TREIZIÈME COLLOQUE GRETSI - JUAN-LES-PINS DU 16 AU 20 SEPTEMBRE 1991

Techniques des moindres carrés et segmentation de corpus de parole multilingue

Guy Pérennou, Harouna Kabré et Nadine Vigouroux

Institut de Recherche en Informatique de Toulouse
URA CNRS 1399, Toulouse, France

Résumé. Nous présentons une méthode d'analyse temporelle du signal vocal basée sur une combinaison des techniques des moindres carrés et des filtres morphologiques. Cette méthode a été appliquée au prétraitement du taux de passage à zéro et à l'amplitude du signal pour la segmentation de corpus multilingues.
Les résultats obtenus sur les corpus anglais, français et italien de EUROM-0 —la base de données européenne des sons [4]— montrent l'efficacité et la robustesse des paramètres prétraités par notre méthode par rapport aux mêmes paramètres non prétraités.

Abstract. This is presenting a method for Temporal Analysis of speech signal, based both on Least Square Technics and morphological filtering. This method has been applied to Zero Crossing Rate and to signal amplitude for a multi-lingual segmentation of speech signal.
The results secured over English, French and Italian corpora of EUROM-0 —the European Speech Database [4]— demonstrate the efficiency and the robustness of the filtered parameters when comparing them to the none filtered ones.

1. INTRODUCTION

Nous nous intéressons à l'étiquetage automatique du signal vocal. Dans une première approche nous avons fait appel aux paramètres spectraux [1]. Nous présentons ici une nouvelle approche utilisant les paramètres temporels.
La plupart des méthodes temporelles d'analyse du signal vocal sont fondées sur la recherche d'un ensemble restreint de propriétés de l'onde vocale caractérisant au mieux un segment du signal. Pour des applications comme le décodage acoustico-phonétique ou l'étiquetage automatique, il s'agit de trouver une bonne adéquation entre les ruptures essentielles du signal et les frontières effectives des événements phonétiques.
Dans une telle perspective, les techniques des moindres carrés (MC) constituent un outil intéressant, mais la non stationnarité du signal vocal rend leur application très délicate. De ce fait, il est difficile de modéliser à la fois correctement les parties stables et les parties très instables du signal sauf au prix d'une complexité algorithmique accrue.
La méthode temporelle que nous présentons est basée sur une combinaison des techniques des moindres carrés et de filtres morphologiques. Elle permet de tenir compte de la non-stationnarité de l'onde vocale. Cette propriété est à nos yeux fondamentale car elle nous permet d'obtenir une meilleure cohérence avec la segmentation donnée par les étiqueteurs manuels.
Après avoir rappelé quelques transformations morphologiques de base, nous décrivons les principaux éléments de notre méthode. La dernière partie est consacrée à l'évaluation de notre méthode appliquée aux paramètres amplitude et taux de passage à zéro (TPZ) utilisés dans la segmentation de corpus de parole continue multilingue.

1.1 Transformations morphologiques

Dans le cas de la parole, deux transformations de base dites d'érosion et de dilatation sont utilisées pour la construction des filtres morphologiques [2].
Soit un signal échantillonné $S = \{S_k / k=1,2,\dots\}$. Nous introduisons une fenêtre glissante de largeur F=L+M+1 qui à l'instant k est l'intervalle [k-L,k+M]. Pour les opérateurs liés à de telles fenêtres on suppose que le signal est prolongé à gauche par des zéros.
L'opérateur E(./L,M,h) (ou, plus brièvement, E) se définit alors comme une transformation de S en un autre signal

$$E(S/L,M,h) = Y = \{ y_k/k=1,2,\dots\},$$

sous-échantillonné dans un rapport de 1/h, où :

$$y_k = \min(S_{kh-L},\dots,S_{kh+M}).$$

L'opérateur E est l'analogue de l'opérateur d'érosion de la morphologie mathématique.
On définit de même l'opérateur D (analogue de la dilatation en morphologie mathématique) par :

$$D(S/L,M,h)=Z=\{z_k/k=1,2,\dots\},$$

où :

$$z_k=\max(S_{kh-L},\dots,S_{kh+M}).$$

On notera la propriété élémentaire

$$D(S/L,M,h) = -E(-S/L,M,h).$$

et le fait que D(./L,M,1) anticipe de L les plages de croissance du signal et produit un retard de M sur les

plages de décroissance. En ce sens cet opérateur dilate les impulsions positives (rectangulaires, triangulaires...). L'opérateur E(./M,L,1) (dual du précédent) a un effet inverse qui sera exploité plus loin. Lorsque h≠1 les résultats précédents sont combinés avec le sous échantillonnage de rapport 1/h. Si l'on compare le signal initial, également sous-échantillonné dans ce rapport, les phénomènes d'anticipation et de retard s'observent encore mais ils sont divisés par h.
Au paragraphe suivant on écrit plus brièvement D_2 ou E_2 quand h>1 et E_1 et D_1 quand h=1.

