to keep p equal to 16.

The most important parameter is the threshold D_0 ; as we can see on the figure 4, except for abrupt changes between voiced and unvoiced signals, the statistics $(D_n(n-L))_n$ varies slowly when a jump happens. This results in some difficulties to fix the threshold. To avoid omissions between two phonemes as |a| and |n|, it is necessary to reduce D_0 to 70, with the risk of increasing the amount of over-segmentations.

II.2 - THE INSTRUMENTAL METHOD [4]

a. The model

Each unit of the signal (y_n) is assumed to be modeled by an autoregressive model with moving average excitation, described by the following equations :

$$y_n - \sum_{i=1}^{p} a_i \, y_{n-i} = \sum_{j=0}^{q} b_j \, e_{n-j} \qquad (5)$$

where $(e_n)_n$ is a standard white noise.

The vector $A^T = [a_1,\ldots,a_p]$ is the vector of the AR parameters and the vector $B^T = [b_0,b_1,\ldots,b_q]$ is the vector of the MA parameters. This model allows a better approach of the speech production in a nasal-context: for nasal sounds, zeros (or anti-resonances) appear in the vocal-tract transfer function. But we consider that all the information relevant to the segmentation problem is contained in the vector A, and that the MA parameters can be considered as nuisance parameters for the test.

b. The test

Here we follow a model validation approach. That is to say we assume a nominal model $A = A_0$ is available (which plays the role of a reference model), and we want to decide whether or not the current part of the signal (y_n) is still in accordance with this nominal vector.

Thus we assume that a sample $(y_1,\ldots,y_n)$ is given and we will decide between the two hypotheses :

H_0 : $A = A_0$ (no change)

H_1 : $A \neq A_0$ (a change happens)

Following a local asymptotic approach [4], we suppose that under a change, A is of the form $A_0 + \delta A/\sqrt{n}$, where δA is an unknown direction of change.

Let us consider the following statistics :

$$U_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (y_k - \sum_{i=1}^{p} a_i \, y_{k-i}) \; Z_k \qquad (6)$$

where

$$Z_k^t = [y_{k-q-1}, \ldots, y_{k-q-p}]$$

is the so-called instrumental variable.

Then [16], for n large, we have :

$$\sum_n^{-1/2} U_n \sim N(0,I) \quad \text{under } H_0$$

$$\sum_n^{-1/2} |U_n - H_n \delta A| \sim N(0,I) \quad \text{under } H_1$$

where N(0,I) is the standard vector gaussian distribution $\sum_n$ is the covariance matrix of U_n and H_n is the p x p Hankel matrix, given by :

$$H_n = \begin{pmatrix} r_{q-p+1}(n) & r_{q-p+2}(n) & \cdots & r_q(n) \\ r_{q-p+2}(n) & & & \\ \vdots & & & \\ r_q(n) & & & r_{q+p-1}(n) \end{pmatrix}$$

$$r_k(n) = \frac{1}{n} \sum_{m=1}^{n} y_{k+m} \, y_m$$

In other words, the detection of a change in the AR vector reduces to a detection of a change in the mean value of the allmost gaussian statistics U_n.

The likelihood ratio yields the χ^2 test statistics

$$T_n = U_n^T \sum_n^{-1} U_n \tag{7}$$

c. The practical implementation

On a growing block Y_0 of signal (Figure 5), the AR reference vector A_0 is identified by the Instrumental Variable method [9], which consists in solving for A the delayed Yule Walker equations :

$$\sum_{k \in Y_0} \lambda^{n-k} Z_k (y_k - \sum_{1=1}^{p} a_i y_{k-i}) = 0 \tag{8}$$

where λ is a forgetting factor.

The statistics U_n (6) is then computed on a delayed block Y_1. But, instead of (7), we use a simplified version of it, namely :

$$T_n = \frac{U_n^T U_n}{\sigma_n} \tag{9}$$

where

$$\sigma_n = \frac{1}{n} \sum_{k \in Y_1} (y_k - \sum_{1=1}^{p} a_i y_{k-i})^2 \quad x \sum_{k \in Y_1} y_k^2$$

d. The first experiment

The first experiments are performed under the same conditions as for Brandt's method, but the problems are different.

The test statistics exhibits a poor behaviour in the neighbourhood of the transitions between voiced and unvoiced segments. To detect this event, the ratio of energy between the two blocks Y_0 and Y_1, is computed simultaneously, denote it R_n. Hence forward, a jump is found when one of the following conditions is satisfied' :

a . $R_n < \frac{1}{\lambda_2}$ (transition from a voiced segment to an unvoiced one)

b . $R_n > \lambda_2$ (transition from an unvoiced segment to a voiced one)

c . $$\begin{cases} \frac{1}{\lambda_1} < R_n < \lambda_1 \\ T_n > T_0 \end{cases}$$

with $1 < \lambda_1, \lambda_2$.

When the signal energy varies slowly (condition c) ,the statistics T_n increases abruptly when a spectral change happens (Figure 6), therefore the threshold T_0 may be chosen large enough without risking omissions.

The preliminary experiments leads us to fix T_0 equal to 30 and λ_1, λ_2 respectively to 2 and 3. The shift between the two windows is chosen equal to 100 to have a good initialisation of the nominal vector A_0. For the same reasons as above, the order of the AR vector is great, $p = 16$, whereas it is sufficient to take an order of the MA vector equal to 1. A greater MA order give the same good results.

Two other experiments have been performed, identifying the AR vector A_0 in (8) without the forgetting factor $\lambda(\lambda=1)$.

The block Y_0 has become sliding and the two windows have been either overlopping or disconnected (Figure 7). As the results are the same, only the first version of the implementation is kept.

II.3 - THE DIVERGENCE TEST [3]

a. The model

Again each homogeneous unit of signal $(y_n)_n$ is assumed to be described by an autoregressive model of order p, denoted by $M(A,\sigma)$, i.e. :

$$\begin{cases} y_n = \sum_{i=1}^{p} a_i \, y_{n-i} + e_n \\ \\ \mathrm{var}\,(e_n) = \sigma^2 \end{cases} \qquad (10)$$

with the same notations as before (1).

b. The test

Let us denote by $g_0(y_n|Y^{n-1})$ and $g_1(y_n|Y^{n-1})$ the conditional densities corresponding to two AR models $M_0(A_0,\sigma_0)$ and $M_1(A_1,\sigma_1)$. The test statistics W_n is a distance measure which involves the cross entropy between these two probability distributions.

The cumulative sum is given by

$$W_n = \sum_{k=1}^{n} w_k \qquad (11)$$

where :

$$w_k = \int g_0(y|Y^{k-1}) \text{ Log } \frac{g_1(y|Y^{k-1})}{g_0(y|Y^{k-1})} \, dy - \text{Log } \frac{g_1(y_k|Y^{k-1})}{g_0(y_k|Y^{k-1})}$$

In the AR gaussian case :

$$w_k = \frac{1}{2} \quad 2 \frac{e_k^0 \; e_k^1}{\sigma_1^2} - (1 + \frac{\sigma_0^2}{\sigma_1^2}) \; {\frac{e_k^0}{\sigma_0^2}}^{2} - (1 - \frac{\sigma_0^2}{\sigma_0^2}) \qquad (12)$$

with

$$e_k^j = y_k - \sum_{i=1}^{p} a_i^j \, y_{k-1}$$

$$A_j^T = (a_1^j, \ldots, a_p^j) \; , \; j = 0,1$$

Under the hypothesis

$$H_0 \; : \; M(A,\sigma) = M_0(A_0, \sigma_0),$$

the statistics (W_n) has a zero drift, and under the hypothesis

$$H_1 \; : \; M(A,\sigma) = M_1(A_1, \sigma_1)$$

the statistics (W_n) has a negative drift, equal to the opposite of the Kullback's divergence [14]. The implementation of the test is based on this behaviour of (W_n).

c. The practical implementation

A long time model $M_0(A_0,\sigma_0)$ is sequentially identified on a growing window by an approximate least squares algorithm in lattice form, the algorithm of Burg [5], and a short time model $M_1(A_1,\sigma_1)$ is identified on a sliding window of fixed length L, by the autocorrelation method [16]. The two windows are overlapping (Figure 8).

