UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ
À L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN GÉNIE ÉLECTRIQUE

PAR
ROBERTO CHIODI

DÉTECTION D'ACTIVITÉ VOCALE BASÉE SUR LA TRANSFORMÉE EN ONDELETTES

SEPTEMBRE 2010

Résumé

Détecter de la voix dans un signal audio peut s'avérer être une tâche difficile, surtout lorsque la voix est soumise à un environnement bruyant, ou encore lorsque le signal audio passe au travers un système non-linéaire. De nombreuses méthodes utilisant des techniques différentes de traitement numérique de signal se sont penchées sur la conception de détecteurs d'activité vocale efficaces dans des niveaux de bruit élevé. La présente étude porte sur l'utilisation de la décomposition en ondelettes pour identifier et délimiter la voix dans un signal audio. L'objectif est de démontrer l'efficacité des ondelettes pour détecter la voix en présence de bruit ainsi qu'en présence de déformations liées au passage dans un système d'écoute non linéaire.

Après avoir passé en revu certaines méthodes de détection d'activité vocale retrouvées dans la littérature, nous proposons une méthode basée sur la transformée en ondelettes par paquet. Cette méthode exploite certains sous-signaux issus de la décomposition en ondelettes par paquet, pour obtenir une courbe qui suit fidèlement l'allure de la voix. À partir de cette courbe, on applique un seuil de décision final pour départager la voix du bruit. Notre méthode proposée est comparée avec une méthode récente basée sur les ondelettes (méthode de Wu et Wang), ainsi qu'avec la méthode traditionnelle G729-B. Ces

trois détecteurs d'activité vocale sont soumis à des signaux audio contenant des voix d'hommes et de femmes auxquels sont ajoutés des bruits divers à des niveaux différents. Afin de démontrer l'efficacité de ces détecteurs d'activité vocale en présence de non-linéarités, les signaux d'entrée sont préalablement passés dans une fonction non-linéaire pour simuler un système d'écoute non linéaire.

Les résultats montrent que les méthodes basées sur les ondelettes sont plus efficaces que le détecteur d'activité vocale G729-B en présence de bruit élevé. Notre méthode proposée s'avère être aussi efficace que la méthode de Wu et Wang tout en étant moins complexe en termes de calculs. Il a également été démontré que les détecteurs d'activité vocale basés sur les ondelettes conservent leurs performances en cas de non linéarité, contrairement à la méthode G729-B.

Remerciements

Je tiens vivement à remercier mon directeur de recherche, le professeur Daniel Massicotte, qui a supervisé ce travail de recherche. Son soutien, sa disponibilité, et la confiance qu'il m'a apportés tout au long de mon étude ont fait en sorte que je puisse accomplir ce projet de recherche dans les meilleures conditions. Ses connaissances et ses jugements m'ont permis d'acquérir des compétences essentielles en recherche.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

AMR	Adaptive Multi-Rate
AWGN	Additive White Gaussian Noise
CELP	Code Excited Linear Prediction
CNG	Confort Noise Generator
CSS	Clean Speech Signal
DAV	Détecteur d'Activité Vocale
DFT	Discrete Fourrier Transform
DOA	Direction Of Arrivals
DSP	Digital Signal Processor
DTX	Discontinuous Transmission
DWT	Discret Wavelet Transform
ETSI	European Telecommunications Standards Institute
FIR	Finite Impulse Response
GSM	Global System for Mobile communications
GSM	Global System for Mobile communications
ITU	Internationnal Telecommunication Union

LPC	Linear Prediction Coefficients
LSF	Line Spectral Frequencies
PWPT	Perceptual Wavelet Paquet Transform
ROC	Receiver Operating Characteristic
RBF	Radial Basis Function
SAE	Speech Average Envelope
SNR	Signal to Noise Ratio
SVAD	Spatial Voice Activity Detection
TO	Transformée en ondelettes
TEO	Teager Energy Operator
UIT	Union Internationale des Télécommunications
VAD	Voice Activity Detector
VAS	Voice Activity Shape
VAS	Voice Activity Shape
WPT	Wavelet Paquet Transform
WT	Wavelet Transform
WVAD	Wavelet Voice Activity Detection

Liste des symboles

f	fréquence
n	indice de temps pour les signaux discrets
t	indice de temps pour les signaux continus
Ts	période d'échantillonnage
Fs	fréquence d'échantillonnage
ms	millisecondes
dB	décibels
s(n)	signal discret représentant le signal recueilli par le microphone et transmis au VAD (signal d'entrée du VAD)
$s_{clean}(n)$	signal discret représentant un son contenant uniquement de la voix
$\hat{S}(n)$	signal $s_{clean}(n)$ passé dans un canal non-linéaire
$s_T(n)$	portion du signal discret s(n) découpé sur une trame T de longueur N
λ_V	Seuil de décision
λ_V^{opt}	Seuil optimal appliqué au VAS
N	nombre d'échantillons par trames

Chapitre 1
Introduction

La détection d'activité vocale se définit de manière générale comme étant le processus qui consiste, à partir d'un signal sonore, à différencier les portions qui contiennent de la voix à celles qui n'en contiennent pas. L'acronyme VAD[1] signifie « *Voice Activity Detector* », ou détecteur d'activité vocale, et définit le système qui exécute le processus de détection d'activité vocale. En principe le VAD fournit en sortie un signal binaire ('0' → absence de voix et '1' → présence de voix) sur la détection de la voix à partir d'un signal appliqué à son entrée.

Le VAD s'applique dans de nombreux systèmes de communications de la voix actuels. Parmi eux, citons la téléphonie mobile, la téléphonie par Internet, les systèmes de communications sans fils et la reconnaissance vocale. Il s'agit d'un problème non trivial qui explique l'existence d'une multitude de méthodes pour développer un VAD. Chaque méthode possède ses qualités et ses limites, et des recherches dans le domaine apportent régulièrement des résultats de plus en plus performants.

[1] L'acronyme anglophone VAD, plus largement utilisé comme terme technique, sera utilisé dans le mémoire

1.1 Problématique

On rencontre plusieurs problématiques lors de la conception et la fabrication d'un détecteur d'activité vocale. Celles que l'on rencontre le plus souvent sont reliées au bruit.

Un niveau de bruit environnant trop élevé noie la voix qu'on désire écouter, ce qui empêche le VAD de bien fonctionner. Pour cette raison, plusieurs méthodes ont été développées pour essayer d'améliorer la qualité du VAD dans des environnements bruyants tels que les mines, le trafic routier, une arrivée de train en gare, ou une usine. Parmi ces méthodes, citons les réseaux de neurones [KIM05], les méthodes basées sur les ondelettes [CHE05], [CHE02], [WUW06] les outils statistiques [SOH99], ou encore des méthodes basées sur l'utilisation de plusieurs microphones [RUB07], [GSC01], très efficaces pour cibler la source de la voix et éliminer le bruit.

Le bruit non-stationnaire, ou, en d'autres termes, un bruit dont la variance varie dans le temps (à l'opposé d'un bruit blanc), peut nuire à la performance d'un VAD. On rencontre ce genre de bruit dans la plupart des situations réelles (exemple: bavardages, usines, mines, chantiers, forages, etc.). Certains VAD sont moins efficaces lorsque la voix est dans un environnement avec un ou plusieurs bruits non-stationnaires. De plus, le nombre de sources de bruits influe sur la qualité du VAD. Il est donc intéressant de se pencher sur cette problématique si l'on veut concevoir un VAD robuste et efficace. Certaines méthodes, telles que les réseaux de neurones (RBF – *Radial Basis Function*) [KIM05], des techniques avec deux microphones [RUB07], et des méthodes combinant plusieurs techniques déjà existantes mais inefficaces dans des bruits non-stationnaires [TAN00], se sont penchées sur le problème du bruit non-stationnaire.

Le signal audio, qui contient la voix, peut également être soumis à un autre type de déformation : la non-linéarité. Lorsque le signal audio passe dans un système d'écoute non linéaire, il en résulte une distorsion du signal proportionnelle à la dureté du canal. On rencontre, par exemple, ce genre de problème dans les communications spatiales. À notre connaissance, aucune méthode de VAD présentée dans la littérature ne s'est penchée sur cette problématique moins connue.

Lorsque l'interférence provient d'une tierce personne, cela amène le problème suivant : le VAD considère cette voix, non désirée, comme étant la voix principale (ou voix d'intérêt). Quelques méthodes basées sur plusieurs microphones prétendent régler ce problème, appelées SVAD (*Spatial Voice Activity Detector*) [GSC01].

La position de l'interlocuteur constitue également un problème. Lorsque la personne qui parle se déplace autour des microphones qui captent le signal audio pour le transmettre au VAD, on peut avoir des données erronées. Des méthodes qui utilisent plusieurs microphones pour capter la voix prétendent améliorer ce problème [GSC01], [RUB07].

Il est pertinent de noter que l'application pour laquelle on conçoit un VAD va changer notre manière de le concevoir. Par exemple, un VAD conçu pour une application de type « téléphone cellulaire », en plus de transmettre la voix, peut se permettre de laisser passer des signaux correspondant à de la musique, ou à une autre source d'information sonore. Dans le cas d'une application de type « reconnaissance vocale », uniquement la parole doit être interprétée.

Il est important de distinguer la parole de la voix : la voix se définit comme tout son qui sort de la bouche, alors que la parole est un ensemble de sons sortant de la bouche qui forme un message servant à la communication. Dans le cas de la reconnaissance vocale, on

ne peut pas se contenter de recueillir la voix : il faut aller chercher les sons correspondant à de la parole. Par exemple, un cri ou une onomatopée est considéré comme de la voix mais non comme de la parole.

Par ailleurs, dans le cas d'un VAD appliqué à la vidéo conférence, l'utilisateur se promène autour du microphone. Sa position par rapport au microphone, qui enregistre le signal à traiter, change, contrairement à une application où le microphone est placé à une distance fixe de la bouche (exemples : téléphone cellulaire, reconnaissance vocale). En quelque sorte, les performances d'un VAD dépendent fortement de la problématique d'application. Par exemple, un VAD développé pour les cellulaires n'offrira pas les performances attendues dans une application minière.

En choisissant une application concrète, on se limite à certaines particularités de la problématique. La variété des problèmes rencontrés rend la conception d'un VAD difficile. Pour cette raison, il faut cibler une application pour notre VAD. Cette démarche va permettre de se concentrer sur certains problèmes uniquement. Dans le présent projet, nous avons choisi l'application du téléphone cellulaire qui, d'une manière générale, s'applique à tout système de communication où le microphone est placé près de la bouche (voix d'intérêt). Elle constitue par ailleurs l'application principale dans la plupart des articles récents traitant de la détection d'activité vocale.

Le problème principal nuisible à la détection d'activité vocale appliqué aux téléphones cellulaires est la présence de bruit de fond. Qu'il s'agisse de bruit élevé ou de bruit non-stationnaire, le bruit fait l'objet principal de la grande majorité des articles traitant des VAD. De plus, le téléphone cellulaire étant mobile comme application ce qui implique des environnements avec des bruits de fond changeant et évoluant dans le temps. Il est donc

impératif que le VAD puisse s'adapter continuellement à n'importe quel bruit, et cela même si le niveau varie.

Les problèmes liés à la position (direction ou proximité du microphone) de la personne utilisant un cellulaire est moins gênant vu que la personne a toujours le microphone placé devant la bouche. La voix provient toujours du même endroit. Par contre, la problématique du bruit qui provient d'une tierce personne s'applique puisque le type de bruit et son niveau par rapport à la puissance du signal d'intérêt auront une incidence certaine sur la performance du VAD. Par exemple, certains environnements peuvent s'avérer très difficile : lieu public avec beaucoup de gens autour, à l'intérieur d'un véhicule, ou encore au restaurant.

Cependant, il est difficile de concevoir un seul VAD capable de répondre à tous ces critères. Nous nous intéresserons plus particulièrement à concevoir un VAD capable de s'adapter à des niveaux de bruit élevés et de nature différente (exemple : bruit de mines, d'avion et de rue). De plus, une originalité de notre étude portera sur la non-linéarité du signal reçu. Comme peu de méthodes se sont orientées sur cette problématique, il serait intéressant de voir comment se comporte un VAD lorsque le signal d'entrée a été passé dans un système d'écoute non linéaire. La non-linéarité ne s'applique pas aux téléphones cellulaires mais plutôt, par exemple, aux communications dans l'espace entre astronautes.

Finalement, nous porterons aussi intérêt à la mise en pratique du VAD dans une technologie d'intégration à très grande échelle telle que les processeurs en traitement numériques de signaux (DSP – *Digital Signal Processing*) et les réseaux prédiffusés programmable par l'utilisateur (FPGA – *Field Programmable Gate Array*).

1.2 Objectifs

L'objectif principal de ce projet de recherche est de concevoir une méthode de détection d'activité vocale qui répondra au mieux aux problématiques reliées à l'intensité et au type de bruit, ainsi qu'aux problématiques liées à la non linéarité.

Les sous-objectifs permettant de rencontrer l'objectif principal sont énoncés ci-dessous :

1. Répertorier et étudier les méthodes de détection d'activité vocale performante dans des environnements difficiles (bruit élevé, bruit non-stationnaire). Nous porterons plus d'intérêt aux méthodes récentes et dont les résultats semblent pertinents à notre problématique d'intérêt. Par ailleurs, nous porterons un intérêt plus particulier aux méthodes VAD basées sur la transformée en ondelettes (TO) car cette technique a des propriétés intéressantes pour ce type d'application. De plus, des travaux basés sur la TO ont été réalisés au LSSI [LMG01], [LGM01], [LMG00], ce qui nous incite à mettre à profit cette piste.
2. Dans les méthodes répertoriées, choisir les méthodes qui vont servir de références dans les études comparatives de performances.
3. Proposer une méthode basée sur la transformée en ondelettes tenant compte des performances (bruit, linéarité et non linéarité) ainsi que le niveau de complexité d'implémentation.
4. Réaliser une étude comparative des performances de chacune des méthodes. La performance d'un VAD se traduit par sa capacité de détecter de la voix. Pour mesurer la performance, chaque méthode sera testée en utilisant de la voix

enregistrée dans différents environnements plus ou moins bruyants, avec et sans passage dans un système d'écoute non linéaire.

Le résultat espéré de ce projet est de montrer :

- que le VAD proposé est supérieur au VAD traditionnel G729B dans des situations de bruit et de non linéarité.
- la performance des VAD qui utilisent la technique des ondelettes dans des situations d'un système d'écoute non linéaire.

1.3 Méthodologie

Une recherche bibliographique permettra de définir et de mieux comprendre les différents algorithmes de détection d'activité vocale. Une recherche plus avancée sera faite sur la détection d'activité vocale basée sur la transformée en ondelettes. On retiendra les méthodes les plus pertinentes et les plus récentes.

Une fois les méthodes déterminées, nous validerons par des simulations leur fonctionnement puis pour comparer leurs performances selon des critères prédéfinis. La simulation s'effectue par la programmation. Avant de programmer une méthode, il faut comprendre et maîtriser l'algorithme de la méthode. On se basera sur les explications fournies dans la littérature, en particulier dans les articles correspondant à la méthode qu'on développe. Pour l'étape de programmation, on utilisera MATLAB® de la compagnie MathWorks, outil logiciel très répandu dans le milieu du génie électrique, qui offre un langage de programmation souple, facile et agréable.

Le sous-objectif #4 consiste à tester les méthodes pour valider leur fonctionnement. Pour comparer les performances de ces méthodes, il faudra établir une base de données de

tests. Cette base de données sera constituée de signaux sonores différents, chaque signal sonore correspondant à une ou plusieurs personnes (homme ou femme) parlant dans un certain environnement et sous des conditions de bruits plus ou moins difficiles. Par exemple, un signal sonore pourrait correspondre à une femme parlant dans une rue avec beaucoup de circulation de voitures. On utilisera une base de données composée de signaux sonores enregistrés et d'échantillons à partir d'une base de donnée d'une étude exhaustive, tel que la base de données AURORA2 [AUR00]. Il s'agit d'une base de données traditionnelle et très utilisé pour montrer la performance des VADs et de systèmes d'annulations de bruit. Cette base de données constituera un outil important pour comparer les méthodes entre elles. Pour plusieurs signaux sonores représentant chacun un environnement et un niveau de bruit différent, on établira une procédure d'évaluation quantitative pour établir une comparaison de performance entre les méthodes en fonction de paramètres tel que le niveau de bruit, la nature du bruit, et la non linéarité.

1.4 Organisation du mémoire

Pour commencer, le chapitre 2 traite des bases de la détection d'activité vocale. Les algorithmes traditionnels sont étudiés à la section 2.2, suivi d'algorithmes plus récents à la section 2.3. Des comparaisons entre les méthodes traditionnelles et les plus récentes sont effectuées à la section 2.4.

Le chapitre 3 traite de la transformée en ondelettes appliquée à la détection d'activité vocale. Premièrement, les principes de bases de la transformée en ondelettes sont expliqués à la section 3.1. Par la suite, la section 3.2 fait l'étude de deux méthodes de détection d'activité vocale utilisant la transformée en ondelettes. La section 3.3 explique

notre méthode proposée. Pour finir, la section 3.4 parle des méthodes de seuils applicables aux méthodes aux sections 3.2 et 3.3.

Le chapitre 4 porte sur les résultats de comparaison entre différents VAD considérés. La procédure de tests et de validations est expliquée à la section 4.1, la section 4.2 décrit soigneusement les paramètres de mesures utilisés pour les tests et la section 4.3 présente les résultats obtenus. Ce chapitre se termine par une synthèse des résultats à la section 4.4.

Ce mémoire prend fin au chapitre 5 par une conclusion générale sur le travail effectuée, les résultats obtenus et nos contributions scientifiques.

En Annexe, nous présentons une implémentation de VAD réalisé sur DSP (*Digital Signal Processor*) et une publication d'article scientifique

Chapitre 2
Méthodes de détection d'activité vocale

Depuis une trentaine d'années, de nombreuses méthodes pour développer des VAD ont été développées pour différentes applications. Ces méthodes sont basées sur des algorithmes de traitement numérique du signal. Dans ce chapitre, nous commencerons à présenter le principe général de la détection d'activité vocale (section 2.1). Ensuite, nous présenterons quelques méthodes traditionnelles utilisées comme références dans la littérature (section 2.2) ainsi que des méthodes plus récentes (section 2.3). Le but de présenter ces méthodes est de faire ressortir les procédures générales de fonctionnement d'un VAD, ainsi que les avantages et limites des différents VAD. L'étude de ces méthodes permet également de mieux comprendre les problématiques reliées aux VAD, ainsi que de comparer les méthodes entre elles pour mieux analyser le processus général d'un algorithme de VAD. On finira le chapitre par une conclusion sous forme d'analyse qui résumera ce qui a été vu dans le chapitre.

2.1 Principe du VAD

On définit détection d'activité vocale comme le processus qui consiste, à partir d'un signal audio, à différencier les portions qui contiennent de la voix à celles qui n'en

contiennent pas. Par convention, la sortie d'un VAD est '1' si on a activité de voix, ou '0' si on n'a pas d'activité de voix. Dans la littérature, on parle parfois de région active (région qui contient de la voix) et région inactive (région sans voix). À titre d'exemple, un résultat de VAD pour un extrait de parole est présenté à la figure 2.1.

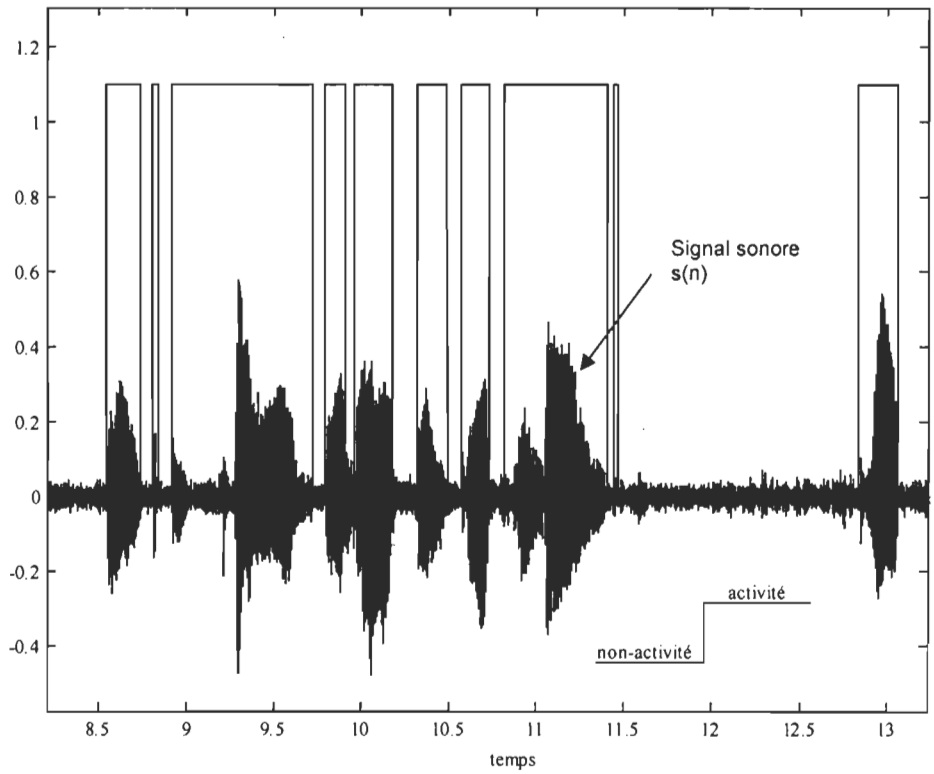


Figure 2.1 Exemple illustrant le principe de la détection d'activité vocale.

La grande majorité des VAD découpent le signal audio d'entrée en trames. Une trame se définit comme une portion de signal de longueur fixe et de l'ordre de quelques millisecondes (habituellement de l'odre de 10 à 50 ms). Le VAD va décider pour chaque trame si elle est active ou inactive. On va donc, dans le résultat du VAD, parler de trame active et de trame inactive.

Les trames peuvent être découpées avec chevauchement (*overlapping*) ou sans chevauchement. La figure 2.2 illustre le principe de découpage du signal audio *s(n)* avec plusieurs types de découpages correspondant chacun à un pourcentage de chevauchement. Un découpage à 0% correspond à un découpage sans chevauchement. Le signal s(n) est décomposé en trames de longueur égale (N échantillons par trame) appellés $s_T(n)$, où T est l'indice de trame.

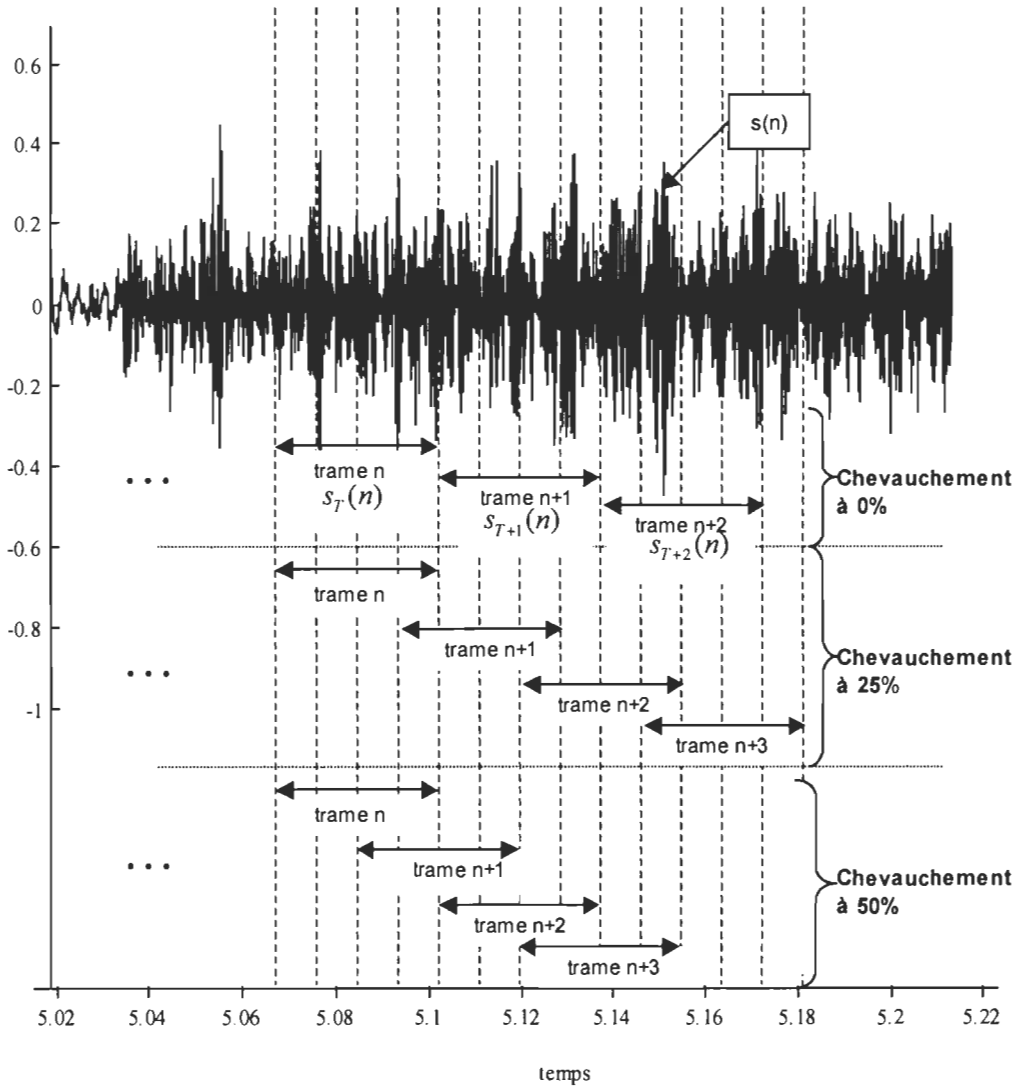


Figure 2.2 Découpage du signal audio à trois taux de chevauchement différents.

Le chevauchement peut avoir une incidence sur la précision du VAD mais aussi sur la complexité de calcul. Effectivement, un chevauchement à 75% va demander plus de calcul qu'un chevauchement à 25%, mais peut, selon la méthode utilisée, fournir plus d'informations et ainsi accroître la performance d'un VAD.

Quel que soit le type de VAD, d'algorithmes et de techniques utilisés, on retrouve une forme générale applicable à l'ensemble des VAD, représentée à la figure 2.3. À partir d'une portion de signal audio, on extrait un ensemble de paramètres. Ces paramètres peuvent être calculés dans le domaine temporel (nombre de passages par zéro, niveau d'énergie, coefficients d'autocorrélation, etc.) ou à partir du domaine spectral (DFT, analyse de forme spectrale, DOA, ondelettes, etc.). À partir de ces paramètres, on doit décider si la sortie vaut '0' ou '1' (respectivement inactivité ou activité vocale). Dans la plupart des cas, on utilise des algorithmes basés sur des règles de seuils. Ces seuils peuvent être fixes ou encore s'adapter en fonction, par exemple, du niveau de bruit. D'autres méthodes plus complexes de décision ont été utilisées comme par exemple les réseaux de neurones (cf. section 2.3) ou la logique floue (*fuzzy logic*) [FUZ98].

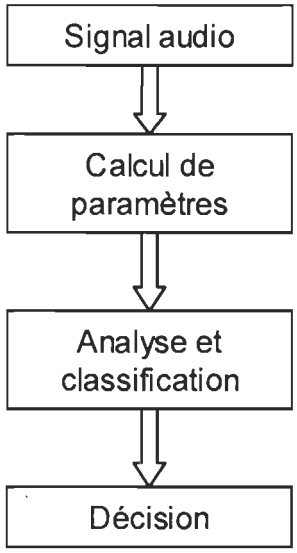


Figure 2.3 Processus général d'un algorithme de détection d'activité vocale.

Voici certains points qui caractérisent une méthode de détection d'activité vocale :

- La performance : la capacité à détecter l'activité vocale dans différents milieux sonores et dans des conditions plus ou moins difficiles.
- La complexité : plus l'algorithme demande de calculs mathématiques lourds et complexes, plus difficile sera son implémentation sur circuit numérique en temps réel.
- Sa capacité à s'adapter à des changements (niveau de bruit, type de bruit)
- Le nombre de paramètres fixes (constantes) à déterminer initialement, ainsi que des paramètres que le VAD doit connaître à priori pour fonctionner (par exemple la connaissance rapport signal sur bruit (SNR – *Signal to Noise Ratio*). Plus le nombre de paramètres est grand et moins le VAD sera autonome et donc moins apte à

s'adapter à des changements. Inversement, un algorithme qui requiert peu de paramètres initiaux à déterminer de manière empirique pourra s'adapter plus facilement et conviendra mieux à une implémentation en temps réel.

2.2 Méthodes VAD traditionnelles

2.2.1 L'algorithme VAD G729 Annexe B

Le G729 [UIT96] est un algorithme de compression audio standardisé par l'Union Internationale des Télécommunications (ITU - *International Telecommunication Union*). L'annexe B du G729 comprend des algorithmes de VAD, de transmission discontinue (DTX – *Discontinuous Transmission*), et de génération de bruit de confort (CNG - *Confort Noise Generator*). Ces algorithmes ont été conçus pour réduire le débit de transmission pendant les pauses de la parole. Considérant que cet algorithme est souvent utilisé comme méthode de référence traditionnelle dans la littérature, nous allons décrire brièvement l'algorithme du VAD [UIT96].

Dans cet algorithme, le signal audio est décomposé en trames de 10ms. Pour chaque trame, un ensemble de paramètres est extrait du signal vocal :

- l'énergie de la pleine bande de fréquences (équation (2.1)),
- l'énergie dans la bande des basses fréquences (équation (2.2)),
- l'ensemble des fréquences de raies spectrales (LSF – *Line Spectral Frequencies*), et
- le nombre de passages par zéro (équation (2.3)).

L'énergie dans la pleine bande de fréquences Ef est le logarithme du premier coefficient d'autocorrélation normalisé $R(0)$:

$$Ef = 10 \cdot \log_{10}\left[\frac{1}{240} R(0)\right] \tag{2.1}$$

L'énergie dans la bande de fréquences basse E_l mesurée sur la bande de 0 à F1 (Hz) est calculée comme suit :

$$E_l = 10 \cdot \log_{10}\left[\frac{1}{N} h^T \cdot R \cdot h\right] \tag{2.2}$$

h étant la réponse impulsionnelle d'un filtre FIR dont la fréquence de coupure est de F1, R étant la matrice d'autocorrélation *Toeplitz* avec les coefficients d'autocorrélation sur chaque diagonale, et N la longueur de la trame.

Le nombre de passages par zéro (ZC – *Zero Crossing*) normalisé pour chaque trame est calculé par:

$$ZC = \frac{1}{2N} \sum_{i=0}^{N-1} \left|\operatorname{sgn}\left(x(i)\right) - \operatorname{sgn}\left(x(i-1)\right)\right| \tag{2.3}$$

où $x(i)$ est le signal d'entrée et N le nombre d'échantillons par trames.

À chaque trame, des variables appelées « moyennes glissantes » sont calculées. Ces variables sont obtenues à partir des valeurs des quatre paramètres (Ef, El, ZC, LSF) obtenus lors des trames précédentes. Ces variables sont mises à jour uniquement pendant les périodes de non-activité vocale.

À chaque trame, quatre mesures sont obtenus selon des calculs de différence entre les paramètres et les moyennes glissantes: la distorsion spectrale, la différence d'énergie dans la pleine bande de fréquence, la différence d'énergie dans la bande de basses fréquence, et la différence des nombres de passage par zéro.