1.2 Eléments de notre méthode

Notre méthode temporelle est une combinaison des transformations précédemment définies et des techniques de moindres carrés.
Cependant l'application des moindres carrés aux signaux non stationnaires se heurte à deux types de problèmes : d'une part, il faut choisir une fenêtre d'une taille convenable pour l'approximation, et d'autre part, il faut choisir un modèle qui permet de mettre en évidence les événements essentiels qui apparaissent dans le signal vocal. Dans le cas de la parole il faut pouvoir aussi bien modéliser des passages perturbés tels que les réalisations de fricatives et d'occlusives que les zones stables des noyaux vocaliques.
Le choix d'un modèle pour l'approximation —simple droite ou polynôme de degré supérieur— détermine la complexité de ce dernier : un modèle complexe conduit à un nombre important de paramètres à estimer pour chaque segment de signal traité. De plus il est souhaitable d'utiliser des fenêtres et/ou des modèles dynamiques pour prendre en compte la non stationnarité du signal de parole. L'inconvénient c'est que de tels choix entrainent un volume de calcul important.
Pour parvenir à une solution efficiente —aussi bien en temps de calcul qu'en qualité du point de vue des contrastes préservés dans la courbe obtenue par les MC— nous avons choisi une simple droite comme modèle pour l'approximation et ceci sur des fenêtres de l'ordre de 32ms avec recouvrement toutes les 4ms. Ce traitement est suivi d'une optimisation dont le but est de rechercher parmi tous les modèles fournis par le module des MC, celui qui approxime le mieux un segment du signal donné.
Ceci nous a conduit à la décomposition de notre méthode sous forme d'opérateurs schématisée à la Fig.1. Elle s'apparente à celle préconisée par Oppenhein [5] pour les systèmes homomorphiques dans lequel A et B sont interprétés comme des opérateurs morphologiques.

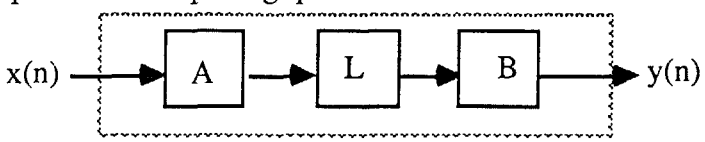

Fig.1 - Principe de la méthode, A et B sont des opérateurs morphologiques, et L un opérateur linéaire.

Plus précisément $A=E_1(D_2-E_2)$ suivant les notations du paragraphe 1. Elle consiste à prendre d'abord la différence entre l'enveloppe positive (founie par D_2) et l'enveloppe négative (fournie par E_2) du signal d'entrée avec un facteur de sous échantillonnage de h=64 —voir figure 1 pour l'effet de la transformation d'un segment de signal par l'opérateur (D_1-E_1)— puis à appliquer l'opérateur E_2, dont le rôle est de corriger la dilatation introduite par la tranformation (D_1-E_1) [3].
Les autres opérateurs de la figure 1 sont :
- B, un opérateur qui renvoie le modèle ayant la plus faible variance [3],
- et L, un opérateur linéaire réalisant l'approximation par les MC sur des fenêtres glissantes.
On notera qu'iI est possible d'envisager l'utilisation d'opérateurs autres que ceux qui ont été définis au paragraphe 1.

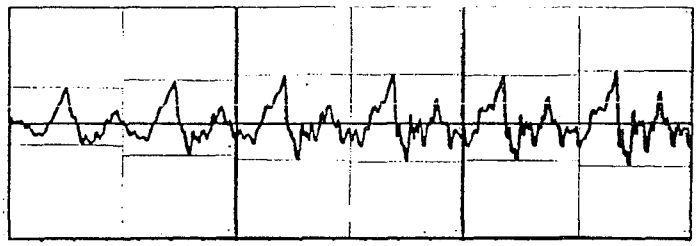

Fig.2 - Effet des transformations D_1 et E_1 sur une portion de signal vocal. Toutes les 4ms E1 retourne la valeur minimale de la trame courante du signal tandis que D1 donne la valeur maximale.

