Application de l'analyse multi-résolution à la segmentation de corpus de parole dédiés à la synthèse vocale

Safaa Jarifi, Dominique Pastor, Olivier Rosec

Application de l'analyse multi-résolution à la segmentation de corpus de parole dédiés à la synthèse vocale

Safaa Jarifi[1], Dominique Pastor[1], et Olivier Rosec[2]

[1] École Nat. Sup. des Télécommunications de Bretagne,
Département Signal et Communication,
Technopôle Brest-Iroise, CS 83818, 29285 Brest Cedex, France.
Tél : int+ 33 2 29 00 14 87, Fax : int+ 33 2 29 00 10 12
{safaa.jarifi,dominique.pastor}@enst-bretagne.fr

[2] France Telecom, Division R&D TECH/SSTP/VMI,
2, avenue Pierre Marzin, 22307 Lannion Cedex, France
Tél : int+33 2 96 05 20 67, Fax : int+33 2 96 05 35 30
olivier.rosec@rd.francetelecom.com

Résumé Cette étude s'insère dans la conception et l'analyse d'algorithmes de segmentation phonétique et automatique de grands corpus de parole dédiés à la synthèse vocale . Dans cet article, nous introduisons un algorithme de segmentation acoustico-phonétique.

Cet algorithme est une extension simple de l'Analyse Multi-Résolution (AMR), car il consiste en une AMR de l'enveloppe complexe (AMREC) du signal associée à une fréquence arbitrairement choisie. Grâce à cette analyse du signal de parole autour de cette fréquence, on peut alors différencier les phénomènes basses fréquences (*BF*) des phénomènes hautes fréquences (*HF*). En choisissant judicieusement une fréquence *BF* et une fréquence *HF* pour l'analyse, la présence ou non d'énergie autour de chacune de ces fréquences permet de segmenter le signal de parole en certaines classes acoustico-phonétiques.

A partir de quelques exemples d'applications à la parole, nous proposons une approche utilisant l'AMREC pour application à la synthèse vocale. Cette approche a pour objectif d'affiner les marques de segmentation produites par l'algorithme décrit dans [4].

Mots clés Algorithme de Mallat, Analyse Multi-Résolution, Détection énergetique, Segmentation acoustique, Synthèse vocale, HMM, GMM.

1 Introduction

Nous nous intéressons à la segmentation automatique de grands corpus de parole continue pour application à la synthèse vocale pr concaténation. Dans l'approche classique, cette segmentation est automatisée par HMM : après apprentissage de ces modèles sur l'ensemble du corpus de parole, la segmentation consiste à apposer les frontières de phones à l'aide d'un alignement forcé à partir d'une phonétisation supposée exacte. Cette technique offre des résultats acceptables dans la mesure où 85% des marques de segmentation du HMM se situent à moins de 20 ms de celles de la segmentation manuelle. Cependant, la précision de la segmentation reste insuffisante car elle entraîne des imperfections en sortie du système de synthèse. Des étapes de vérification, fastidieuses et coûteuses, demeurent donc indispensables. Dans le but de les réduire, nous cherchons à mettre en oeuvre des algorithmes de segmentation automatique dont la précision approche celle de la segmentation manuelle.

Avec l'algorithme de segmentation phonétique "HMM amélioré" décrit dans [4], on obtient un gain de 3%, toujours pour une tolérance de 20 ms. Cet algorithme consiste à affiner les marques obtenues par HMM via une modélisation des frontières par des modèles GMM (Gaussian Mixture Model), ces derniers étant appris sur un petit corpus segmenté manuellement. Pour notre corpus comportant environ 140000 frontières, 88% de marques correctes obtenu par HMM amélioré signifie localiser, vérifier et corriger quelques 16800 marques de segmentation, ce qui reste encore trop lourd. Nous pensons qu'au moyen de certains algorithmes de segmentation acoustico-phonétique, les marques du HMM amélioré peuvent être affinées significativement.

De nombreux algorithmes de segmentation acoustico-phonétique sont proposés dans la littérature. Nous étudions certains de ces algorithmes et analysons comment les combiner avec le HMM amélioré. A ce titre, le lecteur trouvera dans [1] une analyse de l'algorithme de Brandt et du test de divergence, qui sont des algorithmes de détection de rupture de stationnarité ([5]).

