ETUDE DES SEGMENTS TRANSITOIRES EN PAROLE A L'AIDE DE MODELES AR EVOLUTIFS ET DU CRITERE D'AKAIKE

Régine André-Obrecht*, Bernard Delyon**

*IRISA/CNRS
**IRISA/Centre INRIA de Rennes
Campus de Beaulieu - 35042 Rennes Cedex - France

RESUME

Une segmentation automatique du signal de parole, basée sur la divergence de Kullback permet de mettre en évidence trois différents types d'unités acoustiques :

- des zones quasi-stationnaires
- des zones transitoires
- des zones "événementielles".

Alors que les premières correspondent à la réalisation de phonèmes-cibles, et les dernières sont liées à un changement brusque du système de production vocale (telles les variations de l'excitation du conduit vocal, son occlusion...), les zones transitoires sont révélatrices d'une modification articulatoire plus lente synonyme d'une évolution formantique. La reconnaissance phonétique en parole continue passe naturellement par leur détection et leur interprétation. Seule leur localisation fait l'objet de ce papier.

L'étude de cette évolution nous a conduit à supposer que chaque segment non événementiel pouvait être modélisé par un modèle autorégressif gaussien satisfaisant à l'une des contraintes suivantes :

- les coefficients autorégressifs sont indépendants du temps, le modèle est stationnaire et le segment est stable ;
- les coefficients autorégressifs dépendent du temps, le modèle est évolutif et le segment est "transitoire".

La méthode consiste à identifier sur chaque segment deux modèles autorégressifs, un stationnaire et un évolutif. L'identification du modèle évolutif est effectuée à l'aide des méthodes développées par Y. Grenier à propos de modélisation de signaux non-stationnaires ; le critère d'information d'Akaïke est calculé pour chaque modèle ; la plus faible des deux valeurs valide l'hypothèse correspondante.
Les expérimentations ont été réalisées sur des phrases phonétiquement équilibrées et différentes possibilités de mise en oeuvre ont été étudiées, notamment :

- l'ordre de chaque modèle,
- le nombre de fonctions sur la base considérée,
- un possible fenêtrage du segment d'analyse,
- l'utilisation d'un critère théoriquement plus robuste tel que celui de Rissanen.

La lecture en parallèle des résultats de ces tests et des spectrogrammes montre la réelle interprétation des segments détectés évolutifs : ils correspondent aux transitoires entre phonèmes voisées, aux noyaux vocaliques durant lesquels la cible phonétique n'a pu être atteinte ou aux voyelles nasales.

ABSTRACT

An automatic segmentation of the continuous speech signal, based on the Kullback divergence test, provides three different types of acoustic units :

- stationary segment
- transient segments
- short segments

associated with three phonetic events :

- production of a target phoneme
- continuous change in articulatory configuration
- abrupt transition (closure, burst...).

The aim of this paper is to develop an algorithm which discriminates between the first and the second class (segments of the third one are easily detected, see [3]). This discrimination is based on two modelisations of the speech signal, one for each class :

- a standard autoregressive model
- a time-varying autoregressive model (the autoregressive coefficients are time dependent).

An identification of both models is carried out on each segment, the second one being done by using Yves Grenier's method ; for each model, Akaike's information criterion, involving the number of parameters in the model and the likelihood (the residual error energy), is calculated in order to select the best hypothesis, i.e. the one with the lowest value of the criterion.

Experimentations have been done on phonetically balanced sentences with different implementations taking into account the following issues :

- model order
- number of functions in Grenier's model
- pre-windowing of segment
- utilization of Rissanen's criterion

The confrontation of the results with spectrograms shows the right interpretation of transitory segments : they correspond to transitions between voiced phonemes, to vocalic nuclei for which the phonetic target has been missed, and to nasal vowels.

INTRODUCTION

La réalisation d'un système de reconnaissance automatique de parole continue se heurte actuellement au problème fondamental de segmentation et d'identification d'un continuum acoustique en unités d'ordre phonétique. La source du problème est la profonde variabilité du signal de parole. Cette variabilité contextuelle a des causes multiples dont les principales sont sans doute [8] :

- la vitesse d'élocution
- le "phénomène d'assimilation" dû au recouvrement de mouvements articulatoires,
- l'effet coarticulatoire dû aux variations continues des articulateurs.

