DETECTION DES DEBUT ET FIN DE PAROLE EN ENVIRONNEMENT DIFFICILE

Régine ANDRÉ-OBRECHT, Jean Baptiste PUEL,
Serge EICHENE

IRIT-URA 1399 CNRS-UPS
118, route de Narbonne, 31062 Toulouse, France

RÉSUMÉ

La détection des début et fin de parole est un traitement stratégique dans un système de reconnaissance automatique de la parole, en milieu bruité. Nous proposons trois nouvelles approches dont l'originalité commune repose sur le couplage entre l'extraction de paramètres et une segmentation statistique du signal acoustique. Elles diffèrent de par la nature des paramètres (temporels ou fréquentiels) et la mise en oeuvre de ce couplage. Des expériences sont faites sur la base de données enregistrée au cours du projet ESPRIT ARS : des mots isolés ont été enregistrés dans une voiture aux vitesses de 90km/h et 130 km/h. Les résultats montrent qu'une parfaite discrimination entre segment "bruit" et segment "parole" est obtenue ; de plus une meilleure précision est atteinte quant à la localisation de ces début et fin de parole.

ABSTRACT

The detection of speech endpoints is a strategic process for speech recognition systems in adverse conditions, but it remains a rather delicate problem. We introduce three signal processing methods that offer a good robustness without requiring high level informations about the signal. The first approach uses temporal parameters, the two other frequential ones : a parametric method and a non-parametric one. We discuss and compare their performances using the ARS ESPRIT database (isolated words pronounced in a car). We show that these methods coupled with a statistical segmentation offer very good discrimination between noisy segments and speech segments, and a better precision for locating the speech boundaries.

1. INTRODUCTION

A l'heure actuelle, de nombreux systèmes de reconnaissance automatique de parole (RAP) obtiennent dans le cadre d'applications simples (mots isolés-monolocuteur) des résultats excellents en ambiance dite propre (type laboratoire). Malheureusement une dégradation drastique des performances se produit en conditions réelles d'application, dès que le milieu est dit "bruité" [Juang 91]. Le bruit ambiant provoque non seulement une distorsion additive du signal, mais il est à l'origine de modifications au niveau de la production même des sons (effet Lombard [Lombard 11]). Une des premières manières pour rendre plus robustes les systèmes de RAP, est de détecter l'apparition et la disparition de la parole dans le signal : la localisation du bruit seul permettra une meilleure adaptation des modèles de bruit en cours de reconnaissance ; la localisation correcte de la parole rendra, en phase d'apprentissage, les lois d'observation plus discriminantes et, en phase de reconnaissance, elle conduira à un calcul des vraisemblances plus optimal. Sujet apparemment simple, il reste délicat à traiter selon la connaissance que l'on a du bruit a priori, et fait l'objet de nombreux travaux [Lamel 81], [Junqua 92]. Les méthodes sont basées sur des calculs de paramètres classiques tels que l'énergie du signal ou son énergie résiduelle, en se limitant parfois à certaines bandes de fréquences ; des automates à seuils régissent les décisions. Nous proposons trois nouvelles méthodes de détection de début et fin de parole qui n'utilisent que des connaissances acoustiques grossières et qui exploitent les résultats d'une segmentation automatique robuste au bruit. L'une de ces méthodes opère dans le domaine temporel, les deux autres dans le domaine fréquentiel : une méthode paramétrique et une méthode non paramétrique. L'algorithme de segmentation automatique est la méthode de "Divergence" [André-Obrecht 88], elle permet de localiser les zones quasi-stationnaires du signal acoustique sans cependant pouvoir identifier les segments ; les frontières bruit/parole et parole/bruit sont détectées maiselles ne sont pas reconnues comme telles. Les informations apportées par chacun des trois détecteurs permettent d'une part d'affirmer pour chacun des segments s'il

s'agit d'un segment de parole ou de bruit, et d'autre part d'affiner éventuellement les résultats de segmentation.

