

## Détection des occlusives à l'aide de la transformée en ondelettes

F. MALBOS, M. BAUDRY et S. MONTRÉSOR

CNRS, Université du Maine, Laboratoire d'Informatique, BP. 535, 72017 Le Mans cedex, France

**Résumé :** This paper describes the detection of stop consonants in the French language with the help of the wavelet transform. Our hypothesis is that the burst of stop consonants can be approximated with a pulse. The detection system is based on the study of the correlation functions between the modulus of the impulse wavelet transform and the same transform of the speech signal. As we are using a Gaussian wavelet, the correlation functions have a minimum before each maximum, where that maximum is synchronous with the burst of the stop consonant. Two signals  $S_M$  and  $S_m$  are computed. They are respectively the synchronous and asynchronous sums of all the correlation functions. To localize the local minima of the  $S_m$  function, two criteria are used. The temporal localisation of the pulse is accomplished with certain local maxima of the signal  $S_M$ . The method of detection is tested with a corpus which has 137 unvoiced and 55 voiced stop consonants. The rate of detection is 94% for the unvoiced stops and 75% for the voiced stops. This study shows that for detection, the burst of stop consonants can be modelled with a pulse.

### 1. Introduction

Parmi les phonèmes de la langue française, notre étude porte sur la détection automatique de la barre d'explosion des occlusives. Afin de simplifier la détection, l'impulsion glottique est modélisée par un bruit impulsionnel. La transformée en ondelettes est utilisée en raison de sa minimisation de l'incertitude temporelle en hautes fréquences.

### 2. Les ondelettes

#### 2.1. Présentation

La transformée en ondelettes est une transformation temps-échelle à résolution fréquentielle relative constante. Elle décompose un signal réel en une somme de fonctions élémentaires qui se déduisent les unes des autres par contraction et dilatation d'une fonction prototype appelée ondelette "mère" ou ondelette analysante. Elle associe à un signal réel  $p(t)$  une fonction  $O(b,a)$  (1) complexe et bidimensionnelle dont les arguments  $b$  et  $a$  (a strictement positif) représentent respectivement le temps et le facteur d'échelle. Elle peut être perçue comme un banc de filtres linéaires dont chaque cellule présente une résolution fréquentielle relative constante.

$$O(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} h^* \left( \frac{t-b}{a} \right) p(t) dt \quad (1)$$

$h(t)$  : ondelette mère,  $h^*(t)$  : complexe conjugué de  $h(t)$

$h\left(\frac{t-b}{a}\right)$  : ondelette à l'instant  $b$  avec le facteur d'échelle  $a$

#### 2.2. Localisation temporelle et fréquentielle

En raison de son analyse à résolution fréquentielle relative constante et du principe d'incertitude temps-fréquence dans le plan temps échelle ( $\Delta f \cdot \Delta t \geq 1/2$ ), la transformée en ondelettes offre une résolution temporelle élevée en hautes fréquences. Elle permet donc de localiser avec précision les discontinuités temporelles du signal de parole. L'ondelette de Morlet (enveloppe gaussienne) minimise la relation d'incertitude temps-fréquence offrant ainsi le meilleur compromis entre les localisations temporelle et fréquentielle [1].

**3. Décomposition fréquentielle**

La précision temporelle du système de détection de bruits impulsionnels nécessitant la minimisation du produit  $\Delta f \cdot \Delta t$ , l'ondelette de Morlet est à la base de ce travail. La localisation d'événements impulsionnels à l'aide de la transformée en ondelettes a fait l'objet de différents travaux [2]. Les coefficients d'ondelette sont calculés à l'aide d'un algorithme rapide [3]. Contrairement à l'algorithme "à trous" [4] il permet l'obtention d'un nombre constant de coefficients dans toutes les voies d'analyse. Le choix d'une décomposition des voies d'analyse en quart d'octave sur quatre octaves [372 Hz, 5000 Hz] se justifie de la façon suivante :

- une fréquence d'échantillonnage égale à 10000 Hz et une résolution fréquentielle relative de l'ondelette "mère" située entre le tiers et le quart d'octave,
- une concentration énergétique [5] :  
 dans la bande 500 Hz et au dessus de 4000 Hz pour les dentales,  
 en moyennes fréquences [1500-4000] Hz pour les vélares,  
 dans la bande [500-1500] Hz pour les labiales.

**4. Etude des fonctions de corrélation**

Notre système de détection est basé sur l'étude des fonctions de corrélation normalisées notées  $C_i$  entre les modules de chaque réponse impulsionnelle  $h_i$  du banc de filtres et les modules de la transformées en ondelettes  $O_i$  du signal de parole (2). La normalisation énergétique de la fonction  $C_i$  a pour but de s'affranchir de la différence d'amplitude entre les phonèmes de forte énergie (voyelle) et les occlusives.

$$C_i(p) = \frac{\sum_{n=0}^{n=N_0} |O_i(n)| |h_i(n+p)|}{\sqrt{\sum_{n=0}^{n=N_0} O_i^2(n)} \sqrt{\sum_{n=0}^{n=N_0} h_i^2(n)}} \quad (2)$$

Chaque réponse impulsionnelle du banc de filtres étant décrite par une fonction du type exponentielle, décroissante son module peut être considéré nul pour tous les instants  $t$  supérieurs à une valeur limite  $t_{ij}$ . Cette approximation est capitale et explique la présence pour chaque fonction  $C_i$  d'un minimum local  $m_{ki}$  précédant un maximum local  $M_{ki}$  dont l'amplitude est proche de l'unité (Figure 1). Ce phénomène apparaissant dans toutes les voies  $i$ , il est intéressant d'utiliser la notion de banc de filtres et de construire deux signaux  $S_M$  et  $S_m$  définis de la façon suivante :

- $S_M$  est la somme temporelle de toutes les fonctions  $C_i$  sur les 16 voies de la décomposition fréquentielle (3). Ses maxima sont synchrones avec le bruit impulsionnel recherché dans le signal.

$$S_M(n) = \sum_{i=1}^{i=16} C_i(n) \quad (3)$$

- $S_m$  est la somme asynchrone suivant une loi temporelle de toutes les fonctions  $C_i$  (4). Le décalage temporelle  $n_i$  ( $n_i=f(i)$ ) est inversement proportionnelle aux rapports des facteurs d'échelle. La fonction  $S_m$  offre des minima locaux très prononcés qui sont à la base du critère de détection des occlusives.

$$S_m(n) = \sum_{i=1}^{i=16} C_i(n - n_i) \quad (4) \quad n_i > 0 \text{ pour } i \neq 16 \text{ et } n_i = 0 \text{ pour } i = 16$$

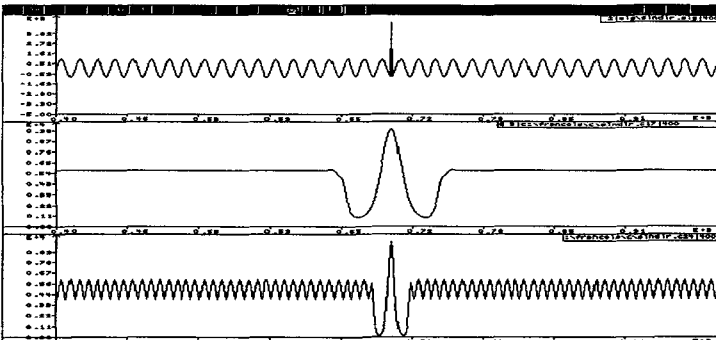


Figure 1

Figure 1a: Insertion d'un pic de Dirac dans un sinus de fréquence 625 Hz

Figure 1b : Fonction de corrélation normalisée pour une fréquence de 1487 Hz

Figure 1c : Fonction de corrélation normalisée pour une fréquence de 5000 Hz

## 5. Critère de détection

Il est intéressant de tracer un histogramme  $D$  représentant pour le corpus la distribution des amplitudes des minima locaux des fonctions  $S_m$  (Figure 2). Cette fonction de type gaussien (modélisée par la fonction  $D'$ ) peut être décomposée en la somme de trois distributions :

- $D_1$  : distribution des occlusives,
- $D_2$  : distribution des voyelles,
- $D_3$  : distribution correspondant aux autres phonèmes.

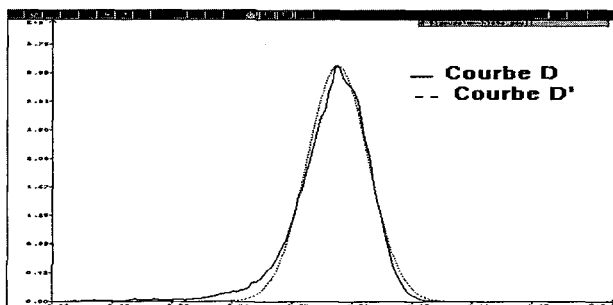


Figure 2

La différenciation de ces trois distributions est relativement délicate. Un premier critère  $C_1$  permettant la détection des occlusives a été implémenté. Il consistait à localiser par une méthode à seuil les minima locaux  $m$  dont l'amplitude était inférieure à  $S_1$  (seuil inférieur à la valeur  $m-2\sigma$ ,  $m=7,6$  et  $\sigma=0,8$  étant respectivement la moyenne et l'écart type de la gaussienne  $D'$ ). La mise en place du critère  $C_1$  a mis en évidence que pour un grand nombre de seuils  $S_1$ , il était difficile de séparer les détections correspondant aux occlusives et aux voyelles (détection des impulsions glottales).

Afin de diminuer le nombre de fausses alarmes correspondant aux voyelles, le critère  $C_1$  a été amélioré en ajoutant un critère temporel  $T_1$  que l'on peut énoncer de la façon suivante, soit :

- $A_m$  l'amplitude du minimum local  $m$  inférieur à  $S_1$ ,
- $A_l$  l'amplitude du minimum local  $l$  précédent  $m$ ,
- $A_k$  l'amplitude du minimum local  $k$  précédent  $l$ ,
- $m_e$  la médiane du signal  $S_m$ , calculée entre deux noyaux vocaliques comprenant le minimum  $m$ .

Les minima ne répondant pas au critère  $T_1$  (5) énoncé ci-dessous seront associés aux voyelles :

$$|m_e - A_m| > 1,6 \left| m_e - \frac{A_l + A_k}{2} \right| \quad (5)$$

D'autres critères complémentaires sont implémentés afin de diminuer le nombre de fausses alarmes :

- Une fonction délimitant le début et la fin de la phrase,
- Le calcul du centre de gravité fréquentiel pour supprimer les fricatives.

Le critère final de détection s'effectue en deux temps :

- La recherche des minima locaux  $m_c$  répondant au critère  $T_1$ ,
- La localisation du maximum local  $M_c$  de la courbe  $S_M$  possédant un minimum antérieur  $m_c$ .

La figure 3 montre les détections des occlusives sonore et sourde en contexte "p/u/b/u" .

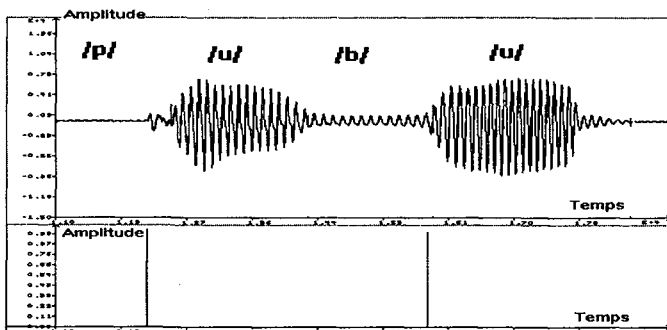


Figure 3

Figure 3a : Représentation amplitude-temps d'un signal de parole contenant une occlusive sourde et sonore

Figure 3b : Localisation temporelle de la barre d'explosion des occlusives

## 6. Synthèse des résultats

Notre système a été évalué à l'aide d'un corpus de 35 phrases présentant 192 occlusives (six hommes, cinq femmes). Les taux de détection et le détail des différentes occlusives sont présentés dans le tableau 1. Ces résultats sont accompagnés de 28 fausses détections parmi lesquelles on peut différencier :

- 21 fausses alertes dues à l'attaque d'une voyelle après un silence. Ce problème a déjà été relaté par d'autres équipes utilisant une méthode de détection différente [6].
- 7 fausses alertes apparaissent dans des situations délicates à supprimer (3 fricatives sourdes, 3 liquides, 1 consonne nasale).

Tableau 1

	Occlusives	Nombre total d'occlusives	Taux de détection	
Sourdes	p	49	92%	94%
	t	58	95%	
	k	30	97%	
Sonores	b	20	75%	75%
	d	29	69%	
	g	6	100%	

## 7. Comparaison des résultats avec d'autres méthodes

Cette comparaison est basée sur un article regroupant les résultats de trois équipes différentes [6]. La détection des occlusives sourdes se présente pour chaque équipe comme une étape précédant le décodage acoustico-phonétique. Le corpus BDFON est à la base de ces travaux. Une présentation succincte des indices utilisés ainsi que le pourcentage de détection de chaque équipe est proposée ci-dessous :

- I.C.P : Etude de l'évolution temporelle de différents paramètres (énergie, centre de gravité, densité des passages par zéro, ...). Cette équipe présente un score de détection de 94,3 %.
- C.N.E.T : La détection s'appuie sur l'évolution temporelle de l'énergie, la recherche d'une zone de stabilité précédant une barre de forte instabilité. Le taux de détection est égal à 95%.
- G.I.A : La localisation s'effectue au niveau de l'évolution de certains paramètres par la recherche de formes caractéristiques. Le pourcentage d'événements détectés est de 89 %.

## 8. Conclusion

L'utilisation de la transformée en ondelettes semble être un outil performant pour la détection des occlusives. Dans cette optique de détection, il est mis en évidence que la barre d'explosion des occlusives sourdes et sonores peut être modélisée par un bruit impulsif. Notre système de localisation est basé sur l'étude des fonctions d'intercorrélation normalisées entre les modules de chaque réponse impulsif du banc de filtres et le module de la transformée en ondelettes du signal de parole. Testé sur un corpus présentant 137 occlusives sourdes et 55 occlusives sonores, il offre un taux de détection respectif de 94% et 75%. Les performances de notre système de conception simple sont sensiblement équivalentes à celles proposées par d'autres équipes. Une étude postérieure à l'aide de la transformée en ondelettes devra conduire à la différenciation des occlusives et des attaques de voyelles. Afin d'améliorer notre système de détection, celui-ci devra être testé sur un corpus comportant un plus grand nombre d'occlusives.

## 9. Références

- [1] Dorize C. and Gram-Hansen K., : "Related positive time-frequency energy distributions", Wavelets and applications (Editor Y Meyer, Masson-Spinger Verlag, 1989) pp. 77-86.
- [2] Montessor M., Valiere J.C., Baudry M., "Détection et suppression de bruits impulsifs appliqués à la restauration d'enregistrements anciens", Premier Congrès de la Société Française d'Acoustique Volume II, P. Filippi and M. Zakharia Ed., Lyon France Avril 1990 (Les Editions de Physique, Les Ulis, 1990) pp. 761-764.
- [3] Barrat M. and Lepetit O., Traitement du signal VIII France 1 (1992) 43-49.
- [4] Holschneider M., Kronland Martinet R., Morlet J., Tchamitchian Ph., A real-time algorithm for signal analysis with the help of the wavelet transform, (Wavelets, time frequency methods and Phase Space, Springer-Verlag, 1989), pp. 286-304.
- [5] Halle M., Hugues G.W., Radley J.-P.A., J. Acoust. Soc. Am. 29 USA 1 (1957) 107-116.
- [6] Caelen J., Tattegrain H., Meloni H. Bulot R., Mercier G., Bonneau A. "Une base de règles pour le décodage acoustico-phonétique : le cas des occlusives sourdes", Actes du séminaire GRÉCO-PRC - Université de Nancy France 1988 pp. 51-63.