La décision finale est prise en fonction de ces quatre paramètres de différences qui sont comparés à des constantes. La sortie du module VAD est soit 1 ou 0, indiquant respectivement la présence ou l'absence d'activité vocale.

Le VAD G729-B est utilisé dans la grande majorité des méthodes de VAD comme la méthode traditionnelle de référence la plus connue. Elle constitue une bonne méthode de comparaison et de référence dans la plupart des articles scientifiques traitant de VAD.

2.2.2 VAD-AMR

L'ETSI (*European Telecommunications Standards Institute*) fournit un VAD conçu pour la téléphonie mobile GSM (*Global System for Mobile communications*). Plus précisément, ce VAD sert à la transmission discontinue (DTX, *Discontinuous Transmission*) pour les encodeurs de type AMR (*Adaptive Multi-Rate*) [ETS99].

Il existe deux algorithmes différents, appelés le « AMR option 1 » et le « AMR option 2 ». Le choix de l'algorithme dépend de l'infrastructure et du matériel utilisé dans le réseau de communication [ETS99]. Nous nous attarderons uniquement sur le « VAD-AMR option1 » pour la simple raison que les deux algorithmes se ressemblent et que ne désirons pas nous attarder trop longtemps sur les méthodes traditionnelles.

Le rôle du VAD-AMR est de décider si la trame contient des signaux qui doivent être transmis (voix, musique, ou autres tonalités susceptibles de contenir de l'information pertinente). La sortie du VAD est donc '1' si elle contient ce genre de signaux, '0' dans le cas contraire.

Le « VAD-AMR option 1 » décompose le signal en trames de 20 ms. À chaque trame, plusieurs paramètres sont extraits de cette portion de signal audio de 20 ms. En premier lieu, le signal est divisé en neuf sous-signaux, chaque sous-signal correspondant à un intervalle de fréquence. Le niveau de chaque sous-signal est ensuite calculé. Parallèlement, un autre paramètre appelé *pitch detection* permet de localiser les signaux périodiques qui correspondent souvent à des voyelles. Le résultat est booléen pour indiquer présence ou absence de *pitch*. Puisque la détection de *pitch* ne peut pas toujours détecter certaines tonalités à caractère informatif, un paramètre appelé détection de ton (*tone détection*) est utilisé. Ce paramètre est calculé à partir d'autres paramètres utilisés dans l'encodeur AMR. Le résultat est '1' s'il y a détection de tonalité, sinon '0'. Un dernier paramètre, *Complex Signal Analysis*, permet d'indiquer si on a affaire à un signal complexe comme de la musique. Ce paramètre est calculé à partir d'autres paramètres d'analyse de *pitch* fournis par l'encodeur.

La différence entre les niveaux de volume du signal d'entrée et le bruit de fond est calculée comme suit :

$$snr\_somme = \sum_{k=1}^{NBss} MAX\left(1.0, \frac{level(k)}{bckr\_est(k)}\right)^2 \qquad (2.4)$$

où *level(k)* est le niveau du signal pour le sous-signal *k*, *bckr_est(k)* est l'estimé du niveau du bruit de fond pour le sous-signal *k*, et *NBss* le nombre total de sous-signaux (*NBss* = 9).

Pour prendre une décision finale de détection d'activité vocale, la différence *snr_somme* est comparée à un seuil (*vad_seuil*). Une moyenne des niveaux de bruits de fond de chaque sous-signal est calculée pour permettre la mise à jour du seuil :

$$noise\_level = \sum_{k=1}^{NBss} bckr\_est(k) \tag{2.5}$$

$$vad\_seuil = VAD\_SLOPE \cdot (noise\_level - VAD\_P1) + VAD\_THR\_HIGH \tag{2.6}$$

où *VAD_SLOPE, VAD_THR_HIGH* et *VADP1* sont des constantes prédéfinies.

2.3 Méthodes modernes pour les VAD

D'autres méthodes, basées sur des techniques plus modernes de traitement numérique du signal, sont apparues dans la dernière décennie. Compte tenu du nombre important de ces méthodes, nous limiterons notre étude à deux méthodes utilisant des techniques de traitement de signal différentes. Suite à notre recherche bibliographique, nous présentons dans cette section deux méthodes VAD sur lesquelles nous avons porté une plus grande attention.

2.3.1 VAD utilisant un réseau de neurones RBF

Le VAD ici proposé est basé sur un réseau de neurones RBF (*Radial Basis Function*) [KIM05]. Le schéma général du VAD est présenté à la figure 2.4. Le réseau de neurones utilise trois entrées, soient trois paramètres utilisés par le codec CELP (*Code excited linear prediction*), un algorithme de codage de la voix utilisé dans le traitement de la voix. Ils sont

calculés à partir du signal d'entrée : le *short-time average power parameter (E)*, le *zero-order most likehood parameter* (Z) et le *pitchperiod difference parameter (P)*.

$$E = 10\log\left[\frac{1}{N}\sum_{n=m-N+1}^{m} s(n)^2\right] \tag{2.7}$$

avec N la longueur de la trame en terme d'échantillons et *s(n)* le signal d'entrée du VAD.

$$Z = \log\sum_{i=0}^{p}\alpha_i^{\,2} \tag{2.8}$$

où α_i sont les coefficients LPC (*Linear Prediction Coefficients*) et p l'ordre de prédiction.

$$P = \max_{\tau}\left[\frac{\log\sum_{n=0}^{N-1} r(n)r(n-\tau)}{\log\sum_{n=0}^{N-1} r(n)r(n)}\right] \tag{2.9}$$

où $20 \le \tau \le 160$ et *r(n)* est le signal d'erreur prédit linéairement (*linear prediction error signal*) au temps *n*.

Ces paramètres fonctionnent très bien sous des bruits de fond élevés. Lors de périodes de voix, E est élevé, Z est élevé, et P est stable. Dans les cas d'inactivité de voix, ces caractéristiques sont complètement opposées.

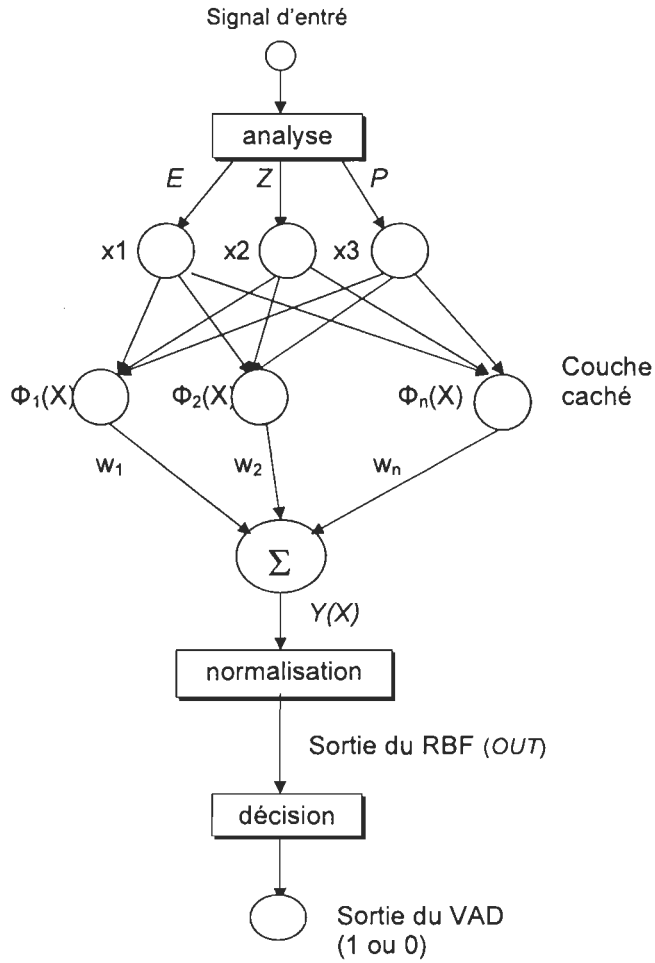


Figure 2.4 Structure du VAD-RBF

La structure de base du réseau de neurones RBF est présentée à la figure 2.4. Il s'agit d'un réseau de neurones à deux couches, dont une couche cachée. Chaque neurone de la couche cachée calcule la distance euclidienne entre son vecteur de centres et le vecteur de longueur trois correspondant à chacune des entrées *E*, *Z* et *P*. La distance au carrée est divisée par un paramètre ρ puis le résultat passe dans une fonction non linéaire, une fonction gaussienne $\Phi(\bullet)$:

$$\phi(x) = \exp\left(-\frac{(x-c)^2}{\rho}\right) \tag{2.10}$$

où c est le centre. Le paramètre ρ détermine la largeur de la gaussienne et est déterminé de manière empirique. Plus les centres sont rapprochés de l'entrée, plus la sortie du neurone va s'approcher de 1. Chaque sortie est ensuite multipliée par son poids, w_i, correspondant. La sortie du RBF est définie par :

$$y(X) = \sum_{i=1}^{N} w_i \cdot \exp\left(-\frac{\|X - ci\|^2}{\rho_i}\right) \tag{2.11}$$

Les centres de la couche cachée sont mis à jour par l'algorithme *k-means clustering* [HAY99], qui est un algorithme d'apprentissage non-supervisé performant et utilisé dans les réseaux de neurones. Les poids, w_i, à la sortie des neurones de la couche cachée sont mis à jour par l'algorithme de rétropropagation [HAY99]. Les algorithmes *k-means clustering* et de rétropropagation sont utilisés pour mettre à jour continuellement les paramètres du réseau de neurones.

La sortie y est ensuite passée dans la fonction sigmoïde, pour normaliser la sortie entre '0' et '1' :

$$OUT = \frac{1}{1+\exp^{-y}} \tag{2.12}$$

La sortie OUT du réseau de neurone est comparée à un seuil pour déterminer si la trame contient de la voix ou non. Le seuil est déterminé de manière empirique, en tenant compte du fait que si sa valeur est trop grande, le VAD va considérer trop souvent des trames

actives comme des trames inactives. À l'inverse, si la valeur de seuil est trop basse, le VAD va considérer trop souvent des trames inactives comme des trames actives. Pour diminuer le nombre d'erreurs, on utilise le *hangovers sum* (*HOsum*) qui est la somme des *m* dernières sorties.

$$HOsum(n) = \sum_{i=n-m}^{n} OUT(i) \tag{2.13}$$

Le paramètre *m* doit être déterminé de manière empirique. Si la sortie est plus grande que le seuil de décision, alors la trame est définie comme active. Pour détecter plus précisément les trames de voix, si la sortie est inférieure au seuil, la sortie n'est pas immédiatement considérée inactive. Une règle de décision basée sur deux seuils fixes, le *silence hangover threshold (HOSilence)* et le *voice hangover threshold (HOVoice),* est appliquée : si la trame précédente est active, on compare *HOSum* au seuil *HOVoice*, et si la trame précédente est du silence, on compare *HOSum* au seuil *HOSilence*. Si *HOSum* est supérieure au seuil (*HOvoice* ou *HOSilence*, selon le cas), la trame est alors déterminée comme active, sinon comme inactive.

2.3.2 VAD basé sur l'homogénéité des estimés DOA

Le présent VAD [RUB07] utilise deux microphones comme entrée. Le schéma général du VAD est illustré à la figure 2.5. Cette méthode présente un bon compromis entre complexité et performance. Effectivement, deux microphones sont plus faciles à appliquer qu'une matrice complète de microphones [GSC01]. La force de ce VAD réside dans le fait que l'utilisation de deux microphones permet à l'utilisateur de se promener autour sans que

la performance du VAD n'en soit notablement affectée.

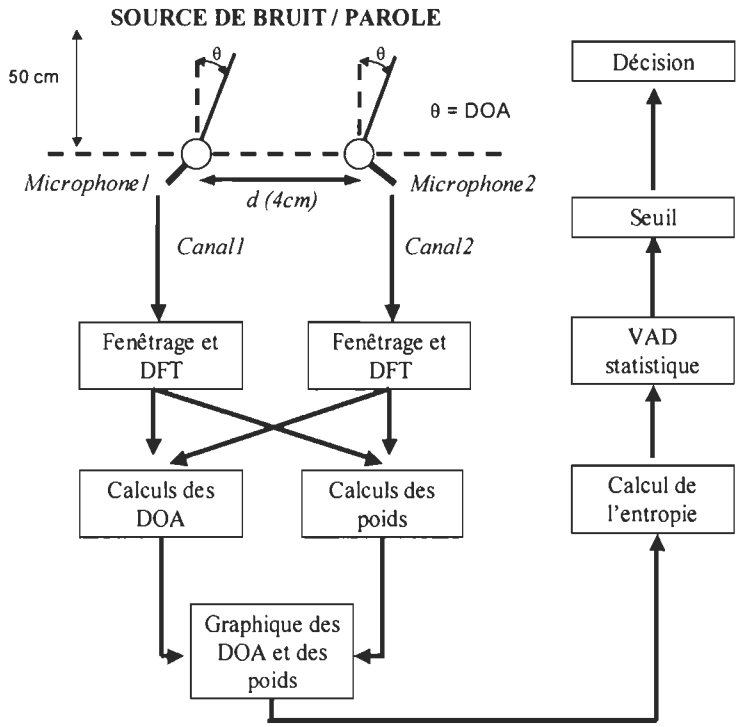


Figure 2.5 Schéma de la méthode DOA

Deux signaux sont respectivement recueillis par deux microphones espacés d'une distance fixe *d*. Chaque signal est divisé en trames de longueur *N*. Chaque trame est multipliée par une fenêtre de *Hamming* puis une transformée discrète de Fourier (DFT – *Discrete Fourrier Transform*) [PRO96] est effectuée sur la trame. On obtient les coefficients *x* dans le domaine des fréquences :

$$x_m(f,t) = \rho_m(f,t) \cdot e^{j\varphi_m(f,t)} \tag{2.14}$$

où $\rho(f,t)$ est l'abscisse et $\Phi(t,f)$ la phase au temps t et à la fréquence f, et m l'indice du canal (entrée microphone 1 ou entrée microphone 2). À partir des données spectrales, x_m, on

calcule les estimés DOA pour chaque échantillon de fréquence :

$$\theta(f,t) = \frac{\arcsin\left(vs \cdot \left(\varphi_1(f,t) - \varphi_2(f,t)\right)\right)}{2\pi \cdot d \cdot f} \tag{2.15}$$

où *vs* est la vitesse du son et *d* la distance entre les deux microphones. Au même moment, on calcule les poids correspondants à chaque θ à l'aide de la relation suivante :

$$W(f,t) = \rho_1^2(f,t) + \rho_2^2(f,t) \tag{2.16}$$

Le but est de trouver une région dans la région temps-fréquence où les DOA sont homogènes. Pour cela, on établi un histogramme pour chaque trame pour analyser la distribution des DOA. L'entropie de cette distribution est ensuite calculée selon la relation :

$$H(f,t) = -\sum_{b=1}^{B} pb(f,t) \cdot \log_2(pb(f,t)) \tag{2.17}$$

où *pb(f,t)* est la probabilité du DOA pour l'incrément *b*. À partir de l'entropie, on calcule l'homogénéité DOA $\Delta(f,t)$ en inversant et normalisant l'entropie selon l'équation:

$$\Delta(f,t) = \frac{(1 - H(f,t))}{H_{\max}} \tag{2.18}$$

Ce résultat est un graphique temps-fréquence, où chaque rang varie de 0 (bruit, processus aléatoire), à 1 (voix). $\Delta(f,t)$ suit une loi normale où la variance devient plus grande en présence de bruit. À partir de ce résultat, on calcule le *log-likehood ratio* tel qu'utilisé dans [SOH99]. Ce dernier est comparé à un seuil fixe pour finalement décider si on a absence ou présence de voix dans la trame.

2.4 Conclusion

En conclusion, nous avons exposé un certain nombre de méthodes pour les VAD. Nous avons pu noter que le VAD G729-B est souvent utilisé comme VAD de référence dans le cadre d'une étude comparative de performance. Les autres méthodes de la section 2.3 ont été programmées et évaluées dans des contextes restreints afin de répondre aux points suivants: i) bien comprendre les propositions de ces méthodes, ii) voir si la méthode décrite est complète et répétable tel que proposé par les auteurs, iii) enrichir nos connaissances sur les méthodes VAD, iv) analyser la complexité des calculs, et v) répondre aux objectifs de notre projet tel la non-linéarité du canal.

Finalement, nous avons pu constater que les méthodes basées sur la transformée en ondelettes nous paraissent plus intéressantes du point de vue performance et complexité d'implémentation. Pour cette raison, le chapitre suivant y est consacré.

Chapitre 3
VAD basé sur la décomposition en ondelettes

L'étude d'un signal dans le domaine des fréquences permet d'aller chercher de nombreuses informations, non accessibles dans le domaine temporel. Dans de nombreuses applications, l'analyse fréquentielle a une place importante dans le domaine du traitement du signal. Dans le cas de la détection d'activité vocale, de nombreuses méthodes vont traiter de l'information à partir des fréquences du signal pour détecter la présence de voix. Ce mémoire porte une attention particulière sur une technique d'analyse temps-fréquence : la transformée en ondelettes.

Nous avons vu au chapitre 2 quelques méthodes de VAD qui se différencient par leurs techniques utilisées. Dans ce mémoire, il est question d'étudier plus en profondeur la transformée en ondelettes (TO) afin d'élaborer une méthode de VAD basée sur cette technique et d'en étudier les performances dans différent contextes. Ce chapitre décrit les principes de base de la transformée en ondelettes, son application dans les VAD, décrit précisément deux méthodes de VAD qui l'utilisent, puis nous terminerons par la description de notre proposition de VAD à base de la TO.

3.1 La transformée en ondelettes

3.1.1 La transformée en ondelettes : généralités

La transformée en ondelettes (TO), contrairement à la transformée de Fourrier, ne se limite pas à une technique d'analyse fréquentielle. En appliquant la transformée en ondelettes à un signal, on peut observer son comportement dans le domaine à la fois des fréquences et du temps. Cette analyse temps-fréquence la mène à appartenir au groupe d'analyse de méthodes multi-échelle tel que la transformée de Fourier à fenêtre glissante et la transformée en cosinus. Le principe de base consiste à convoluer le signal analysé avec une fonction appelée ondelette (*wavelet*). Une ondelette Ψ est une fonction de moyenne nulle [MAL98] :

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0 \tag{3.1}$$

qui peut être dilatée par un paramètre d'échelle s et translatée de u :

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi(\frac{t-u}{s})dt \tag{3.2}$$

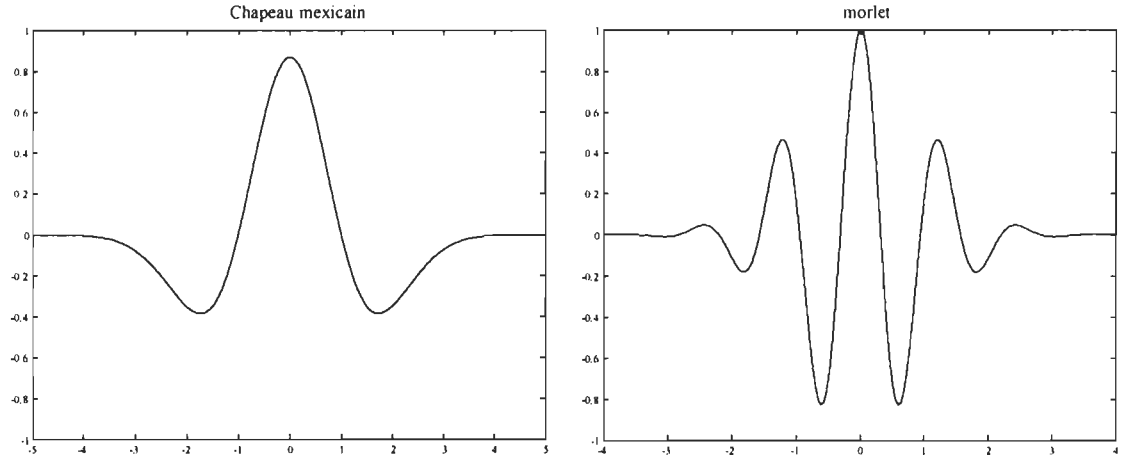
L'ondelette Ψ, appellee ondelette mère, produit une base orthonormée de fonctions appelées ondelettes filles ou plus simplement ondelettes. La transformée en ondelettes de f à une échelle s et une position u est obtenue en corrélant f avec l'ondelette :

$$Wf(u,s) = \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\psi^*(\frac{t-u}{s})dt \tag{3.3}$$

$Wf(u,s)$ est appelé coefficient d'ondelettes à l'échelle s et à la position u de la fonction f.

Le résultat d'une transformation en ondelettes est présenté dans un espace temps-fréquence, avec *u* l'abscisse et l'échelle *s* comme ordonnée [MAL97].

Il existe différentes ondelettes de forme différentes. Le choix d'une ondelette va dépendre du type d'application. La figure 3.1 montre quelques ondelettes en guise d'exemple.



(a) (b)

Figure 3.1 Exemples d'ondelettes connues : (a) Chapeau mexicain (*mexican hat*) (b) Morlet.

3.1.2 La transformée en ondelettes discrète

La transformée discrète en ondelettes (DWT – *Discrete Wavelet Transform*) est l'application numérique de la TO. Son utilisation est populaire, car elle peut s'implémenter facilement sur des circuits numériques (FPGA, DSP). La DWT adopte une technique de fenêtrage (*windowing technique*) qui consiste à travailler le signal morceau par morceau. Le principe général de la DWT consiste à décomposer un signal en plusieurs sous-signaux. En 1989, Mallat [MAL89] découvrait une approche efficace pour implémenter la DWT en utilisant des banques de filtres (*filter banks*).

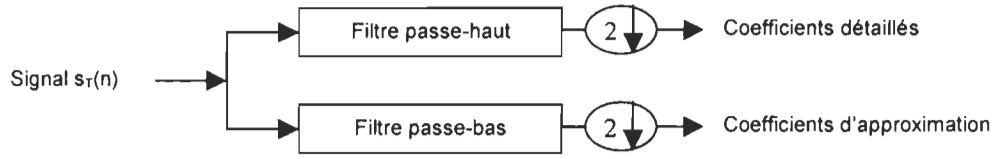


Figure 3.2 Principe de décomposition en ondelettes du signal $s_T(n)$

La figure 3.2 explique le principe. Le signal de base $s_T(n)$ (le signal à traiter) est passé dans un filtre passe-haut et parallèlement dans un filtre passe-bas. Ce signal discret est dyadique c'est-à-dire qu'il est composé de 2^k échantillons où k est un entier. Il est à noter que les coefficients des filtres passe-haut et passe-bas sont identiques. Seul leur ordre est inversé, à savoir, le premier coefficient du filtre passe-bas correspond au dernier coefficient du filtre passe-haut, et ainsi de suite. Les coefficients des filtres sont définis à partir de la nature de l'ondelette mère (voir [MAL97] chapitre 5 pour plus de détails ; aux étapes de simulations du chapitre 4, nous utiliserons la fonction MATLAB du « Wavelet Toolbox » *wpdec(x j, « type ondelette mère »)* qui permet de décomposer le signal dyadique *x* en *j* étages selon l'ondelette mère définie). À la sortie de chaque filtre, une opération de sous-échantillonnage (*downsampling*) par 2 est effectuée. On obtient les coefficients A et D, appelés respectivement coefficients d'approximation (*approximated coefficients*) et coefficients des détails (*detail coefficients*). Les équations correspondantes vont comme suit :

$$a(k) = \sum_{n=1}^{N} g(n-2k)s_T(n) \tag{3.4}$$

$$d(k) = \sum_{n=1}^{N} h(n-2k)s_T(n) \tag{3.5}$$

où g(n) et h(n) correspond au filtre passe-bas et passe-haut, respectivement. En effectuant l'opération présentée à la figure 3.2, on passe du niveau *j* au niveau *j+1*. Le signal de base étant conventionnellement au niveau j=0, les coefficients A et D du niveau 1 s'écrivent A1 et D1. Pour généraliser, Aj et Dj correspondent respectivement aux coefficients d'approximations et de détails pour le niveau j ($0<j<\log_2(N)$, N étant la longueur de $s_T(n)$). À partir des coefficients A_j, on peut obtenir, selon l'opération de la figure 3.2, les coefficients A_{j+1} et D_{j+1}. On poursuit ainsi de suite jusqu'au niveau désiré. La figure 3.3 présente une structure de DWT décomposée en 3 niveaux, obtenue en implémentant ces filtres en cascade.

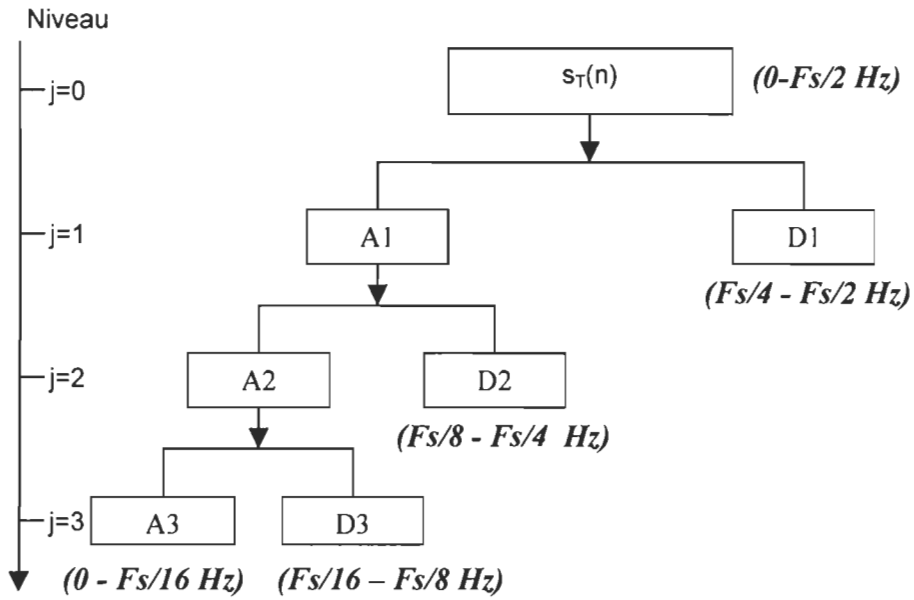


Figure 3.3 Exemple de décomposition en ondelettes

Comme on peut voir sur la figure 3.3, le signal de base est décomposé en deux sous-signaux, qui sont à leur tour décomposés en d'autres sous-signaux jusqu'au nombre d'étages désirés. *Fs* correspond à la fréquence d'échantillonnage du signal de base. D'après

le théorème de Shannon [PRO96], le spectre de fréquence du signal s'étend de 0 à *Fs/2* Hz. Lorsqu'on effectue un filtrage de type passe-bas, on recueille à l'étage suivant deux signaux dont l'intervalle de fréquences se situe entre 0 et *Fs/4* et *Fs/4* et *Fs/2*. Plus on « monte » dans les étages et plus l'intervalle de fréquence diminue. D'après la figure 3.3, on peut déduire que les coefficients correspondant aux fréquences plus basses sont situées du côté gauche de l'arbre, alors que les fréquences plus hautes du côté droit de l'arbre.

Comme il a été mentionné auparavant, le signal traité (signal de base) est découpé en trames de taille égales. Chaque trame est donc traitée successivement. La longueur de la trame doit être dyadique. Sa dimension dépend de l'information qu'on veut recueillir dans le domaine fréquentiel. Plus on désire de l'information précise dans une large gamme de fréquence, plus la trame doit être grande. On aura donc plus d'échantillons par trame. À l'inverse, si on veut une analyse plus axée sur les hautes fréquences, il suffira de prendre moins d'échantillons par trame. Évidemment, il faudra choisir Fs en fonction des hautes fréquences d'intérêts.

Pour résumer, la DWT permet une analyse multi-frequencielle. Son principe de décomposition permet d'aller chercher de l'information dans plusieurs gammes de fréquences. De plus, l'approche par filtrage la rend attrayante pour la simulation sur des plateformes logicielles et pour l'implémentation sur des processeurs numériques.

3.3.3 La transformée en ondelettes par paquets

La transformée en ondelettes par paquets (WPT – *Wavelet Paquet Transform*) se définit en quelque sorte par une forme générale de la transformée en ondelettes. À chaque fois qu'on descend d'un niveau, au lieu de décomposer seulement les coefficients

d'approximation, on décompose également les coefficients de détail. Il en résulte un arbre où chaque sous-signal (ou paquet de coefficients) correspond à une zone de fréquence. Il est évident que l'on peut y recueillir plus d'informations. Un exemple d'arbre de décomposition d'ondelettes par paquet est présenté à la figure 3.4. Les techniques utilisées pour décomposer les signaux sont identiques à celles de la décomposition en ondelettes de la section 3.1.2.

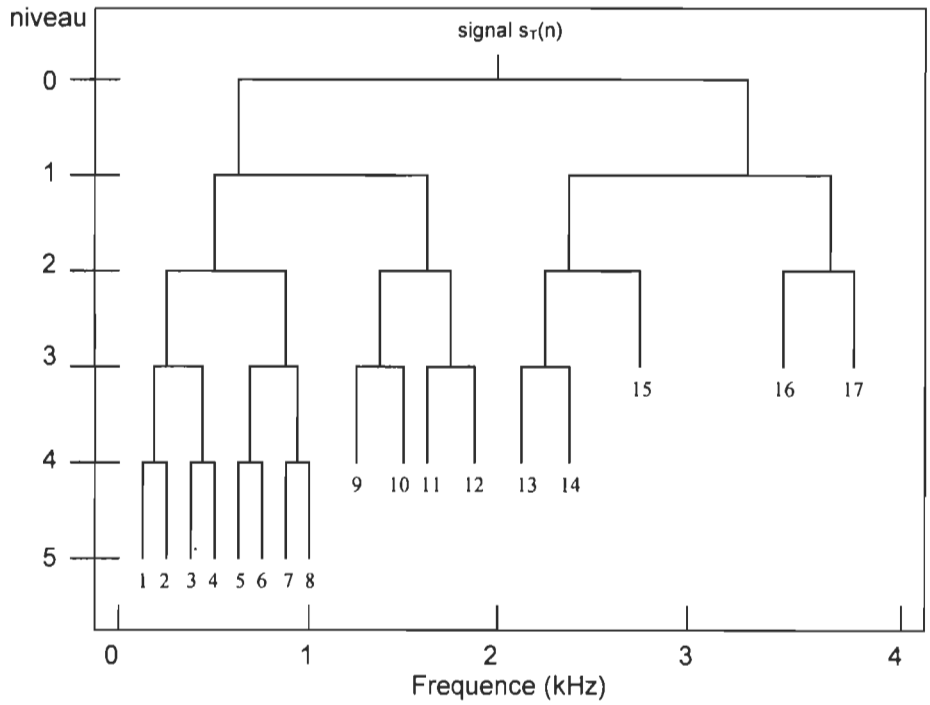


Figure 3.4 Décomposition du signal $s_T(n)$ en 17 sous-signaux par la transformée en ondelettes par paquets.