As we indicated above, a spectral jump is detected when the drift of (W_n) becomes negative and so when the statistics crosses a threshold. In order to reduce the detection delay and to obtain a good estimate of the change time, Hinkley's cumulative sum test [11] is applied to detect this change drift ; it consists in fixing a minimal size of drift $2\delta > 0$ to be detected, and in computing :

$$W_n = \sum_{k=1}^{n} (w_k + \delta) \tag{13}$$

The observation of the deviation of this cumulative sum with respect to its maximum (Figure 9) gives :

. the detection time n_D corresponding to the crossing of a threshold λ by the quantity

$$\max_{1 \leqslant m \leqslant n} W_m - W_n > \lambda \tag{14}$$

. the estimated jump time r which is the argument of

$$\max_{1 \leqslant m \leqslant n_D} W_m \tag{15}$$

d. A first experiment

The length L (Figure 8) required to have a good identification of the sliding model by the autocorrelation method is 256 (20ms) and the order p of this model required to avoid omissions is 16.

The first results show that the behavior of the statistics (W_n) is different for voiced and unvoiced segments (Figure 10). Hence the energy of the signal is computed and gives coarsely the decision "voiced-unvoiced", and the statistics (W_n) is updated with :

$$\begin{cases} (\delta_v, \lambda_v) = (\ 0.2,\ 40) & \text{for voiced segments} \\ (\delta_b, \lambda_b) = (\ 0.8,\ 80) & \text{otherwise} \end{cases}$$

Even with these two pairs of drift and threshold, some omissions may occur. Because of the dissymetry of the test, a soft transition from a high frequency segment to a low-frequency segment may be missed while the reverse transition might be detected (Figure 11). Thus if a voiced segment is judged too long, say longer than Lmin, it is processed backward with the same statistics. In practice, let us denote by $[y_{n_0},\ldots,y_{n_D}]$ the currently detected segment; if it is voiced and

$$n_D - n_0 > L_{min}$$

the backward test is fired. Two cases may happen (Figure 12).

1. no jump is detected and the process goes on from the sample y_{n_D}

2. jump are detected; if $n_{D'}$ is the detection time the nearest to n_0, the segment $[y_{n_0},\ldots,y_{n_D}]$ is forgotten, the new segment $[y_{n_0},\ldots,y_{n_{D'}}]$ is validated and the process goes on forward again from the sample $y_{n_{D'}}$.

   This resulting procedure will be referred as the forward-backward divergence test.

See a first result on the Figure 13

II.4 - THE PULSE METHOD [2]

As the divergence statistics is sensitive to the glottal pulses (as can be seen on figure 10).

A new method is examinated to eliminate this influence; it concerns only the voiced signals. So a preliminary detection between voiced and unvoiced signals is necessary (see the next paragraph).

a. The model

The presence of the glottal impulse is taken into account in the excitation of the acoustic channel, hence the model $M(A,\sigma)$ is as follows :

$$y_n = \sum_{i=1}^{p} a_i \, y_{n-i} + \sigma_n \, e_n \qquad (16)$$

where (e_n) is a standard white noise and the variance σ_n^2 is periodic, piecewise constant, with values σ^2 and $\varepsilon^2\sigma^2$ (Figure 14).

We first assume that the values T, l_p and the locations of the pulses are known.

To derive simple and robust testing and identification procedures, we proceed as follows :

. design a test or identification procedure for a given ε

. let ε go to zero, and use the resulting limit procedures for the implementation.

Let P be the whole set of pulse time intervals, the study of the maximum likelihood of the observations $(y_1,\ldots,y_n)$ shows that the limit result $M^*(A^*,\sigma^*)$ is obtained by minimizing the expression :

$$L_n = \sum_{\substack{m \notin P \\ m \leqslant n}} \frac{\left(y_m - \sum_{i=1}^{p} a_i \, y_{m-1}\right)^2}{\sigma^2} + (N - l_{pitch}) \operatorname{Log} \sigma^2 \qquad (17)$$

where :

$$1_{\text{pitch}} = \sum_{m \leqslant n} 1_p(m) \;;\; 1_A(x) = 1 \text{ if } x \in A \;;\; 0 \text{ otherwise.}$$

b. The test

The same statistics as before, namely the divergence (équation 11-12); gives the following test after approximation :

$$W'_n = \sum_{\substack{k \notin P \\ k \leqslant n}} w_k + |p| \left(1 - \frac{\sigma_0^2}{\sigma_1^2} \right) \tag{18}$$

Hence the resulting design methodology follows for the identification as well as testing procedures :

When the pulse set P is known, delete the terms corresponding to the indices belonging to P in the equations and apply the previous testing methods (this does not correspond to consider the signal $(y_n)_{n \notin P}$ as the new signal to be monitored).

c. The practical implementation

Three tasks have to be performed :

1. identification of the pulse set P
2. identification of the models
3. computation of the test

1. Identification of the pulse set P : the problem of the locations of the pulse is also a jump detection problem ; we consider the length 1_p of every pulse as constant, and we detect on line a jump in the variance of the excitation.

Assuming that the AR parameters are known, a standard Page-Hinkley stopping rule gives the classical decision statistics [12] :

$$P_n = \sum_{n \leqslant k \leqslant n+l_p} \frac{(y_k - \sum_{i=1}^{p} a_i\, y_{k-i})^2}{\sigma^2} \qquad (19)$$

The maxima of (P_n) give the locations of the beginning of the pulse intervals.

To initialize this search, the AR parameters are computed on a short window with $P = \emptyset$; then two maxima of (P_n) separated by a minimum length T_{min} are searched inside a large window of length T_{max} (Figure 15). A_n estimate of the pitch T is given by difference.

To take into account the fact that the pitch period varies slowly in practice, the statistics P_n is then monitored only inside a "search window", the position of which is determined from the estimate T and the last detected pulse (Figure 15).

2.Identification of the models : The models $M_0^*(A_0^*,\sigma_0^*)$ and $M_1^*(A_1^*,\sigma_1^*)$ are located as in the divergence test (Figure 8). As soon as a new pulse and therefore a new "search window" are located, the growing window model and the sliding window model are updated by solving the equations (16).

3.Computation of the test : A simplified version of the test statistics is computed with Hinkley's stopping rule, i.e.

$$W_n = \sum_{\substack{k \leqslant n \\ k \notin p}} (w_k + \delta) \qquad (20)$$

The identification and testing procedures are monitored in parallel.

d. A first experiment

Exact least-squares algorithms used for the identification allow to reduce the length of the sliding window to 150 samples, for a model of order 16 and a pulse length l_p equal to 8 samples.

The same remarks as for the divergence test hold for the behavior of this statistics : the drift and the threshold are fixed respectively to 0.5 and 40 for voiced segments.

Long voiced segments are also processed by a backward procedure which uses only the divergence test without pulses. This resulting procedure is named the forward-backward pulse test (Figure 16).

The parameters T_{min}, T_{max}, clearly depend on the speaker and are the most difficult to tune : if the choice is poor, the estimate T can be equal to a multiple of the pitch period and the interest of the method is last. As we can see on the figure 16, the estimate T is the double of the pitch period during the first segment of $|\tilde{\varepsilon}|$ and is good during the second.

If we want to implement this procedure in a speech recognition system, this problem might be solved by coupling the segmentation procedure to a pitch detector, a good estimation of T would be found and allows to tune T_{min} and T_{max}.