Les taux de reconnaissance phonétique obtenus lors de la lecture de spectogrammes par des experts phonéticiens (> 85%, [6]) laissent supposer que ces informations sont présentes dans le signal acoustique. Les performances des systèmes experts développés récemment, valident cette affirmation, sans pour autant donner la solution.

Plusieurs investigations sont actuellement tentées afin d'introduire cette information dite "dynamique" ; elles se regroupent en deux tendances :

- la recherche d'une segmentation robuste au sens acoustique [1,9]; les zones sont infra-phonétiques et nécessitent une nouvelle définition des lexiques
- la recherche de nouvelles paramétrisations de type "régression de coefficients" [4].

Les performances de ces systèmes semblent nettement améliorées ; elles montrent que de réels progrès peuvent être obtenus en décodage acoustico-phonétique, par ce biais.

Notre étude consiste à faire coopérer ces deux approches en discriminant, après segmentation automatique, les zones évolutives des zones stables par un critère statistique, le critère d'information d'Akaïke. Sa mise en oeuvre et ses résultats font l'objet de ce papier.

II. MODELES EVOLUTIFS ET CRITERE D'AKAIKE

II.1 - Nature des segments

Dans notre approche, le premier traitement effectué sur le signal de parole est une segmentation automatique basée sur un test statistique : l'algorithme de divergence "forward-backward" [1]. Cette méthode permet de localiser précisément les changements dans les caractéristiques spectrales, révélateurs d'événements articulatoires. Une rupture peut être un changement brusque comme le début ou la fin d'un voisement, d'une friction, d'une occlusion ou une modification du mode ou lieu d'articulation en période voidée.

La segmentation conduit donc à trois types d'unités (figure 1) :

- des zones courtes dites "événementielles",
- des zones transitoires durant lesquelles se produit une variation spectrale lente ; il peut s'agir d'une transition entre deux phonèmes, de la réalisation incomplète d'une voyelle, ou de la production des voyelles fortement évolutives telles que les semi-voyelles ou les voyelles nasales,
- des zones quasi-stables qui correspondent à la réalisation d'une cible qu'elle soit voyelle ou consonne.

Alors que la détection des zones événementielles par des méthodes classiques [3] ne pose pas de problème a priori, la distinction entre les deux autres types d'unités est un sujet plus difficile.

L'idée principale est d'essayer de représenter ces segments de parole par des modèles évolutifs et de mesurer statistiquement l'information qu'un tel modèle apporte par rapport à la modélisation stationnaire classique. Cette approche a été envisagée pour reconnaître les transitions consonnes-voyelles [2]. Notre but ultérieur est d'étendre cette caractérisation.

II.2 - Modèles évolutifs

Nous faisons l'hypothèse qu'au cours d'un segment de parole la structure formantique varie linéairement en fonction d'un indice temps τ, $\tau \in [0, T]$.

Notons $F^{(i)}$ le ième formant ; nous avons :

$$F^{(i)}(\tau) = F_0^{(i)} + \Delta F^{(i)} \tau$$

pour $\tau \in [0,T]$, $i = 1,...q$.

Si de plus, nous supposons que le signal peut être représenté à chaque instant par un modèle autorégressif :

$$Y_n = \sum_{i=1}^{p} a_i(\tau)\, Y_{n-i} + e_n$$

$$\text{var}\, e_n = \sigma^2$$

chaque coefficient autorégressif apparait comme combinaison linéaire des fonctions

$$\cos\left[\sum_{i=1}^{q} \varepsilon_i \Delta\omega^{(i)}\right]\tau$$

$$\sin\left[\sum_{i=1}^{q} \varepsilon_i \Delta\omega^{(i)}\right]\tau$$

où $\varepsilon_i \in \{0, 1, -1\}$

$\Delta\omega^{(i)} = 2\pi\, \Delta F^{(i)} / (Fech * T)$

Fech, la fréquence d'échantillonnage

Une première approche consiste à identifier sur chaque segment un modèle évolutif en développant chaque coefficient sur la base de fonctions de Fourier. Dans un deuxième temps, nous supposons que les variations formantiques sont faibles par rapport à la fréquence d'échantillonnage ; un développement limité nous conduit à décomposer les coefficients autorégressifs sur la base de Legendre.