3.1.4 Synthèse

Nous avons vu que la transformée en ondelettes décompose un signal en un ensemble de sous-signaux. Inversement, il est possible de reconstituer le signal de base initial $s_T(n)$ par plusieurs opérations de filtrage inversé. Ce processus s'appelle la synthèse [MAL97]. L'équation 3.6 montre la forme générale d'une étape de recomposition où a_j et d_j sont respectivement les coefficients d'approximation et de détails obtenus par décomposition (éq. 3.4 et 3.5) du signal a_{j-1}. g_0 *et* g_1 représentent les coefficients des filtres inversés :

$$a_{j-1}(n) = \sum_k \left\{ g_0(2k-n)a_j(k) + g_1(2k-n)d_j(k) \right\} \tag{3.6}$$

La figure 3.5 illustre l'étape de recomposition de l'équation 3.6 :

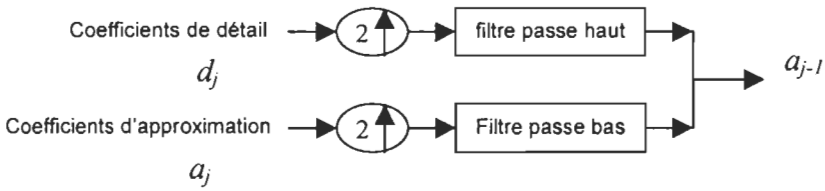


Figure 3.5 Principe de synthèse

Par étapes successives de recomposition, on peut ainsi remonter l'arbre de la figure 3.3 ou de la figure 3.4 à partir des sous-signaux pour retrouver le signal de base.

3.2 VAD basé sur la DWT

On retrouve dans la littérature plusieurs VAD à base de la TO [CHE05], [CHE02], [WUW06], [SHF04], [SHR97]. Cette section décrit deux méthodes récentes de VAD qui utilisent la transformée en ondelettes. La première est basée sur la DWT tel que décrite à la

section 3.1.2. La deuxième méthode utilise la WPT, décrite à la section 3.1.3. Les deux méthodes utilisent des techniques pour extraire de l'information des ondelettes afin d'obtenir un signal appelé VAS (*Voice Activity Shape*). L'amplitude de ce signal augmente dans les zones actives et diminue dans les zones inactives. Elle décrit en quelque sorte les zones d'activités vocales. Au cours de ce mémoire, nous utiliserons l'acronyme WVAD (*Wavelet Voice Activity Detector*) pour désigner un VAD basé sur la transformée en ondelettes.

3.2.1 VAD utilisant le Teager Energy Operator (TEO) et la fonction d'auto-corrélation

La périodicité est une des propriétés des signaux de voix. Les sons de voix vont contenir plus de périodicité que les sons dépourvus de voix tels les bruits aléatoires [RAB93]. En tenant compte de cette propriété, l'idée principale de la méthode de Wu et Wang [WUW06] est d'analyser le taux de périodicité de certains sous-signaux issu de la décomposition en ondelettes pour déterminer avec précision l'activité vocale. Comme les intervalles de fréquences fondamentales de la voix sont situés dans les basses fréquences ([85-155Hz] pour les hommes, [165-255Hz] pour les femmes), il faut extraire avec une bonne résolution les sous-signaux correspondant aux basses fréquences. Pour cela, uniquement les sous-signaux de basses fréquences sont extraits. La figure 3.3 présentée à la section 3.1 présente la structure de décomposition en trois niveaux de l'arbre utilisé dans la méthode. Une décomposition en trois niveaux décompose le signal de base en quatre sous-signaux, où chaque sous-signal correspond à un intervalle de fréquence. La méthode prétend que cette structure de décomposition en quatre sous-signaux peut être utilisée pour obtenir la périodicité la plus significative dans le domaine des sous-signaux. Comme pour

la grande majorité des VAD, le signal de base est découpé en trames. La décomposition présentée à la figure 3.3 est ainsi effectuée à chaque trame. À chaque trame, il résulte donc quatre sous-signaux de longueur différente appelés $w_m^j(k)$ où j est le niveau de décomposition, *m (1 à 4)* l'indice arbitraire du sous-signal et k l'indice de temps discret:

$$w_m^j(k) = DWT\{s_T(n),3\}\big|n=1,2,...N\big|m=1,2,3,4\big|k=1...\frac{N}{2^j}\big|j=1,2,3 \qquad (3.7)$$

avec N la longueur de la trame et $DWT\{s_T(n),3\}$ dénote l'opération de décomposition en ondelettes du signal de base $s_T(n)$ en 3 niveaux.

Le choix des filtres d'ondelettes ou ondelette mère a une influence sur la sélection des fréquences ainsi que sur la résolution des ondelettes. Dans cette méthode, l'ondelette mère proposée par Daubechies est utilisée. Cette dernière préserve la sélectivité des fréquences lorsque le niveau de décomposition augmente. Ceci est dû à leur propriété régulière [SIN93]. Il existe plusieurs ordres pour les filtres de Daubechies. Plus l'ordre du filtre est grand, plus la sélectivité des fréquences pour un sous-signal sera grande, mais plus la complexité de calcul sera plus élevée. La figure 3.6 illustre deux séries de coefficients de Daubechies pour deux ordres (p) différents. Ici, l'ordre choisi est de 10 car il correspond à un bon rapport entre complexité et précision.

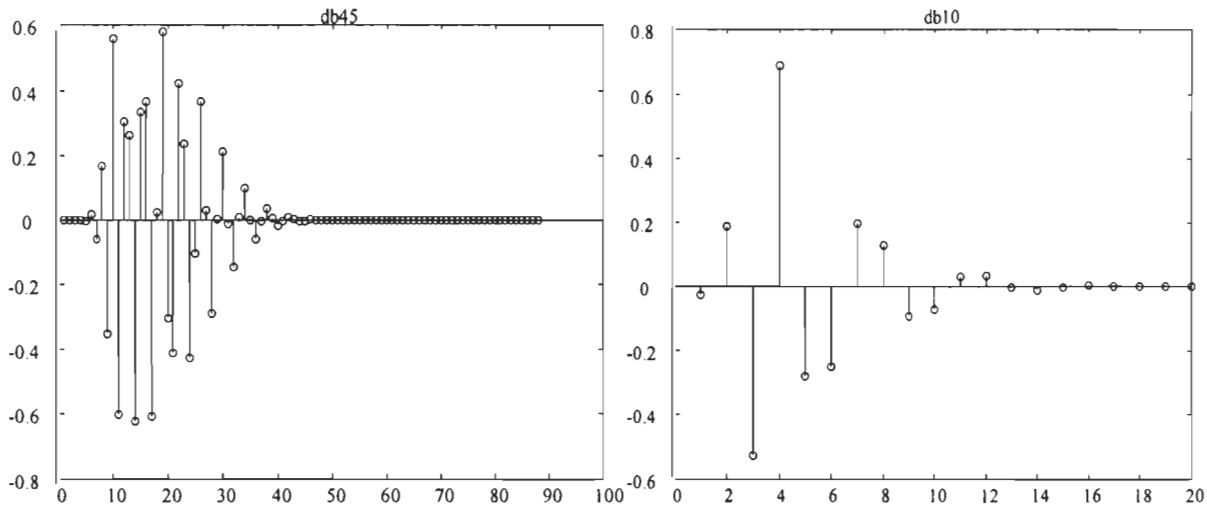


Figure 3.6 Coefficients d'ondelettes Daubechies d'ordre 10 (a) et 45 (b).

L'étape suivante est de déterminer la périodicité contenue dans chaque sous-bande. Pour cela, la méthode utilise une série de calculs successifs représentés à la figure 3.7.

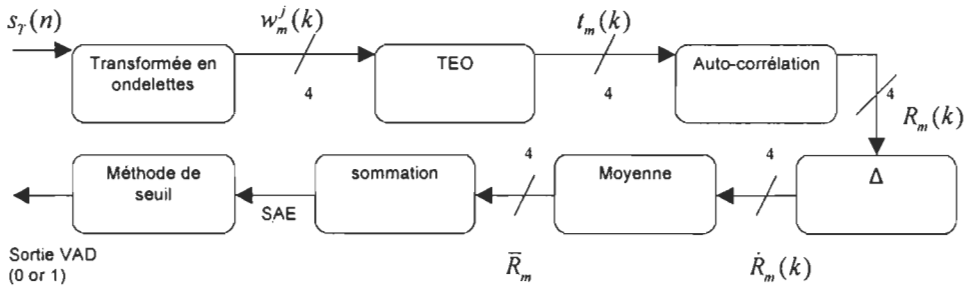


Figure 3.7 Structure générale de la méthode de Wu et Wang [WUW06].

En premier lieu, le *Teager Energy Operator* (TEO), est une fonction qui permet de mieux différencier la voix du bruit. En effet, le TEO accentue les composantes de la voix tel que la périodicité et rejette les composantes de bruit [JAB99], [KAI90]. Dans le domaine discret, le TEO est défini par la fonction $\psi(\bullet)$ qui suit :

$$\psi[x(n)] = x(n)^2 - x(n+1)x(n-1) \tag{3.8}$$

où *x(n)* est un signal discret. Contrairement à certaines fonctions dans le domaine fréquentiel utilisées pour différencier le bruit de la voix, cette fonction s'implémente facilement dans le domaine temporel. On obtient donc quatre (4) signaux issus de chaque opération de TEO :

$$t_m(k) = \Psi\left[w_m^j(k)\right] | m = 1...4 \tag{3.9}$$

Pour encore mieux déterminer l'intensité de la périodicité, on utilise la fonction d'autocorrélation [PRO96] :

$$r(l) = \sum_{n=-\infty}^{\infty} x(n)x(n-l) \tag{3.10}$$

La fonction d'autocorrélation est utilisée dans plusieurs domaines pour déterminer si un signal contient des périodicités. Le résultat *r(l)* est un vecteur de longueur *2N-1, N* étant la longueur du signal discret *x(n)*. On applique la fonction d'autocorrélation à chaque sous-signal t_m, et on obtient quatre (4) signaux (correspondant à chaque sous-signal) nommés SSACF (*Subband Signal Auto-Correlation Function*) :

$$R_k(k) = R[t_m(k)] \tag{3.11}$$

où *R* dénote l'opération d'autocorrélation.

Par la suite, on applique la méthode de la moyenne des Deltas (*Mean-Delta method*). Une mesure similaire à une évaluation de spectre de delta (*delta cepstrum evaluation*) est utilisée pour estimer l'intensité périodique de chaque SSACF. Cela consiste à appliquer la fonction appelée « Delta Subband Signal Auto-Correlation Function (DSSACF) » :

$$\dot{x}_M(n) = \frac{\sum_{l=-M}^{M} l \cdot x(n+M)}{\sum_{l=-M}^{M} l^2} \tag{3.12}$$

où $\dot{x}_M$ est appelé le DSSACF sur un voisinage de M échantillons. M est une valeur fixé initialement (à titre d'exemple M=8 dans [WUW06]). Appliqué à $R_m(k)$ on obtient

$$\dot{R}_m(k) = \Delta\left[R_m(k)\right] \tag{3.13}$$

où $\Delta[\bullet]$ représente la fonction DSSACF

Finalement, on calcule la moyenne pour chaque signal $\dot{R}_m(k)$:

$$\overline{R}_m = \frac{1}{Nb}\sum_{k=0}^{Nb-1}\left|\dot{R}_m(k)\right| \tag{3.14}$$

où Nb est la longueur du sous-signal $\dot{R}_m(k)$. On obtient ainsi quatre valeurs pour chaque décomposition. Les quatre valeurs sont ensuite sommées :

$$SAE = \sum_{m=1}^{4} \overline{R}_m \tag{3.15}$$

Le résultat final est un scalaire appelé SAE (*Speech Activity Envelope*). Pour chaque trame, et donc à chaque décomposition, une valeur de SAE est calculée. La courbe du SAE en fonction du temps a la caractéristique d'augmenter dans les périodes d'activité de voix et de baisser dans les zones de non-activité. À partir de cette courbe, on applique une méthode de seuil de décision permettant de détecter les portions qui contiennent de la voix (cf. section 3.4).

3.2.2 VAD basé sur la transformée en ondelettes par paquet

La méthode de VAD développée par Chen et Wang [CHE02] utilise la transformée en ondelettes par paquet. Un schéma de la méthode est présenté à la figure 3.8.

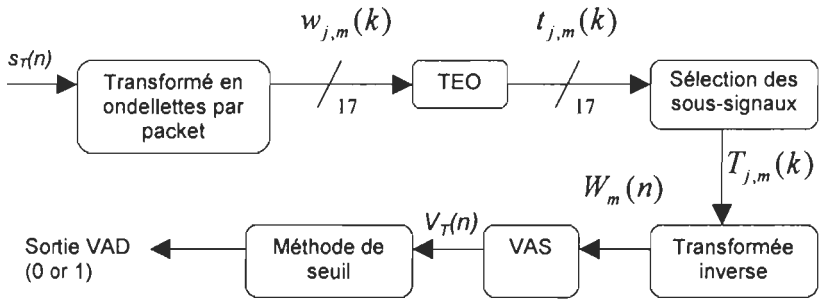


Figure 3.8 Structure générale de la méthode de Chen et Wang [CHE02].

En premier lieu, le signal d'entrée $s_T(n)$ est décomposé à l'aide de la transformée en ondelettes par paquet. Contrairement à la méthode de Wu et de Wang [WUW06] qui décompose le signal en 4 sous-signaux, le signal est ici décomposé en 17 sous-signaux avec une décomposition en 5 niveaux par paquet. L'arbre de décomposition est présenté à la figure 3.4 à la section 3.3.3. Comme on peut voir, on a 8 sous-signaux au niveau 5, 6 sous-signaux au niveau 4 et 3 sous-signaux au niveau 3, pour un total de 17 sous-signaux. Le choix des sous-signaux est effectué en fonction des fréquences associées à chaque sous-signal. Les niveaux de fréquences ont été choisis d'après une étude portée sur les composantes spectrales de la voix ainsi que sur la fréquence d'échantillonnage *Fs* et sur la longueur de la trame [RAB93]. Comme la méthode de Wu et Wang et pour les mêmes raisons expliquées à la section 3.2.1, un filtre de type Daubechies d'ordre 10 est utilisé.

Après avoir décomposé le signal $s_T(n)$ en 17 sous-signaux selon les équations (3.4) et (3.5), chaque sous-signal est passé par la fonction TEO (éq. 3.8). Comme pour la méthode précédente, le TEO est utilisé pour mieux détecter la périodicité :

$$t_{j,m}(k) = \Psi\left[w_{j,m}(k)\right] \Big| \; k = 1, 2, \ldots \frac{N}{2^j} \tag{3.16}$$

où $t_{j,m}(k)$ représente les sous-signaux avec j le niveau de décomposition ($3 \leq j \leq 5$), m l'indice du sous-signal ($1 \leq m \leq 17$), N la longueur de la trame, $\Psi(\bullet)$ la fonction TEO et $t_{j,m}(k)$ le sous-signal obtenu.

L'étape suivante consiste à rejeter les sous-signaux non pertinents pour la détection d'activité vocale. Pour cela, la variance de chaque sous-signal est comparée à un seuil λ dépendant du niveau de décomposition. Si la variance est supérieure au seuil, on garde le sous-signal, sinon on le rejette en mettant tous ses coefficients à zéro :

$$T_{j,m}(k) = \begin{cases} t_{j,m}(k), & si \;\; \text{var}\left\{t_{j,m}(k)\right\} \geq \lambda_j \\ 0 & si \;\; \text{var}\left\{t_{j,m}(k)\right\} < \lambda_j \end{cases} \tag{3.17}$$

où $\text{var}\left\{t_{j,m}(k)\right\}$ est la variance de $t_{j,m}(k)$ et λ_j est défini par :

$$\lambda_j = \sigma_j \sqrt{2\log(N)} \tag{3.18}$$

avec σ_j l'écart type de l'ensemble des sous signaux de niveau j.

L'étape suivante consiste à recomposer le signal de base initial à partir des sous-signaux sélectionnés à l'étape précédente

$$V_T(n) = \sum_{m=1}^{17} W_m(n) \tag{3.19}$$

où $W_m(n)$ est la transformée inverse de $T_{j,m}(k)$. $W_m(n)$ est obtenu selon l'équation (3.6). Chaque signal $T_{j,m}(k)$ sélectionné est reconstitué par étapes de recomposition (cf. section 3.1.4, *synthèse*) jusqu'au niveau j=0. En sommant tous ces signaux, on obtient V_T dont l'allure dépend de l'activité vocale. L'ensemble de tous les V_T obtenus à chaque trame représente la forme de l'activité vocale (*VAS -Voice Activity Shape*). Dans les zones actives, le VAS est d'amplitude plus élevée que dans les régions inactives. L'étape finale consiste à détecter les régions de voix à partir du VAS. Pour cela une méthode de seuil adaptative est proposée par Chen et Wang [CHE02] appelé AWT (*Adaptive Weighted Threshold*). Il s'agit de déterminer un seuil à partir des informations sur des calculs de moyennes et de seuils par rapport aux valeurs précédentes du VAS. Cependant, nous n'avons pas pu mettre en application ce type de procédure par manque d'information peu explicite dans l'article nous permettant d'assurer son fonctionnement.

3.3 Méthode VAD proposée

La méthode proposée est une variante de la méthode développée par Chen et Wang [CHE02]. On utilise la transformée en ondelettes par paquets en décomposant le signal en 17 sous-signaux. Après une opération de TEO à chaque sous-signal, la somme des variances est calculée pour obtenir un VAS. À partir du VAS, une simple méthode de seuil est appliquée pour déterminer les portions d'activité vocale. La méthode est présentée à la figure 3.9.

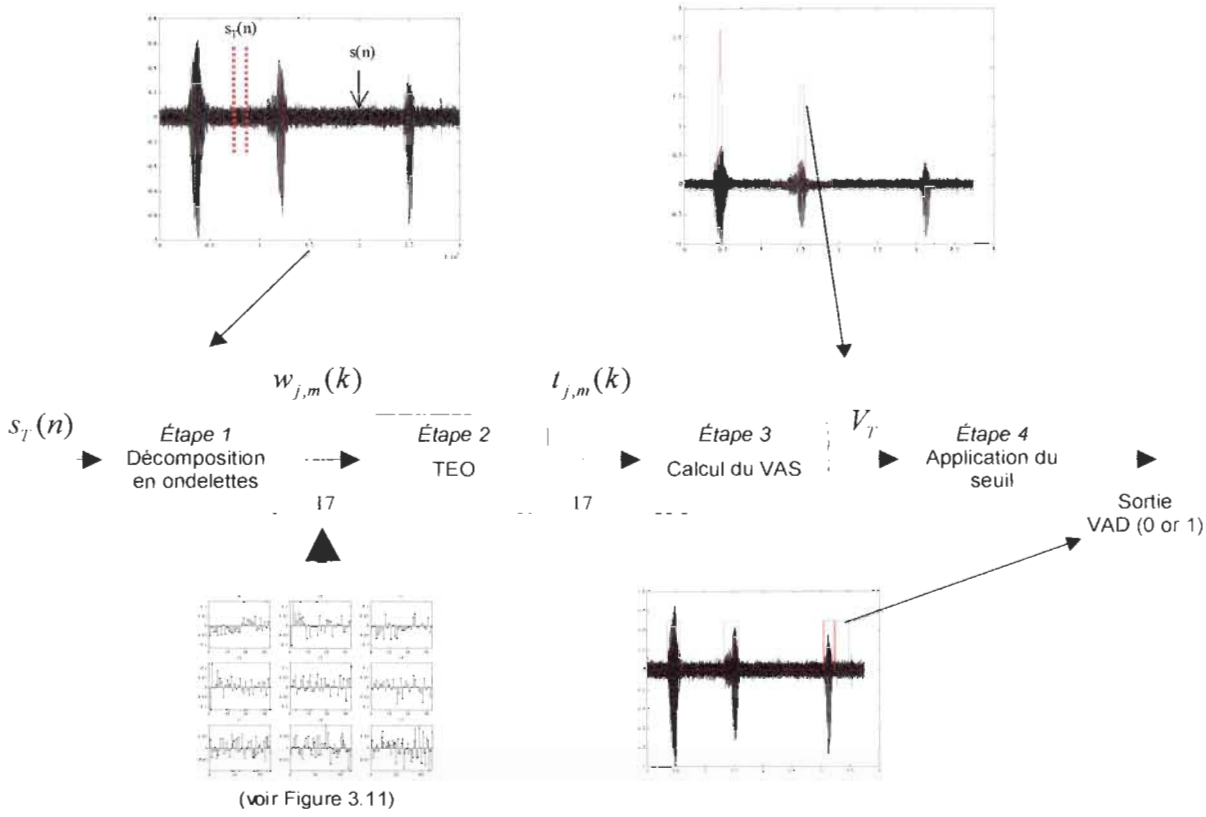


Figure 3.9 Structure générale de la méthode proposée.

En observant l'allure des coefficients d'ondelettes obtenus pour des signaux de voix et des signaux de bruits, nous avons pu constater une différence dans l'amplitude de certaines ondelettes ainsi que dans leur variations. La figure 3.10 montre les 17 sous-signaux obtenus selon [CHE02] pour un signal entaché de bruit blanc (SNR 15dB) à deux trames différentes : une trame active et une trame inactive.

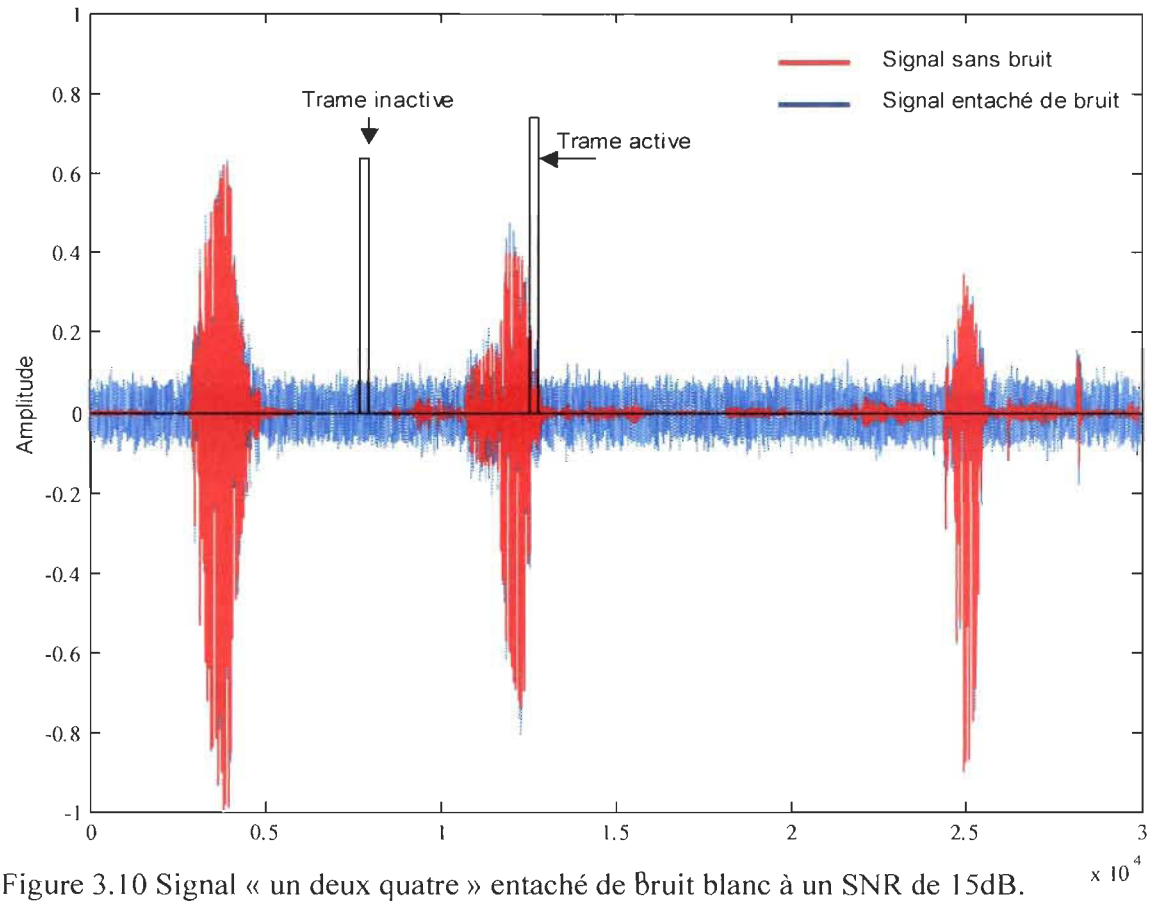


Figure 3.10 Signal « un deux quatre » entaché de bruit blanc à un SNR de 15dB.

La figure 3.10 montre les trames utilisés pour obtenir les 17 séries de coefficients d'ondelettes présentés à la figure 3.11. Lorsqu'on observe ces coefficients, on s'aperçoit que l'amplitude et la variation des coefficients pour certaines séries de coefficients sont plus élevés. Cette observation explique notre calcul de variance de l'étape 3 de notre méthode.

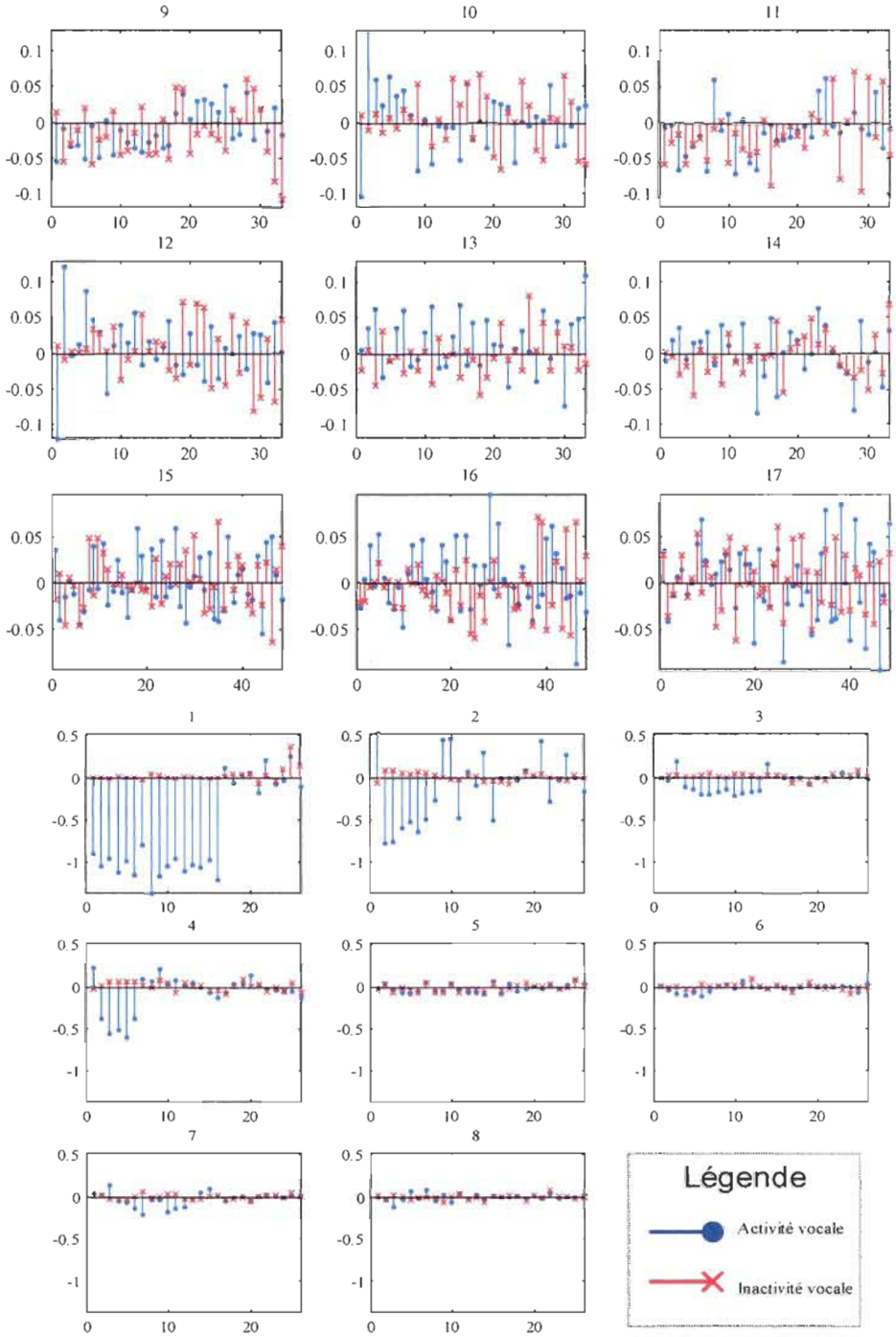


Figure 3.11 Coefficients d'ondelettes calculés selon [CHE02] pour a) une trame active et b) une trame inactive.

3.3.1 Étape 1 : décomposition en ondelettes

De la même manière que la méthode de Chen et Wang [CHE02], on décompose le signal en 17 sous-signaux dont l'arbre est présenté à la figure 3.4. Comme il a été expliqué à la section 3.2.2, ce type d'arbre de décomposition permet de bien exploiter la détection d'activité vocale. Les formules utilisées pour obtenir les 17 sous-signaux sont présentés à la section 3.1 équations (3.4) et (3.5).

$$w_{j,m}(k) = DWT\{s_T(n)\} \tag{3.20}$$

où $DWT\{\bullet\}$ dénote ici l'opération de décomposition par paquet d'ondelettes selon l'arbre de la figure 3.4, $s_T(n)$ le signal d'entrée, j est le niveau de décomposition, m l'indice de sous-signal. De la même manière que les deux méthodes présentées à la section 3.2, la nature des filtres est Daubechies d'ordre 10. Ce choix s'explique par la précision des coefficients obtenus dans le domaine fréquentiel, tout en gardant une complexité acceptable.

3.3.2 Étape 2 : TEO

Comme pour les deux méthodes d'ondelettes présentées précédemment, l'opération de TEO permet de mieux dissocier le bruit de la voix.

$$t_{j,m}(k) = TEO\{w_{j,m}(k)\} \tag{3.21}$$

où $TEO\{\bullet\}$ représente l'opération de TEO selon l'équation (3.8).

3.3.3 Étape 3 : calcul du VAS

Notre variante proposée pour notre VAD comparativement à la méthode de Chen et Wang [CHE05] est basée sur le calcul des variances de $t_{j,m}(k)$ pour calculer le VAS. On obtient ainsi 17 valeurs, chaque valeur étant associée à un sous-signal $t_{j,m}(k)$. Toutes les variances sont ensuite sommées pour obtenir V_T :

$$V_T = \sum_{m=1}^{17} \text{var}\left\{t_{j,m}(k)\right\} \tag{3.22}$$

3.3.4 Étape 4 : seuil

L'étape finale consiste à appliquer une méthode de seuil de décision binaire sur le VAS. Il s'agit d'une étape critique dans la performance du VAD puisqu'il s'agit d'une prise finale de décision sur l'activité ou non activité vocale de la trame analysée. Cette étape est donc décrite à la section 3.4 traitant d'une proposition de méthode de seuil applicable à notre méthode proposée.

3.4 Méthode de seuil de décision

En parcourant les deux méthodes étudiées à la section 3.2 ainsi que notre méthode proposée, on finit par obtenir une courbe qui est supposée suivre l'allure de la voix. Cette courbe est appelé SAE (méthode de Wu et Wang) ou encore VAS (méthodes de Chen et Wang et proposée). À partir de cette courbe, il s'agit de trouver une méthode de seuil qui va prendre la décision '1' ou '0' (activité ou non activité vocale),

$$VAD(n) = \begin{cases} 1 & \text{si } VAS(n) > \lambda_V \\ 0 & \text{si } VAS(n) < \lambda_V \end{cases} \tag{3.23}$$

où λ_V est le seuil de décision soit une constante à ajuster pour maximiser la performance du VAD.