Nous consacrons le présent papier à l'analyse multi-résolution de l'enveloppe complexe (AMREC) et son application à la segmentation automatique des grands corpus. L'AMREC est une amélioration de l'algorithme "Mallat-Modulation" proposé dans [3]. Dans la section 2, nous commençons par décrire cet algorithme. La section 3 est consacrée à l'application de l'AMREC à l'analyse du signal de parole. Nous constatons que cet algorithme est adapté à la segmentation de certaines classes acoustico-phonétiques. Aussi, dans la section 4, nous décrivons comment cet algorithme peut être utilisé pour affiner la segmentation produite par le "HMM amélioré".

2 Description de l'AMREC

L'AMR usuelle et la décomposition de Mallat n'analysent que les basses fréquences du signal. D'autre part, les décompositions en paquets d'ondelettes et en ondelettes de Malvar ne sont pas adaptées à une analyse locale autour d'une fréquence arbitrairement choisie par l'utilisateur. Il est pourtant simple d'étendre l'AMR de manière à analyser un signal et de mettre en évidence la présence ou non d'énergie autour d'une fréquence f_0 arbitraire. Il suffit en effet d'appliquer une AMR sur l'enveloppe complexe du signal associée à f_0. Par construction, les coefficients en sortie de cette AMR sont des valeurs complexes. Soit h_0 le filtre passe-bas associé à la fonction d'échelle de l'AMR. Un premier filtrage de l'enveloppe

complexe par h_0 récupère la tendance globale du signal. Après sous-échantillonnage par 2, un deuxième filtrage par h_0 est effectué sur cette tendance et ainsi de suite jusqu'à ce que la résolution choisie soit atteinte. Les paramètres de cette méthode sont donc la fréquence f_0, la résolution (ou niveau de décomposition) p, et la régularité r du filtre passe-bas. Le fonctionnement de l'AMREC est résumé sur le schéma 1.

La fréquence f_0 dépend des phénomènes qu'on veut détecter. Bien entendu, on choisit f_0 inférieure ou égale à la moitié de la fréquence d'échantillonnage.

La régularité ne doit être ni trop faible, ni trop forte. En effet, si r est faible, le filtre h_0 sera peu sélectif en fréquence ; si r est grand, le filtrage risque d'amalgamer des phénomènes physiques qu'on souhaiterait justement séparer. Par exemple, pour les filtres de Daubechies, la longueur L de la réponse impulsionnelle est égale à $2r + 1$.

Plus la résolution est élevée, meilleure est la localisation fréquentielle. Bien évidemment, une bonne localisation fréquentielle entraîne une mauvaise localisation temporelle selon le principe classique d'*Heisenberg*. Le choix de f_0, p, et r résulte donc d'un compromis lié à l'application.

L'AMREC est aussi une extension de l'algorithme "Mallat-Modulation" décrit dans [3]. En effet, Mallat-Modulation analyse seulement la partie réelle de l'enveloppe complexe et ne prend pas en compte la partie imaginaire. Pour conserver toute l'information sur le signal, nous préférons travailler sur l'enveloppe complexe. Cependant comme pour "Mallat-Modulation", cette analyse est de complexité linéaire et facile à mettre en oeuvre. De plus, les propriétés de l'analyse "Mallat-Modulation" pour une large classe de bruits corrélés (voir [6]) restent aussi vraies pour l'AMREC. Autrement dit, les coefficients à la sortie de l'AMREC d'un bruit coloré tendent à être décorrélés lorsque la résolution augmente.