III. EXPERIMENTAL RESULTS

Three different databases are used to evaluate the four methods. They consist of three different vocabularies. The first database is generated by one male speaker and is composed of two lists of ten phonetically balanced sentences. For the second, sequences of digits are pronounced by ten speakers (four male and six female). The sampling rate of these signals is equal to 12 800 Hz. The third database whose sampling rate is equal to

10KHz is composed of "logatomes", words of two syllables without signification but extracted from a carrier sentence.

III.1 - NATURE OF THE OBTAINED UNITS

All the results are compared to a hand-made segmentation which is performed on the corresponding sonagrams (Figure 17).

The segments which are obtained correspond to homogeneous parts of the signal, but it is more interesting to define them by characterizing the type of the boundaries. A jump time always corresponds to an articulatory or an acoustic change : it can be a voice or frication onset or termination, a release, a closure which we call an "event" [1] or a change of the formantic structure. So most segments are indeed stationary parts of phonemes, but sometimes segments can also be either slow transitions during which the formantic variations are monotonous or short "events".

We can see on the figure 13, the phoneme |i| decomposed into three units :

. the first one is the onset of the new formantic structure of an |i| under a nasal context, it is a slow transition,

. the second one is the stable part of the phoneme

. the third one starts with the beginning of noise and ends with the loss of the formantic structure (a characterization via "events").

An other example is exhibited in the unvoiced plosive |t| on the figure 17 :

. the first unit begins with the loss of the formantic structure and ends with the voice termination

. the second one is a silence

. the third one is the burst : frication onset, release and expiration are gathered in this unit. These events are too close each other to be distinguished.

The next paragraph gives indications about the results of each method in comparison with this notion of acoustic or articulatory changes.

III.2 - COMPARATIVE STUDY

The four methods give comparable results but each one have advantages and drawbacks.

The forward-backward pulse test does not exhibit the best performances. Even though the locations of the pulses are exact (Figure 18) the learning of some parameters depends on the speaker (Figure 16); formantic changes between two vowels such as |yi| or transitions in a nasal context such as |ã n| are missed.

The instrumental method provides in unvoiced areas over-segmentations which we cannot explain with our notion of unit. Moreover some boundaries are not well located. However, the behavior of this test is excellent in a nasal context (Figure 6 |ãny|).

The two best procedures are certainly Brandt's test and the forward-backward divergence test (Figure 4 and 13). Each detection correspond to a real phonetic change, except some ones of Brandt's test which may be considered as over-segmentations. Of course, all the "events" or formantic changes we can find by hand are not detected; but these omissions are rare and not dramatic for an phonemic identification : for example, the burst of a plosive may form a single segment with an unvoiced fricative |ks|. A comparison between the results of these two last methods (Figure 4 and 13) shows that the divergence test has a more regular behavior than Brandt's test and a change is detected without ambiguity : we can see during the unit |a| and |i| (Figure 4) some troubles of Brant's test statistics. For these reasons and due to its lower computational cost, the divergence test appears as the best method to preprocess on line the speech signal in a recognition system (its computational cost is mainly due to the parallel processing of the sliding autocorrelation method and the sequential Burg method).

Let us emphasize that each of the three last methods is speaker independent.

IV. CONCLUSION

A new approach has been presented for speech automatic segmentation, and four algorithms based on statistical techniques are compared.

The main conclusion is that the four methods lead to the some kind of units : these segments are characterized by their boundaries, articulatory or acoustic events or formantic changes.

One of the procedures, the forward-backward divergence test, is robust and precise enought to be chosen as a preprocessing of the speech signal, It results in a more releable representation of each unit, and in some segmental information.

An interesting extension of this approach would be the simultaneous detection and labelling of the changes in terms of acoustico-phonetic features : work is in progress towards this goal.

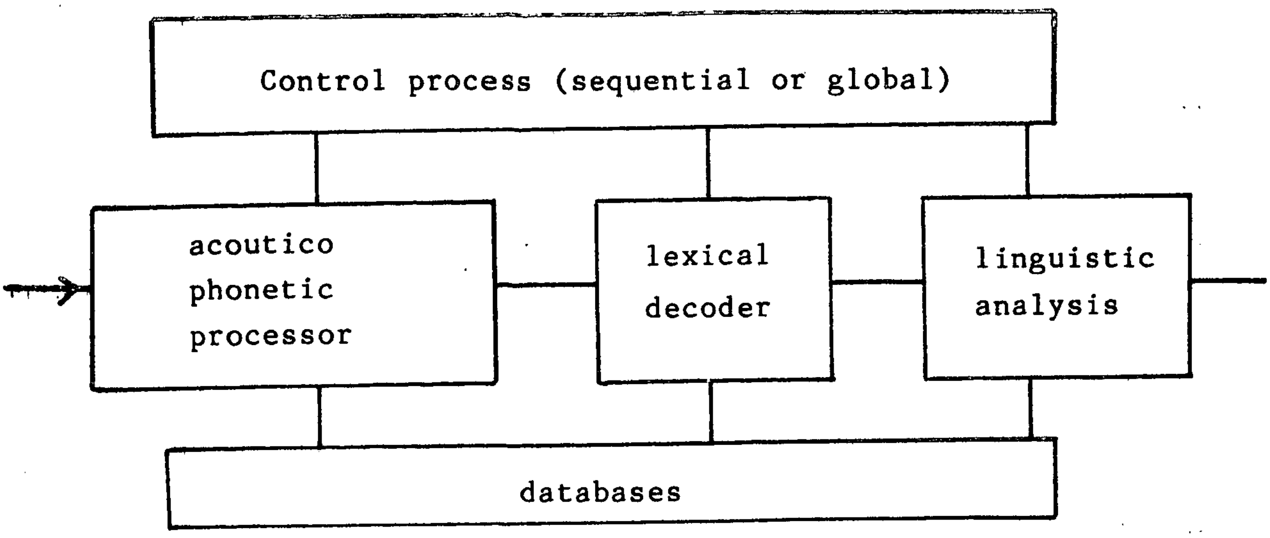
Control process (sequential or global)
acoutico phonetic processor
lexical decoder
linguistic analysis
databases

Figure 1 : An analytic recognition system

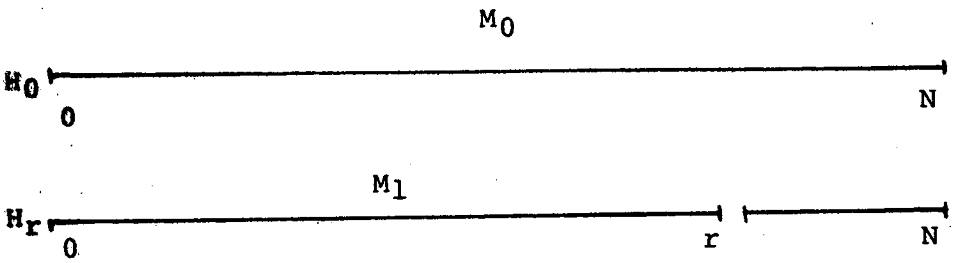

figure 2: Locations of the models by the method of Brandt

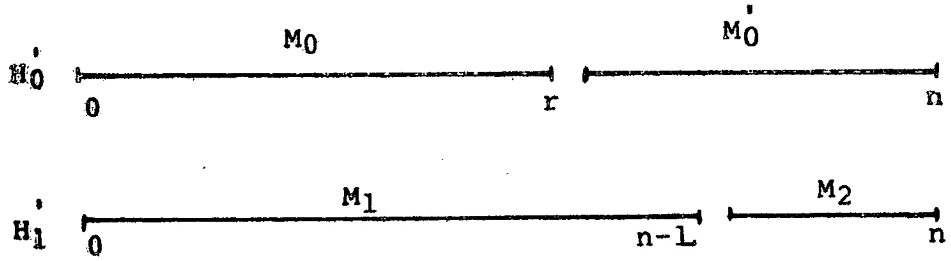

figure 3: Locations of the models during the change time estimation by the method of Brandt