L'étape de seuil de décision est très importante car elle a une grande influence sur la performance du VAD. Même si le VAS obtenu trace fidèlement les zones d'activité vocale, une mauvaise méthode de seuil n'exploitera pas adéquatement la qualité de la méthode. Pour cela, chaque VAD décrit à la section 3.2 a développé sa propre méthode de seuil. Cependant, le manque d'information et la description peu explicite des méthodes proposées par les auteurs nous a poussé à développer notre propre méthode de seuil. Celle-ci consiste à trouver le seuil qui minimise simultanément l'erreur de mauvaise acceptation et l'erreur de mauvais rejet (cf. section 4.2 pour plus d'explications).

La figure 3.10 montre l'influence du seuil sur l'erreur. La définition de l'erreur de mauvais rejet et de l'erreur de mauvaise acceptation est expliquée en détail au chapitre 4, à la section 4.2.1. Pour le moment, simplifions nous à dire que plus le seuil augmente, moins le VAD va laisser passer de la voix, ce qui augmente l'erreur de mauvaise acceptation et diminue de l'erreur de mauvais rejet. Inversement, plus le seuil diminue, plus le VAD va considérer comme actives des trames non-actives, ce qui diminue l'erreur de mauvaise acceptation et augmente l'erreur de mauvais rejet.

Dans un compromis de clarté d'écoute, nous avons constaté que le seuil recherché est celui pour lequel les deux courbes se croisent, c'est-à-dire pour lequel l'erreur de mauvaise acceptation et l'erreur de mauvais rejet ont la même valeur. Pour cette méthode de seuil, la

connaissance du VAD idéal est nécessaire. Le seuil final est calculé par itérations successives. Une fois le seuil déterminé, on compare à chaque échantillon le VAS avec le seuil. Si le VAS est supérieur au seuil, la sortie du VAD est 1, si le VAS est inférieur au seuil la sortie du VAD vaut 0. Cette méthode ne s'apprête pas pour un VAD commercial, puisqu'en pratique on ne connaît pas le VAD idéal. Cependant, elle va permettre de comparer équitablement les méthodes de VAD présentées à la section 3.2 avec notre VAD proposé (cf. section 3.3).

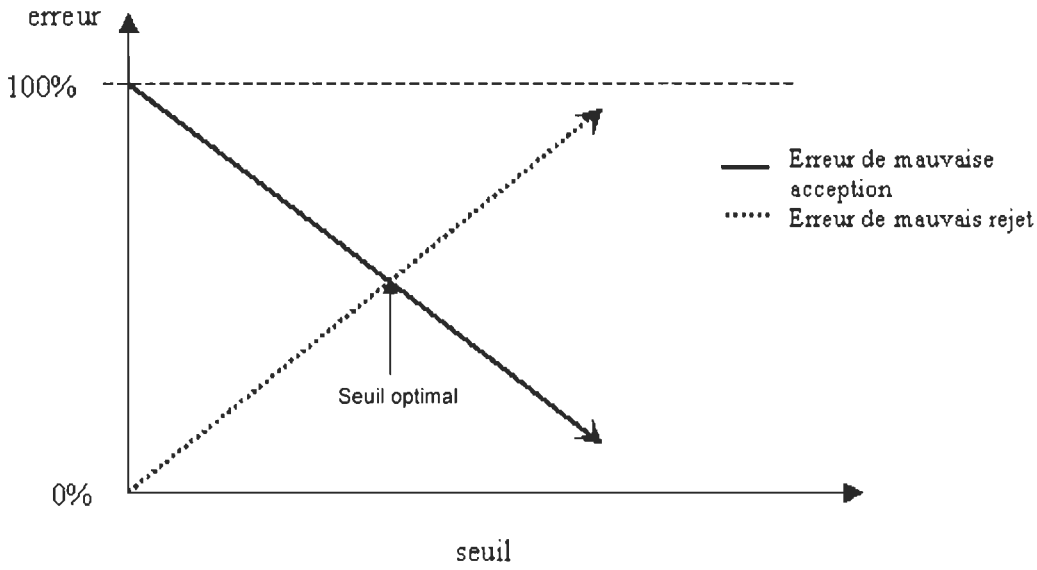


Figure 3.12 Courbes des erreurs en fonction du seuil.

3.5 Analyse de la complexité

La complexité peut jouer un rôle déterminant dans la comparaison de méthodes de VAD. Un VAD très efficace mais qui demande beaucoup de calculs sera difficile à implémenter sur un composant électronique de type DSP ou FPGA. Dans la conception d'un VAD, il s'agit donc de faire attention à l'aspect de la complexité et la régularité de

calcul. Cette section vise à analyser la complexité entre les trois méthodes de VAD basé sur la TO et présentées :

- WuVAD : méthode de Wu et Wang [WUW06] (section 3.2),
- ChenVAD: méthode de Chen et Wang [CHE02] (section 3.2), et
- Propose VAD: méthode proposée (section 3.3).

L'étape de calcul de seuil n'est pas tenue en compte dans cette section. La complexité est calculée jusqu'à la fin de l'étape de calcul du VAS (ou SAE, pour méthode de Wu-Wang).

Afin de mieux comparer la complexité de calcul arithmétique entre chaque méthode, nous allons établir le nombre grossier des additions et des multiplications nécessaires à chacune des méthodes. Ces nombres permettront de définir les ressources nécessaires pour une implémentation sur silicium (DSP, FPGA). Pour commencer, le tableau 3.1 résume l'ensemble des étapes de calcul pour chaque méthode.

Tableau 3.1 Étapes de calcul des trois méthodes de VAD

Wu et Wang [WUW05]	**Chen et Wang [CHE02]**	**Méthode proposée**
Décomposition en ondelettes en trois niveaux (éq. 3.4 et 3.5)	Décomposition en en ondelettes en 17 sous signaux (éq. 3.4 et 3.5)	Décomposition en en ondelettes en 17 sous signaux (éq. 3.4 et 3.5)
Calcul du TEO (éq. 3.8)	Calcul du TEO (éq. 3.8)	Calcul du TEO (éq. 3.8)
Auto-correlation (éq. 3.10)	Calcul de variance	Calcul de variance
Méthode de la moyenne des deltas (éq. 3.12)	Calcul de 3 seuils et comparaison (éq.3.17 et 3.18)	Sommation (3.22)
Calcul de moyenne (éq.3.14)	Synthèse (éq. 3.6 et 3.19)	-
Sommation (éq.3.15)	-	-

Les étapes de calcul du tableau 3.1 sont décrite à l'annexe C. Pour chaque étape de calcul, un nombre d'additions et de multiplications est déterminé.

Le tableau 3.2 récapitule l'ensemble des formules décrites dans l'annexe C et donne la complexité de chaque méthode en termes de multiplications et d'additions. Les opérations sont données pour une seule trame de longueur N.

Tableau 3.2 Résumé de la complexité de calcul des trois méthodes en terme d'opérations d'addition, de multiplication et division

Opération	Wu et Wang	Chen et Wang	Méthode proposée
Décomposition	$\frac{10}{4}N(d\otimes+(d-1)\oplus)$	$\frac{59}{8}\cdot N(d\otimes+(d-1)\oplus)$	$\frac{59}{8}\cdot N(d\otimes+(d-1)\oplus)$
TEO	$(N-6)\cdot(2\cdot\otimes+1\cdot\oplus)$	$(N-34)\cdot(2\cdot\otimes+1\cdot\oplus)$	$(N-34)\cdot(2\cdot\otimes+1\cdot\oplus)$
Auto-corrélation	$\left(\frac{11}{16}N^2-\frac{N}{16}\right)(\oplus+\otimes)$		
Méthode de la moyenne des deltas	$(2N-3)\cdot[M\oplus+M\otimes]$ $+1\cdot\otimes$		
Variance		$(3N-2)\oplus$ $+34\cdot divisions$	$(3N-2)\oplus$ $+34\cdot divisions$
Calcul de 3 seuils		$3\cdot\begin{bmatrix}(3N-2)\oplus\\ +2divisons\end{bmatrix}$	
Moyenne	$(2N-6)\oplus+4\cdot division$		
sommation	$4\cdot\oplus$		$17\cdot\oplus$
recomposition		$\frac{31}{8}\cdot N(d\otimes+(d-1)\oplus)$	

Nombre de $\otimes$	$\frac{11}{16}N^2 + \left(\frac{10}{4}d + 2 - \frac{1}{16} + 2M\right)N - 12 - 3M + 1$	$\left(\frac{90}{8}d + 2\right)N - 12$	$\left(\frac{59}{8}d + 2\right)N - 12$
Nombre de $\oplus$	$\frac{11}{10}N^2 + \left(\frac{10}{4}d + \frac{7}{16} + 2M + 2\right)N - 3M - 8$	$\left(\frac{90}{8}(d-1) + 13\right)N - 14$	$\left(\frac{59}{8}(d-1) + 4\right)N + 9$
Nombre de divisions	4	40	34

Notre méthode proposée et celle de Chen et Wang se comparent bien lorsqu'on regarde le nombre d'opérations pour chacune d'elles. On n'a pas besoin de faire de démonstration mathématique pour voir que la méthode de Chen et Wang demande plus de multiplications, d'additions et de divisions que notre méthode proposée, et ce quel que soit les valeurs de N et d.

La comparaison de la méthode de Wu et Wang avec les deux autres méthodes est moins évidente dû à la forme différente des équations et aux nombre de variables (N, d, et M). Pour évaluer la complexité, le tableau 3.3 présente les résultats en faisant varier le nombre d'échantillons par trame N, pour d=10 (ordre des filtres) et M=5. Ces valeurs sont celles utilisées pour les simulations des méthodes [WUW06, CHE02].

Tableau 3.3 Comparaison de complexité pour les trois méthodes pour d=10 (ordre des filtres) et M=5

Opération	Wu et wang			Chen et Wang			Méthode proposée		
	N=256	N=512	N=1024	N=256	N=512	N=1024	N=256	N=512	N=1024
Nombre de $\otimes$	54 486	199 110	758 694	29 244	58 556	117 180	19 324	38 716	77 500
Nombre de $\oplus$	$8.2 \cdot 10^4$	$3{,}1 \cdot 10^5$	$1.2 \cdot 10^6$	$2.9 \cdot 10^4$	$5.8 \cdot 10^4$	$1.2 \cdot 10^5$	$1.8 \cdot 10^4$	$3.6 \cdot 10^4$	$7.2 \cdot 10^4$
Nombre divisions	4			40			34		

D'après le tableau 3.3, on peut voir que les méthodes de Wu et Wang et de Chen et Wang s'équivalent en complexité pour N=256. Lorsque N augmente, la méthode de Wu et Wang devient la méthode la plus complexe.

Le tableau 3.3 vient appuyer l'hypothèse que notre méthode est moins complexe que celle de Chen et Wang. On peut également observer que notre méthode est moins complexe que celle de Wu et Wang. Les nombres de multiplications et d'additions pour notre méthode proposée sont toujours les plus bas quel que soit N. Par rapport à [CHE02], le nombre de divisions pour notre méthode est plus élevé, mais la différence de 30 opérations de divisions entre les deux méthodes est négligeable devant la différence de nombre d'opérations de multiplications et d'additions.

Chapitre 4
Résultats de simulations

Dans ce chapitre, les performances de certains VAD étudiés dans les chapitres précédents sont évaluées. Deux VAD à base de la transformée en ondelettes sont ici mis en comparaison, soit le VAD de Wu et Wang [WUW06] (section 3.2.1) et le VAD proposé (section 3.3). Le VAD de Wu et Wang a été choisi comme VAD de comparaison car il est le plus récent algorithme qui traite sur la détection d'activité vocale basée sur la TO. Les VAD sont comparés à un VAD traditionnel, soit le G729 Annexe B [UIT96] développé par l'Union Internationale des télécommunications (UIT), et utilisé dans la téléphonie mobile. Le G729 Annexe B est utilisé dans la majorité des études de VAD à des fins de comparaison de performance.

La section 4.1 expose la banque des données vocales utilisée pour l'évaluation des VAD étudiés. La section 4.2 se penche sur les éléments de mesure quantitative qui vont permettre de comparer la performance des VAD. Dans la section 4.3, les résultats obtenus seront présentés. Une synthèse des résultats est présentée à la section 4.4.

4.1 Base de données

4.1.1 Fichiers audio

Afin d'évaluer les performances des VAD, on utilise une base de données de fichiers audio qui contient des voix d'hommes et de femmes parlant français ou anglais. Une partie des fichiers proviennent de la base de données AURORA 2 [AUR00]. Cette base de données est très populaire dans le domaine du traitement du signal vocal. L'autre partie des fichiers constitue des sons de voix que nous avons nous même enregistrées. Pour simuler des conversations dans des environnements bruyants, nous avons également utilisé des fichiers audio correspondant à du bruit, provenant de la base de données NOISEX-92 [NSX92].

Pour tous les fichiers de la base de données, la fréquence d'échantillonnage Fs est de 8000 Hz et la résolution 16 bits. Afin de suivre les standards de communications. Le pourcentage d'activité vocale pour chaque fichier audio est situé dans un intervalle de 40% à 60%, ce qui représente le taux d'une conversation standard.

Concernant les sons qui nous avons enregistrés, le microphone a toujours été placé à proximité de la bouche, afin de simuler un téléphone cellulaire ou autre appareil de communication où le microphone est placé près des lèvres. Les phrases ont été prononcées de manière naturelle dans un français international. Les sons ont été enregistré à partir d'un ordinateur, à l'aide du logiciel Sound Forge [SFS10].

4.1.2 Ajout de bruit

Pour analyser l'influence du bruit sur la performance des VAD, du bruit est ajouté au signal source de voix propre, s_{clean}, (CSS - *Clean Speech Signal*):

$$s(n) = s_{clean}(n) + \eta(n) \tag{4.1}$$

où $s_{clean}(n)$ représente le signal de voix 'sans bruit', η(n) le bruit et s(n) le signal vocal entaché de bruit qui va être traité par le VAD.

Les VAD sont testés à des niveaux de bruit différents. Pour mesurer le niveau de bruit, on utilise la formule du rapport signal sur bruit (SNR – *Signal to Noise Ratio*) qui donne une mesure en décibel (dB) [PRO96] :

$$SNR = \log 10\left(\frac{\|s\|_2^2}{\|\eta\|_2^2}\right) \tag{4.2}$$

où $\|\bullet\|_2^2$ représente la norme en base 2 du signal. Le signal du bruit est atténué ou amplifié selon le SNR désiré.

Compte tenu de l'infinité de scénario bruit que l'on peut rencontrer, nous nous limiterons dans notre étude aux trois bruits suivants :

- bruit blanc gaussien (AWGN - *Additive White Gaussian Noise*): bruit obtenu artificiellement à l'aide de Matlab® et dont les valeurs suivent une distribution Gaussienne de moyenne nulle.
- bruit de "machine" : bruit réel enregistré dans une mine souterraine. Le bruit est celui d'une machine de style 'foreuse', provoquant un bruit répétitif et très agressif.

- bruit de "gare" : son enregistré dans une gare ferroviaire. Ce bruit contient des personnes qui parlent et qui marchent, avec plusieurs bruits ambiants différents en arrière fond.

Le choix de ces trois bruits s'explique par leur nature différente, ce qui permettra une meilleure comparaison de la performance des VAD étudiés. Afin de mieux démontrer les différences entre ces bruits, des spectrogrammes réalisés avec l'instruction *fft()* de MATLAB© sont présentés à l'annexe B. Le bruit AWGN, bien qu'on ne le retrouve dans aucun environnement naturel, est souvent utilisé dans la littérature comme bruit de référence dans les évaluations des VAD. Le bruit de "machine" et le bruit de "gare" sont des bruits communs qui permettent de recréer des environnements dans lesquels un téléphone cellulaire ou autre appareil de communication pourrait se retrouver.

4.1.3 Système d'écoute non linéaire

Dans notre étude, nous portons une attention particulière aux non-linéarités qui peuvent subvenir dans certains systèmes d'écoutes. Les systèmes de communications spatiales par exemple peuvent engendrer des non linéarités au signal appliqué à l'entrée du VAD. Pour simuler certaines non linéarités rencontrées dans ces systèmes d'écoute, nous avons fait passer le signal $s_{clean}(n)$ de l'équation (4.1) dans une fonction synthétique non linéaire, choisie arbitrairement (figure 4.1):

$$F_{nl}(n) = 0.5x(n) - 0.25x(n-1)^3 \tag{4.3}$$

où *x(n)* est un signal discret quelconque. Nous constatons que la non linéarité est appliquée sur le signal $s_{clean}(n)$ résultant du signal source de voix sans bruit ajouté. Ce modèle très simplifié, correspond par exemple à l'utilisation d'un microphone de piètre qualité impliquant une non-linéarité des signaux captés.

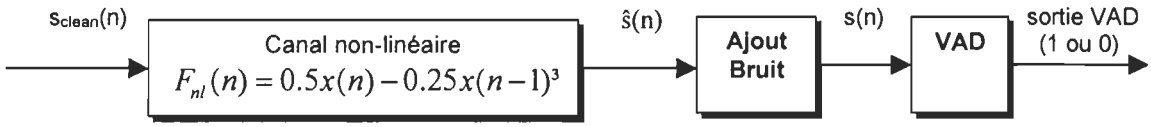


Figure 4.1 Schéma fonctionnel d'un système d'écoute non-linéaire.

4.2 Paramètres de mesure de la qualité de performance d'un VAD

La qualité d'un VAD se mesure par sa capacité à distinguer de la voix et à ce qui n'en n'est pas. Évaluer la performance d'un VAD n'est pas une tâche facile due à la subjectivité des résultats. Le contenu du message vocal doit pouvoir être compréhensible.

On distingue principalement deux types de paramètres : les paramètres objectifs et les paramètres subjectifs. Les paramètres objectifs sont basés sur des formules mathématiques. Utilisés dans toutes les méthodes d'évaluation, ils permettent de donner une bonne évaluation quantitative et de comparer les VAD entre eux. Les paramètres subjectifs, basé sur des tests d'écoute, donnent une évaluation qualitative et ils sont peu utilisés dû à leur coût et le temps qu'ils demandent. L'obtention des paramètres subjectifs est habituellement appliquée pour une phase finale et exhaustive d'évaluation comparative de VAD. Dans la littérature, la plupart des études portées sur les VAD se limitent aux tests objectifs. Pour pouvoir évaluer notre VAD, il faut le comparer à un VAD idéal. Ainsi, pour chaque fichier de voix, nous avons créé un fichier correspondant de VAD idéal, fait à l'aide

du logiciel *Sound Forge* [SFS10] en signalant manuellement les zones actives et inactives du signal.

4.2.1 Tests objectifs

La sortie d'un VAD est de nature binaire. Deux types d'erreurs sont alors possibles :

- l'erreur de mauvaise acceptation (ou erreur de fausse alarme) : le VAD considère comme actif une portion de signal qui ne contient pas de voix. Le terme « portion » désigne au minimum un échantillon de signal audio (figure 4.2).
- l'erreur de mauvais rejet: le VAD considère comme inactif une portion de signal qui contient de la voix (figure 4.2).

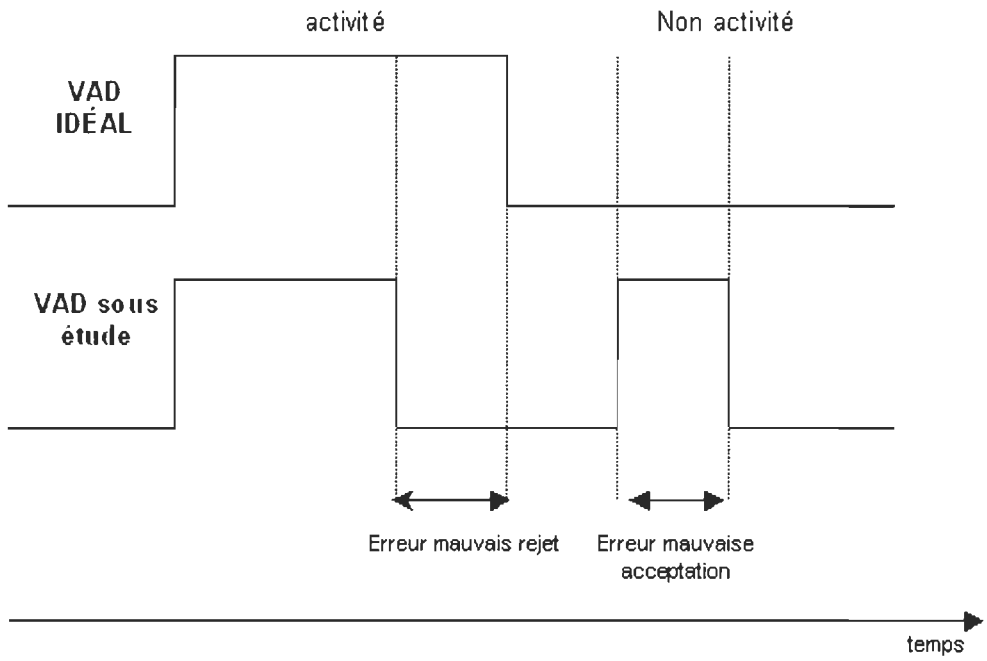


Figure 4.2 Illustration des erreurs de mauvais rejet et de mauvaise acceptation.

4.2.1.1 Paramètres Pd et Nd

Dans notre étude, nous allons nous baser sur deux paramètres pour mesurer l'efficacité de nos VAD :

- Pd : probabilité de bonne décision dans les régions actives (ZA).
- Nd : probabilité de bonne décision dans les régions inactives (ZNA).

$$Pd = \frac{Nbre[\text{bonne décision ZA}]}{Nbre[\text{activité}]} \tag{4.4}$$

$$Nd = \frac{Nbre[\text{bonne décision ZNA}]}{Nbre[\text{non-activité}]} \tag{4.5}$$

où *Nbre*[bonne décision ZA] désigne le nombre (en terme d'échantillons) de bonnes décisions dans les zones actives, *Nbre*[bonne décision ZNA] le nombre (en terme d'échantillons) de bonnes décisions dans les zones non-actives, *Nbre*[activité] désigne la durée (en terme d'échantillons) d'activité de voix, et *Nbre*[non activité] désigne la durée (en terme d'échantillons) de non-activité de voix.

Plus Pd est grand et plus le VAD va laisser passer de la voix. Plus Nd est grand et moins le VAD va laisser passer le bruit et autres signaux non vocaux. Donc plus Pd et Nd s'approchent de 100%, meilleure est la performance du VAD. À l'inverse, un Pd et/ou un Nd faible indiquent un VAD de mauvaise qualité. Pour un VAD idéal, les valeurs de Pd et Nd sont donc à 100%.

4.2.1.2 Exemple de calcul pour Pd et Nd

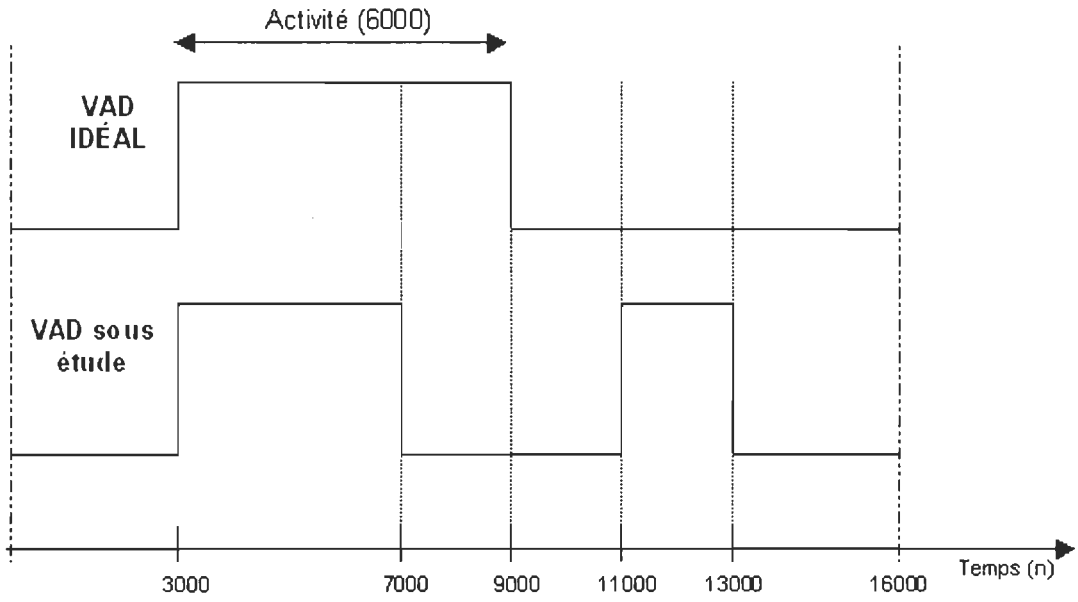


Figure 4.3 Exemple illustrant Pd et Nd.

Dans l'exemple de la figure 4.3, on a un signal de deux secondes échantillonné à 8kHz (ce qui correspond à 16 000 échantillons). Le temps est ici exprimé en termes d'échantillons. On a respectivement un temps d'activité et de non activité de 6 000 et de 10 000, respectivement. Le temps de bonne détection dans les zones actives est de 4 000 et le temps de bonne détection dans les zones inactives est de 8 000. À partir de ces données on peut calculer Pd et Nd :

$$Pd = \frac{Nbre[\text{bonne décision ZA}]}{Nbre[\text{activité}]} = \frac{7000-3000}{6000} = \frac{4000}{6000} \simeq 68\% \qquad (4.6)$$

$$Nd = \frac{Nbre[\text{bonne décision ZNA}]}{Nbre[\text{non-activité}]} = \frac{3000+2000+3000}{3000+7000} = \frac{8000}{10000} = 80\% \qquad (4.7)$$

D'après les résultats obtenus, la probabilité d'avoir une décision correcte dans les zones actives est de 68%, et la probabilité d'avoir une bonne décision dans les zones inactives est de 80%.

4.2.1.3 Interprétation subjective de Pd et Nd

Afin de mieux interpréter les résultats des sections précédentes, le tableau 4.1 présente une correspondance entre le résultat en terme de pourcentage de *Pd* et le jugement subjectif basé la compréhension du message vocal (par 'message vocal' on définit le sens de la voix). Les jugements ont été portés suite à l'écoute de plusieurs scénarios. Cette classification est approximative et subjective, mais permet une meilleure interprétation du pourcentage *Pd*.

On n'a pas présenté de tableau de correspondance pour *Nd*, vu que sa valeur est associée aux zones inactives. Contrairement aux zones actives qui contiennent un message vocal, les zones inactives ne contiennent rien qui doit être interprété. On ne peut donc pas associer *Nd* avec un jugement de perception. Cependant, on peut directement associer le pourcentage de *Nd* avec le taux de bruit qu'on entend dans les zones inactives (exemple : Nd de 50% signifie que la moitié des zones inactives sont considérés comme actives ; on va donc entendre du bruit à 50% dans les zones inactives).

Tableau 4.1 Correspondance entre différentes valeurs de *Pd* et la perception auditive.

Pd	**Commentaires**
Pd >90%	excellente qualité, message vocale très compréhensible.
80% < Pd < 90%	message compréhensible, excepté quelques débuts ou fin de mots

	coupés qui peuvent nuire légèrement à la compréhension.
65% < Pd < 80%	message plus difficile à saisir dû à certaines trames manquantes.
Pd < 65 %	message très difficile à saisir, message incompréhensible.

4.2.2 Tests subjectifs

Bien que les tests objectifs présentent une façon simple et efficace de montrer l'efficacité d'un VAD, ils comportent des limites. Par exemple, la distribution des erreurs n'est pas considérée dans le calcul de *Pd* et *Nd*. Ainsi, les résultats des deux VAD à la figure 4.4 sont donc considérés équivalents en termes de *Pd* et *Nd*. Bien que *Pd* et *Nd* soient identiques pour les deux signaux, il se peut que le sens du message vocal pour l'un soit compréhensible alors que l'autre ne l'est pas du tout.

Des tests subjectifs, qui consistent à écouter le résultat du VAD et d'en porter un jugement, pourraient donc être requis pour une meilleure comparaison. L'ITU a proposé une technique permettant d'encadrer des tests subjectifs [ITU96]. Ces tests ont fait l'objet d'une étude de comparaison entre trois VAD différents [BCR02]. Dans notre étude, nous limiterons les tests subjectifs seulement pour savoir à quoi correspondent les pourcentages de *Pd* et *Nd* en termes de compréhension du message de la voix (cf. section 4.2.1.3). Cette démarche permettra de mieux interpréter les résultats en terme de pourcentage présentés à la section 4.3.

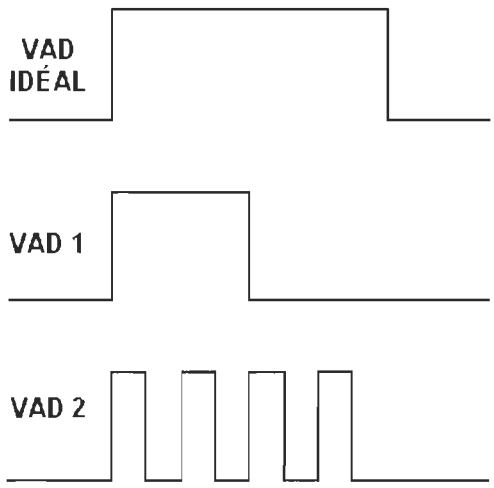


Figure 4.4 Exemple de deux distributions d'erreur différentes.

4.3 Résultats comparatifs d'évaluations des VAD

Cette section présente les résultats obtenus pour les trois VAD mentionnés dans l'introduction du chapitre (G729-B, VAD de Wu et Wang [WUW06], et notre VAD proposé). Nous avons décomposé les résultats en trois parties :

1. Le signal d'entrée $s_{clean}(n)$ est soumis à trois types de bruit différents et à des intensités de bruit variables.
2. On refait les mêmes tests sauf que $s_{clean}(n)$ est passé dans un système d'écoute non linéaire avant que du bruit ne soit rajouté.
3. Un échantillon de conversation provenant d'une navette spatiale à travers un réseau cellulaire de la troisième génération est utilisé pour mettre à l'épreuve nos VAD étudiés.

4.3.1 Ajustement du seuil de décision finale du VAD

Afin de démontrer l'influence du seuil de décision, λ_V de l'équation (3.23), la figure 4.5 montre *Pd* et *Nd* pour différentes valeurs de seuils, $0<\lambda_v<0.0006$. Les deux courbes de *Pd* et *Nd* ont été obtenues à partir d'un signal vocal de la base de données (cf. section 4.1.1) ajouté de bruit blanc à 10dB. Pour chaque seuil appliqué au VAS, on calcule Pd et Nd.