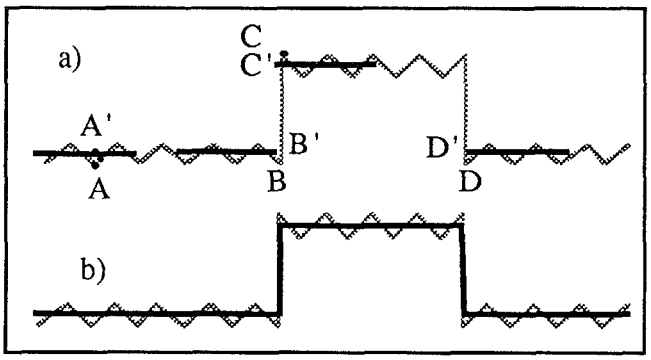

Fig.3 - Toutes les 32ms on recherche pour tous les points contenus dans une trame donnée du signal, quel est le meilleur modèle au sens de la variance et on le remplace par sa représentation dans ce modèle ; ainsi A devient A'. On voit que le point B aura pour meilleur modèle celui qui correspond au segment le plus à gauche le contenant et sera approximé par B'— commentaires analogues pour C et D [6].

2. IMPLEMENTATION

La méthode présentée ci-dessus est implémentée suivant une structure hiérarchisée à deux niveaux :
- une micro-modélisation non linéaire de l'onde considérée correspondant à l'implémentation des opérateurs A et L. Cette étape génère un ensemble de micro-modèles —correspondant aux modèles fournis par l'approximation au sens des moindres carrés— candidats pour modéliser un segment de signal,
- une optimisation sur une fenêtre glissante correspondant à l'implémentation de l'opérateur B. La

figure 3 illustre le choix d'un modèle optimal dans un segment de signal donné.
L'interêt de cette hiérarchisation est qu'elle permet d'envisager le calcul des paramètres des micro-modèles en parallèle. Il suffit d'utiliser un tampon du type registre à décalage en entrée de chacun des deux modules pour mémoriser les échantillons —ceci permet d'obtenir le recouvrement des différentes fenêtres utilisées— et un circuit du type ligne à retard pour synchroniser les différents paramètres en sortie de chaque module. Les opérations élémentaires à effectuer se réduisent alors à des décalages et à des additions. Les tests qui sont faits lors de la recherche du micro-modèle optimal peuvent être simplifiés et accélérés en mémorisant les résultats partiels des tests précédents. La taille des données mémorisables relève bien sûr d'un compromis entre la vitesse d'exécution souhaitée et la taille mémoire disponible.
Ces deux étapes sont précédées d'un prétraitement dont le rôle est de nous affranchir des conditions particulières d'enregistrement : correction d'un biais éventuel du signal introduit par le matériel d'acquisition (par exemple si le microphone introduit une composante continue, elle est retranchée), réglage du rapport signal sur bruit, préamphase.
Nous appelons paramètres —amplitude et/ou TPZ— non corrigés les sorties de l'opérateur A. Les paramètres qui résultent de l'ensemble des transformations de la figure 1 sont dits paramètres corrigés.
Pour la détection des discontinuités majeures en vue de la segmentation automatique, on calcule aussi les dérivées premières et secondes des paramètres corrigés et non corrigés. Tous ces paramètres sont normalisés entre 0 et 1 [3].

3. RESULTATS

Nous avons appliqué l'ensemble de ces traitements à l'amplitude et au taux de passage à zéro du signal vocal afin de mesurer l'apport de notre méthode en termes de précision des frontières automatiques par rapport à celles d'un étiquetage manuel.
Le principe de segmentation mis en œuvre est une segmentation à seuil. Dès que la dérivée seconde d'un des deux paramètres (amplitude ou TPZ) dépasse un seuil une frontière de segmentation est positionnée.
Pour évaluer cette segmentation un coefficient de qualité de segmentation a été introduit comme suit :

$$CQ(\delta,PAR,\theta) = n(\delta,PAR,\theta) / N$$

où N est le nombre de frontières manuelles et $n(\delta,PAR,\theta)$ le nombre de frontières manuelles approximée par une frontière manuelle à δ ms près, en utilisant le paramètre PAR et le seuil θ.
Suivant ce coefficient, deux segmentations ne sont comparables que si elles ont le même taux de sursegmentation. Ceci nous a conduit à introduire le paramètre TSS (Taux de SurSegmentation) défini de la manière suivante :

$$TSS(PAR,\theta) = n\ total(PAR,\theta) / N,$$

où n total (PAR,θ) est le nombre total de frontières obtenues par une segmentation automatique basée sur le paramètre PAR au seuil θ.