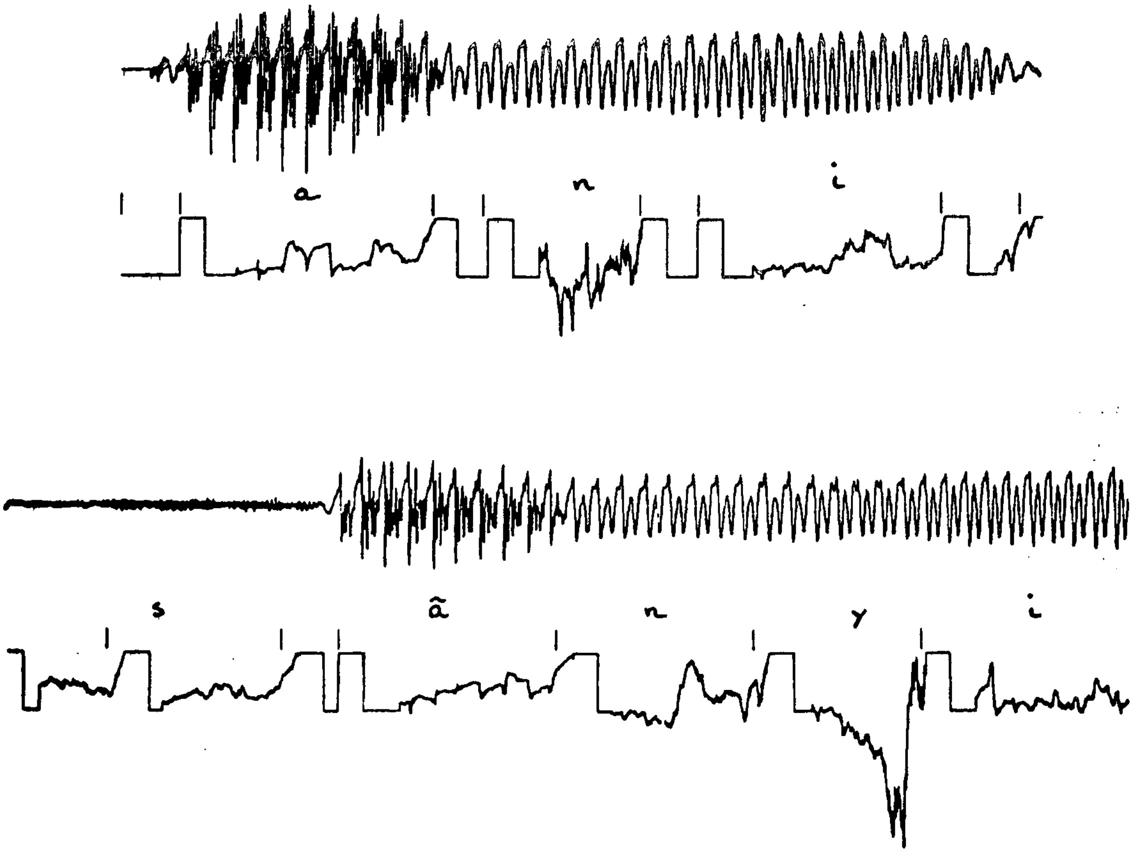

Figure 4 : Results of Brandt's test (D_n) on the sentence "annie s'**ennuie**..."

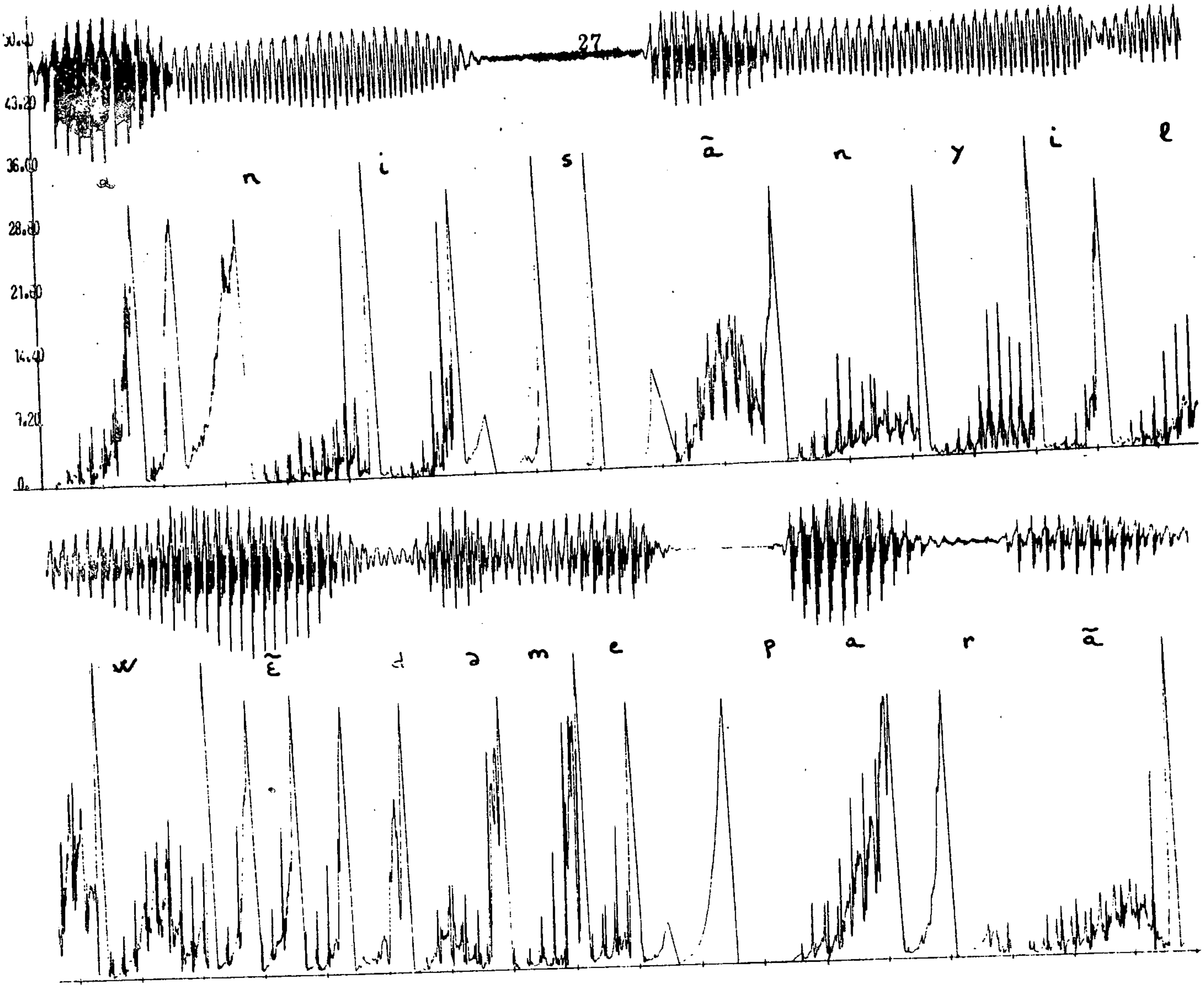

Figure 6 : Results of the instrumental method on the sentence "annie s'ennuie loin de mes parents"
(vertical lines correspond to a jump detected par (T_n) and spires to a jump detected par (R_n))

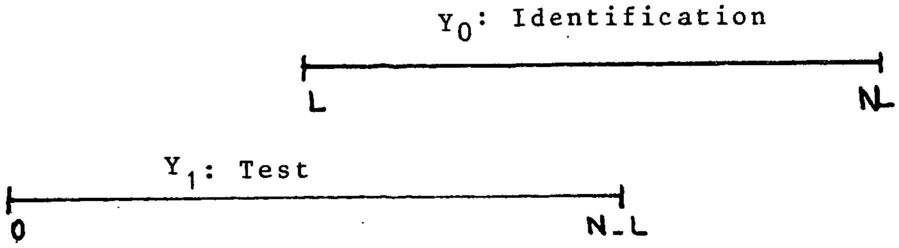

Figure 5: Locations of the signal blocks for the Instrumental method.