La méthode d'autocorrélation étudiée par Y. Grenier [5] est utilisée pour l'identification de chaque modèle.

II.3 - Critères d'Information

Afin de quantifier la nature évolutive d'un segment, il convient de pouvoir mesurer l'adéquation des modèles évolutifs face aux modèles autorégressifs stationnaires classiques.

De nombreux critères ont été étudiés en vue de sélectionner un modèle statistique [7]. Les plus classiques dérivent de la mesure de l'information de Kullback, et peuvent se mettre sous la forme

$$-l_n(\hat{\theta}) + \alpha_n$$

où $\hat{\theta}$ est le vecteur de paramètres estimés caractérisant le modèle et $l_n(\hat{\theta})$ le log de la vraisemblance du modèle sur n observations ; α_n mesure le coût de l'accroissement en taille du modèle. Les deux principales estimations de α_n donnent le critère d'information d'Akaïke

$$AIC = -2\, l_n(\hat{\theta}) + 2P$$

et le critère de Rissanen :

$$RIC = -2\, l_n(\hat{\theta}) + 2P \log n$$

Nous utilisons ces critères en supposant que chaque segment de parole peut être représenté par un modèle autorégressif gaussien satisfaisant à l'une des contraintes suivantes :

- H_0 les coefficients du modèle $(a_i)_{i=1,P}$ sont indépendants du temps [cas stationnaire],
- H_1 les coefficients du modèle $(a_i)_{i=1,P}$ dépendent du temps [cas évolutif]