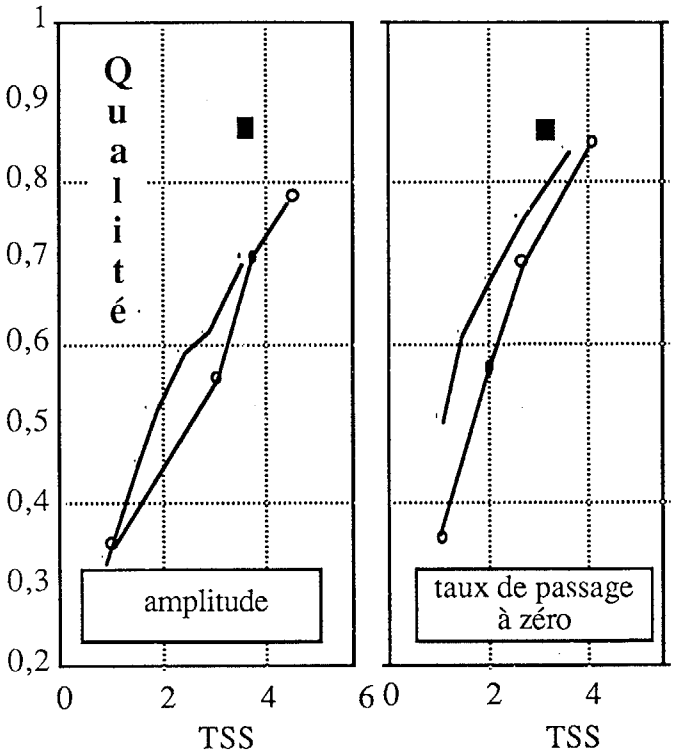

Fig.4 - Coefficient de qualité de la segmentation avec l'amplitude et le taux de passage à zéro.

3.1 Comparaison des paramètres corrigés et non corrigés

La figure 4 donne un exemple de résultats obtenus par deux segmentations :
- la première correspond à une segmentation sur l'amplitude corrigée et non corrigée,
- et la seconde à une segmentation sur le TPZ corrigé et non corrigé.
Les courbes obtenues montrent une amélioration de l'ordre de 10% avec les paramètres corrigés par rapport aux paramètres non corrigés pour un taux de sursegmentation voisin de 2. Pour l'interprétation des résultats, on précise que ces courbes ont été obtenues sur deux locuteurs fançais de EUROM-0 —la base européenne des sons [4]— sur un total de 200 phonèmes.
Ces résultats de segmentation sont encore améliorés dans nos applications d'étiquetage lorsque nous utilisons deux critères de segmentations à la fois : la première basée sur le seuillage de la dérivée seconde et la seconde sur les changements significatifs entre macro-classes [3]. Des résultats obtenus suivant ce critère sont donnés à la figure 6 et discutés dans la suite.

3.2 Comparaison de l'amplitude corrigée avec l'énergie du signal

La figure 5 montre en 5c, 5d deux paramètres caractérisant l'amplitude. La comparaison de 5c et 5d

montre que notre prétraitement préserve bien les événements essentiels, autrement dit les discontinuités et les contrastes restent présents sur l'amplitude corrigéee en 5c. En revanche les variations aléatoires sont atténuées. Ceci confirme l'efficacité de notre méthode. On notera que l'énergie moyennée (sur 8 ms avec recouvrement toutes les 4 ms) du signal vocal en 5d a fait disparaître les frontières dans la séquence phonétique [nym] du mot 'numéro'.