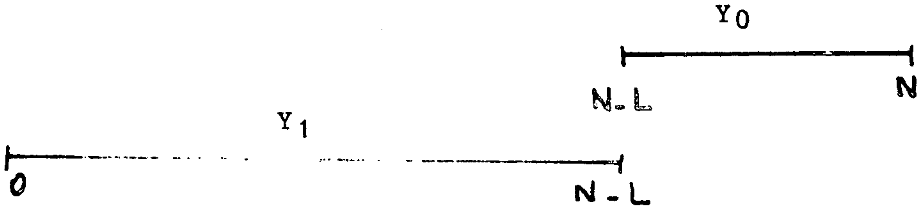

Figure 7: Locations of the models in an experiment of the Instrumentl Method.

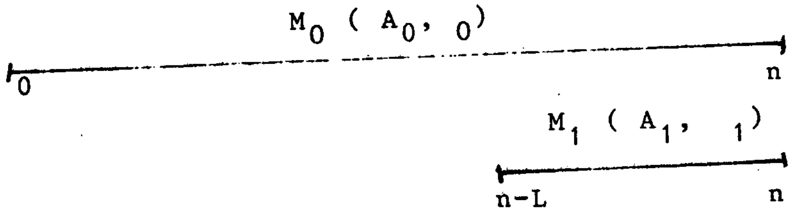

Figure 8: Locations of the models for the divergence test.

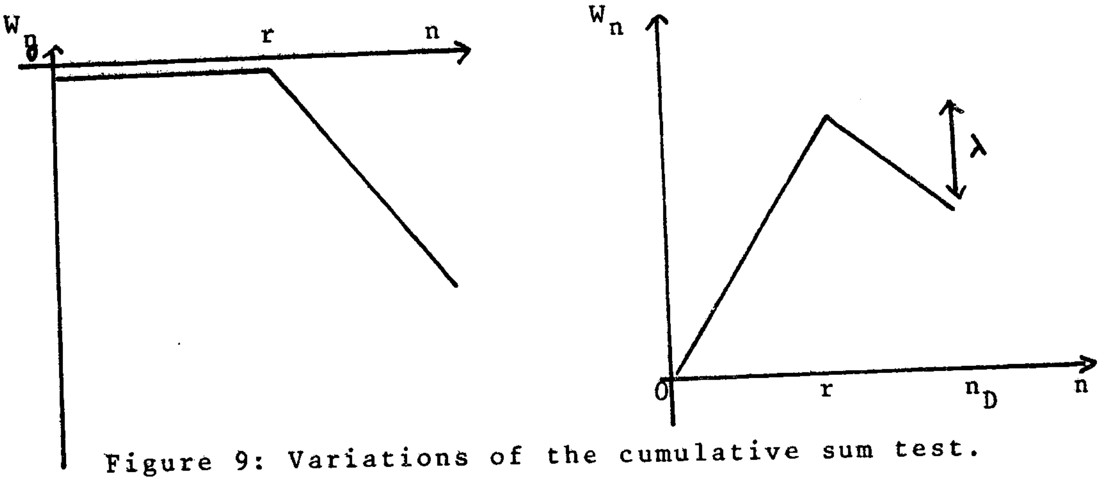

Figure 9: Variations of the cumulative sum test.

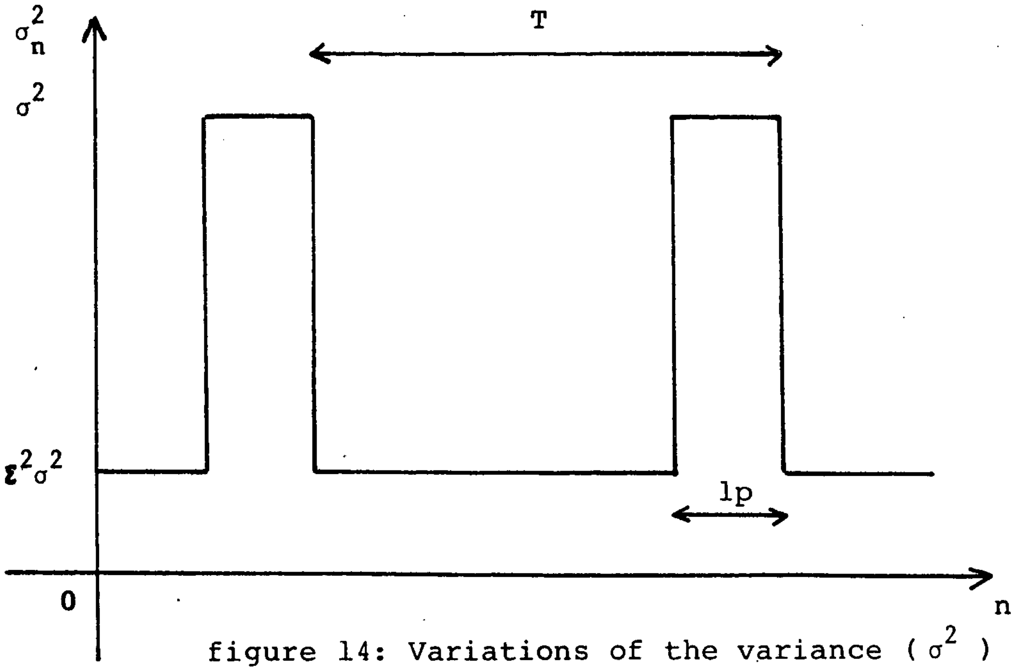

figure 14: Variations of the variance (σ^2) in the pulse model (T is the pitch period and lp the pulse length)

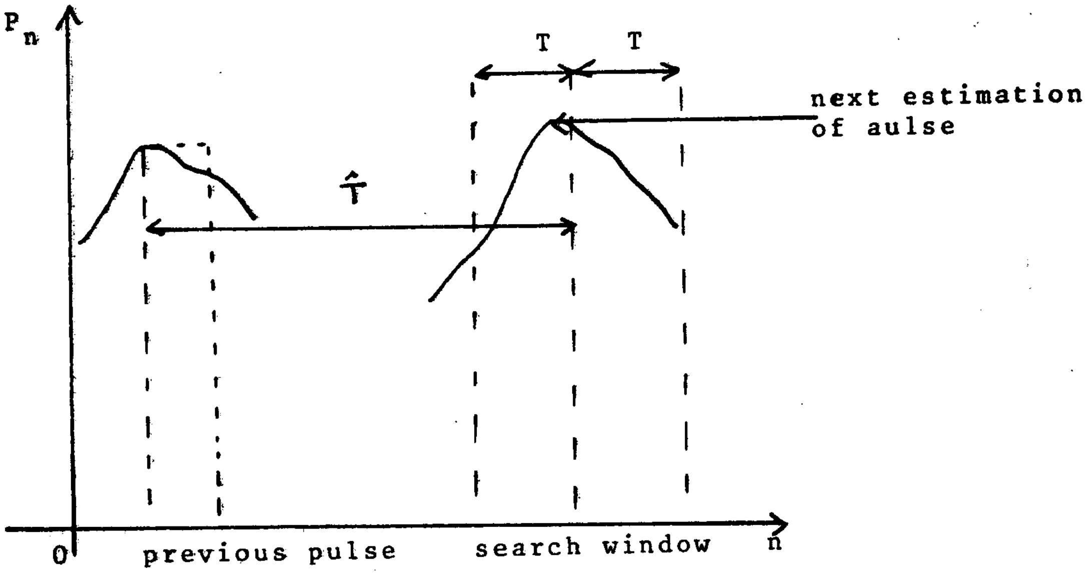

Figure 15: " Search window" of the next pulse.