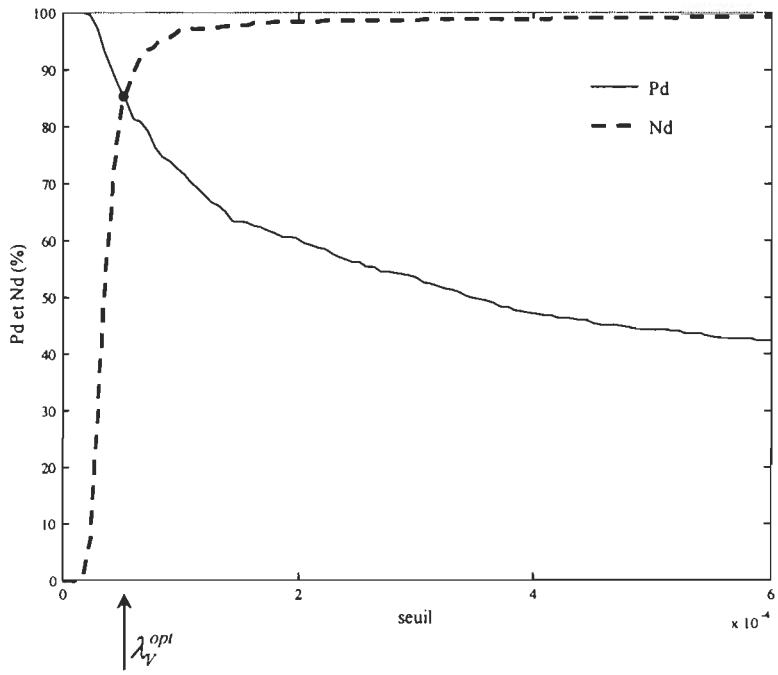


Figure 4.5 Un exemple de Pd et Nd pour différentes valeurs de seuil λ_V.

On observe que lorsque le seuil est proche de 0, Pd égal à 100% et Nd vaut 0%. Plus le seuil augmente, plus Pd diminue et Nd tend vers 100%. Cela se traduit par le fait que plus le seuil est bas, plus le VAD a de probabilités de considérer des trames inactives comme actives. Inversement, plus le seuil augmente et plus le VAD a de probabilités de considérer des trames actives comme inactives. La figure 4.5 montre l'importance de la valeur du seuil de décision. La sensibilité de Pd au seuil est très forte dans la région $[0,10^{-4}]$.

Une légère variation de seuil dans cette région a une grande influence sur Pd et Nd. Cependant, ceci est vrai dans l'exemple illustré mais ne peut absolument pas être généralisé. La sensibilité de *Pd* et *Nd* au seuil est fonction des conditions du signal étudié *s(n)*.

Dans notre étude, nous appliquerons comme méthode de détermination de seuil optimal, λ_v^{opt}, un seuil constant sur toute la durée du signal s(n) étudié qui maximise le point de croisement des courbes de *Pd* et *Nd*.

Les résultats de *Pd* et *Nd* qui seront présentés aux sections suivantes pour les deux VAD d'ondelettes ont été obtenus en ajustant le seuil selon cette procédure.

4.3.2 Synthèse des conditions de simulation

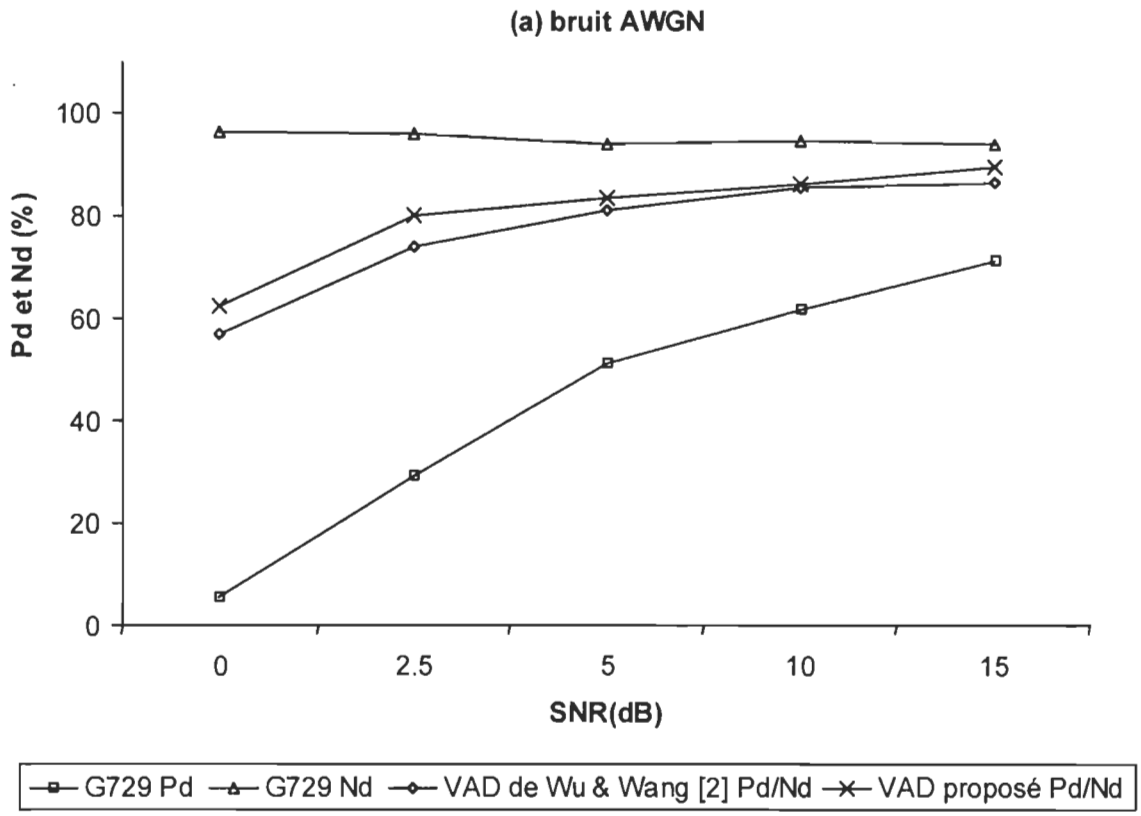
Le tableau 4.2 présente une synthèse des conditions et caractéristiques de simulation de l'étude des VAD.

Tableau 4.2 Synthèse des conditions et caractéristiques de simulation

Caractéristiques	Définition
VAD étudiés	G729-B Wu-VAD [WUW06] (section 3.2.1) VAD proposé (section 3.3)
Type de voix	Extraits de conversations vocales
Type de bruits	AWGN, "machine", "gare"
Système d'écoute	Linéaire et non linéaire (Eq. (4.3))
Mesure de performance	Nd (4.4) et Pd (4.5), courbes ROC
Longueur de la trame N	N=256
Chevauchement de trame	25% pour le WuVAD et VAD proposé 0% pour le G729
Fréquence d'échantillonnage, Fs	8kHz
Paramètres d'influences	SNR de 0dB à 15dB

4.3.3 Résultats pour un système d'écoute linéaire

La figure 4.6 présente les courbes de résultats pour trois types de bruits différents (bruit AWGN, "machine", "gare"). Les résultats sont exprimés en fonction de Pd et Nd. Pour les VAD à base de la transformée en ondelettes, comme il a été dit à la section 4.3.1, les résultats de Pd et Nd ont été obtenus avec λ_v^{opt}, c'est-à-dire que *Pd=Nd.*



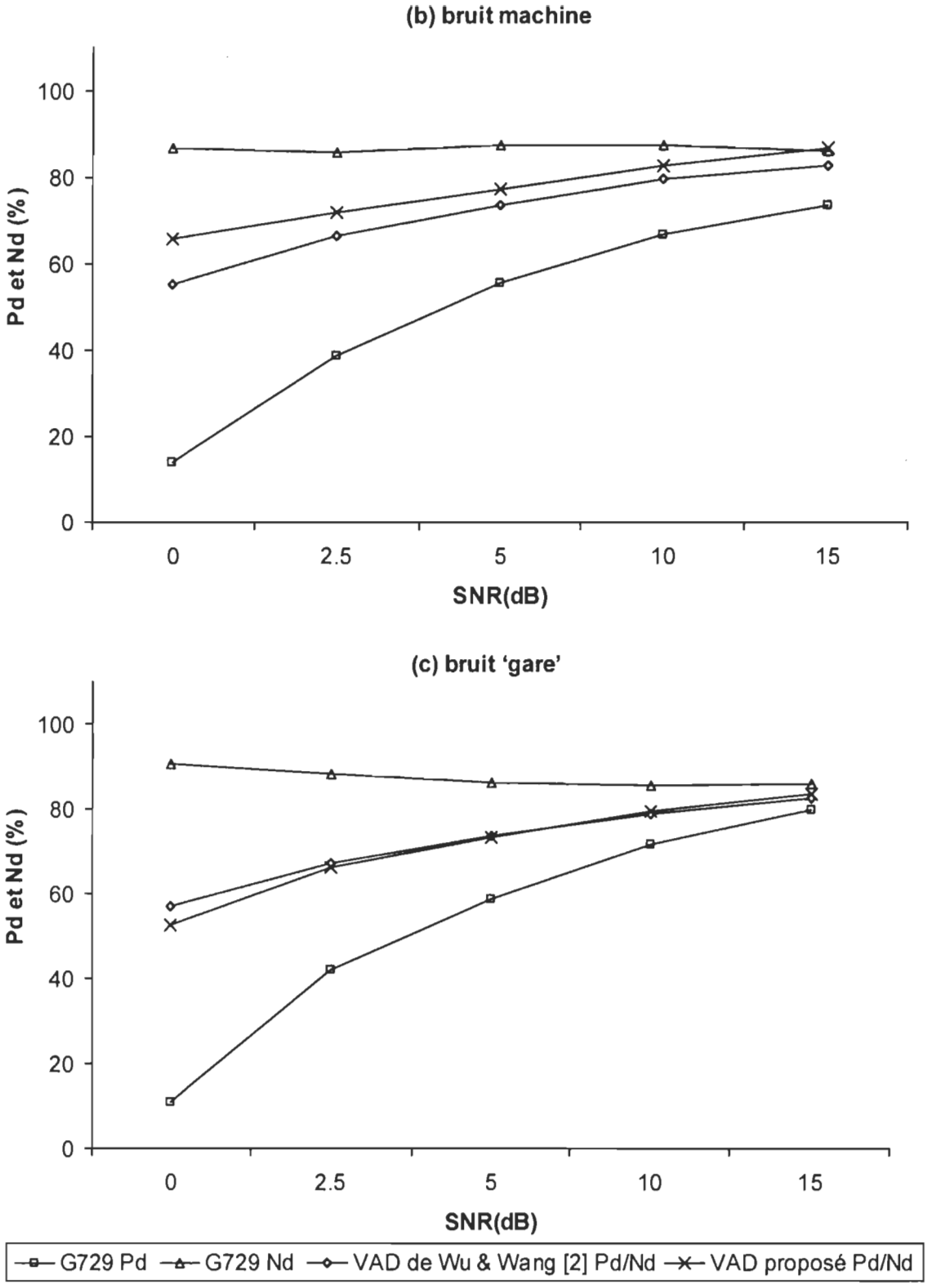


Figure 4.6 Résultats des VAD avec ajout de bruit à différents SNR pour trois types de bruits différents : (a) bruit blanc AWGN, (b) "machine", (c) "gare".

Lorsqu'on observe les trois graphiques de la figure 4.6, on s'aperçoit que quel que soit le type de bruit et le type de VAD, *Pd* et *Nd* diminuent proportionnellement au SNR. Cela signifie que plus le niveau de bruit augmente et plus le VAD a du mal à distinguer la voix.

Lorsqu'on compare les trois VAD entre eux, le VAD de Wu et Wang (*Wu-VAD*) et notre méthode proposée présentent des résultats similaires, alors que les résultats du G729 sont bien différents. Dans les trois cas de bruit et pour n'importe quel SNR, le Nd du G729 est supérieur aux Nd des deux autres VAD. Cependant chaque Pd correspondant est beaucoup plus faible, ce qui veut dire que le G729 considère beaucoup de trames comme inactives. Cela entraîne un bon résultat en terme de Nd mais un résultat médiocre en terme de Pd. Comme il a été mentionné dans les chapitres précédents, la qualité du VAD se traduit par un équilibre entre *Pd* et *Nd*. Ainsi, un *Pd* faible, associé à un *Nd* élevé, ne correspond pas à un bon VAD. Inversement, un *Nd* faible associé à un *Pd* élevé représente un VAD qui considère tout le signal *s(n)* comme de la voix. Contrairement au G729, les deux VAD d'ondelettes présentent des Pd et Nd similaires, ce qui les rend plus stables que le G729 quel que soit le bruit.

Nous notons une légère amélioration de notre VAD comparativement au Wu-VAD pour le bruit blanc (AWGN) et le bruit "machine", et cela quel que soit le SNR. Cependant, le Wu-VAD est légèrement supérieur que le nôtre pour un bruit "gare", et pour les niveaux de bruits inférieurs à 5dB. Ce qui est remarquable, c'est que notre méthode, comparativement au Wu-VAD, a un gain en dB qui croît lorsque le niveau de bruit augmente, ce qui indique que nous sommes plus performant que le Wu-VAD à fort bruit. En effet, nous gagnons

près de 10dB de gain à un SNR de 0dB dans une zone considérée critique (voir Table 4.1 pour la compréhension du message).

Le type de bruit ne semble pas vraiment différencier la performance des VADs. En effet, l'allure des courbes est similaire pour les trois bruits. On observe le même degré de pente lorsque le SNR diminue.

Pour mieux analyser la sensibilité du seuil de décision et l'effet du SNR sur le Wu-VAD et le VAD proposés, les figures 4.7 et 4.8 montrent des résultats obtenus en faisant varier le seuil de décision λ_v . Ce type de courbe est utilisé dans la littérature (e.g. [RUB07]) et porte le nom de courbe de ROC (*Receiver Operating Characteristic*). Pour chaque seuil de décision on calcule Pd et Nd sur l'ensemble du signal étudié s(n). On obtient ainsi une courbe qui permet de donner un jugement plus exact de la méthode. Plus la courbe tend vers le point (0,0) (VAD idéal), plus la méthode est performante.

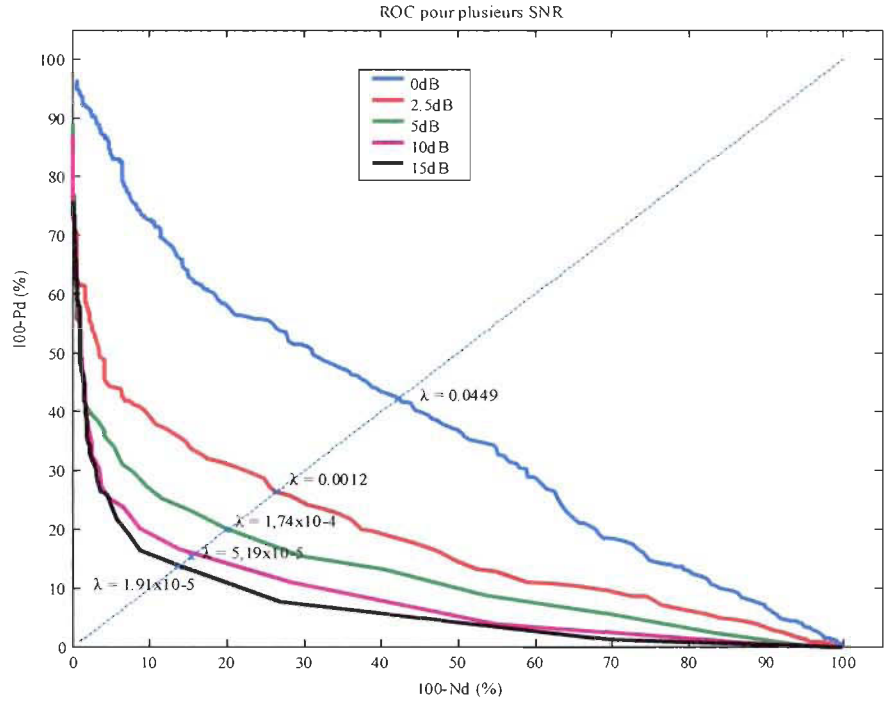


Figure 4.7 Courbes ROC du Wu-VAD pour un bruit blanc AWGN à différents SNR.

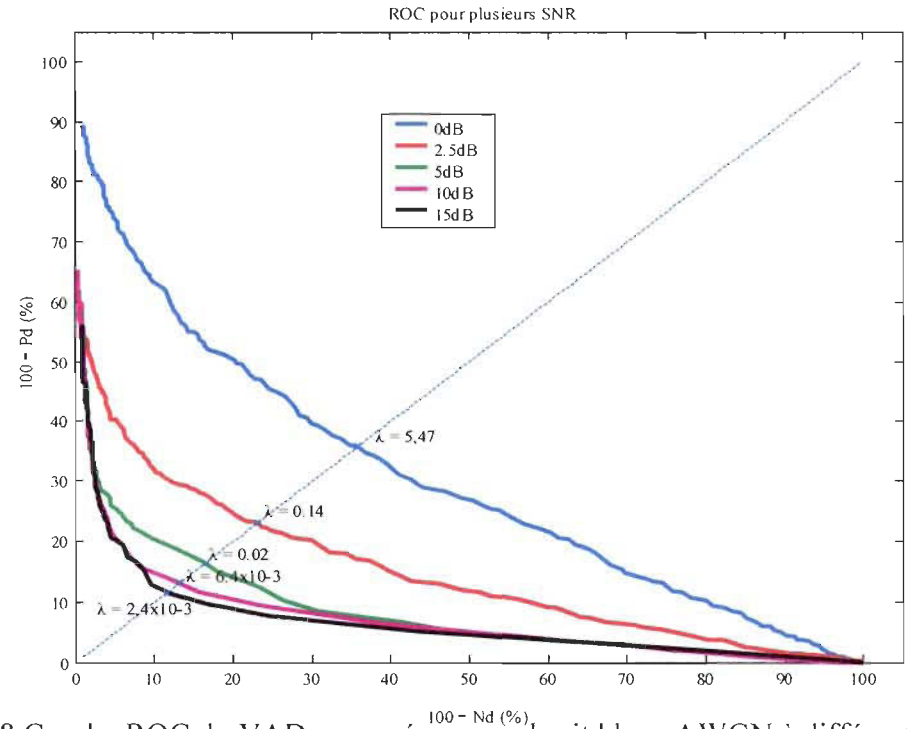


Figure 4.8 Courbe ROC du VAD proposé pour un bruit blanc AWGN à différents SNR.

Les résultats des figures 4.7 et 4.8 ont été obtenues à partir d'un extrait de la base AURORA [AUR00] auquel on a ajouté un bruit AWGN. On peut voir clairement sur ces courbes que plus le SNR diminue, plus la courbe tend vers le point de coordonnées (0,0) à *Pd=Nd*. Inversement, plus le SNR augmente et plus la courbe s'éloigne de l'idéal (0,0), c'est à dire que plus la détection vocale devient difficile.

La figure 4.9 présente la sortie du VAS, $V(n)$, du VAD proposé en comparaison avec le SAE (équivalent au VAS, seule l'appellation change) du Wu-VAD pour différents niveaux de bruit "machine". Le signal $s(n)$ correspond au message « un-deux-quatre ». Pour des soucis de clarté et pour une meilleure comparaison, les courbes de VAS ont été normalisées. C'est-à-dire que nous avons divisé chaque courbe $V(n)$ par son amplitude maximale

$$V_{normalisé}(n) = \frac{V(n)}{\max\left(V(n) \middle| n = 0,1,2,...,N\right)}$$

D'après la figure 4.9 on remarque que la sortie du VAS suit moins bien la voix lorsque le SNR diminue. Les deux VAS retracent partiellement la voix dans le cas du bruit très élevé (SNR=0dB). Les deux courbes de VAS se suivent et ont à peu près la même allure à faible bruit. Ceci vient appuyer les résultats de la figure 4.6 qui démontrent des performances similaires.

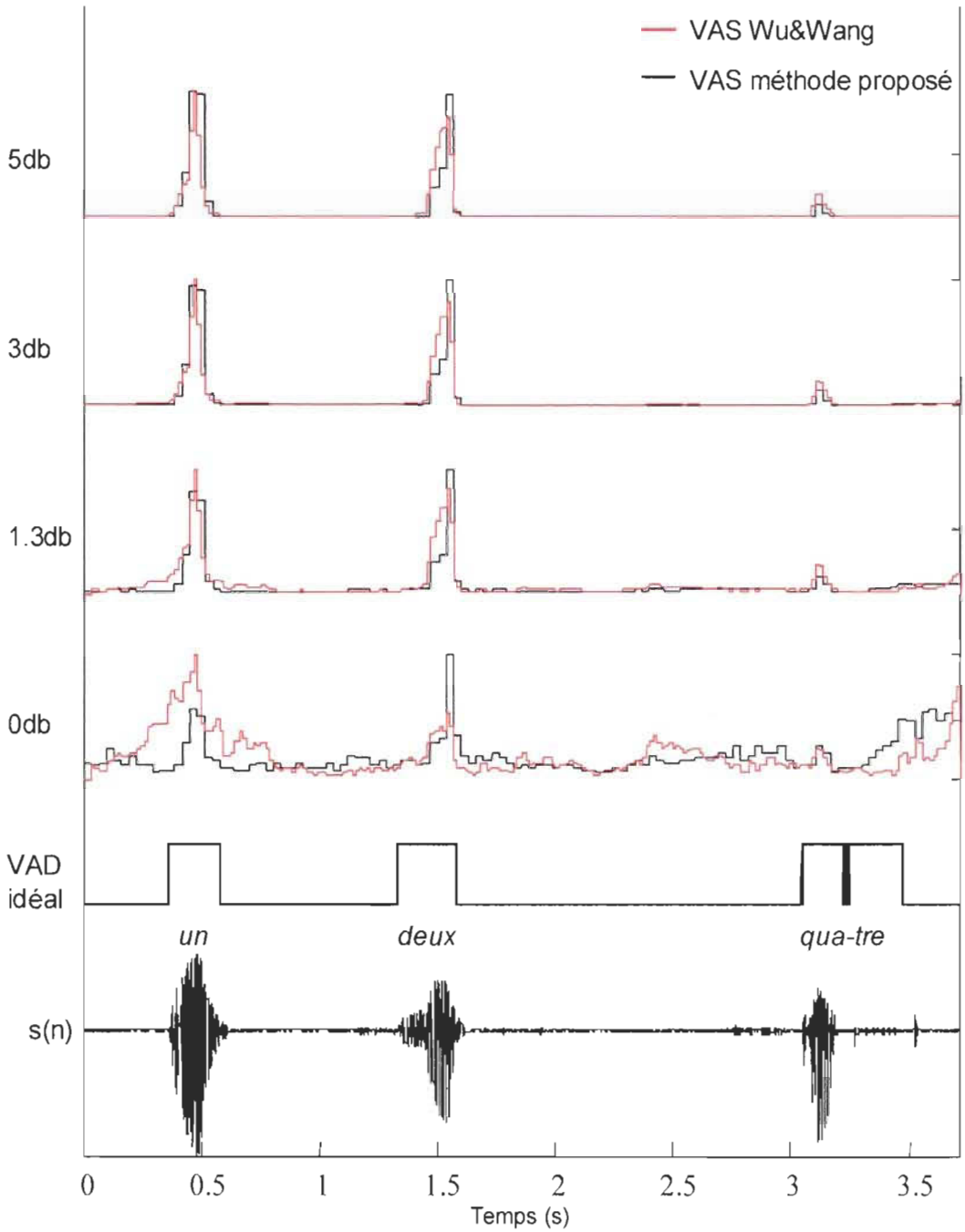


Figure 4.9 Comparaison des VAS (système d'écoute linéaire).

L'exemple de la figure 4.9 montre que les VAD présentent plus de difficultés pour certaines consonances de fin de phrase tel le « *tre* » de « *quatre* », alors que le son « [K] » ressort plus. Cet exemple montre que certaines consonances sont plus faciles à détecter que d'autres.

La figure 4.10 montre un exemple typique de résultat pour les trois VAD étudiés. Pour cet exemple, le signal s(n) provient de la base de données AURORA [AUR00]. Les résultats ont été obtenus pour un bruit de "machine" de SNR 5dB. On peut voir que la voix est presque totalement noyée dans le bruit. On constate que le G729B a du mal à retracer la voix. Les VAD basés sur la TO performent bien malgré certaines portions. Les courbes de voix montrent la dégradation du bruit sur le VAD.

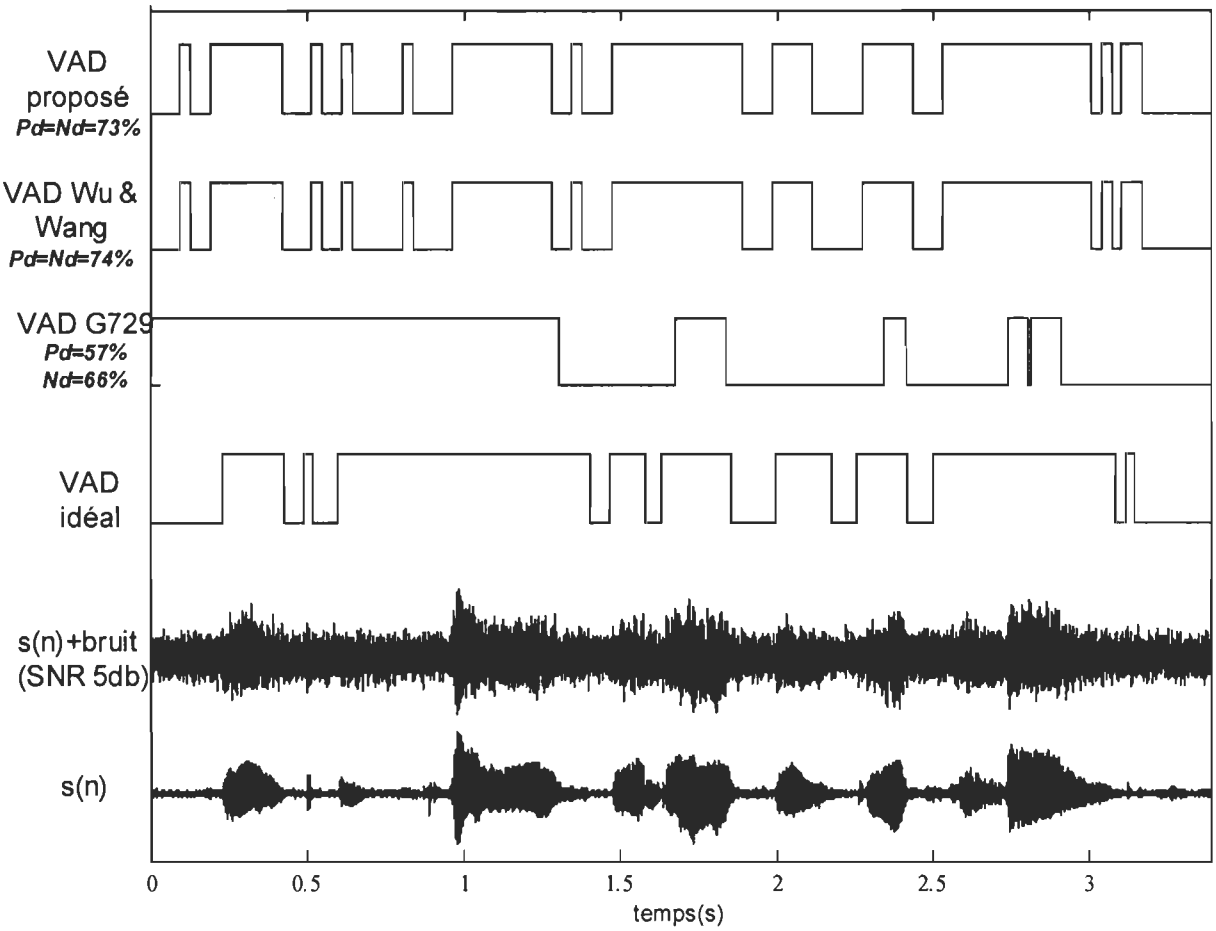


Figure 4.10 Comparaison des VAD pour les trois algorithmes.

4.3.4 Résultats pour un système d'écoute non-linéaire

Dans cette section, nous évaluerons les performances des VAD étudiées dans le cas d'un système d'écoute non linéaire. Le signal s(n) est passé dans la fonction non-linéaire $f_{NL}(\bullet)$ (cf. équation (4.3)). La figure 4.8 montre l'effet de la non linéarité sur le signal audio. On aperçoit une diminution des crêtes et une certaine déformation non linéaire du signal. Cependant, on note que la périodicité reste la même.

Dans la suite du chapitre, nous reprenons les mêmes conditions de simulations selon les mêmes étapes que le cas d'un système linéaire de la section précédente.