Les trois méthodes sont expérimentées sur un corpus formé de mots isolés prononcés dans une voiture roulant aux vitesses de 90km/h et 130km/h (corpus délivré par Matra Communication et enregistré dans le cadre du projet ESPRIT ARS) ; il s'en suit que certaines des heuristiques utilisées pour la mise en oeuvre des détecteurs sont liées à la forme du spectre du bruit de voiture.

2. DETECTEUR TEMPOREL

Même si l'énergie du signal de parole n'est pas un paramètre robuste en milieu bruité, il n'en demeure pas moins que les noyaux vocaliques correspondent toujours à des maxima d'amplitude et que l'abcisse curviligne du signal temporel doit accuser une croissance plus rapide dans les zones de parole de type voisé que dans les zones de bruit.

Pour quantifier ce phénomène, nous introduisons les paramètres temporels suivants :

si

- $s(t)$ représente l'abcisse curviligne du signal de parole $y(t)$,
- t désigne l'indice temporel et L un nombre d'échantillons fixé a priori,

nous posons :

- $S(n) = s(nL) - s((n-1)L)$
- $DS(n) = S(n) - S(n-1)$

$S(n)$ représente une valeur moyenne de la "longueur de la courbe" par unité de temps (ou trame) et $DS(n)$ sa dérivée.

Sous l'hypothèse que le bruit ambiant est stationnaire par morceaux, la fonction S doit peu varier dans les zones de bruit, croître rapidement en début de parole et décroître en fin de parole ; en d'autres termes, la dérivée de cette fonction, DS, par ses brusques variations, doit indiquer l'apparition et la disparition de la parole dans le signal. Afin de détecter ces variations, nous introduisons deux seuils λ_1 et λ_2 :

- si $DS(n) \leq \lambda_1$, la trame n correspond à du bruit,
- si $DS(n) \geq \lambda_1$, la trame n correspond à de la parole bruitée.

Lors de la mise en oeuvre de ce détecteur, nous couplons cette décision aux résultats de segmentation obtenus a priori par la méthode "Divergence" selon la procédure suivante :

pour tout segment $[t_0, t_f]$,

- si la moyenne des valeurs prises par $DS(n)$ pour n vérifiant $t_0 \leq nL \leq t_f$, est inférieure à λ_1, le segment est classé "bruit",
- si cette même moyenne est supérieure à λ_2, le segment est classé "parole",
- dans tous les autres cas, le segment est classé en fonction des décisions prises sur les segments adjacents et en tenant compte de leur durée.

Le seuil λ_1 est déterminé automatiquement à partir de la valeur moyenne de DS calculée sur les premières trames de signal considérées comme du bruit ; le seuil λ_2 lui est proportionnel. Ce calcul est donc indépendant du bruit mais nécessite une réelle ambiance bruitée. La valeur de L est égale à 4 ms, soit 32 échantillons pour une fréquence d'échantillonnage de 8 kHz.

3. DETECTEUR BASE SUR LA DERIVEE SPECTRALE

Dans le cas d'une voiture, le spectre du bruit peut être vu comme une courbe monotone de dérivée très faible alors que celui de la parole bruitée accuse une forte dynamique. L'idée intuitive consiste à calculer, à chaque instant, le cumul des variations d'énergie entre bandes de fréquence adjacentes et élementaires. Les maxima de cette courbe correspondent aux noyaux vocaliques au sein de la parole et les minima à la présence de bruit seul.

Le détecteur se décompose alors en trois modules principaux (figure 1) :

- la localisation de frontières primaires liées à la détection de maxima,
- la segmentation du signal de parole,
- le module de décision.