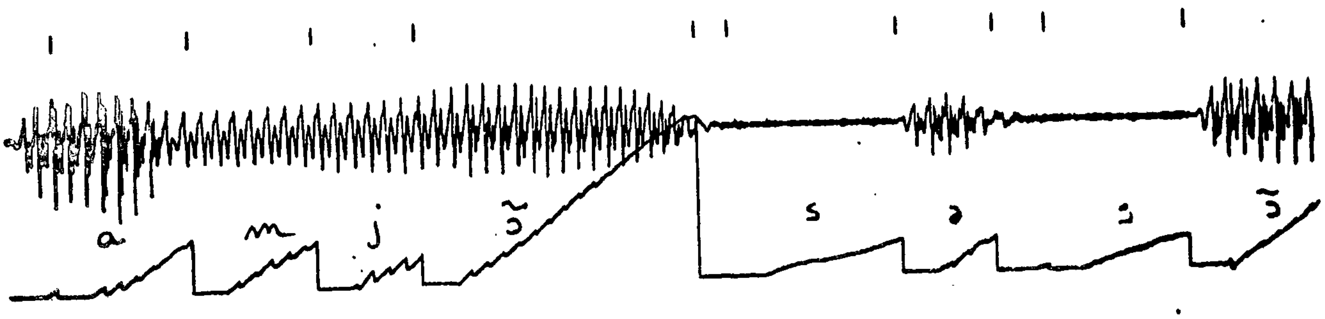

Figure 10 : Different behavio rs of the divergence test

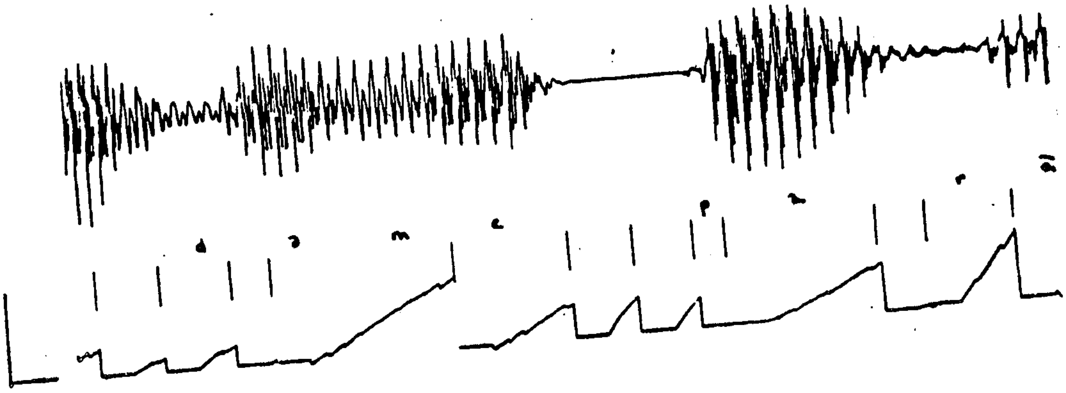

Figure 11 : A forgotten jump between |ə| and |m|, because of the dissymmetry of the test (W_n)

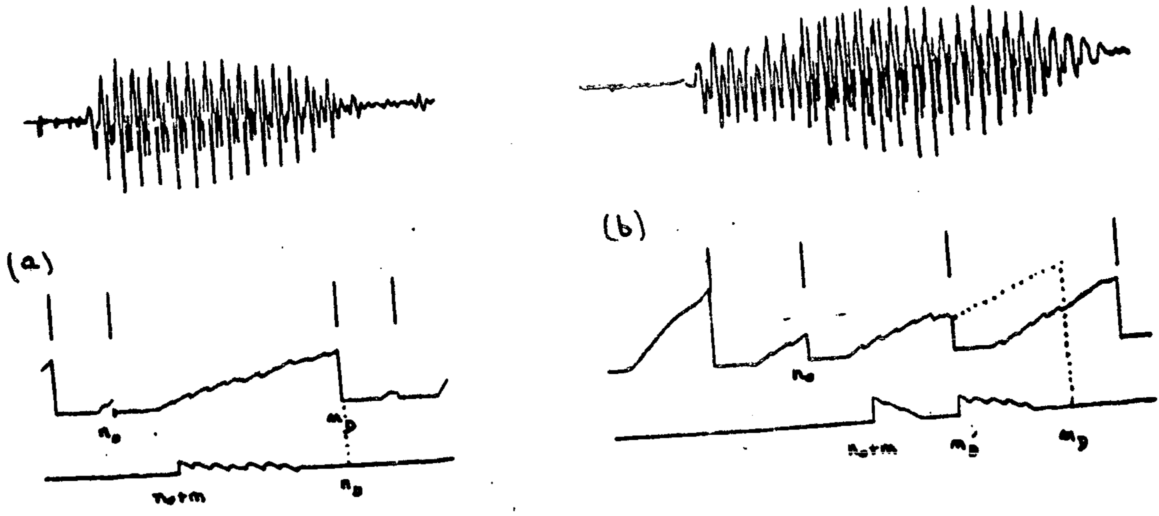

Figure 12 : The forward-backward divergence test
(a) no new jump
(b) a new jump is detected