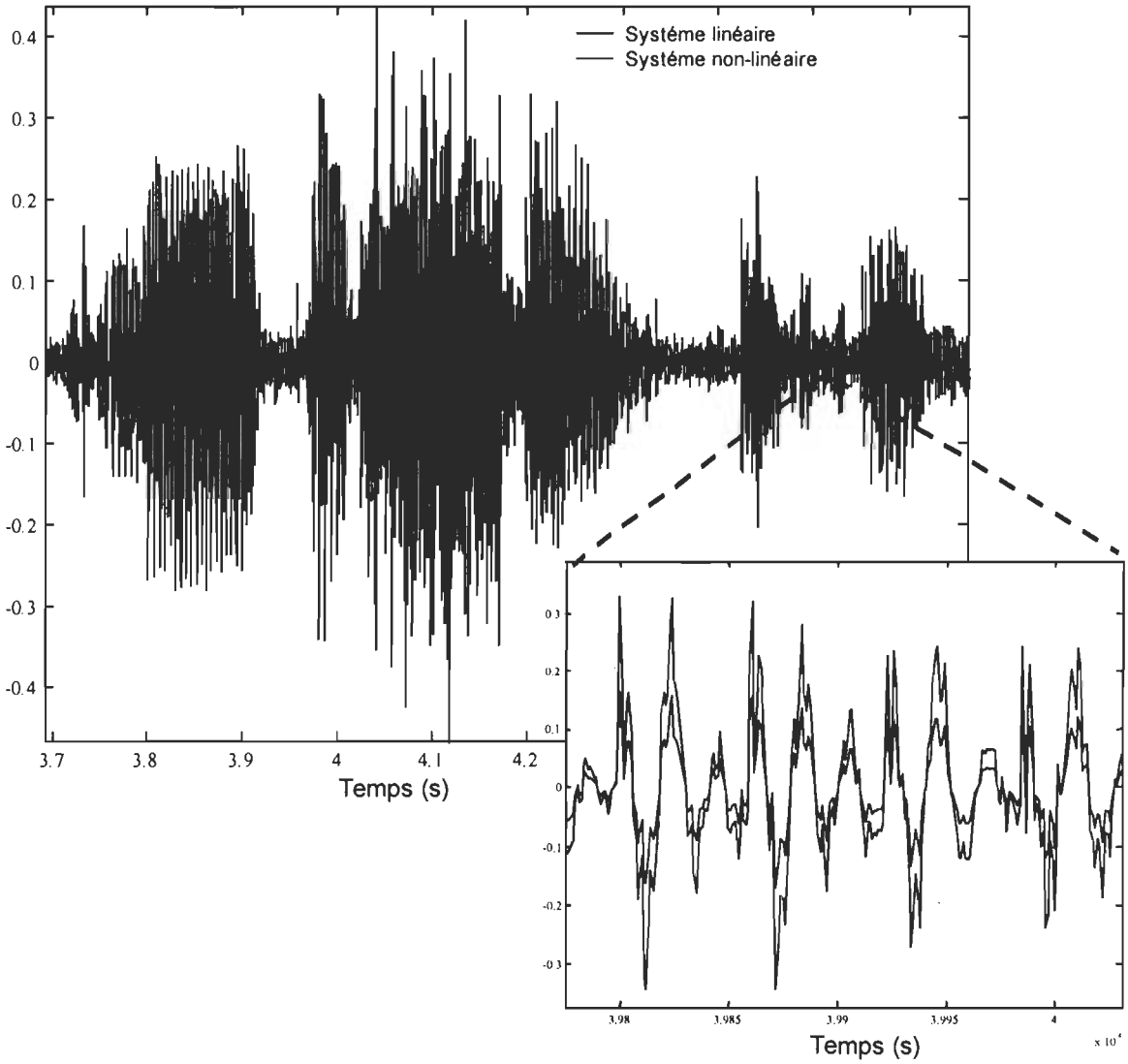
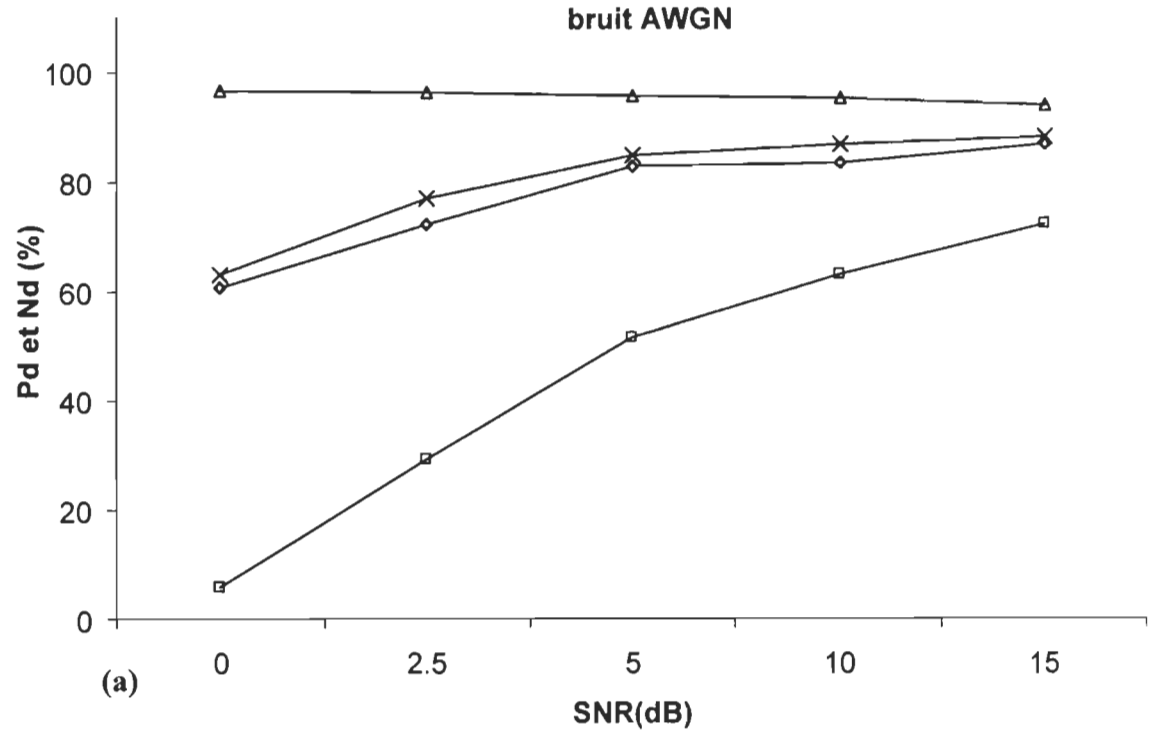
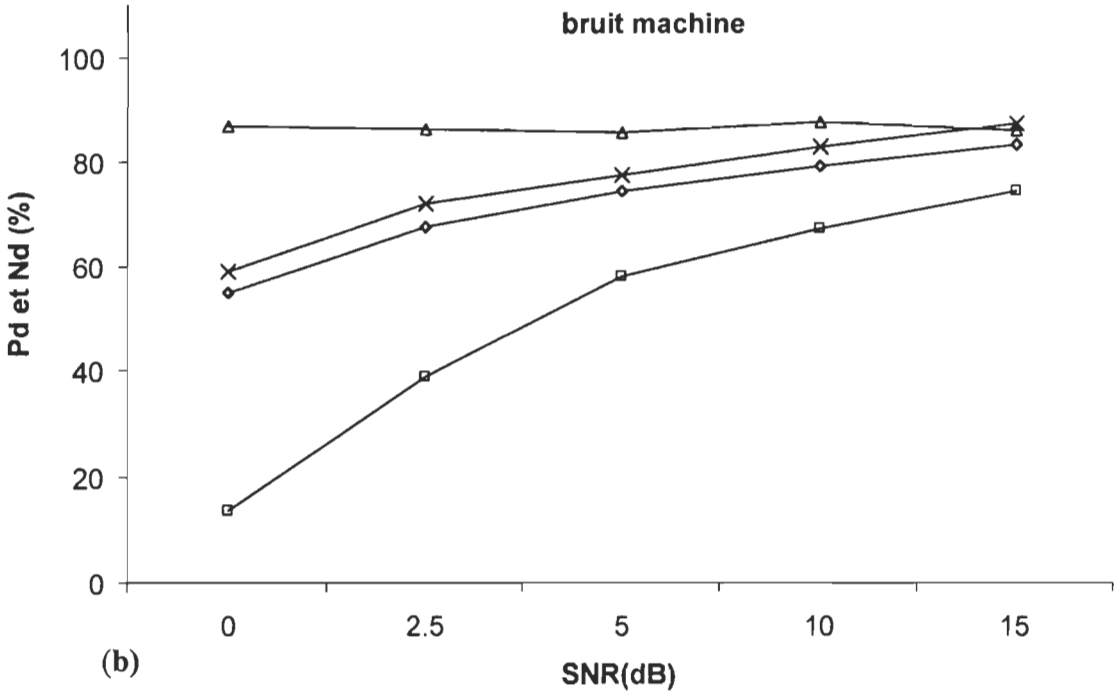


Figure 4.11 Signal avant et après le passage dans le système d'écoute non-linéaire.

De la même manière qu'à la figure 4.6, la figure 4.12 présente les résultats obtenus en termes de *Pd* et *Nd* pour les trois bruits (AWGN, "machine" et "gare") et pour les trois

VAD (G729B, Wu-VAD et le VAD proposé). La différence est qu'on fait passer le signal d'entrée $s_{clean}(n)$ dans la fonction non linéaire avant d'ajouter le bruit.





G729 Pd — G729 Nd — VAD de Wu & Wang [2] Pd/Nd — VAD proposé Pd/Nd

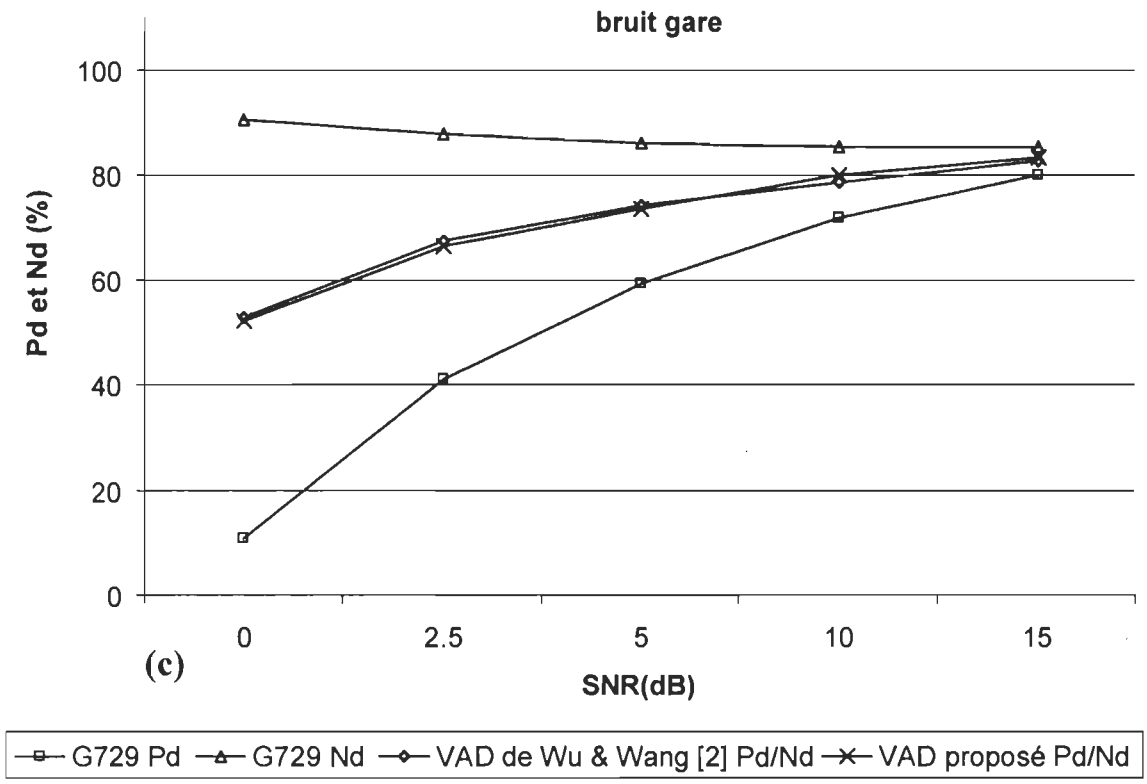


Figure 4.12 Résultats des VAD dans le cas d'un système d'écoute non-linéaire pour trois bruits différents : bruit blanc AWGN (a), bruit "machine" (b), et bruit de "gare" (c).

D'un point de vue général, si l'on compare les résultats obtenus à la figure 4.6 et ceux de la figure 4.12, on observe les mêmes caractéristiques. Plus le SNR diminue et plus *Pd* et *Nd* diminuent. Pour les trois types de bruits, le G729B performe moins bien que les Wu-VAD et le VAD proposé lorsque le niveau de bruit augmente. Les deux WVAD présentent à peu près les mêmes performances que dans le cas d'un système linéaire. Pour chaque simulation des figures 4.6 et 4.12, le seuil de décision est ajusté à λ_{opt} (*Pd=Nd*). Ceci permet une comparaison équitable entre les différentes simulations.

Pour mieux analyser la différence entre un système d'écoute linéaire et non linéaire, la figure 4.13 montre des courbes obtenues en faisant varier le seuil de décision (courbes ROC) tel que présenté aux figures 4.7 et 4.8.

À la figure 4.13, les courbes ont été obtenues pour son de la base de données AURORA auquel on a ajouté un bruit AWGN de 5dB. On voit que les deux courbes (linéaire et non-linéaire) correspondant à notre VAD proposée sont plus proches de (0,0) que les deux courbes du Wu-VAD dans la région d'intérêt (Pd et Nd $\approx<$ 60%). On peut ainsi dire que dans la région d'intérêt, notre méthode est plus performante. En dehors de cette région, *Pd* ou *Nd* sont trop faibles pour pouvoir être considérés pour un jugement pertinent de la méthode.

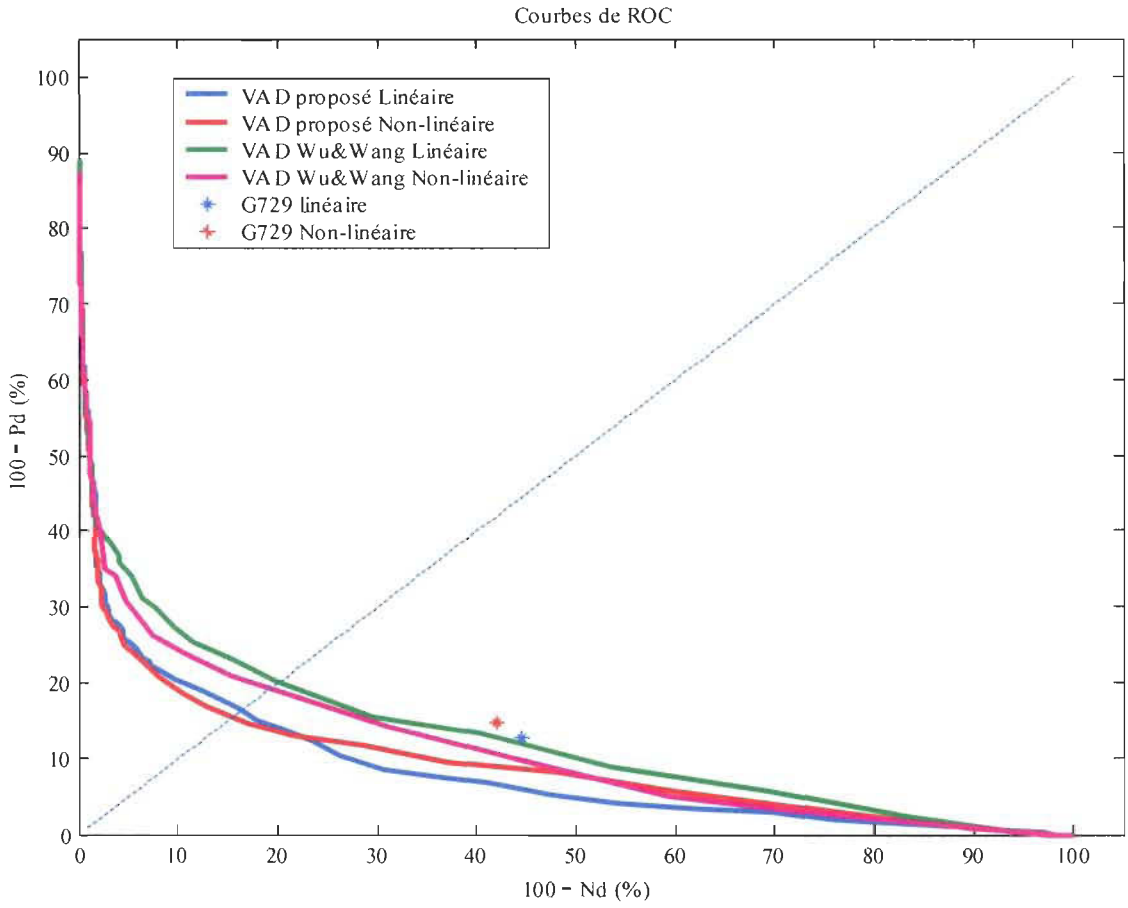


Figure 4.13 Courbes de ROC pour avec et sans linéarité pour un bruit AWGN de 5dB.

Si on compare les courbes 'linéaire' avec 'non-linéaire', on s'aperçoit que pour chaque méthode, les deux courbes se suivent de près et se chevauchent. D'après cette constatation, on peut dire que la non-linéarité ne semble pas affecter la performance des deux VAD.

La figure 4.14 présente les VAS des deux WVAD à différents niveaux de bruit pour un bruit de machine. Lorsqu'on compare avec la figure 4.9 (même figure mais dans le cas linéaire), on observe que les courbes ont la même allure. Les VAS suivent la voix de la même manière, avec les mêmes imperfections au niveau de « tre » de « quatre ». D'après cette figure, la non linéarité n'affecte pas les VAS, et donc les VAD.

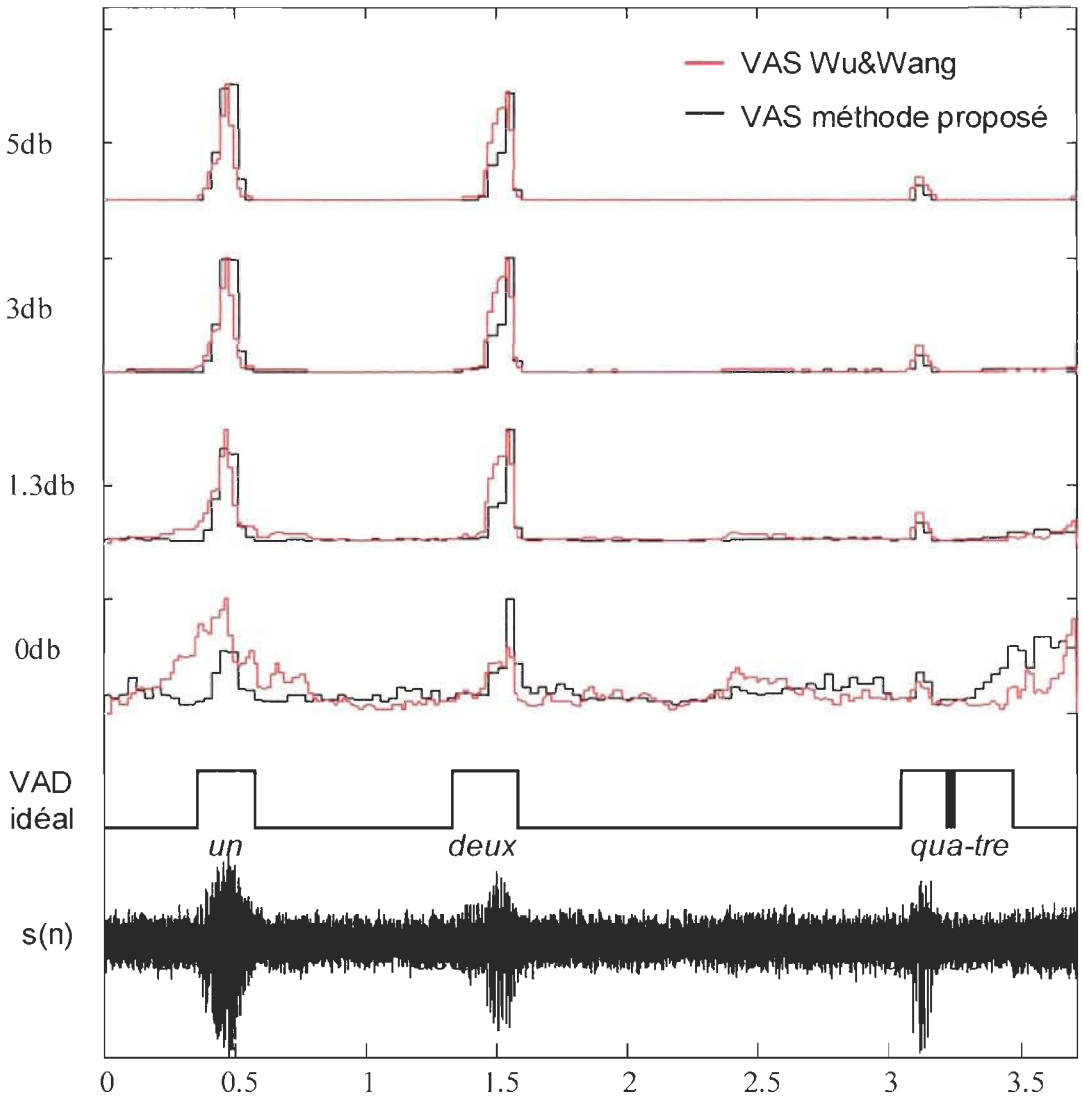
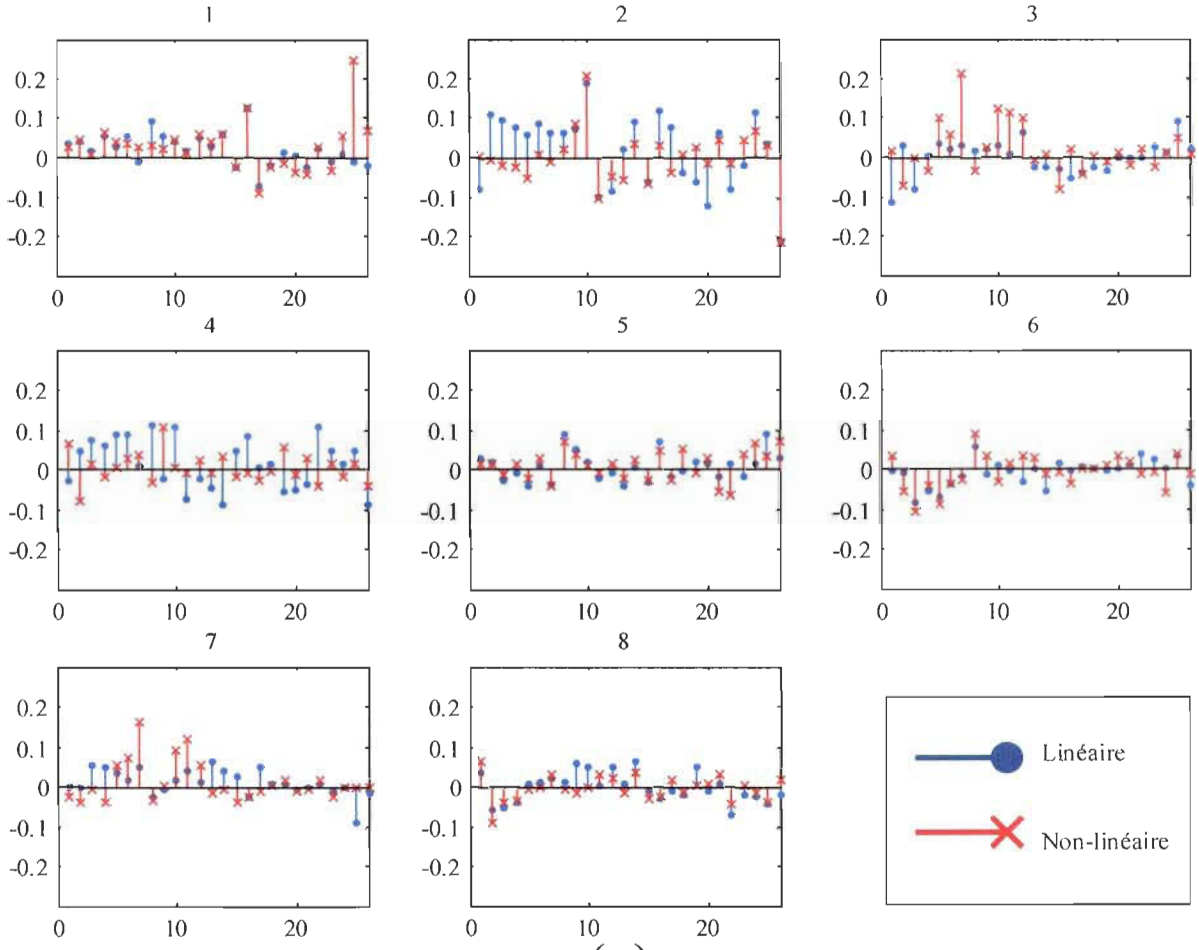


Figure 4.14. Sortie du VAS pour le Wu-VAD et notre VAD dans le cas d'un système d'écoute non-linéaire.

Dans le but de mieux comprendre les résultats obtenus, nous avons voulu observer l'effet de la non-linéarité au niveau des ondelettes pour notre méthode proposée. La figure 4.15 présente les coefficients d'ondelettes pour une trame contenant de la voix avec et sans non-

linéarité du système. L'extrait provient d'un son de la base de données AURORA auquel on a ajouté un bruit AWGN de 15dB. On peut voir que bien que les 17 sous-signaux soient différents, l'amplitude et la périodicité restent équivalentes. La variance de chaque sous-signal obtenue dans le cas linéaire est donc presque identique à la variance obtenue dans le cas d'un système d'écoute non linéaire. Ceci explique que les VAS – somme des variances de chaque série de coefficients - dans les cas linéaires et non linéaires soient équivalents.



(a)

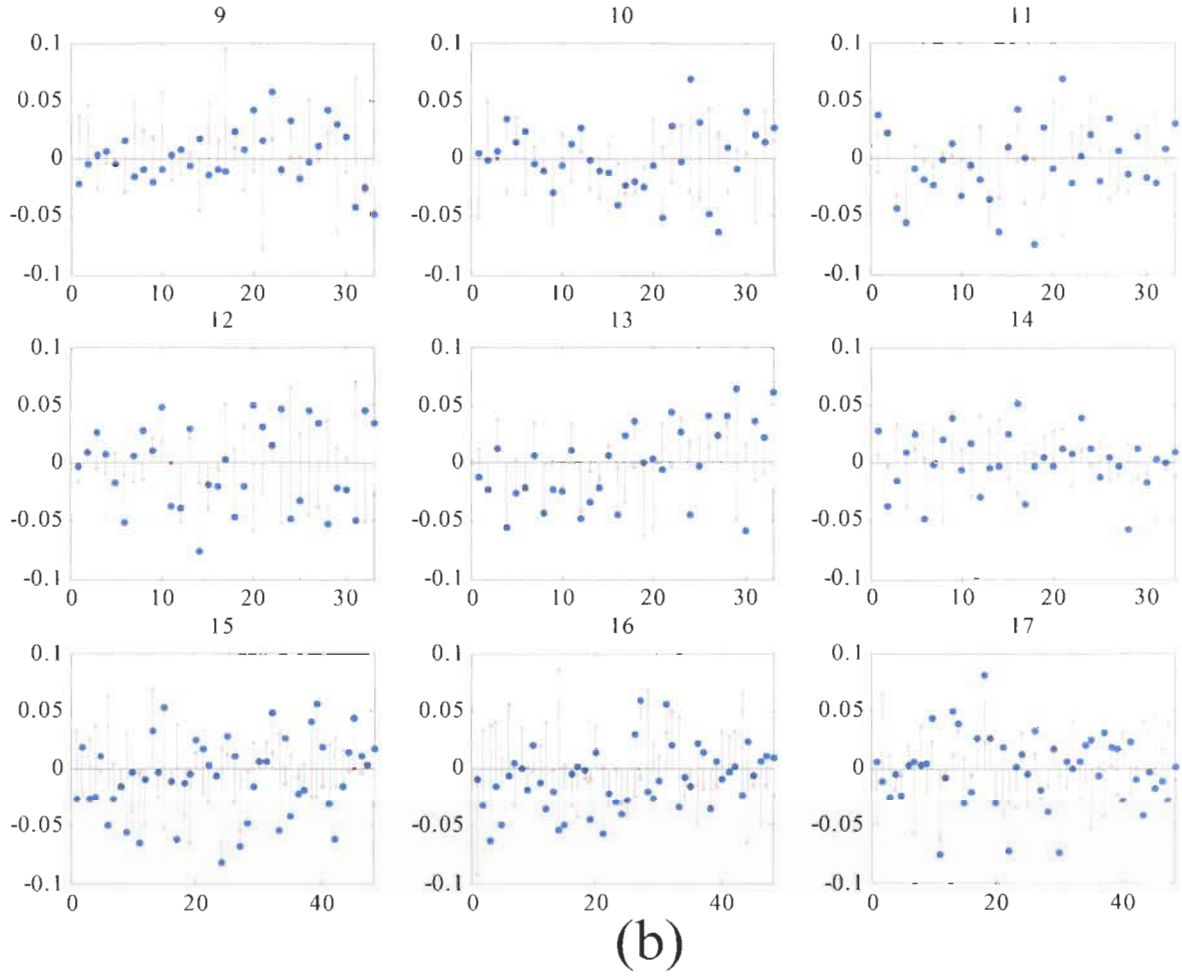


Figure 4.15 Coefficients d'ondelettes dans le cas d'un système d'écoute linéaire et non linéaire : (a) : des coefficients des sous-signaux 1 à 8; (b) : coefficients des sous-signaux 9 à 17.

Les coefficients de 1 à 8 correspondent à ceux de l'étage j=5, les coefficients 9 à 14 à l'étage 4 et les coefficients 15 à 17 sont associés à l'étage j=3 (cf. figure 3.4).

4.3.5 Résultats expérimentaux d'un message en provenance de l'espace

Le tableau 4.3 présente les résultats obtenus pour un son provenant d'une communication spatiale pour le VAD proposé et le G729. Il s'agit d'un son correspondant à des astronautes communiquant avec la station terrestre. Le message est très difficile à comprendre, voire

inaudible. Ce son a été choisi car on assume qu'il contient des non-linéarités en plus du bruit élevé et de la saturation. Le son a été enregistré à la réception du signal, à la station spatiale. Le pourcentage d'activité vocale est de 89%, selon le VAD idéal fait à la main, ce qui est un pourcentage très élevé.

Tableau 4.3 Résultats pour un son provenant d'une communication spatiale

VAD	*Pd*	*Nd*
G729	90.7	22.3
Proposé	89.9	4.6

Nous avons simulé notre VAD proposé après avoir simulé le G729. Pour des résultats équitables, nous avons ajusté le seuil de manière à ce que Pd du G729 soit équivalent au *Pd* du VAD proposé. On s'aperçoit que les deux valeurs de Pd sont très hautes mais les pourcentages de *Nd* sont extrêmement bas. Un Pd très élevé et un Nd très faible signifie que le VAD considère comme actif la grande majorité des trames, ce qui n'est pas la caractéristique d'un bon VAD. Le *Nd* du G729 est plus élevé que notre *Nd*. Cependant, le taux d'inactivité sur l'ensemble du message étant très faible, un pourcentage de *Nd* très faible n'aura presque pas ou très peu d'incidence à l'écoute du message. La qualité du son est trop mauvaise, et le pourcentage d'inactivité trop fort pour pouvoir comparer adéquatement nos VAD et en tirer une conclusion.

4.4 Discussion et Conclusion

Il a été observé que lorsque le niveau de bruit augmente, quelque soit le type de bruit, la performance de tous les VAD décroît. Le manque de performance se traduit par des morceaux de voix coupés, la plupart des cas en début et en fin de mots. Dans des cas de bruit très élevés, cela se traduit par une incompréhension de la voix. Dans l'ensemble des résultats obtenus, on peut noter qu'en dessous d'un SNR d'environ 3 dB, il devient très difficile pour les VAD de détecter la voix. Les VAD vont laisser passer seulement quelques portions de voix (souvent les « pics » les plus intenses du mot) ce qui va rendre le message incompréhensible.

Les trois VAD étudiés présentent des résultats similaires en cas de linéarité et de non-linéarité. On peut observer que dans certains cas, les VAD étudiés à base de la TO performent mieux en présence de non linéarité.

Dans l'ensemble des résultats, notre VAD présente à peu près les mêmes résultats que le VAD de Wu et Wang. Dans quelques cas de bruit, notre VAD proposé offre un léger gain de performance qui peut s'avérer d'intérêt dans les zones critiques de Pd ($65\%<Pd<80\%$). Les deux VAD présentent des performances supérieures au G729B observables lorsque le bruit augmente. Ces résultats montrent l'intérêt d'utiliser la technique des ondelettes dans des situations de bruit et de non linéarité.

L'importance de l'influence du seuil sur la performance du VAD a également été démontrée. Cela montre l'importance d'avoir une méthode de seuil efficace.

Chapitre 5
Conclusion

Dans un système de communication utilisant la détection d'activité vocale, le signal recueilli à partir de la voix peut contenir du bruit (lorsque la voix est dans un environnement bruyant) et peut être soumis à des distorsions liées à des non linéarités lorsque le signal est passé dans un canal de communication. Ce genre de déformation du signal fait décroître les performances des VAD. Plusieurs méthodes basées sur des techniques différentes se sont penchées sur le problème du bruit. Compte tenu de la multitude de techniques de VAD, nous avons choisi de cibler notre étude sur les VAD basés sur les ondelettes car ils présentent des bonnes performances en présence de bruit. Comme la non-linéarité n'a jamais été traitée dans les VAD, nous avons également étudié la capacité des VAD basés sur les ondelettes à repérer la voix dans un signal contenant de la non-linéarité. Nous avons également proposé une nouvelle méthode basée sur les ondelettes, que nous avons testée et comparée à deux méthodes déjà existantes (méthode de Wu et Wang et G729).

Les méthodes ont été testées à partir de signaux audio de voix d'hommes et de femmes parlant dans un microphone, simulant une application où le microphone est enregistré près de la bouche. Différents types de bruit ont été ajoutés à plusieurs SNR. La

linéarité a aussi été ajoutée pour simuler un système d'écoute non linéaire. Pour évaluer la performance des VAD, les pourcentages de bonnes décisions dans les zones actives et inactives ont été calculés à partir de courbe de VAD idéal. Pour les méthodes à base d'ondelettes, un courbe appelée VAS qui suit le signal de voix est également utilisée pour fins d'étude.