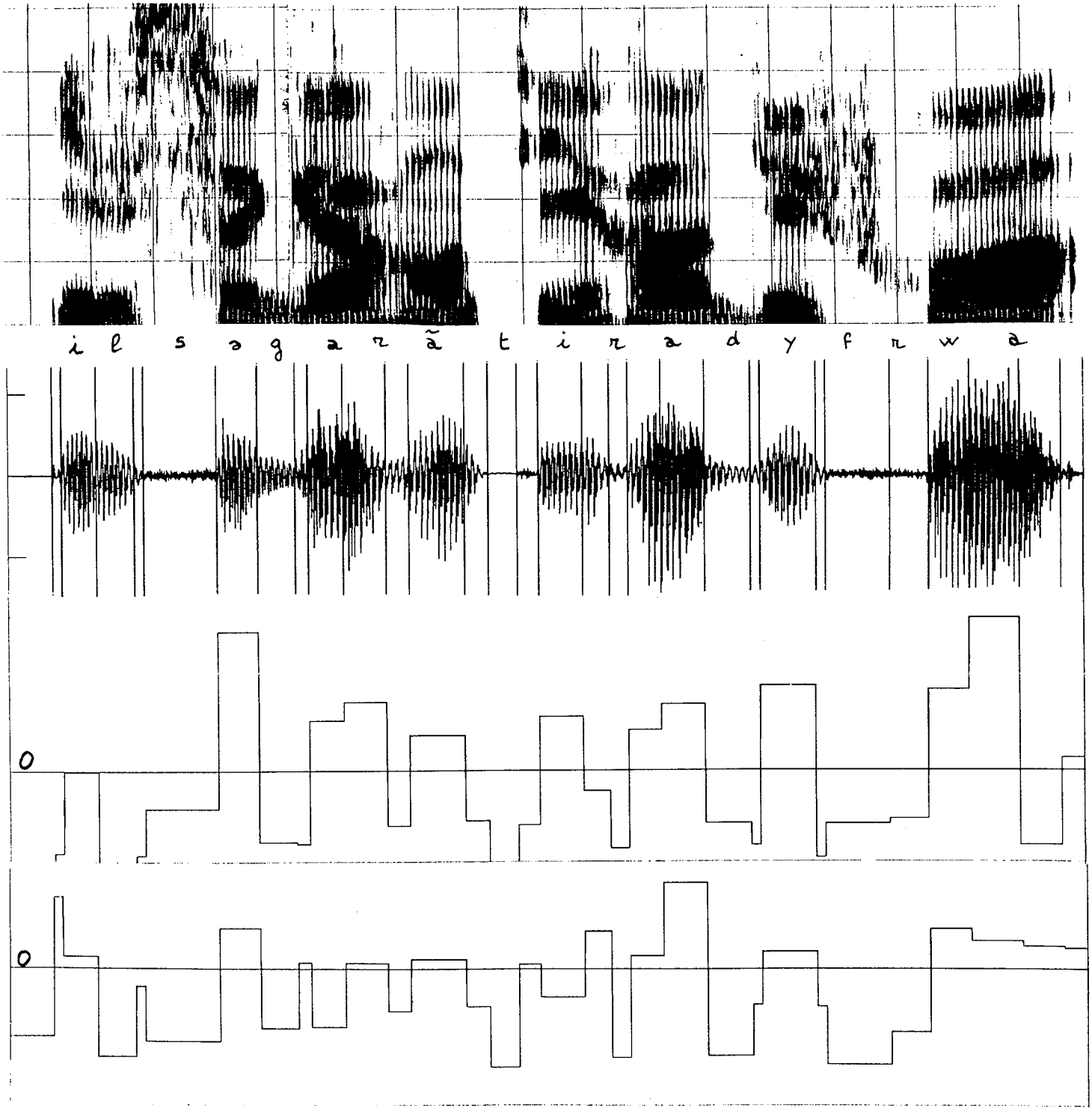

Figure 1 : Sonagramme et signal de la phrase
"il se garantira du froid"

Figure 2 : Valeurs du critère "Rissanen-Legendre"

Figure 3 : Valeurs du critère "Akaïke-Fourier"

L'identification de deux tels modèles nous conduit aux valeurs respectives du critère d'Akaïke :

$$A\,IC(0) = n \log \sigma_0^2 + 2\,P_0$$

$$A\,IC(1) = n \log \sigma_1^2 + 2\,P_1$$

où σ_i est l'énergie résiduelle pour chacun des modèles,
n est la longueur du segment considéré,
P_i est le nombre de paramètres libres de chacun des modèles.

Théoriquement, le segment est évolutif dès que :

$$A\,I\,C(1) < A\,I\,C(0)$$

ou

$$R\,I\,C(1) < R\,I\,C(0)$$

III - EXPERIMENTATIONS ET RESULTATS

Les expériences sont réalisées sur une liste de dix phrases phonétiquement équilibrées, et permettent de préciser les paramètres de l'algorithme.

III.1 - Choix des paramètres

Pour pouvoir représenter l'évolution globale d'un segment, il est souhaitable de faire l'identification des modèles en prenant comme fenêtre d'analyse la totalité des observations.

Nous verrons que cette option ne permet pas d'atteindre les segments où les cibles sont atteintes et nous envisageons aussi l'analyse sur une fenêtre de longueur fixe. L'expérience aurait dû, de plus, mettre en concurrence toutes les combinaisons possibles quant au nombre de fonctions de base et à l'ordre. Nous nous sommes limités et les expérimentations nous ont conduits à proposer deux solutions.

- les deux modèles sont identifiés sur la totalité du segment et comparés à l'aide du critère de Rissanen ; les coefficients évolutifs sont décomposés sur les trois fonctions de Legendre

$$1$$

$$2\tau - 1$$

$$6\tau^2 - 6\tau + 1$$

- les identifications sont réalisées sur une fenêtre de longueur fixe centrée sur le segment (40ms) et la décision est prise à l'aide du critère d'Akaïke ; les coefficients évolutifs sont décomposés sur les 5 fonctions de Fourier

$$1$$

$$\cos m\,\pi\,\tau$$

$$\sin m\,\pi\,\tau \quad , \quad m = 1,2$$

Dans tous les cas, l'ordre des modèles est l'ordre théorique calculé à partir de la fréquence d'échantillonnage et la longueur du conduit vocal, soit 12. Pre-emphase et fenêtrage (Hamming) n'ont pas apporté d'amélioration.