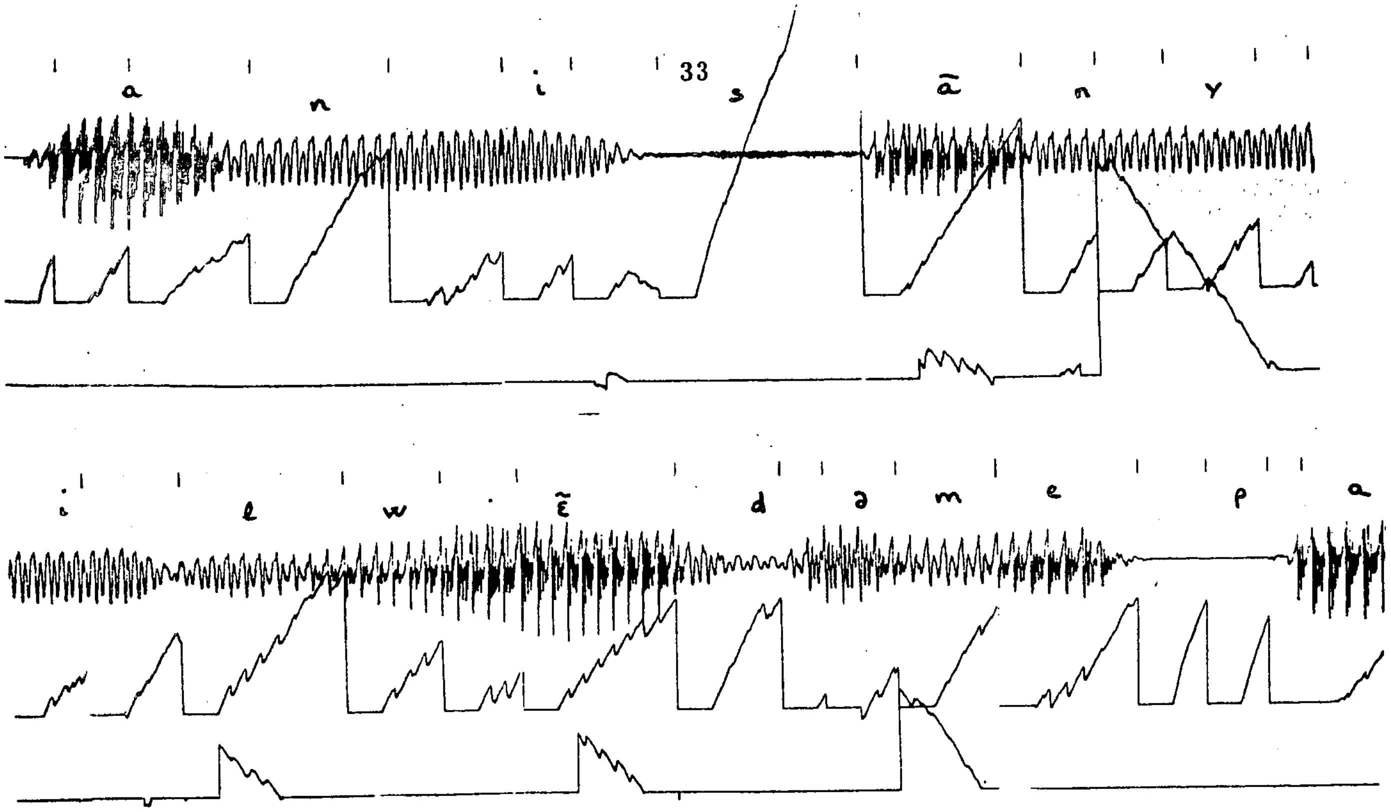

Figure 13: Results of the divergence test W_n forward-backward on the sentence "annie s'ennuie loin de mes parents"

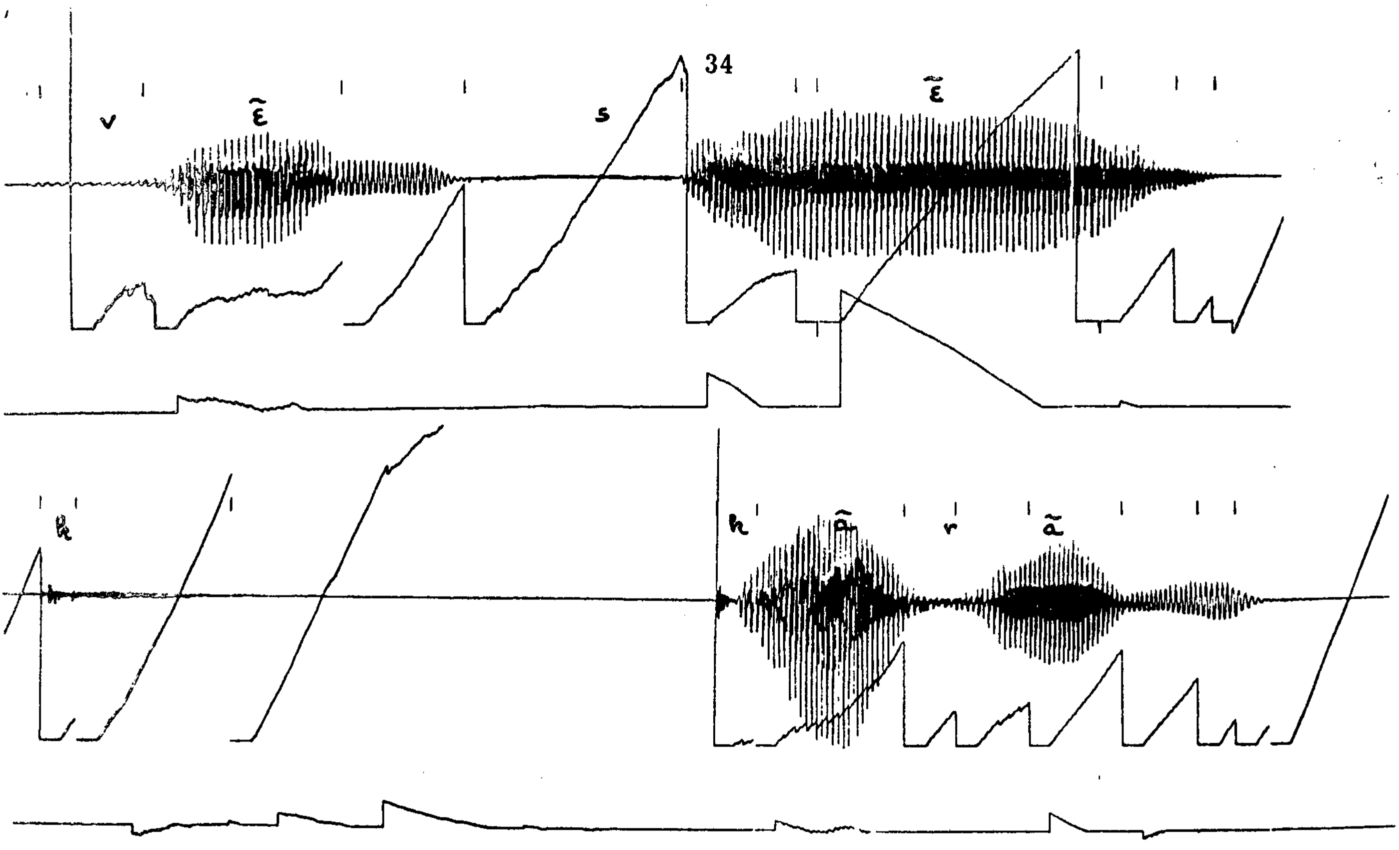

Figure 16 : Results of the pulse method forward-backward on the sequence "25" pronounced by a female speaker