Dans l'ensemble des résultats, il a été observé que les WVAD performent beaucoup mieux que le G729, particulièrement lorsque le niveau de bruit augmente. Le type de bruit affecte très peu les résultats. Le VAD proposé suit à peu près les mêmes résultats que le VAD de Wu et Wang [WUW06], et ce quel que soit le type et le niveau de bruit. Par contre, nous avons démontré que notre algorithme est moins complexe que le VAD de Wu et Wang. Cet élément important vient justifier la proposition de notre nouveau VAD.

Les résultats ont montré que la non-linéarité n'affecte pas la performance des VAD à base d'ondelettes. Pour pousser l'étude plus loin, des tests ont été faits sur un son provenant de l'espace où on prétendait une non-linéarité additionnée de bruit à la réception. À cause de la dureté du canal de transmission et des nombreuses interférences présent dans le son, ainsi que du très haut taux d'activité vocale, l'interprétation des résultats est très difficile. On ne peut pas clairement dire quel VAD performe mieux. Ce son nous amène à nous conscientiser sur les difficultés de communication que peuvent engendrés interférences et canaux de communications.

On a pu noter dans cette étude la difficulté à évaluer un VAD. Les tests objectifs donnent une bonne approximation de la performance des VAD, mais des tests subjectifs s'avèreraient nécessaires pour de meilleurs résultats. Cependant, la difficulté à effectuer ces tests rend leur pratique difficile. De plus, les résultats des VAD sont comparés à des

résultats supposés « idéaux » faits à la main, à base d'écoute, ce qui ne donne pas lieu à une évaluation strictement exacte. On peut rajouter que nous n'avons pas pu obtenir la base de données AURORA. Cette base de données de sons, souvent utilisée comme référence dans des tests de reconnaissance vocale et détection d'activité vocale, aurait peut-être permis à une meilleure évaluation.

Aucune méthode de seuil n'a été appliquée sur les VAD ondelettes. Une courbe appelée VAS a été obtenue à chaque cas et un seuil qui minimise les deux erreurs a été calculé pour chaque cas. Pour calculer le meilleur seuil, le VAD idéal est nécessaire ce qui ne rend pas notre VAD applicable pour des applications pratiques. Dans des travaux futurs, une recherche sur la mise en place d'une méthode de seuil efficace pourrait être envisageable.

Pour la première fois, la non linéarité a été mise en avant dans les VAD. Ce phénomène étant présent dans certains canaux de communications, cette étude pourrait être le début d'autres recherches plus poussées sur la non-linéarité chez les VAD. Par exemple, nous avons étudié la non-linéarité en imposant une fonction non linéaire arbitraire. Une étude pourrait être suggérée sur plusieurs types de non linéarités.

Finalement, un article a été publié sur la méthode proposée mettant en avant le problème de la non linéarité [CHI09] (cf. appendice 2).

Bibliographie

[AUR00] G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems unders Noisy Conditions", Proc. of Interspeech, Pékin, Chine, vol.1, 2000, pp.341-344.

[BCR02] F.Beritelli, S.Casale, G. Ruggeri, S.Serrano, "Performance Evaluation and Comparaison of G729, AMR, Fuzzy VAD", IEEE Signal processing letters, vol.9, No.3, Mars 2002, , pp. 85-88.

[CHE05] S.-H. Chen, H.-T. Wu, C.-H. Chen, J.C Ruan, T.K. Truong, "Robust Voice Activity Detection Algorithm Based on The Perceptual Wavelet Packet Transform", IEEE Int. Symp. on Intelligent Signal Processing and Communication Systems, Hong Kong, Chine, Dec. 2005, pp.45-48.

[CHE02] S.-H. Chen, J.-F. Wang, "A Wavelet-Based Voice Activity Detection Algorithm in Noisy Environments", IEEE Int. Conf. on Electronics, Circuits and Syst., Dubrovnik, Croatie, vol.3, Sept. 2002, pp.995-998.

[CHI09] R. Chiodi et D. Massicotte, "Voice Activity Detection Based on Wavelet Packet Transform in Communication Nonlinear Channel", SPACOMM, Colmar, France, 2009.

[ETS99] ETSI, "Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels", ETSI EN 301-708, 1999.

[FUZ98] F. Bertelli, S. Casale, and A. Cavallero, "A robust voice activity detector for wireless communications using soft computing", IEEE J.Select. Areas Commun., vol. 16, Dec. 1998, pp. 1818-1829.

[GSC01] M. W. Hoffman, Z.Li, D. Khataniar , "GSC-Based Spatial Voice Activity Detection for enhanced speech coding in the presence of competing Speech", IEEE Transactions on speech and audio processing, vol. 9, no. 2, mars 2001, pp. 175-178.

[HAY99] Simon HAYKIN, *Neural Networks – A comprehensive Foundation.* New Jersey, USA: Prentice Hall International, 2nd Edition, 1999, chap.5, 'Radial Basis Function Networks', pp. 256–317.

[ITU96] ITU-T, "Methods for subjective determination of transmission quality", Rec.p.800 Tech. Rep., aout 1996.

[JAB99] F.A. Jabloun, E. Cetin, and E. Erzin, "Teager Energy Based Feature Parameters For Speech Recognition In Car Noise", IEEE Signal Processing Letters, Vol. 6, no. 10, 1999, pp.256-261.

[KAI90] J.F Kaiser, "On A Simple Algorithm To Calculate The 'Energy' Of A Signal", IEEE Int. Conf. on Acoustic, Speech and Signal Processing, Albuquerque, U.S.A., vol. 1, 1990, pp.381-384.

[KIM05] Kim and Park, "Voice Activity Detection Algorithm Based on Radial Basis Function Network", IEICE Trans. Communication, vol. E88–B, no. 4, avril 2005, pp. 1656-1657.

[LGM01] S. Legendre, J. Goyette, D. Massicotte, "Ultrasonic NDE of Composite Material Structure Using Wavelet Coefficients", NDT&E International-Elsevier, Vol. 34, No. 1, 2001, pp. 31-37.

[LMG01] S. Legendre, D. Massicotte, J. Goyette, "Neural Classification of Lamb Wave Ultrasonic Weld Testing Signals Using Wavelet Coefficients", IEEE Trans. Instrum&Measurement, Vol. 50, No 3, 2001, pp. 672-678.

[LMG00] S. Legendre, D. Massicotte, J. Goyette, T. Bose, "Wavelet-Transform-Based Method of Analysis for Lamb-Wave Ultrasonic NDE Signals", IEEE Trans. on Instrum.&Meas., Juin 2000, pp. 524-530.

[MAL89] Mallat, S., " A theory for multiresolution signal decomposition : the wavelet representation " , IEEE Transactions on Pattern Analysis and machine Intelligence, 11(7), 1989, pp.674-693.

[MAL97] Mallat, S., « A wavelet tour of signal processing », Wiley, 1997.

[NOI92] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, Vol. 12, No. 3, pp. 247 - 251, 1993.

[PRO96] J.G. Proakis, D.G. Manolakis, "Digital Signal Processing – Principles, Algorithms, and Applications", Third Edition, Prentice Hall, 1996.

[RAB93] L. Rabiner and B.H. Juang, "Fundamental of speech recognition", Upper Saddle River, NJ Prentice Hall, 1993.

[RUB07] J.E. Rubio, K. Ishizuka, H. Sawada, S. Araki, T. Nakatani, and M. Fujimoto "Two Microphones Voice Activity Detection Based on The Homogeneity of

The Direction of Arrival Estimates", IEEE Int. Conf. on Acoustic, Speech and Signal Processing, Honolulu, USA, avril 2007, pp. 385-388.

[SFS10] Sony Creative Sound Forge,

URL: *www.sonycreativesoftware.com/soundforgesoftware*

[SHF04] Jiang Shaojun, Guo Haitao, Yin Fuliang, "A new algorithm for voice activity detection based on wavelet transform", International sympsosium on intelligent multimedia, video and speech processing, Hong Kong, Chine, octobre 2004.

[SHR97] Joachim Stegmann, Gerard Schröder, "Robust voice activity detection based on the wavelet transform", IEEE workshop on Speech coding for telecommunications proceeding, 1997, pp.99-100.

[SIN93] D. Sinha and A. Tewfit, "Low bit rate transparent audio compression using adapted wavelet", IEEE Trans. On Signal Processing, vol.41, Dec. 1993, pp.1170-1183.

[SOH99] J. Sohn, N.S.Kim, W.Sung, "A statistical Model-Based Voice Activity Detection", IEEE Signal Processing Letters, Vol.6, No.1, janvier 1999, pp.1-3.

[TAN00] S.G. Tanyer, H.Ozer "Voice Activity Detection in NonStationary Noise", IEEE Transactions on speech and audio processing, vol.8, No.4, juillet 2000, pp.478-482.

[UIT96] ITU-T Recommendation G.729 "Annex B: A Silence Compression Scheme for Use with G.729 Optimized for V.70 Digital Simultaneous Voice and Data Applications", IEEE Communications Magazine, 35(9), 1997, pp.64-73.

[WUW06] B.-F. Wu, K.-C. Wang, "Voice Activity Detection based on Auto-Correlation Function Using Wavelet Transform and Teager Energy Operator",

Computational Linguistics and Chinese Language Processing, vol. 11, no. 1, mars 2006, pp.87-100.

Annexe A

Implémentation d'un VAD sur DSP

Cette annexe présente un travail qui a été réalisé conjointement dans le cadre d'un partenariat industriel. Le projet demandé consiste à implémenter un VAD sur un DSP en temps réel. L'algorithme de VAD n'a pas été décrit ni mentionné dans le mémoire. Pour cette raison, une description de l'algorithme sera présentée dans cette annexe. Nous l'avons appellé 'Jaber-VAD' en référence à M. Marwan Jaber, fondateur de la méthode. Par la suite, nous allons décrire l'environnement dans lequel a été implémenté cet algorithme (DSP, carte DSP, outils logiciels). Pour finir nous présenterons les résultats de simulations obtenus avec Matlab et sur DSP.

1. Description de l'algorithme

Le signal d'entrée *s(n)* est découpé en trames de longueur de 1024 échantillons (N=1024). À chaque trame, trois paramètres sont calculés :

- L'énergie de la trame (E):

$$E = 10\log_{10}\left(\sum_{n=0}^{N-1} s_k^2(n) + \varepsilon\right) \qquad (1.1)$$

où $s_k(n)$ est le signal d'entrée s à la trame k, ε est une constante dont la valeur est $2.2204 \cdot 10^{-16}$, et N la longueur de la trame.

- Le nombre de passage par zéro (Z) :

$$Z = \sum_{n=1}^{N-1} \left|\mathrm{sgn}\left(s_k(n)\right) - \mathrm{sgn}\left(s_k(n-1)\right)\right| \qquad (1.2)$$

où *sgn* correspond à la fonction signe.

- Le coefficient d'autocorrélation (R)

$$R = \sum_{l=-N}^{N} \left[\sum_{n=0}^{N-1} s_{k-1}(n) s_k(n-l) \right] \qquad (1.3)$$

Ces trois paramètres sont calculés à chaque trame à partir desquels deux seuils sont calculés. Les paramètres sont comparés à des seuils qui sont mis à jour regulièrement. Une décision finale est prise après cette série de calculs. La figure A.1 montre en détail l'algorithme. Le tableau A.1 présente les calcul des variables présenté à la figure A.1.

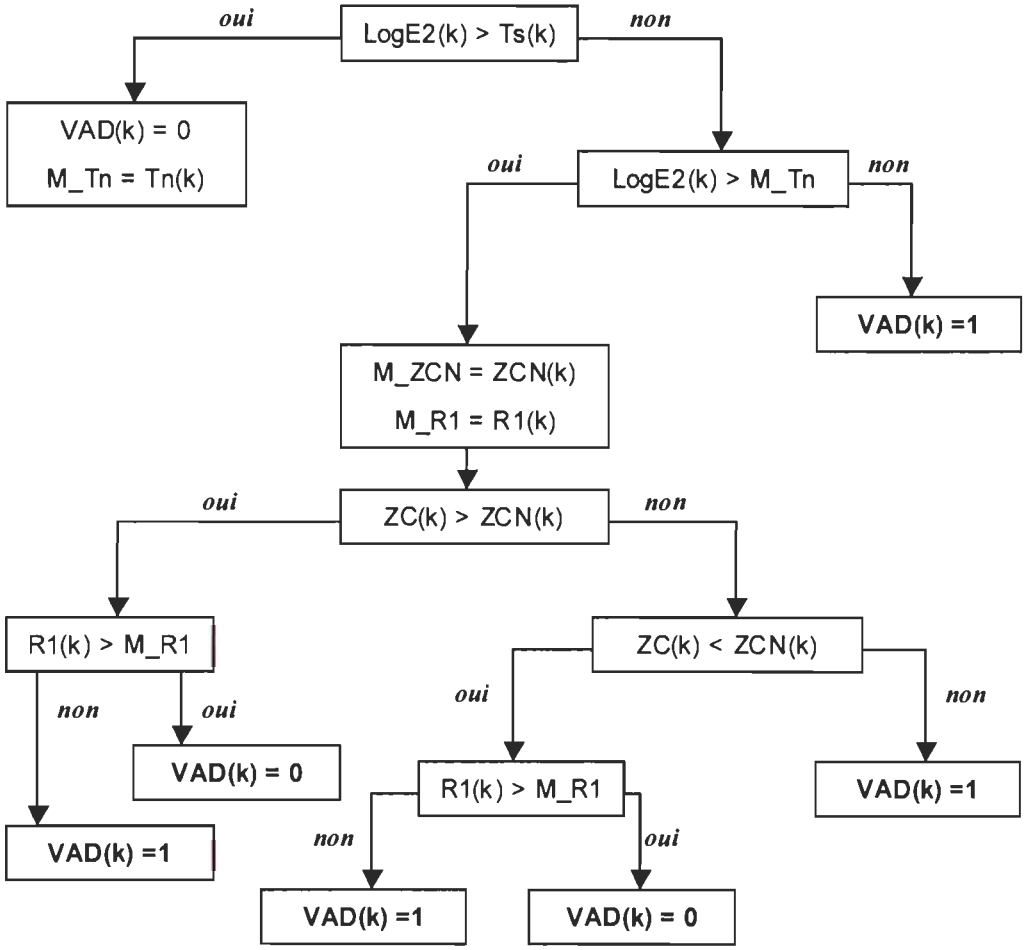


Figure A.1 Algorithme détailé du Jaber-VAD.

Tableau A.1. Équations des paramètres présentés à la figure A.1.

LogE1	$LogE1(k) = 10\log_{10}\left(\sum_{n=0}^{N-1} s_{k-1}^2(n) + \varepsilon\right)$
Ts	$Ts(k) = \sigma(LogE1)\cdot 0.3 + \mu(LogE1)$ où $\sigma(LogE1)$ et $\mu(LogE1)$ designent respectivement l'écart type et la moyenne des M dernières valeurs de *LogE1*. M est une constante determiné de manière empirique (M=10).
Tn	$Tn(k) = \sigma(LogE1)\cdot 0.1 + \mu(LogE1)$
ZCN	$ZCN(k) = \frac{ZC(s_{k-1}(n))}{256}$ où ZC désigne l'opération de 'zéro crossing'
R1	$R1(k) = \sum_{l=-N}^{N}\left[\sum_{n=0}^{N-1} s_{k-1}(n)s_k(n-l)\right]$
ZC	$ZC(k) = \frac{ZC(s_k(n))}{256}$
LogE2	$LogE2(k) = 10\log_{10}\left(\sum_{n=0}^{N-1} s_k^2(n) + \varepsilon\right)$

2. Description de l'environnement

2.1 Description du DSP

Le DSP sur lequel l'algorithme a été implémenté est un TMS320VC5416 à point fixe, fabriqué par la compagnie Texas Instrument. Il fait partie de la famille des C5000, située entre la famille C2000 et C6000. Les C5000 ont été conçu pour une haute efficacité, à des coups acceptable, tout en gardant une taille assez petite et une consommation faible pour permettre leurs applications dans de nombreux systèmes de communications actuels (Kit

mains libres, consoles de jeux, eBooks, PDAs, modems, etc.). Ces types de DSP sont conçu pour effectuer le plus de nombre d'opérations possible en un cycle d'horloge. L'utilisation du pipeline permet au processeur d'exécuter une instruction par cycle d'horloge. La vitesse d'horloge du DSP peut aller jusqu'à 160Mhz. Il possède 128Kmots de mémoire de données. D'autres caractéristiques de ce DSP sont présentés sont présentés au tableau A.2.

Tableau A.2. Caractéristiques du DSP 5416

CARACTERISTIQUES DU DSP5416	
Format des données	Point fixe
Mémoire de données	128kwords
Mémoire de programme	16Kwords
Vitesse d'horloge maximale	160Mhz
Vitesse d'instructions	120-160 MIPS
Puissance consommée	90mW (à 160 MHZ)
Prix	22.5 - 25$
Périfériques :	
Contrôleur DMA 6 canaux	
McBSP (Multi chanel Buffer Serial Port) 3 ports	
Timer	
HPI 8/16 bits	

2.2 Description de la carte

Le DSP est installé sur une carte DSK (DSP Starter Kit) spécialement conçue par Texas Instrument pour apprendre à démarrer avec les DSP. La communication entre l'ordinateur et la carte se fait à travers le port USB. Code Composer Studio (CCS) communique avec le DSP à l'aide d'un émulateur JTAG, intégré sur la carte et relié au port USB. La carte

possède des connecteurs audio d'entrée sortie (dont une entrée microphone) pour le traitement audio en temps réels. Des DEL et des interrupteurs sont également mis au profit de l'utilisateur.

2.3 Implémentation avec Code Composer Studio

L'étape d'implémentation consiste à programmer notre algorithme sur le DSP. Pour cela, on utilise Code Composer Studio, qui est un compilateur en langage C spécialement conçu pour programmer les DSP de Texas Instrument. La version de Code Composer Studio utilisée ici est une version spécialement destiné à la carte DSK5416. Code Composer Studio contient toutes les librairies nécessaires pour la programmation de la carte. Une fois le programme codé en langage C, on compile et on envoi le programme sur le DSP de la carte DSK5416, connectée à l'ordinateur par le port USB.

Le PCM3002 est une interface sur la carte DSK qui sert à communiquer entre les ports d'entrée et sortie audio et le DSP. Nous utiliserons les fonctions disponibles du codec relatifs au PCM3002 et fourni par CCS pour pouvoir utiliser les ports audio.

Pour gagner du temps de calcul, nous utiliserons certaines fonctions de la librairie DSPLIB. Cette librairie fournit une série de fonctions utiles dans le traitement numérique de signal (FFT, Convolution, FIR, autocorrélation, logarithme, exponentielle,etc.).

3. Simulation et résultats

Comme il s'agit d'une implémentation en temps réel, les résultats ont été obtenus en faisant jouer un son de 30 secondes de la base de données AURORA à l'aide d'un lecteur MP3 branché à l'entrée du port audio de la carte DSK. La figure A.2 montre le montage.

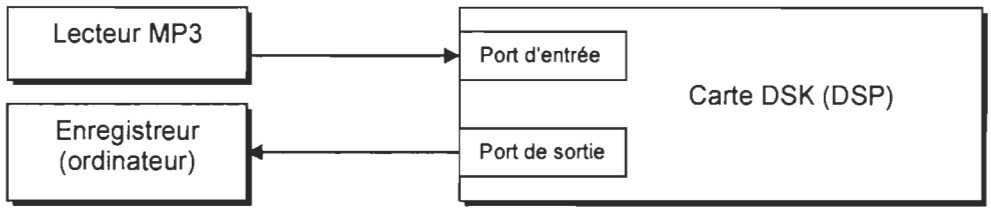


Figure A.2. Schéma du montage pour les simulations en temps réel.

La sortie a été enregistrée à partir du port de sortie audio pour pouvoir obtenir le VAD calculé par le DSP. Ainsi, le VAD obtenu avec MATLAB et le VAD obtenu avec le DSP sont comparés entre eux et avec le VAD idéal correspondant. Les tableaux A.3 et A.4 montrent l'ensemble des résultats et la figure A.3 montre un extrait du signal.

Tableau A.3. Pd et Nd obtenus avec Matlab et obtenus avec le DSP

	Pd (%)	**Nd (%)**
Matlab	85.73	81.26
DSP	89.15	78,03

Tableau A.4 Comparaison entre Matlab et DSP en terme de pourcentage d'erreur

Erreur dans les zones actives	Erreur dans les zones inactives
4,46 %	3,97

D'après le tableau A.3, les résultats pour le VAD-DSP sont très bons. Le *Pd* est élevé et *Nd* a un pourcentage très acceptable. Les résultats montrent que le VAD-DSP présente des performances similaires aux résultats obtenus avec Matlab. Le léger écart d'environ 4% entre les *Pd* et *Nd* (cf. tableau A.4) peut s'expliquer par le fait que l'algorithme est déjà en cours de processus lorsqu'on fait jouer le son avec notre lecteur MP3. Les variables initiales telles que *Ts* et *Tn* sont ainsi dejà mis à jour, contrairement dans le cas de la simulation Matlab où l'algorithme commence son processus au premier échantillon de *s(n)*.

En conclusion, on peut dire que l'implémentation sur DSP d'un algorithme de VAD a été réalisée. Cette première étape peut amener à d'autres études visant à implémenter des méthodes de VAD plus complexes dans le futur.

Annexe B

Spectogrammes des sons de bruits

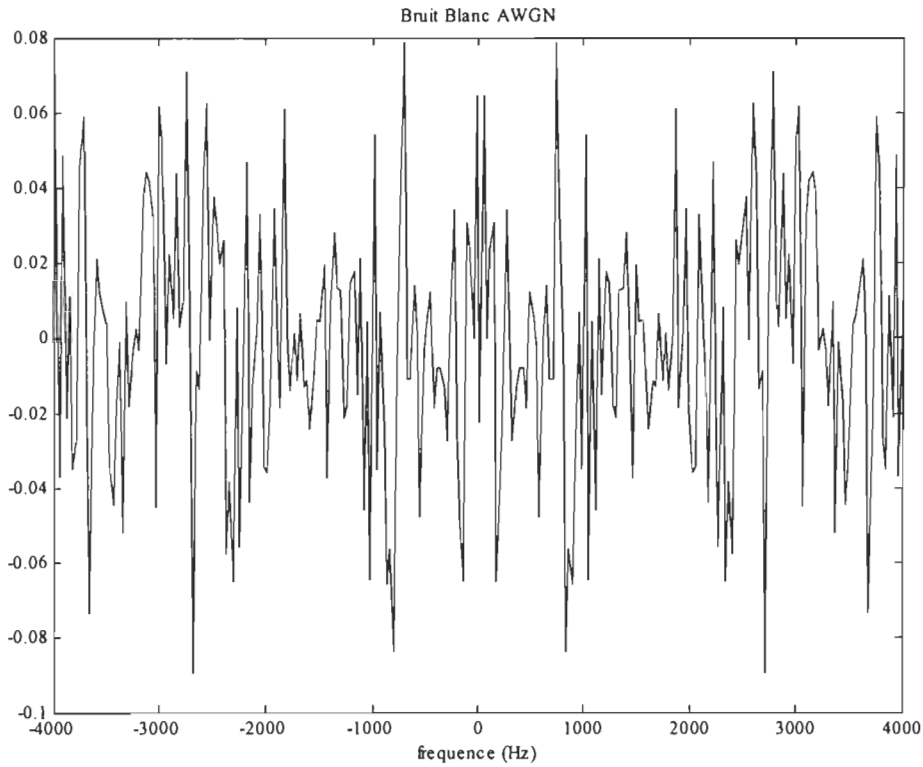


Figure B.1 Spectogramme du son « Bruit Blanc AWGN »

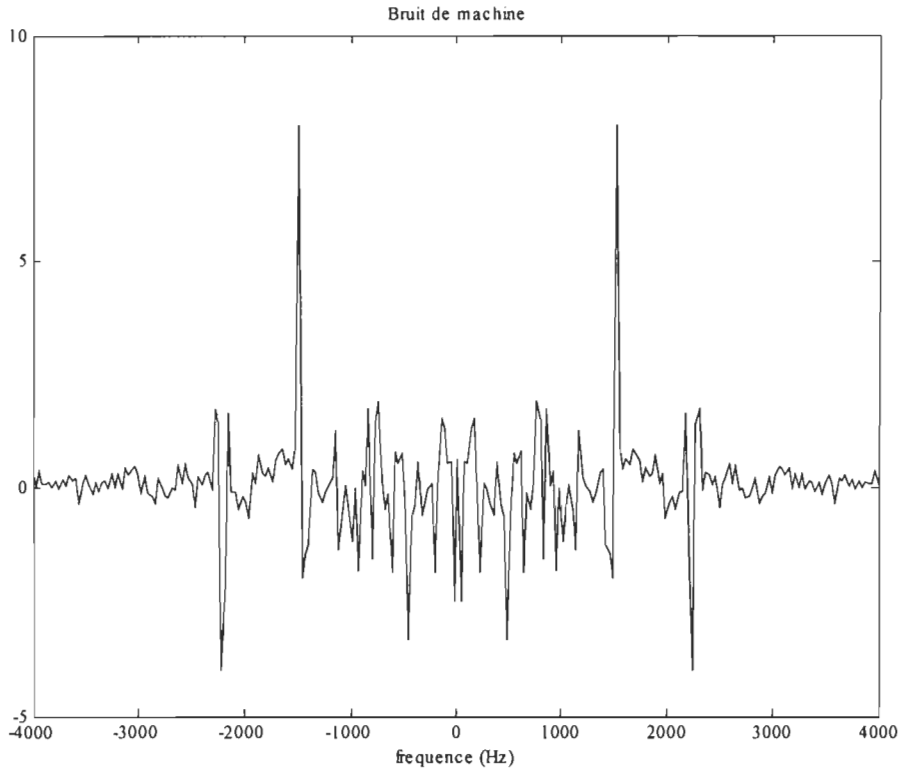


Figure B.2 Spectogramme du son « Bruit machine».

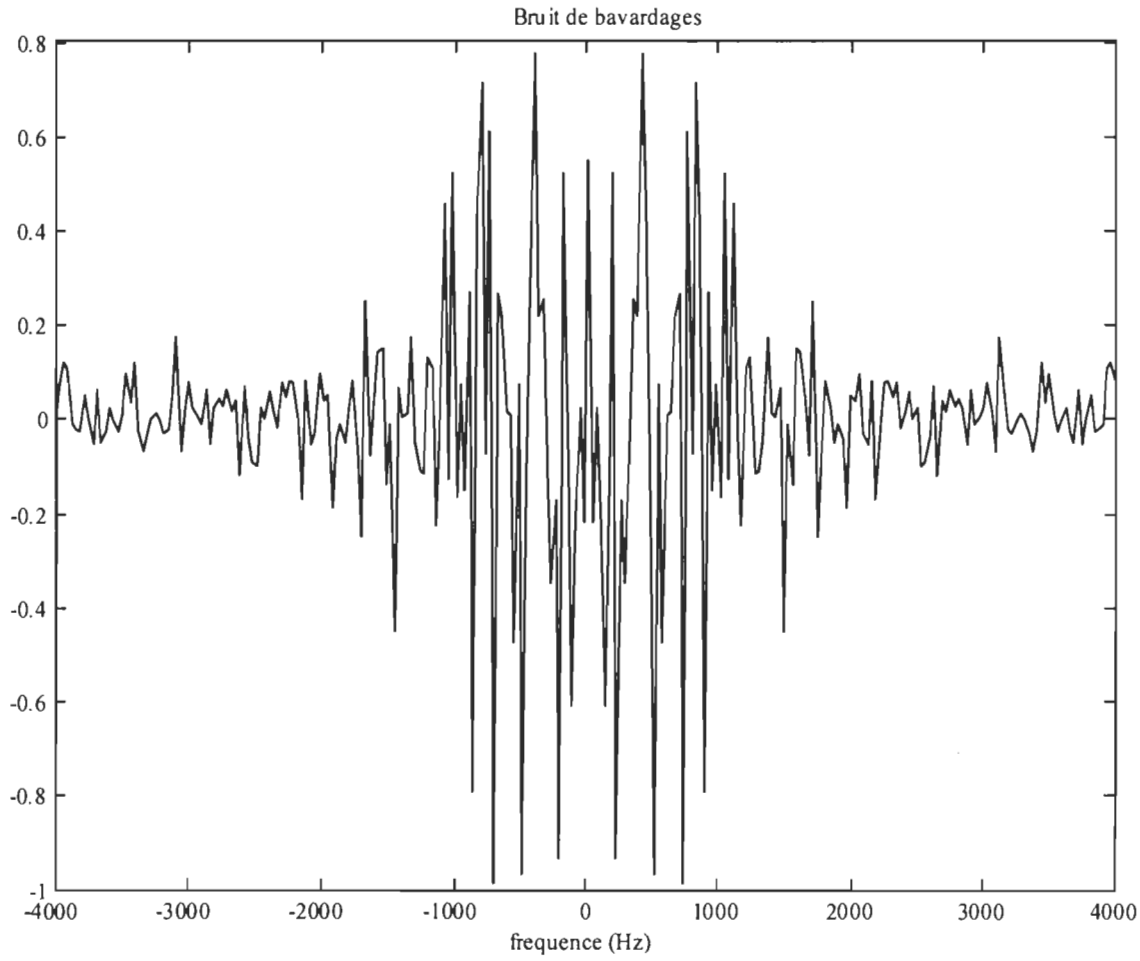


Figure B.3 Spectogramme du son « Bruit de bavardages » (« *Babble* »).

Annexe C

Analyse de la complexité

Dans cette annexe, chaque étape de calcul du tableau 3.1 est analysée. Le but est d'obtenir par une série de calcul mathématiques le nombre d'additions et le nombre de multiplications nécessaires à une étape de calcul. Ces nombres permettront de mieux analyser la complexités des méthodes décrites dans la section 3.5.

L'étape de décomposition en ondelettes se résume à une série de filtres FIR en cascades [MAL97]. L'équation suivante est un exemple de calcul de FIR d'ordre 3 :

$$y(n) = a_1 x(n) + a_2 x(n-1) + a_3 x(n-2) \tag{6.1}$$

avec $x(n)$ le signal d'entrée du filtre, $y(n)$ la sortie, et a_1, a_2, et a_3 les coefficients. Pour cet exemple, on remarque qu'on a 3 multiplications et 2 additions. Si l'on généralise en posant *d* comme étant l'ordre du filtre, on obtient *d* multiplications et *d-1* additions.

La complexité d'une décomposition va ainsi dépendre du nombre de niveaux de décomposition ainsi que de l'ordre des coefficients d'ondelettes (exemple : Daubechies d'ordre 10, 20 ou 45). Pour le VAD de Wu et Wang, on a une décomposition en trois niveaux. Pour le premier niveau, on a un filtrage, suivi de deux filtrages pour le deuxième niveau puis encore deux autres filtrages pour le troisième niveau. Si l'on pose *d* l'ordre des coefficients d'ondelettes, et *N* le nombre d'échantillons par trame, on obtient le nombre de multiplications et d'additions pour une décomposition pour une trame :

$$\begin{aligned}
&\frac{N}{2^0}(d\otimes+(d-1)\oplus)+2\cdot\frac{N}{2^1}(d\otimes+(d-1)\oplus)+2\cdot\frac{N}{2^2}(d\otimes+(d-1)\oplus)\\
&=\left(\frac{N}{2^0}+2\cdot\frac{N}{2^1}+2\cdot\frac{N}{2^2}\right)(d\otimes+(d-1)\oplus)\\
&=\frac{10}{4}N(d\otimes+(d-1)\oplus)
\end{aligned} \tag{6.2}$$

où $\otimes$ désigne une opération de multiplication et $\oplus$ une opération d'addition. À chaque fois qu'on descend d'un niveau, le nombre de données à filtrer est divisé par deux (cf. section 3.1.2 pour plus de détails).