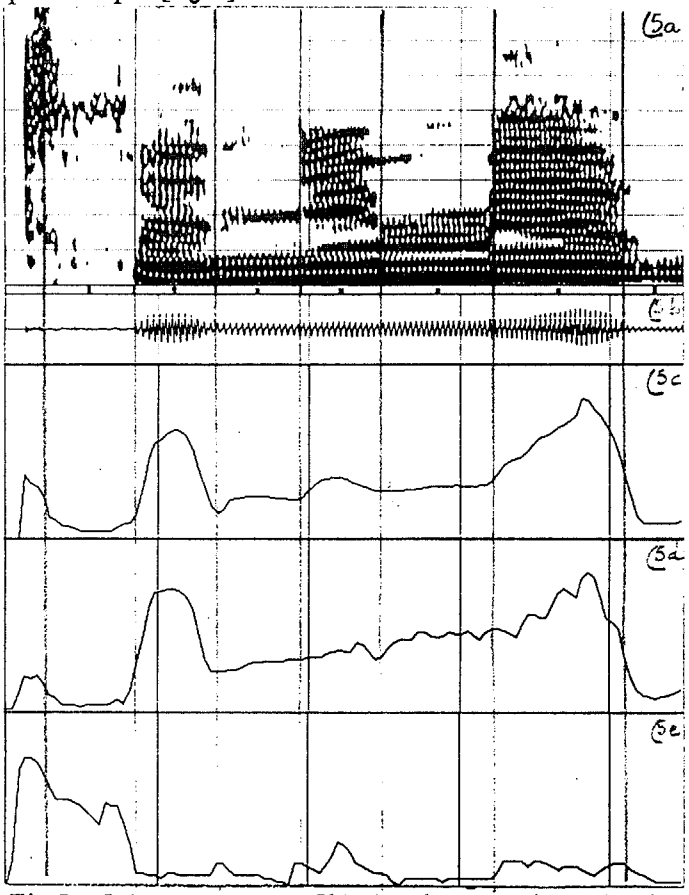

Fig.5 - 5a) spectrograme 5b) signal temporel 5c) Amplitude corrigée 5d) Energie moyennée 5e) TPZ corrigé.

Langage	Locuteurs	TSS	CQ	nb. ph.
Anglais	4	1.85	0.91	4476
Français	2	1.78	0.92	2909
Italien	4	1.68	0.94	5261

Fig.6 - Qualité de la segmentation sur trois langues (anglais, français, italien) à $\delta = \pm 20$ *ms. La dernière colonne donne le nombre de phonèmes des corpus.*

La figure 6 montre les résultats de l'évaluation de la segmentation mise en œuvre pour trois langues (anglais, français et italien). On constate qu'en moyenne pour les trois langues, un coefficient de qualité voisin de 92% est obtenu pour un taux de sursegmentation de 1.8. La stabilité des résultats s'explique par la faible sensibilité aux seuils de segmentation des paramètres corrigés.

Cette méthode constitue actuellement le module d'analyse acoustique de notre système d'étiquetage automatique en cours d'évaluation dans le cadre du projet SAM [6]. Cependant l'implémentation sur un compatible PC ne nous a pas encore permis d'exploiter le parallélisme des opérations suggéré au paragraphe 2. Une solution envisagée pour parvenir à ce résultat est de transférer les traitements acoustiques sur un processeur spécialisé de traitement du signal.

4. CONCLUSION

Nous avons présenté une méthode temporelle qui combine à la fois des transformations morphologiques et des techniques des moindres carrés.
Un des objectifs de notre système est l'analyse de la structure syntagmatique du signal vocal en terme de succession d'événements. Cela suppose que l'on puisse détecter les fontières essentielles (et non celles qui résultent de fausses alarmes déclenchées par le bruit et/ou certaines évolutions chaotiques du signal).
C'est l'introduction d'opérateurs morphoplogiques, d'approximation au sens des moindres carrés et d'une fonction d'optimisation située en aval qui a permis de résoudre cette difficulté.
Au plan phonétique ce traitement du signal, nous a permis de montrer qu'une segmentation automatique en événements, sans apprentissage, indépendante du corpus, du locuteur et des langues était possible.
Nos travaux futurs iront, d'une part, vers une expérimentation de cette méthode sur plus de corpus et sur plus de langues, et, d'autre part, vers la reconnaissance de la parole par HMM (Hidden Markov Model) où sont exploitées les informations phonétiques issues de l'analyse syntagmatique du signal vocal.

5. REFERENCES

1] C. Dours M. De Calmès, H. Kabré, J. M. Pécatte G. Pérennou, N. Vigouroux, "A Multi-Level Automatic Segmentation : SAPHO and VERIPHONE", Proc. of EUROSPEECH 89, Paris, France, Sept. 26-28 1989, pp.83-86, Vol 2.
[2] A. Ben Slimane et B. Zouabi, "Première Approche de Segmentation par Filtrage Morphologique", 16è JEP, Hammamet, 5-9 Oct. 1987.
[3] H. Kabré, G. Pérennou et N. Vigouroux, "A Non Linear Filtering Method Applied to Automatic Segmentation of Multilingual Speech Corpora", EUROSPEECH, Genova, Italy, 1991.
[4] SAM—Multi-lingual Speech Input/Output: Assessment, Methodology and Standardisation, Extension Phase, Final Report, 1 April 1988-28.
[5] A.V Oppenheim, R.W Schaffer, T.G Stockham, "Non-Linear Filtering of Multiplied and Convolued Signals", Proc.IEEE, Vol.56, No 8, Ang.1968, pp. 1264-1291.February 1989.
[6] H. Kabré, G. Pérennou et N. Vigouroux, "Automatic Labelling of Speech Signal into Events", ICPhS, Aix en Provence 1991.