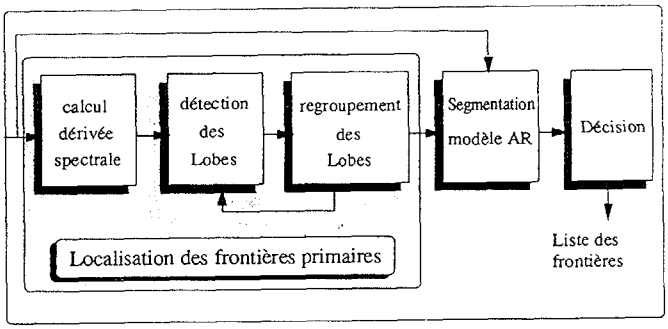

figure 1 : Description du détecteur basé sur la dérivée spectrale.

Les frontières primaires Le signal échantillonné à 8 kHz est analysé toutes les 16ms sur des fenêtres de 32ms (256 échantillons). Après pré-accentuation, et fenêtrage (Hamming), une transformée de Fourier est effectuée et l'énergie est calculée dans 24 filtres triangulaires répartis sur l'axe fréquentiel selon l'échelle Mel, soit $(En_i)_{1 \leq i \leq 24}$. La dérivée spectrale est donnée pour chaque trame de signal par :

$$Dspect_n = \sum_{i=1}^{23} \left(En_{i+1}^n - En_i^n\right)^2$$

Les maxima locaux significatifs de cette courbe (appelés lobes) sont recherchés à partir de franchissement de seuil : le début (resp. la fin) d'un lobe est détecté lorsque la dérivée spectrale franchit un seuil à condition que ce paramètre reste croissant (resp. décroissant) un temps minimum. Les lobes correspondent essentiellement aux noyaux vocaliques et à certaines zones consonantiques voisées.

Pour déceler le début et la fin de parole, une première étape du processus consiste à regrouper les lobes selon des critères de type durée des sons ou instabilité de la dérivée spectrale ; en effet, des zones de bruit peuvent être détectées à l'intérieur même d'une phrase :

- les plosives non voisées se décomposent en plusieurs parties et l'une de ces parties est un silence,
- des fricatives non voisées haute fréquence disparaissent à cause de la bande passante limitée à 3,3kHz.

A l'issue de ce regroupement, sont proposées des frontières primaires qui délimitent la portion de parole comprise entre les noyaux vocaliques extrêmes. Il faut affiner ces frontières pour tenir compte de sons non voisés situés éventuellement en début et fin de parole (remarques précédentes).

Relâchement des frontières à l'aide d'une segmentation statistique L'algorithme de segmentation automatique est activé de part et d'autre de la zone de parole détectée, dans un rayon de 200ms. Les performances de cet algorithme sont telles que même en milieu fortement bruité, les débuts et fins de parole sont détectés et les zones de bruit sont fort peu sursegmentées. Par conséquent on retiendra comme frontières définitives de la zone de parole, les ruptures détectées par l'algorithme de segmentation éloignées de moins de 200ms des frontières primaires (figure 2).

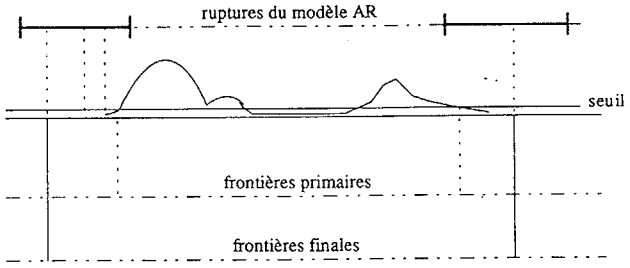

figure 2 : Relâchement des frontières primaires à l'aide de la segmentation automatique.

Signalons que les seuils utilisés dans cette méthode sont adaptés en cours de détection après confirmation de chaque fin de parole et tant qu'un nouveau début de parole n'est pas soupçonné ; ils sont fonction de la valeur moyenne de la dérivée spectrale dans le bruit.