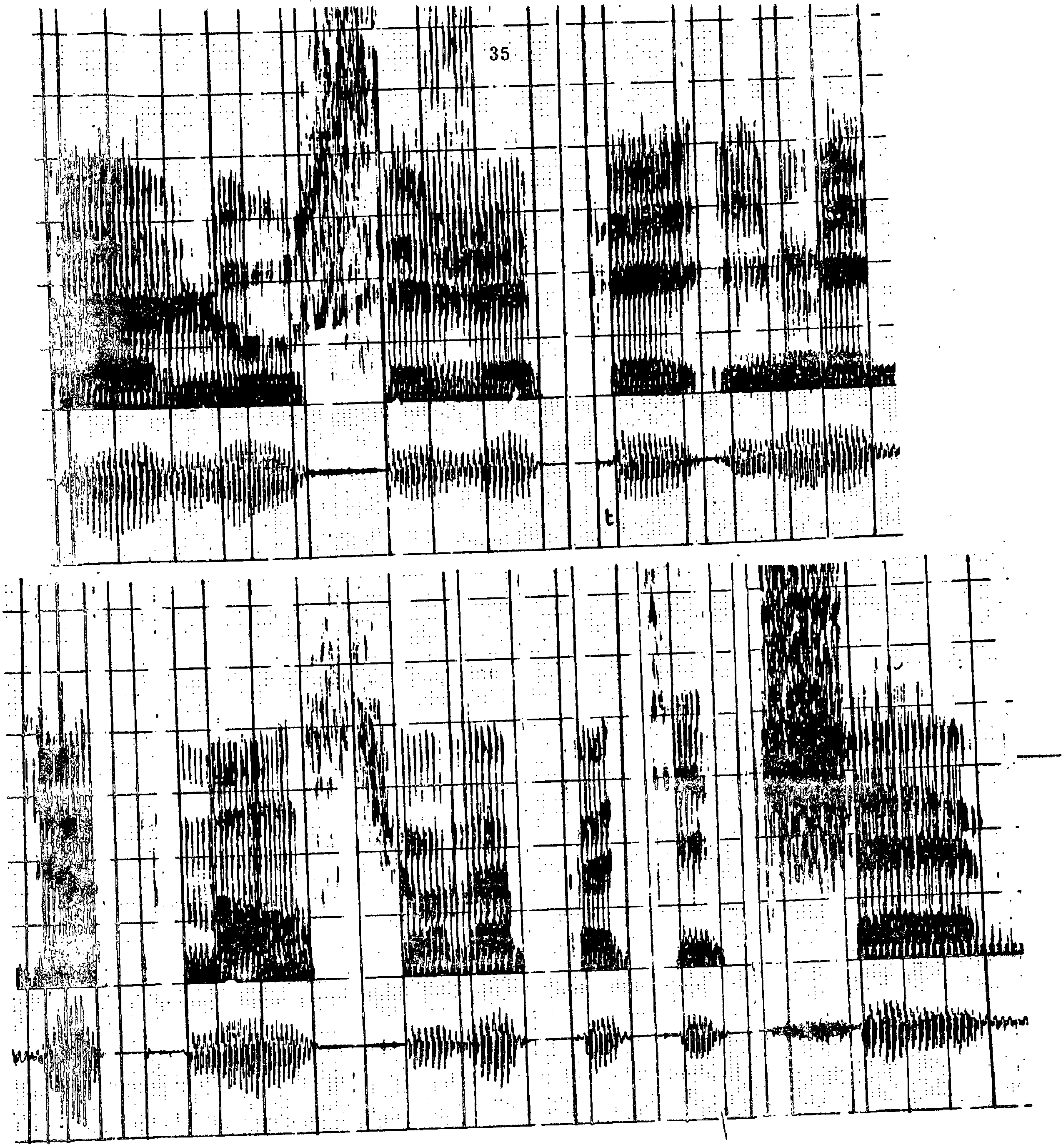

Figure 17 : Results of the divergence test (W_n) forward-backward on the sentence "un loup s'est jeté immédiatement sur la petite chèvre".

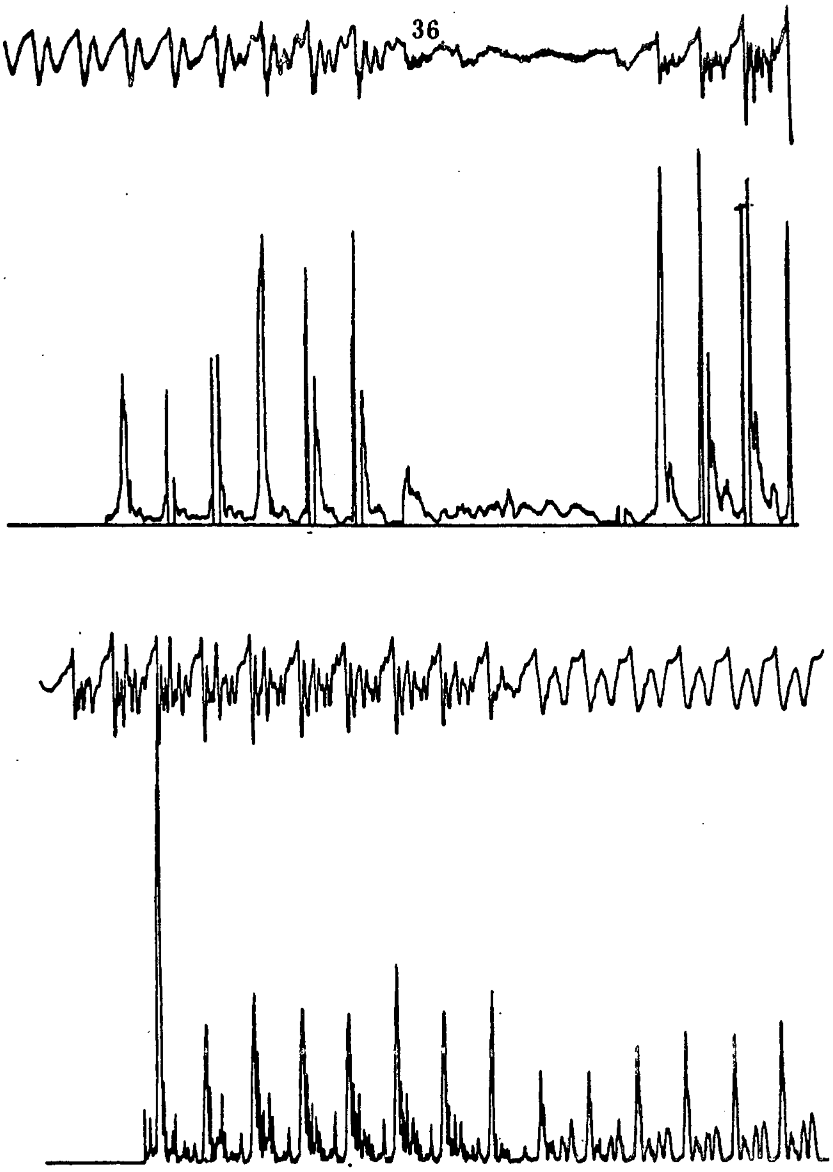

Figure 18 : Behavior of the test (P_n) on the strings |ira| and |ãn|

REFERENCES

|1| C. ABRY : "Organisation segmentale et temporelle du signal de parole en fonction de sa production". **Rapport Institut de Phonétique** de Grenoble. 1984.

|2| R. ANDRE-OBRECHT : "Segmentation automatique du signal de parole". **Thèse 3ème cycle.** Université de Rennes I. Mai. 1985.

|3| M. BASSEVILLE. A. BENVENISTE : "Sequential Detection of abrupt changes in spectral characteristics of digital signals". **IEEE Trans. on Information Theory.** vol.29. n°5 : 703-723. September 1983.

|4| M. BASSEVILLE. A. BENVENISTE. G. MOUSTAKIDES : "Detection and diagnosis of abrupt changes in modal characteristics of non stationary digital signals". To appear in **IEEE Trans. on Information Theory.** May. 1986.

|5| A. BENVENISTE : "Algorithmes simples d'estimation en treillis pour les séries longues". **Outils et modèles mathématiques pour l'automatique. l'analyse des systèmes et le traitement de signal.** vol.2. Editions du CNRS. 1982.

|6| A. VON BRANDT : "Detecting and estimating parameters jumps using ladder algorithms and likelihood ratio test". **ICASSP** Boston. 1983.

|7| A.V. BRANDT : "Modellierung von signalen mit sprunghaft veranderlichem leistungsspektrum durch adaptative segmentierung". **Dissertation.** München. 1984.

|8| R.J. FONTANA. R.M. GRAY. J.C. KREFFER : "A symptotically mean stationary chunnels". **IEEE Trans. on Information Theory.** vol.IT 27. n°3. May. 1981.

|9| B. FRIEDLANDER : "The over determinated recursive instrumental variable method". **IEEE Trans. on Automatic control.** vol.29. n°4. April. 1984.

|10| J.P. HATON. M. LAZREK : "Segmentation et identification des phonèmes dans un système de reconnaissance automatique de la parole continue". **4ème congrès AFCET** "Reconnaissance des formes et intelligence artificielle. January. 1984. Paris.

|11| D.V. HINKLEY : "Inference about the change point from cumulative sum tests". Biometrika. vol.58. n°3: 509-523. 1971.

|12| R.H. JONES, D.H. CROWELL, L.E. KAPUNIAI : "Change detection model for serially correlated multivariate data". **Biometrics**, vol.26, n°2: 269-280, June 1970.

|13| O.H. KLATT : "Scriber and lafs : two new approaches to speech analysis". **Trends in speech recogntion**, Wayne A. Lea, Prentice Hall.

|14| S. KULLBACK, A. LEIBLER : "On information and sufficiency". **A.M.S.**, vol.22: 73-86, 1951.

|15| B. LOWERRE : "The harpy speech understanding system". **Trends in speech recognition**, Wayne A. Lea, Prentice Hall.

|16| J.O. MARKEL, A.H. GRAY : "Linear prediction of speech". Springer-Verlag, N.Y., 1976.

|17| G. MOUSTAKIDES, A. BENVENISTE : "Detecting changes in the AR parameters of a non stationary ARMA processes". To appear in **Stochastics**.

|18| N. VIGOUROUX : "Décodage acoustique phonétique de la parole continue multilocuteur : élaboration d'une base de connaissance". **Thèse de 3ème cycle**, Toulouse, Janvier 1984.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique