ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

COMME EXIGENCE PARTIELLE
À L'OBTENTION DE LA
MAÎTRISE EN GÉNIE ÉLECTRIQUE
M. Ing.

PAR
JULIE SÉRIS

ÉTUDE DE QUELQUES MÉTHODES DE DÉTECTION D'ACTIVITÉ VOCALE DANS DES ENVIRONNEMENTS INDUSTRIELS BRUITÉS

MONTRÉAL, LE 10 AVRIL 2006

CE MÉMOIRE A ÉTÉ ÉVALUÉ
PAR UN JURY COMPOSÉ DE :

M. Christian Gargour, directeur de mémoire
Département de génie électrique à l'École de technologie supérieure

M. Frédéric Laville, codirecteur
Département de génie mécanique à l'École de technologie supérieure

M. Marcel Gabréa, président du jury
Département de génie électrique à l'École de technologie supérieure

M. Jean-Marc Lina, membre du jury
Département de génie électrique à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC
LE 10 MARS 2006
À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ÉTUDE DE QUELQUES MÉTHODES DE DÉTECTION D'ACTIVITÉ VOCALE DANS DES ENVIRONNEMENTS INDUSTRIELS BRUITÉS

Julie Séris

SOMMAIRE

Le travail proposé est la mise au point d'un algorithme de détection d'activité vocale performant pour une utilisation dans des milieux industriels bruités. Pour cela, deux approches ont été abordées. La première a consisté à modifier un détecteur d'activité vocale (DAV) existant afin qu'il soit adapté à ce type d'environnement. La plupart des méthodes proposées dans la littérature ont été développées pour les télécommunications. Nous en avons étudié quelques unes. En examinant la complexité de leur algorithme ainsi que leurs avantages et inconvénients, nous en avons retenu une, à savoir celle utilisée dans le codeur de parole G729. Nous l'avons ensuite adaptée aux milieux industriels. Les performances du DAV ainsi ajusté sont satisfaisantes. La deuxième approche est basée sur la théorie des ondelettes. Nos recherches nous ont permis d'aboutir à un nouvel algorithme de détection d'activité vocale que nous avons appelé le DAV INNES. Il repose sur deux notions fondamentales : la décomposition en paquets d'ondelettes selon l'échelle de Mel et la prise de décision en fonction des valeurs du Paramètre du Seuil de Johnstone et Silverman (PSJS) et des énergies. Une procédure d'ajustement automatique et rapide a été mise au point afin de déterminer pour chaque milieu industriel les meilleures règles de décision et d'avoir ainsi des performances accrues. Les résultats obtenus sont très satisfaisants.

STUDY OF SOME METHODS OF VOICE ACTIVITY DETECTION IN NOISY INDUSTRIAL ENVIRONMENTS

Julie Séris

ABSTRACT

The work suggested here is the development of an effective algorithm of voice activity detection for use in noisy industrial backgrounds. For that, two approaches were used. The first one consisted in modifying an existing voice activity detector (VAD) so that it is adapted to this kind of environment. The majority of the methods which are proposed in the literature were developed for telecommunications. We studied some of them. By examining the complexity of their algorithm as well as their advantages and disadvantages, we retained one of them, namely that used in the speech coder G729. Then, we adapted it to industrial backgrounds. The performances of the system thus adjusted are satisfactory. The second approach is based on the wavelets theory. Our research enabled us to develop a novel algorithm of voice activity detection that we named VAD INNES. It rests on two fundamental concepts: the decomposition in wavelet packets in accordance with the Mel scale and the decision-making according to the values of the parameter of the threshold of Johnstone and Silverman and of the energies. An automatic and fast procedure of adjustment was developed in order to determine for each industrial background the best rules of decision and to have thus increased performances. The obtained results are very satisfactory.

REMERCIEMENTS

Je tiens à remercier mon directeur de recherche, Monsieur Christian Gargour, professeur à l'École de technologie supérieure, pour son écoute, son soutien et ses conseils judicieux tant lors de mes recherches que lors de l'écriture de ce mémoire.

Je souhaiterais remercier mon codirecteur, Monsieur Frédéric Laville, professeur à l'École de technologie supérieure, pour ses commentaires pertinents et son aide lors de la révision du mémoire.