4. COMPARAISON DE MODELES AUTOREGRESSIFS

Ce détecteur repose sur un test d'hypothèses, l'une correspond à la présence de bruit seul, l'autre à celle de parole bruitée. Nous supposons en effet que dans le cadre d'un habitacle de voiture "confortable", le bruit ambiant peut être considéré comme une succession de zones stationnaires, et que chaque zone peut être correctement identifiée à l'aide d'un modèle autorégressif ; les mêmes hypothèses sont faites sur la parole bruitée mais nous supposons que les ordres des modèles sont différents.

Nous pouvons décomposer ce deuxième détecteur en trois étapes :

- le signal acoustique est segmenté à l'aide de la méthode de "Divergence" ; les segments très longs (supérieurs à 500ms) sont d'ores et déjà classés comme des segments de bruit pur ;
- sur chaque segment, deux modèles autorégressifs gaussiens sont identifiés, un modèle d'ordre p, M_p, et un modèle d'ordre q M_q ;
- à l'issue de cette identification, est calculée la vraisemblance de chaque modèle $Vrais_p$ et $Vrais_q$ afin de décider quelle est l'hypoth'ese la plus vraisemblable :

  Idéalement, si le modèle d'ordre p correspond à l'hypothèse "bruit seul" et si $Vrais_q \leq Vrais_p$, le segment est étiqueté bruit, sinon le segment est étiqueté parole.

  En pratique, il est évident qu'un modèle autorégressif d'ordre élevé obtient toujours une vraisemblance plus élevée qu'un modèle d'ordre faible. C'est pourquoi nous avons fait appel au critère d'Akaike pour mettre en oeuvre cette méthode. Au cours de nos expériences, nous avons posé $q = 8$, ordre couramment utilisé pour modéliser la parole, le choix de p est effectué de manière automatique sur les premières trames de signal sachant qu'elles correspondent à du bruit.

5. EXPERIMENTATIONS

Comme nous l'avons dit précédemment, Les trois méthodes sont testées sur une des bases de données enregistrées au cours du projet Esprit ARS : ce corpus est formé de 43 mots isolés (noms propres), prononcés 12 fois par chacun des quatre locuteurs dans une voiture. Quatre répétitions sont prises à 130 km/h, quatre à 90km/h, quatre à 0km/h. Le rapport signal/bruit pour les vitesses de 90km/h et 130 km/h est compris entre 0 et 30 db, la valeur moyenne étant de 15 db. Une segmentation manuelle est fournie avec les données afin de pouvoir réaliser l'apprentissage de modèles de références pour un système de reconnaissance ; dans la mesure où les performances du système seraient fortement dégradées si des mots étaient tronqués, ces frontières manuelles sont fortement relachées. Toutes les frontières de mots trouvées automatiquenment par l'une des trois méthodes proposées, sont à l'intérieur de celles fournies manuellement. Reste à savoir si les mots ne sont pas tronqués !

A la suite d'un examen minutieux des résultats, nous avons constaté que seuls les mots commençant ou finissant par une fricative sourde sont quasiment systématiquement erronés (figure 3).