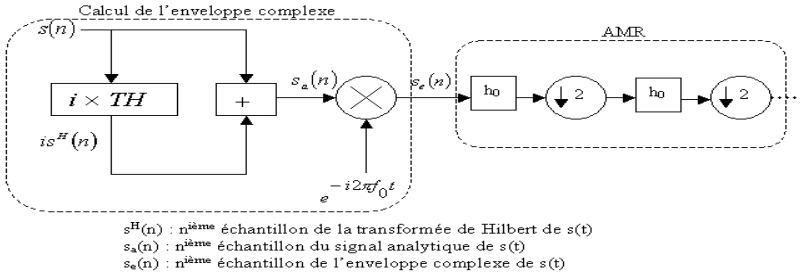

Fig. 1. Fonctionnement de l'AMREC

3 Quelques exemples d'applications de l'AMREC à la parole

Comme nous l'avons mentionné dans la section précédente, l'AMREC est une analyse locale de la présence d'énergie du signal autour d'une fréquence choisie par l'utilisateur. Comme dans [3], l'AMREC peut être utilisée pour analyser le signal de parole dans trois canaux : un canal de basses fréquences, un autre pour les fréquences intermédiaires et un dernier pour les hautes fréquences. A partir des coefficients en sortie de l'AMREC dans ces trois canaux, on peut identifier certains traits acoustiques, segmenter le signal de parole, ou combiner identification et segmentation de manière à reconnaître certains phonèmes.

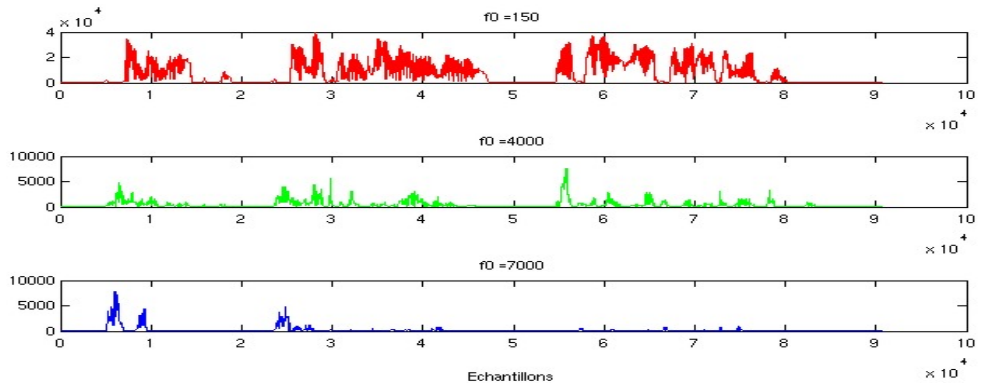

Fig. 2. Modules des coefficients de l'AMREC en BF,HF, et MF

Pour l'identification des traits acoustiques, la répartition énergétique d'une trame donnée dans ces trois canaux permet de déterminer la classe acoustico-phonétique à laquelle cette trame appartient parmi celles du tableau 1. Dans ce tableau, chaque classe est caractérisée par deux indices forts et un indice faible. Seuls les indices forts permettent de prendre une décision quant au phénomène analysé. Les deux indices forts sont l'indice de préférence et l'indice d'exclusion. Par exemple, pour décider que la trame appartient à la classe $\{[s], [\int]\}$, il faut détecter de l'énergie en *HF* (indice de préférence) et ne pas détecter d'énergie en *BF* (indice d'exclusion).

Pour effectuer la segmentation de la parole, les modules des coefficients de l'AMREC sont seuillés dans chacun des canaux. Grâce à ce seuillage, on réalise des segmentations temporelles *HF*, *BF* et *MF*. Ces segmentations sont ensuite fusionnées afin d'obtenir la segmentation acoustico-phonétique du signal de parole.

Par exemple, la figure 2 représente les modules des coefficients à la sortie de l'AMREC d'une phrase de parole échantillonnée à 16 kHz et analysée autour des trois fréquences : $150, 4000$ et 7000 Hz. Ce choix résulte de l'analyse de quelques spectrogrammes de signaux de parole issus de la base de données que nous traitons. Il n'est pas strict et c'est l'ordre de grandeur de chacune de ces fréquences qui importe : les fréquences de l'ordre de 150 Hz sont dédiées à la détection des phénomènes voisés ; la présence d'énergie autour de 7000 Hz révèle les phénomènes fricatifs non voisés ; le canal dont la fréquence centrale est 4000 Hz permet de compléter la segmentation acoustico-phonétique comme nous allons le voir dans cet exemple. La résolution est fixée à 4 et la régularité du filtre à 6 (choix classique dans la parole). On annule les coefficients dont le module est inférieur à un seuil de manière à mettre en évidence les phénomènes significativement énergétiques.