Je remercie aussi la compagnie SONOMAX pour m'avoir donné l'opportunité de participer à ce projet qui m'a fort intéressée. Je remercie les membres du projet ÉTS-SONOMAX-CRSNG et tout particulièrement Jérémie Voix pour m'avoir fourni des bases de données ainsi que pour ses commentaires judicieux tout au long de ce projet.

Je voudrais aussi remercier le CRSNG pour leur soutien financier.

J'ai une pensée toute particulière pour ma famille. Je remercie mes parents pour leur aide et leur support tout au long de cette maîtrise ainsi que ma grande tante et mon grand père pour leur gentillesse et leurs appuis moral et financier.

À mes proches d'ici et d'ailleurs.

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVATIONS ET SIGLES

AMR	Taux multi adaptatif (« Adaptative Multi-Rate »)
AMR1	Détecteur d'activité vocale option 1 pour le codeur de parole AMR
AMR2	Détecteur d'activité vocale option 2 pour le codeur de parole AMR
CAN	Convertisseur Analogique Numérique
CNA	Convertisseur Numérique Analogique
DAV	Détecteur d'Activité Vocale
DSP	Processeur dédié aux traitements des signaux (« Digital Signal Processor »)
ETSI	Institut européen des normes de télécommunication (« European Telecommunications Standards Institute »)
FAV	Forme d'Activité Vocale
GSM	« Global System for Mobile communications »
G729.B	Détecteur d'activité vocale utilisé par le codeur de parole G729
INNES	Nom du détecteur d'activité vocale basé sur les ondelettes mis au point au cours de ce projet de maîtrise (Industriel oNdelettes johNstone silvErman ajuStement)
ITU-T	Union internationale des télécommunications, division standardisation pour les télécommunications (« International Telecommunication Union)
LPC	Méthode de prédiction linéaire (« Linear Predictive Coding »)
MAD	Médiane de la déviation absolue (« Median Absolute Deviation »)
Nor	Bruit enregistré dans une affinerie de cuivre de NORANDA
NPZ	Nombre de Passage par Zéro
Nx	Bruit issu de la base de données NOISEX
PSJS	Paramètre du Seuil de Johnstone et Silverman
RSB	Rapport Signal à Bruit

STFT	Transformée de Fourier à fenêtre glissante (« Short Time Fourier Transform »)
TEO	« Teager Energy Operator »
TF	Transformée de Fourier
TOC	Transformée en Ondelettes Continue
TOD	Transformée en Ondelettes Discrète
VAD	Voice Activity Detector
T_e	Période d'échantillonnage
F_e	Fréquence d'échantillonnage
σ_t	Résolution temporelle
σ_ω	Résolution fréquentielle
$L^2\{\Re\}$	Espace des fonctions réelles continues et de carré intégrable
ψ	Fonction d'ondelette
φ	Fonction d'échelle
a_n^j	Coefficients d'échelle à la résolution j, appelés aussi coefficients d'approximation
d_n^j	Coefficients d'ondelette à la résolution j, appelés aussi coefficients de détails
V_j	Sous-espace de $L^2\{\Re\}$ associé à la résolution j
W_j	Complément orthogonal de V_j dans V_{j-1}
g	Filtre passe-haut utilisé pour l'obtention des coefficients de détails lors d'une analyse multirésolution
h	Filtre passe-bas utilisé pour l'obtention des coefficients d'approximation lors d'une analyse multirésolution
λ	Seuil utilisé pour le débruitage de la parole
s	Trame d'analyse
$r(k)$	Coefficients d'autocorrélation

E_f	Énergie de la trame d'analyse dans toute la bande
E_l	Énergie de la trame d'analyse dans les basses fréquences
ZC	Taux de passages par zéro (« Zero Crossing ») de la trame d'analyse
LSF	Coefficients de raies spectrales (« Line Spectral Frequency ») de la trame d'analyse
$\overline{E_f}$	Énergie de l'estimé du bruit dans toute la bande
$\overline{E_l}$	Énergie de l'estimé du bruit dans les basses fréquences
$\overline{ZC}$	Taux de passages par zéro de l'estimé du bruit
$\overline{LSF}$	Coefficients de raies spectrales de l'estimé du bruit
ΔE_f	Différence de l'énergie dans toute la bande entre la trame d'analyse et l'estimé du bruit
ΔE_l	Différence de l'énergie dans les basses fréquences entre la trame d'analyse et l'estimé du bruit
ΔZC	Différence du taux de passages par zéro entre la trame d'analyse et l'estimé du bruit
DS	Distorsion Spectrale
E_{min}	Énergie minimale
E_{fprec}	Énergie de la trame précédente
rc	Second coefficient de réflexion (« Reflexion Coefficient »)
$decision1$	Décision préliminaire du détecteur d'activité vocale
vad_flag	Sortie du détecteur d'activité vocale, appelée aussi décision finale
RSB_{min}	Rapport signal à bruit minimum
RSB_{max}	Rapport signal à bruit maximum