III.2 - Interprétation des résultats

Sur les figures 2 et 3 sont dessinées respectivement les courbes représentant les test de

Rissanen $\quad R\,I\,C(0) - R\,I\,C(1)$

et Akaïke $\quad A\,I\,C(0) - A\,I\,C(1)$

Toute valeur positive signifie théoriquement que le signal est mieux représenté par le modèle évolutif.

Le test Rissanen-Legendre met en évidence les zones à structure formantique marquée, c'est-à-dire les segments correspondant aux voyelles et semi-voyelles. Ce résultat est attendu dans le sens où les cibles phonétiques sont rarement atteintes en parole continue, particulièrement celles des voyelles. Il reste que certaines zones réellement transitoires ne sont pas détectées comme telles (exemple /i - r/). Ces zones présentent une certaine instabilité qui pénalise la vraisemblance du modèle identifié sur une large fenêtre qu'il soit évolutif ou non.

A l'aide du test Akaïke-Fourier sont détectées naturellement les semi-voyelles /ω/, les voyelles nasales /â/ auxquelles s'ajoutent les segments fortement influencés par le contexte ; citons, comme exemple sur la figure 3, les segments adjacents :

- au phonème /r/ : le deuxième formant chute,
- au phonème /g/ : une pince formantique apparait ; remarquons qu'elle n'a cependant pas été détectée sur un des trois segments du /a/ de /ga/.

L'identification des modèles sur 40 ms permet de localiser les segments où les cibles sont atteintes (/i/).

IV - CONCLUSION

Les résultats obtenus montrent, sur bon nombre de segments vocaliques, la supériorité (au sens d'Akaïke ou de Rissanen) de la modélisation évolutive vis à vis du schéma AR traditionnel. Ceci illustre non seulement le caractère profondément non-stationnaire du signal de parole mais aussi l'adéquation du modèle de Grenier. Si la détection de segments transitoires reste imparfaite, il convient de rappeler que les algorithmes ont tourné avec les valeurs théoriques des paramètres et ne nécessitent le règlage d'aucun seuil.

REFERENCES

[1] R. André-Obrecht : "A new statistical approach for the automatic segmentation of speech signals". IEEE Trans. on ASSP, vol., Janvier 1988.

[2] G. Boulianne, J.P. Tuback, Y. Grenier, G. Chollet : "Recognition of non stationary speech segments using auto-regressive time-dependent models". ICASSP, Tokyo 1986.

[3] B. Delyon, R. André-Obrecht, H.Y. Su : "Expériences en vue du décodage acoustico-phonétique à partir d'une recherche statistique d'événements acoustiques et d'un codage vectoriel". Journal d'Acoustique n°1, Septembre 1988.

[4] Sadaoki Furui : "A V.Q. based preprocessor using cepstral dynamic features for speaker independent large vocabulary word recognition". IEEE Trans. on ASSP, vol.36, n°7, July 1988.

[5] Y. Grenier : "Time dependent ARMA modelling of non-stationary signals". IEEE Trans. on ASSP, vol.31, n°4, Août 1983.

[6] F. Lonchamp : "Reading spectrograms : the view from the expert" in Fundamentals in computer understanding : speech and vision. Ed. J.P. Haton, Cambridge University Press, 1987.

[7] Ritei Skibata : "Criteria of statistical model selection". Rapport de Recherche, Université de Keio-Yokohama, Japon, Août 1986.

[8] J. Vaissière : "Speech Recognition : A tutorial" in Computer Speech Processing, Prentice Hall International.

[9] J.G. Wilpon, B.H. Juang, L.R. Rabiner : "An investigation on the use of acoustic sub-word units for automatic speech recognition". ICASSP, Dallas 1987.