A ce stade là, la difficulté est le choix des seuils. Si le corpus a été enregistré dans des conditions optimales, on peut se permettre de choisir les seuils de manière empirique sur la base de quelques phrases ; on peut aussi envisager un apprentissage sur une petite partie du corpus segmenté manuellement en classes acoustico-phonétiques présentées dans le tableau 1. Si le corpus est bruité, on peut utiliser les méthodes dédiées proposées dans [3].

Notre exemple (figures 2, 3, 4) est issu d'un corpus contenant 7000 phrases et enregistré dans des conditions optimales. Le seuillage empirique est suffisant. Les segmentations de la figure 3 en découlent. Elles représentent simultanément la segmentation en phénomènes *BF*, *MF*, et *HF*. Pour fusionner ces segmentations, on a procédé de la manière suivante. D'abord

on récupère toutes les marques des trois segmentations. Ensuite, pour chaque segment ne dépassant pas 10 ms, si les deux marques en jeu correspondent à des marques *HF* et *BF*, on garde celle obtenue par la segmentation *BF*. Si ces marques sont *HF* et *MF*, on conserve seulement la marque *HF*. Dans le cas d'une marque *BF* et d'une marque *MF*, on garde celle de la segmentation *BF*. En d'autres termes, on considère que les marques *BF* sont plus précises que les marques *HF* et que les marques *HF* sont aussi plus précises que les marques *MF*. Notre démarche se justifie par le fait que les phénomènes *BF* sont plus énergétiques que les phénomènes *HF* qui, eux-mêmes, sont plus énergétiques que les phénomènes *MF*. Le résultat de la fusion est représenté sur la figure 4.

Plusieurs tests de ce type montrent que l'AMREC commet des insertions et des omissions par rapport à la segmentation manuelle acoustico-phonétique. Les marques qui ne sont pas des insertions, sont précises. Étant donné que notre application est dédiée à la synthèse vocale, notre but est d'obtenir une segmentation phonétique sans insertions ni omissions. Pour cette raison, la section suivante décrit comment nous envisageons d'utiliser l'AMREC pour corriger certaines marques de segmentation produites par un algorithme de segmentation phonétique (qui, par construction, ne commet ni insertion ni omission).

groupe	*indice de préférence*	*indice d'exclusion*	*indice faible*
Voyelles	Détection BF ou MF	Détection HF	
$[s]$ $[\int]$	Détection HF	Détection BF	Détection MF
Plosives non voisées			Détection HF, BF ou MF
$[f]$	Détection HF		Détection BF ou MF ou HF
Fricatives voisées	Détection BF		Détection HF ou MF
Plosives voisées	Détection BF	Détection HF	Détection MF
Consonnes sonnantes		Détection HF	Détection BF ou MF

Tab. 1. Indices forts et indices faibles caractérisant les classes acoustico-phonétiques