La méthode de Chen et Wang et notre méthode décomposent le signal de base $s_T(n)$ en 17 sous-signaux (figure 3.4, cf. sections 3.2.2 et 3.3). L'utilisation de la décomposition en ondelettes sur 5 étages rend la décomposition plus complexe en terme de calculs. Selon le principe de FIR et la figure 3.4, le nombre de multiplications et d'additions pour une décomposition en ondelettes selon la méthode de Chen et Wang (ou notre méthode proposée) est défini comme suit :

$$\begin{aligned}
&\left(2\cdot\frac{N}{2^0}+4\cdot\frac{N}{2^1}+8\cdot\frac{N}{2^2}+10\cdot\frac{N}{2^3}+8\cdot\frac{N}{2^4}\right)(d\otimes+(d-1)\oplus)\\
&=\frac{59}{8}\cdot N(d\otimes+(d-1)\oplus)
\end{aligned} \tag{6.3}$$

La synthèse (recomposition) est l'étape inverse de la décomposition en ondelettes. D'un point de vu calcul, le principe de FIR est appliqué comme pour la décomposition sauf que les coefficients des filtres changent. Seul la méthode de Chen et Wang utilise la synthèse. Lorsqu'on regarde l'arbre de la figure 3.4, on a 8 sous-signaux au niveau 5 à recomposer et qui vont former 4 sous-signaux au niveau 4. Par la suite, ces 4 sous-signaux

additionnés au 6 sous-signaux vont donner un total de 10 sous-signaux au niveau 4 à recomposer, pour donner à leur tour 5 sous-signaux au niveau 3, ainsi de suite jusqu'à arriver à un seul signal de longueur N (niveau 0). Le nombre d'additions et de multiplications est donné par l'équation suivante :

$$\begin{aligned}&\left(8\cdot\frac{N}{2^5}+10\cdot\frac{N}{2^4}+8\cdot\frac{N}{2^3}+4\cdot\frac{N}{2^2}+2\cdot\frac{N}{2^1}\right)\left(d\otimes+(d-1)\oplus\right)\\&=\frac{31}{8}\cdot N\left(d\otimes+(d-1)\oplus\right)\end{aligned} \tag{6.4}$$

Il est à noter que les opérations de sous-échantillonnage (*downsampling)* et de sur-échantillonnage (*up-sampling)* ne sont pas comptabilisées, que ce soit pour la décomposition ou la synthèse, pour la simple raison que ces opérations ne nécessitent pas de complexité d'implémentation des calculs. Il faut se garder comme objectif qu'il s'agit d'une étude grossière mais suffisante de la complexité.

Le **TEO** tel que définit dans l'équation 3.8 se traduit en nombre d'additions et de multiplications par la relation suivante :

$$\left(\frac{N}{2^j}-2\right)\cdot\left(2\cdot\otimes+1\cdot\oplus\right) \tag{6.5}$$

où j définit le niveau de l'étage du sous-signal auquel on applique le TEO, et N la longueur de la trame.

Pour la méthode de Wu et Wang on a 4 sous-signaux (un pour le niveau 1, un pour le niveau 2, et 2 pour le niveau 3). On obtient la complexité suivante :

$$\left[\left(\frac{N}{2^1}-2\right)+\left(\frac{N}{2^2}-2\right)+2\left(\frac{N}{2^3}-2\right)\right]\cdot(2\cdot\otimes+1\cdot\oplus)$$
$$=(N-6)\cdot(2\cdot\otimes+1\cdot\oplus) \tag{6.6}$$

Pour la méthode de Chen et Wang et notre méthode proposée, on a 17 sous-signaux soit :

$$\left[3\cdot\left(\frac{N}{2^3}-2\right)+6\cdot\left(\frac{N}{2^4}-2\right)+8\cdot\left(\frac{N}{2^5}-2\right)\right]\cdot(2\cdot\otimes+1\cdot\oplus)$$
$$=(N-34)\cdot(2\cdot\otimes+1\cdot\oplus) \tag{6.7}$$

La complexité du calcul de **moyenne** dépend de la longueur L du signal :

$$(L-1)\oplus+1\cdot\boxed{\div} \tag{6.8}$$

La **variance** est plus complexe que la moyenne car son calcul nécessite le calcul de la moyenne au préalable :

$$(L-1)\oplus+L\cdot\oplus+1\cdot\boxed{\div}+1moyenne$$
$$=(L-1)\oplus+L\cdot\oplus+(L-1)\oplus+2\cdot\boxed{\div} \tag{6.9}$$
$$=(3L-2)\oplus+2\cdot\boxed{\div}$$

Appliqué aux 17 sous-signaux de la méthode de Chen et Wang (et notre méthode proposée), on obtient la complexité suivante pour le calcul des variances des 17 sous-signaux :

$$
\begin{aligned}
&3\left[\left(3\frac{N}{2^3}-2\right)\oplus+2\cdot\boxed{\div}\right]+6\left[\left(3\frac{N}{2^4}-2\right)\oplus+2\cdot\boxed{\div}\right] \\
&+8\left[\left(3\frac{N}{2^5}-2\right)\oplus+2\cdot\boxed{\div}\right] \\
&=(3N-2)\oplus+34\cdot\boxed{\div}
\end{aligned}
\tag{6.10}
$$

La méthode de Chen et Wang calcule la variance pour chaqu'un des 17 sous-signaux puis la compare à un seuil dont la valeur dépend de l'étage (cf. section 3.2.2). À chaque trame, trois seuils sont calculés (niveaux 3, 4, 5) selon l'équation 3.18. Il s'agit d'un calcul de variance de longueur N (longueur de la trame), multiplié par une constante de valeur fixe et égale à $2\log(N)$. Le calcul de cette constante n'est pas comptabilisé dans l'étude de la complexité puisque les constantes sont généralement calculées une seule fois au moment de l'initialisation du programme. L'étape de comparaison entre la variance et le seuil n'est pas non plus comptabilisée compte tenu qu'il s'agit d'une opération à faible complexité.

La méthode de Wu et Wang utilise deux formules qui demandent un certains nombre d'additions et de multiplications supplémentaires : **l'auto-corrélation** (Éq. 3.10) et la **méthode de la moyenne des deltas** (Éq. 3.12).

L'autocorrélation se traduit en termes d'additions et de multiplications par la fonction suivante :

$$
[L\otimes+L\oplus]\cdot(2L-1) \tag{6.11}
$$

avec L longueur du signal d'entrée variable selon le niveau j de décomposition du sous-signal, définit par $L = \frac{N}{2^j}$.

Appliqué aux quatre sous-signaux, on obtient la complexité de calcul suivante :

$$\begin{aligned}
&\left(\frac{N}{2^1}\otimes+\frac{N}{2^1}\oplus\right)\left(2\frac{N}{2^1}-1\right)+\left(\frac{N}{2^2}\otimes+\frac{N}{2^2}\oplus\right)\left(2\frac{N}{2^2}-1\right)+\\
&2\left(\frac{N}{2^3}\otimes+\frac{N}{2^3}\oplus\right)\left(2\frac{N}{2^3}-1\right)\\
&=\frac{11}{16}N^2(\oplus+\otimes)-\frac{N}{16}(\oplus+\otimes)\\
&=\left(\frac{11}{16}N^2-\frac{N}{16}\right)(\oplus+\otimes)
\end{aligned} \tag{6.12}$$

La fonction des deltas (Éq. 3.12) peut se simplifier en prenant compte que la dénominateur est une constante calculée initialement. De plus, en calculant initialement l'inverse de cette constante, on évite une division. On obtient finalement la complexité suivante :

$$(2L-1)\cdot[M\oplus+M\otimes]+1\cdot\otimes \tag{6.13}$$

où M étant une constante définie dans l'équation (3.12), et L la longueur du signal d'entrée, qui correspond au signal de sortie de la fonction d'autocorrélation définie auparavant. Dans le cas d'une fonction d'autocorrélation, la sortie d'un signal d'entrée de longueur L est de longueur *2L-1* [PRO96, *chap. 2*]. La longueur L étant ici $L = \frac{N}{2^j}$, l'entrée de la fonction

delta sera $2\frac{N}{2^j}-1$. Ainsi, la complexité de la fonction delta pour les quatre sous-signaux est définie par :

$$\begin{aligned}&\left[\left(2\frac{N}{2^1}-1\right)+\left(2\frac{N}{2^2}-1\right)+2\cdot\left(2\frac{N}{2^3}-1\right)\right]\cdot\left[M\oplus+M\otimes\right]+1\cdot\otimes\\&=(2N-3)\cdot\left[M\oplus+M\otimes\right]+1\cdot\otimes\end{aligned} \tag{6.14}$$

L'opération suivante est un calcul de **moyenne** des quatre signaux de sortie de la fonction delta. La longueur de sortie pour chaque sous-signal est identique à sa longueur d'entrée soit $2\frac{N}{2^j}-1$. On obtient ainsi la complexité pour le calcul de moyenne des quatre signaux :

$$\begin{aligned}&\left[\left[\left(2\frac{N}{2^1}-1\right)-1\right]\oplus+1\cdot\boxed{\div}\right]+\left[\left[\left(2\frac{N}{2^2}-1\right)-1\right]\oplus+1\cdot\boxed{\div}\right]\\&+2\cdot\left[\left[\left(2\frac{N}{2^3}-1\right)-1\right]\oplus+1\cdot\boxed{\div}\right]\\&=(2N-6)\oplus+4\cdot\boxed{\div}\end{aligned} \tag{6.15}$$

Annexe D

Publications

Voice Activity Detection Based on Wavelet Packet Transform in Communication Nonlinear Channel

Roberto CHIODI and Daniel MASSICOTTE

Université du Québec à Trois-Rivières, Electrical and Computer Engineering Department
Laboratory of Signal and System Integrations, Trois-Rivières, Canada
{roberto.chiodi, daniel.massicotte}@uqtr.ca

Abstract— **This paper presents a voice activity detection (VAD) algorithm based on the Wavelet Packet Transform and the Teager Energy Operation (TEO) processing. The signal is decomposed into subband signals. We used the multi-resolution analysis property of the Wavelet Transform to extract and analyse time-frequency components corresponding to speech. In order to obtain a parameter called Voice Activity Shape (VAS), we used TEO processing to better distinguish subband signals corresponding to speech. The subband variance values of each TEO signal are summed to obtain the VAS, which is higher in speech regions than in non speech regions. Experimental results show that our VAD perform better than the G729B, particularly in difficult noise conditions and also in the case when the speech sound is passed in a nonlinear communication channel. Experimental results are shown in the case of real speech communications from a spaceship to terrestrial 3G cellular network assuming nonlinear interferences.**

Index Terms—**Speech processing, Voice Activity Detection, Wavelet Transform.**

I. INTRODUCTION

Voice Activity Detection (VAD) refers to the process which consist of separating voice to non-voice regions in a speech signals. Voice activity detection is required in many speech processing applications like speech coders used in cellular communications, speech enhancement, or speech recognition. For example, in the case of speech coders, high bit rates are sent during voice activity, while low bit rates are transmit during non-voice periods.

Most of the conventional VAD [3],[15],[16] algorithms use parameters which are extracted and computed from the speech signal. After that, thresholds techniques are used to compare a fixed or adaptive threshold with parameters. These VAD perform well in the case of clean speech signal, but difficulties occur in the case of speech corrupted with noise. In difficult environment with high noisy and non-stationary noise, determining what is speech or not is a difficult task. Another example of a difficult environment: space speech communications. Astronaut have to communicate each other and with the earth station. However, the speech is corrupted by interferences, or echo providing from the astronaut spaceship. Under these conditions, detecting voice in a speech signal is a difficult task.

Recently, different VAD algorithms have been developed : VAD based on neural networks [4], VAD based on fuzzy logic [5], VAD using single or several microphones (microphone arrays) [6],[7]. Wavelet Transform is a signal processing technique based on time-frequency signal analysis [8]. In literature, during the past few years, we find some proposed VAD algorithms based on Wavelet Transform (WT) or Wavelet Packet Transform (WPT) (e.g. [1],[2]). It has been demonstrated that WT is a powerful technique to solve VAD problems.

Generally, VAD based on WT applied the following steps: at each frame, the signal is decomposed in S subband signals by using the WT. For each subband, the Teager Energy Operator (TEO) [10], a powerful non-linear operator, is calculated. After that, Voice Activity Shape (VAS) [1] operator is applied to each TEO signals. The VAS operator is one of the key steps to discriminate the speech and non speech regions. A simple adaptive or non adaptive threshold technique is finally used to correctly select the speech regions.

The present paper proposes i) a VAD algorithm based on the WPT such as [2] where a variant is applied for the VAS, and ii) a nonlinear communication channel is considered to show the robustness of VAD-WT approach. The proposed VAD is compared with another VAD-WT algorithm [2] and the standard ITU VAD G729B [3]. The VADs are tested with sounds providing from the AURORA database [18].

In addition, real speech signals from a spaceship to terrestrial cellular network corrupted with interferences are used to determine which VAD perform better.

The paper is organized as follows; Section 2 described the VAD based on wavelet transform while Section 3 defines the discrete wavelet transform operation. Section 4 provides the proposed VAD method. Section 5 provides experimental results including real world space data communications. Finally, Section 6 reports the conclusions.

II. VAD BASED ON WAVELET TRANFORM

A. Discrete Wavelet Transform

By using Wavelet Transform, it is possible to extract the desired time-frequency components of a signal corresponding to speech. After the wavelet decomposition, some subband signals corresponding to speech frequency domain can be analysed. After that, each chosen subband can be analysed.

The Discrete Wavelet Transform (DWT) is the WT application on discrete signals. In 1989, Mallat [9] developed an efficient way to implement the DWT using filter banks. As shown in the Fig. 1a, mirror filters are used to decompose the signal (analysis WT process). The approximation and detail coefficients are obtained by using high-pass filter and low pass filter, followed with a downsampling operation. Here the discrete equations corresponding to this filter operation [8]:

$$a(k) = \sum_{n=1}^{N} g(n-2k)s(n) \quad (1)$$

$$d(k) = \sum_{n=1}^{N} h(n-2k)s(n) \quad (2)$$

978-0-7695-3694-1/09 $25.00 © 2009 IEEE
DOI 10.1109/SPACOMM.2009.28

where, n is the discrete time, $s(n)$ is the input speech signal, $g(n)$ and $h(n)$ correspond respectively to the low pass and high-pass filter coefficients. The values of the filter coefficients depend of the nature of the wavelet. Most of the VAD algorithms using DWT use the Daubechies wavelet filters because they preserve frequency selectivity as the wavelet decomposition level increases [13].

The DWT is implemented by cascading these conjugate mirror filters (Fig.1a) [8]. A wavelet decomposition tree is presented at Fig. 2 and corresponds to the wavelet packet transform decomposition. This filter approach makes the DWT suitable for real time applications. The inverse DWT consist to synthesis WT process shown in Fig. 1b.

B. *VAD using DWT*

1) *VAD based on auto-correlation function using Wavelet*

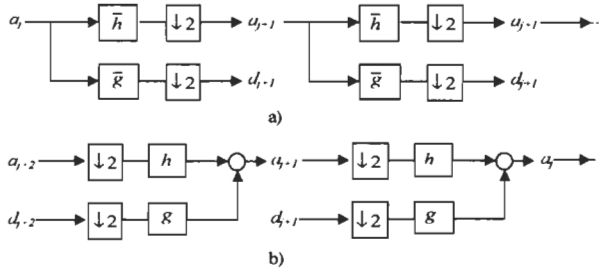


Fig. 1 Discrete Wavelet Transform (DWT) using filter banks for analysis (a) and synthesis (b).

Transform and Teager Energy Operator (TEO)

The VAD algorithm of Wu and Wang [2] decompose into subband signals. From Fig. 2, we considered the right side of tree wavelet decomposition where we keep one wavelet coefficient by level j. Therefore, 4 subband signals are obtained. The goal is to determine the periodicity of each subband because the periodic property is an inherent characteristic of speech signals. To determine the amount of periodicity, the TEO parameter is calculate for each subband signal. The TEO is a powerful non-linear operator [10], [11] which can enhance stable or half stable signal and decay transient or unstable signal. For a discrete signal $y(n)$, the discrete TEO is given by [10]

$$\psi[y(n)] = y^2(n) - y(n-1)y(n+1), \tag{3}$$

where $\psi[y(n)]$ is called the TEO coefficient of $y(n)$.

After that, the autocorrelation function is computed with the delta function [2] as

$$\bar{R}_M(k) = \frac{\sum_{m=-M}^{M} mR(k+m)}{\sum_{m=-M}^{M} m^2}, \tag{4}$$

where $\bar{R}_M$ is the result of the delta function over an M – sample neighbourhoods. The delta function is used to detect the amount of periodicity. The average of each delta function results is computed and finally the 4 values are summed. The result is a time varying curve which grows up in voice periods. An adaptive threshold is applied to the curve to determine the VAD output [2].

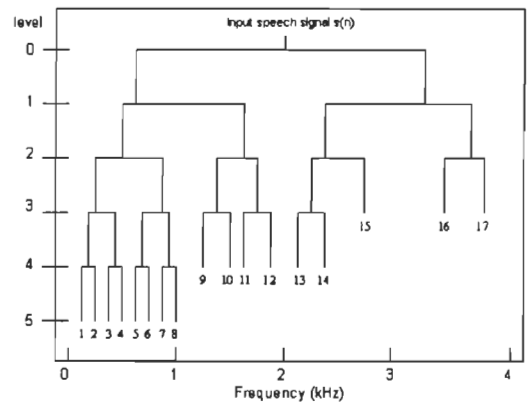


Fig. 2 Decomposition tree structure of the Wavelet Packet Transform for j=1,2,…,5.

2) *VAD using Wavelet Packet Transform*

The WPT is the general form of the DWT. Instead of decomposing only the approximated coefficients, every subband is decomposed. The idea proposed in [1] is to choose special subbands containing voice components. This method decomposes the signal into S subbands [12]. Fig. 2 shows the decomposition tree.

In this example, the signal is decomposed into frames of 256 points. At each frame, 17 subband signals are obtained as shown in Fig. 2. After that, TEO is calculated for each subband signal. The next step is to select the subband signals that contain voice. The variance of each TEO-subband is computed and compared to a level-dependent threshold λ_j:

$$\lambda_j = \frac{\sigma_j \sqrt{2\log(N)}}{N} \sum_{n=1}^{N} \frac{1}{S_n}, \tag{5}$$

where σ_j is the variance of all the subband signals with level j and N the frame size. If the variance is greater than λ_j the signal is selected. An Inverse Wavelet Transform (synthesis) is computed to obtain the Voice Activity Shape (VAS). Finally, an adaptive weighted threshold (AWT) is used to determine which parts corresponds to voice.

III. THE PROPOSED VAD ALGORITHM

The VAD algorithm based on WPT can be decomposed in 4 steps as shown in Fig. 3.

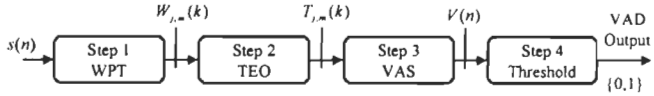


Fig. 3 Scheme of the proposed VAD

Step 1 Wavelet Decomposition

As explained in previous section, the speech signal $s(n)$ is decomposed into frames of 256 samples. This choice of the size depends of which frequency range we want to analyse and the sampling frequency. If we want more information in low frequencies, the frame must be bigger and inversely if we want analyse high frequencies the frame size will be smaller. For

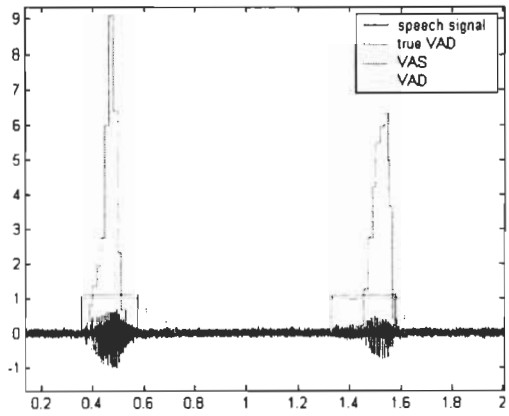


Fig. 4 VAS result (red line) of our VAD method for a speech French sentence '*quatre – cinq*' (blue line) corrupted with a heavy noisy machine at SNR=10dB. The VAD result is compared to the ideal hand-labelled VAD (green line).

speech application 256 is a good range at a sampling frequency of 8kbps.

Using the WPT each frame is decomposed into S subband signals (S=17). As in [1], we use the same decomposition tree (Fig. 2), and the choice of the 'mother' wavelet is a 10 points Daubechies. To decompose the signal, we use the filter banks corresponding to the 10 points-Daubechies, applied to the equations (1) and (2). We obtained 17 signals of different size (depending of the level decomposition) namely $W_{j,m}(k)$ where j is the decomposition level (frequency scale), $j=1,2,...,2^J$ (J=8), m the index of the subband signal ($1 \le m \le S$), and k the index of the coefficients $k=1,2,...,2^j$. The decomposition level, j, represents the frequency ranges of interests to detect the speech or non-speech frame. In our study we have considered $3 \le j \le 5$. Therefore, for j=5 $1 \le m \le 8$, for j=4 $9 \le m \le 14$, and for j=3 $15 \le m \le 17$.

Step 2 TEO operation

The goal of TEO operation is to determine the periodicity of each subband signal. As in [1],[2], the TEO is computed for each subband using the equation (3):

$$T_{j,m}(k) = \psi\left[W_{j,m}(k)\right]. \qquad (6)$$

This operation helps to detect the periodicity shape and decay untransient and unperiodic signals.

Step 3 VAS extraction

After the TEO operation, we propose to compute the variance of each TEO signal $T_{j,m}(k)$ and compute the summation

$$V(n) = \sum_{m=1}^{S} \mathrm{var}\left(T_{j,m}(k) \middle| k = 1,2,...2^j\right), \qquad (7)$$

for n=1,2,...,N and where var(•) is the variance operator. At each frame is assigned a value $V(n)$. The result is a VAS curve whose characterises the speech and non-speech evolution in the observed signal $s(n)$. The value of $V(n)$ is high in voice periods and low in non-voice periods. Based on Eq. (6), our propose VAS extraction has a low complexity level compared to the Wang's method. Wu and Wang's VAD has only three decomposition levels, but the algorithm follows with a autocorrelation and a mean-delta method. VAS output result based on Eq. (6) is shown in Fig. 4. We can observe that the VAS values grow up in speech activity regions.

Step 4 Threshold Decision

To determine speech and non-speech frames, a simple threshold algorithm is used. Considering that the first 10 frames are non-speech signal, an initial threshold is calculated. This initial value corresponds of the maximum of the 10 first VAS values.

IV. Experimental Results

In this section, we was studied the VAD quality proposed in comparison with the well known ITU standard G729B [3] and the Wu and Wang's VAD [2]. We have tested our VAD using linear and nonlinear communication voice channel and real world space communication.

To evaluate the performance of the VAD, the *Pd* and *Nd* parameters are utilized. *Pd* is the probability of correct decision in voice frames and *Nd* is the probability of correct decision in inactive frames. The greater *Pd* is, the more the voice will be understandable. Inversely, a high *Nd* value means that noise has correctly been rejected.

The speech signals are selected from the AURORA database [18], sampled at 8 kHz and 16-bit resolutions. The signals are corrupted with real heavy noisy from manufacture machine, at different signal to noise ratio (SNR) levels. The frame length is fixed at 256 samples per frame, the overlapping size is 128 samples, S=17, and $3 \le j \le 5$ for WT decomposition.

Table 1 compares the performance of the proposed VAD with the reference methods under different specific SNR values. To give a better idea of the VAD, the mean of *Nd* and *Pd* is calculated. From Table 1, it can be seen that in terms of the average of correct and false speech detection probability, our proposed VAD is superior to Wu-Wang's VAD [2] and G729B. Particularly in poor SNR conditions, our VAD outperform the G729 VAD, which considers most of frames as non-active, which explain a high *Nd*.

To evaluate the robustness of VAD methods to nonlinear communication channel, we corrupted the original speech signal, $s(n)$, by passing it through this nonlinear function

$$\tilde{s}(n) = 0.5s(n) - 0.25s^3(n-1), \qquad (8)$$

where $\tilde{s}(n)$ represents the nonlinear speech signal apply to the

TABLE 1 Pd (%) and Nd (%) of the proposed VAD, Wang's VAD, and the standard ITU G729B VAD for different SNRs.

	G729		Wu & Wang [2]		Proposed	
SNR (dB)	*Pd*	*Nd*	*Pd*	*Nd*	*Pd*	*Nd*
0	14	86.62	23.17	83.7	37.75	80.4
2.5	38.6	85.8	53.53	81.24	67.61	78.4
5	55.4	87.3	71.55	78.03	77.5	75.2
10	66.7	87.4	81.21	77.42	86.5	70.76
Average	43.7	86.8	57.4	80.1	67.3	76.2

input of the first step, Wavelet Decomposition step, at Fig. 3. This signal is a saturated speech signal unintelligible.

Table 2 gives results for the speech signal passed through nonlinear function (8). The results for the clean version (without nonlinear function) of the same speech signal are given for comparison at column 'clean speech'. We observe that the VADs based on wavelet (proposed VAD and Wang's VAD) perform very well in the two cases (clean signal and non-linear). This is due to the periodic properties detected by WT approach which are not affected by the distortion and the nonlinearity.

Finally, we have evaluated our proposed VAD using a speech signal of 60 seconds providing from a spaceship to terrestrial 3G cellular network corrupted by a heavy noisy environment assuming nonlinearity. The sound is naturally corrupted by interferences, so it is difficult to understand the voice conversation. As we observed, from Table 3, that the VADs have also difficulties to detect the speech regions. The G729 has a very high *Pd*, but a very low *Nd*, which means that most of frames are considered as speech frames. Our VAD is most stable with a respectable *Pd* and *Nd* values.

TABLE 2 Pd (%) and Nd (%) of the proposed VAD compared to reference methods for speech signal passed through a nonlinear function Eq. (8).

	Linear channel			Nonlinear channel		
VAD	*Pd*	*Nd*	Avg	*Pd*	*Nd*	Avg
G729	92.31	74.78	83.55	59.87	99.19	79.53
Wu & Wang [2]	89.94	81.86	85.9	88.47	91.03	89.75
Proposed	85.71	92.26	88.99	91.33	89.01	90.17

ACKNOWLEDGMENT

The authors would like to thank Mr. Marwan Jaber from JaberTech Canada inc. for technical supports and to provide the space communication data making the project possible, and the Natural Sciences and Engineering Research Council of Canada.

TABLE 3 Pd (%) and Nd (%) of the proposed VAD for a 60 seconds of speech signal from a spaceship to terrestrial 3G cellular network.

VAD	*Pd*	*Nd*	Average
G729	95.15	22.46	58.805
Proposed	76.2	52.09	64.15

V. CONCLUSION

This paper present a VAD algorithm based on the Wavelet Packet Transform. The results show that our VAD perform better than the G729B IUT standard VAD, particularly in environments with high SNRs. When we compared our VAD with the Wu and Wang's Wavelet-based VAD, we observe that our VAD perform better. In terms of complexity, our VAD is less complex. It only needs a wavelet decomposition which can easy be computed with mirror filters followed with a TEO function.

In the future work, a better threshold decision method could be developed. A decision based on the analysis of the VAD line could more accurately detect when speech starts and ends.

REFERENCES

[1] S.-H. Chen, H.-T. Wu, C.-H. Chen, J.C Ruan, T.K. Truong, "Robust Voice Activity Detection Algorithm Based on The Perceptual Wavelet Packet Transform", IEEE Int. Symp. on Intelligent Signal Processing and Communication Systems, Hong Kong, China, Dec. 2005, pp.45-48.

[2] B.-F. Wu, K.-C. Wang, "Voice Activity Detection based on Auto-Correlation Function Using Wavelet Transform and Teager Energy Operator", Computational Linguistics and Chinese Language Processing, vol. 11, no. 1, March 2006, pp.87-100.

[3] ITU-T Recommendation G.729 "Annex B: A Silence Compression Scheme for Use with G.729 Optimized for V.70 Digital Simultaneous Voice and Data Applications", IEEE Communications Magazine, 35(9), 1997, pp.64-73.

[4] Kim and Park, "Voice Activity Detection Algorithm Based on Radial Basis Function Network", IEICE Trans. Communication, vol. E88–B, no. 4, April 2005, pp. 1656-1657.

[5] F. Beritelli, S. Casale, and A. Cavallero, "A Robust Voice Activity Detector For Wireless Communications Using Soft Computing", IEEE J. Select. Areas Communication, vol.16, Dec. 1998, pp. 1818-1829.

[6] M. W. Hoffman, Z.Li, D. Khataniar, "GSC-Based Spatial Voice Activity Detection for Enhanced Speech Coding in The Presence of Competing Speech", IEEE Transactions on speech and audio processing, vol. 9, no. 2, March 2001, pp. 175-178.

[7] J.E. Rubio, K. Ishizuka, H. Sawada, S. Araki, T. Nakatani, and M. Fujimoto "Two Microphones Voice Activity Detection Based on The Homogeneity of The Direction of Arrival Estimates", IEEE Int. Conf. on Acoustic, Speech and Signal Processing, Honolulu, April 2007, pp. 385-388.

[8] S. Mallat, "A wavelet tour of signal processing", Wiley, 1997.

[9] S. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, 1989, pp.674-693.

[10] J.F Kaiser, "On A Simple Algorithm To Calculate The 'Energy' Of A Signal", IEEE Int. Conf. on Acoustic, Speech and Signal Processing, Albuquerque, New Mexico, vol. 1, April 1990, pp.381-384.

[11] F.A. Jabloun, E. Cetin, and E. Erzin, "Teager Energy Based Feature Parameters For Speech Recognition In Car Noise", IEEE Signal Processing Letters, Vol. 6, no. 10, 1999, pp.256-261.

[12] L. Babiner and B.H. Juang, Fundamental of Speech Recognition, Upper Saddle River, Prentice-Hall, 1993.

[13] D. Sinha and A.H. Tewfit, "Low Bit Rate Transparent Audio Compression Using Adapted Wavelet", IEEE Trans. on Signal Processing, vol.41, Dec.1993, pp.1170-1183.

[14] I. Daubechies, Ten Lectures on Wavelets, CBMS, SIAM, publ.,1992.

[15] F. Beritelli, S. Casale, G. Ruggeri, S. Serrano, "Performance Evaluation and Comparaison of G729, AMR, Fuzzy VAD", IEEE Signal Processing Letters, March 2002, pp. 85-88.

[16] K. El-Maleh and P. Kabal, "Comparaison of Voice Activity Detection Algorithms for Personal Communications Systems", IEEE Canadian Conf. on Electrical and Computer Engineering, St-John's, May 1997, pp. 470-473.

[17] S.-H. Chen, J.-F. Wang, "A Wavelet-Based Voice Activity Detection Algorithm in Noisy Environments", IEEE Int. Conf. on Electronics, Circuits and Syst., Dubrovnik, Croatia, vol.3, Sept. 2002, pp.995-998.

[18] G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems unders Noisy Conditions", Proc. of Interspeech, Beijing, China, vol.1, 2000, pp.341-344.