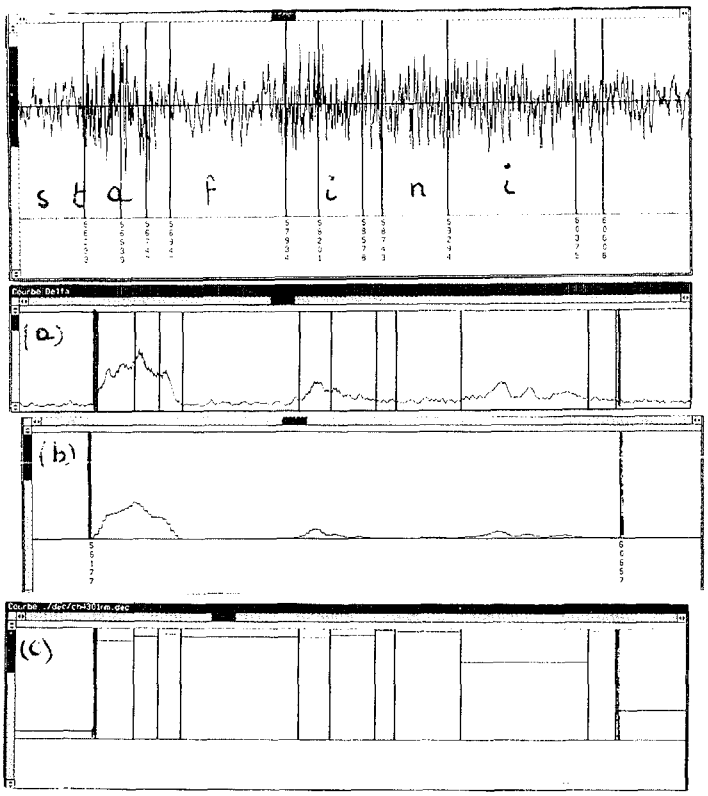

figure 3 : Comparaison des différentes méthodes. Le mot prononcé est "Stafini". La courbe (a) représente la fonction S, la courbe (b) la dérivée spectrale et la courbe (c) le rapport de vraisemblance. Sur le signal sont tracés les segments issus de la méthode de Divergence, sur les courbes (a), (b) et (c), les frontières détectées par chacune des méthodes sont les frontières extrêmes. La fricative /s/ est insoupçonnée.

Une des principales causes de ce type d'omission est liée à un filtrage passe-bas sévère du signal, effectué en amont du pré-traitement ($\leq 3,3kHz$) ; or les fricatives sourdes que sont /s/, / /, sont des sons haute fréquence ($\geq 3,5kHz$) ; ces sons sont donc inexistants sur le signal acoustique numérisé, aucune méthode ne pourra les détecter !

Signalons qu'afin de rendre la méthode temporelle et la méthode paramétrique plus robuste, il suffirait d'appliquer le même type de relâchement de frontières que celui appliqué à la méthode non paramétrique.

6. CONCLUSION

Encouragés par ces résultats, nous poursuivons cette étude de la manière suivante :

- nous effectuons une segmentation manuelle plus précise de chaque enregistrement afin d'évaluer quantitativement chacune des méthodes. Nous préciserons quels sons sont acoustiquement présents et effectivement tronqués et nous chiffrerons les taux d'erreurs.
- après optimisation et/ou couplage des méthodes, sera conçu un système de RAP utilisant au niveau du pré-traitement :
  * la segmenttion automatique du signal,
  * un des détecteurs bruit/parole décrits précédemment,
  * un débruitage de type NSS [Lockwood 92].

  Ce système sera comparé aux approches classiques qui n'utilisent qu'une détection bruit/parole grossière : seule cette comparaison pourra apporter une conclusion définitive sur l'intérêt de tels détecteurs en RAP.

References

[André-Obrecht 88] R. André-Obrecht, "A New Statistical Approach for Automatic Segmentation of Continuous Speech Signals", IEEE Trans. on ASSP, vol. 36 pp 26-40, January 1988.

[Juang 91] B.H. Juang, "Speech recognition in adverse environnement", Computer Speech and Language, Vol 5, pp 275-294, 1991.

[Junqua 92] B. Mak, J.C. Junqua, B. Reaves, "A robust speech/non-speech detection algorithm using time and frequency-based features", ICASSP 1992.

[Lamel 81] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg, J.G. Wilpon "An Improved Endpoint Detector for Isolated Word Recognition", IEEE Trans on ASSP, Vol. 29, pp 777-785, August 1981.

[Lockwood 92] P. Lockwood, J. Boudy, M. Blanchet, "Non-Linear Spectral Subtraction (NSS) and Hidden Markov Models for Robust Speech Recognition in Car Noise Environnements", ICASSP 1992.

[Lombard 11] . Lombard, "Le signe de l'élévation de la voix", Ann. Maladies oreille, larynx, nez, pharynx, 37, pp 101-119, 1911.