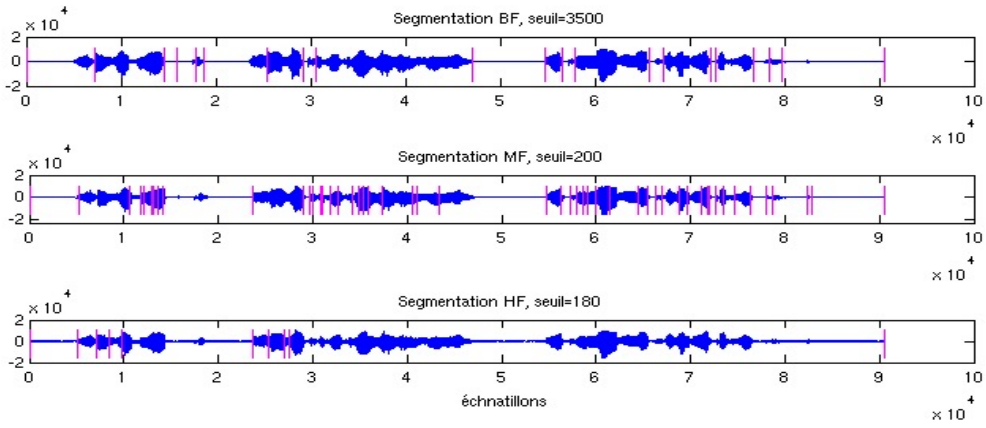

Fig. 3. Segmentations BF, HF et MF d'une phrase avec l'AMREC

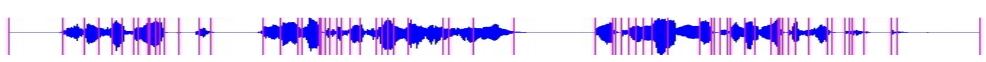

Fig. 4. Segmentation en classes acoustico-phonétiques

4 Perspective d'application de l'AMREC à la synthèse vocale

La méthode consiste à affiner certaines marques de l'algorithme dit "HMM amélioré" [4]. On commence par calculer la vraisemblance généralisée ou GLR (Generalized Likelihood Ratio) pour chaque marque de segmentation issue du "HMM amélioré". Ce GLR mesure la vraisemblance que cette marque soit une rupture de stationnarité du signal au sens de l'algorithme de Brandt décrit dans [5].

Si le GLR dépasse un certain seuil T (à définir), alors on conserve la marque de segmentation du HMM amélioré. Sinon, on utilise la phonétisation dont on dispose pour affiner la marque de segmentation comme suit :

- Si l'un des phonèmes est BF et l'autre est HF ou à la fois HF et BF, alors on effectue une AMREC en HF. Si l'AMREC trouve une seule marque, on remplace la marque initiale par cette marque. Sinon, on conserve la marque du "HMM amélioré".
- Si l'un des phonèmes est HF et l'autre est à la fois HF et BF, alors on effectue une AMREC en BF. Si l'AMREC trouve une seule marque, on remplace la marque initiale par cette marque. Sinon, on conserve la marque du "HMM amélioré".
- Si les deux phonèmes sont uniquement BF ou uniquement HF, alors on utilise l'algorithme de Brandt pour détecter le ou les instants de ruptures, id est les instants où $GLR > T$. Si un seul instant de rupture est détecté par l'algorithme de Brandt, on utilise cette marque. Dans le cas contraire, on garde la marque initiale.

5 Conclusion

Nous avons décrit un algorithme d'analyse multi-résolution qu'on a appelé AMREC. Nous avons aussi décrit les possibilités d'application de cet algorithme à la parole. En particulier, quelques expérimentations très simples nous ont conduit à une stratégie pour la segmentation phonétique de grands corpus. Cette méthode tire parti des avantages de plusieurs algorithmes de segmentation dont l'AMREC. Nous travaillons à déterminer les seuils de l'AMREC par apprentissage sur une petite base de données en utilisant un critère de mesure de performance introduit dans [2].

Références

1. S. Jarifi, D. Pastor, and O. Rosec. Jump and silence/speech detection for automatic continous speech segmentation. *Second International Symposium on Image/Video Communications (ISIVC)*, 2004.
2. S. Jarifi, D. Pastor, and O. Rosec. Brandt's glr method & refined hmm segmentation for tts synthesis application. *13th European Signal Processing Conference (EUSIPCO)*, 2005.
3. C. Lemoine. *Recherche de traits acoustiques de la parole bruité par analyse multi-résolution.* PhD thesis, Université de Bordeaux I, 1998.
4. L.Wang, Y. Zhao, M. Chu, J. Zhou, and Z. Cao. Refining segmental boundaries for tts database using fine contextual-dependent boundary models. *ICASSP*, vol.I :pp.641–644, 2004.
5. R.A. Obrecht. A new statistical approach for the automatic segmentation of continous speech signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.36 :pp.29–40, January 1988.
6. D. Pastor and R. Gay. Décomposition d'un processus stationnaire du second ordre. propriétés statistiques d'ordre 2 des coefficients d'ondelettes de localisation fréquentielle des paquets d'ondelettes. *Traitement du Signal*, vol.12 :pp.393–420, 1995.