INTRODUCTION

Au cours de l'évolution des espèces, l'être humain est le seul être vivant à avoir considérablement développé son expression orale. Cela a finalement abouti à un langage parlé très élaboré qui lui permet aujourd'hui de communiquer avec ses semblables. Il utilise cette faculté pour transmettre des informations, des idées, des pensées ou des sentiments, lorsqu'il en ressent le besoin. Durant une conversation entre deux personnes ou plus, chaque interlocuteur parle et s'arrête pour écouter les propos des autres et créer un échange. De même, lors d'un monologue, le locuteur fait des pauses pour donner un certain rythme à son message et faciliter sa compréhension, comme par exemple pour indiquer la fin d'une idée. À l'intérieur même d'une phrase, il peut y avoir des périodes de silence : entre les mots, pour marquer une ponctuation ou encore lors d'une hésitation. La détection d'activité vocale consiste à distinguer les segments sonores contenant de la parole de ceux qui en sont dépourvus. Cette opération peut être très utile dans de nombreuses applications. Lors du stockage d'un signal de parole par exemple, les segments de silence peuvent être codés différemment, de manière à réduire la taille des données et ainsi économiser l'espace mémoire. Dans un système de traitement de la parole, elle peut éviter le traitement inutile des segments sans voix. Par exemple, dans un codeur de parole pour la téléphonie, le fait d'identifier ces parties dépourvues d'activité vocale permet de ne pas les envoyer dans le canal de transmission et ainsi de le libérer pour d'autres conversations ou applications. Un dernier exemple peut être pour la reconnaissance du locuteur ou des mots. En effet, elle permet alors de mettre en évidence l'information à reconnaître et ainsi de faciliter l'identification.

Objectifs :

Ce projet de recherche fait parti du développement des bouchons d'oreille « intelligents » entrepris par la compagnie SONOMAX. Ce type de bouchons va permettre aux travailleurs d'être protégés du bruit environnant tout en ayant le confort d'entendre les voix des autres travailleurs ainsi que les signaux d'alarme. Le traitement

et le rehaussement de la parole sont les fondements de ce système complexe. Afin que ces opérations soient effectuées correctement et uniquement quand cela est nécessaire, la différentiation des segments sonores composés de bruits industriels avec et sans parole est requise. En effet, le but de ces bouchons d'oreille est à la fois de bloquer le bruit lorsqu'il n'y a pas d'activité vocale et de rendre la parole et les signaux d'alarme intelligibles lorsqu'ils sont présents. Le travail proposé ici consiste donc à mettre au point un détecteur d'activité vocale (DAV) efficace dans les milieux industriels bruités. L'objectif visé est un fonctionnement correct pour des environnements industriels dont le rapport signal à bruit (RSB) est compris entre 5dB et 15dB, et si possible pour des RSB encore plus faibles. On rappelle que plus le bruit est présent, plus le RSB est petit.

Démarche :

Les différentes méthodes de détection d'activité vocale ont été développées pour les télécommunications. Ce projet de recherche contient deux approches. La première consiste à mettre au point un DAV à partir d'une méthode existante afin qu'il soit performant dans nos conditions particulières. Il s'agit ici de prendre connaissance des algorithmes de détection d'activité vocale proposés dans la littérature et d'étudier leur comportement et leurs performances. La procédure qui semble la plus adéquate sera ensuite mise en œuvre. Ayant été développée pour un usage en télécommunications, elle sera modifiée afin de l'adapter aux environnements industriels dont le type de bruits et les rapports signal à bruit sont généralement plus défavorables. La deuxième approche est la réalisation d'un DAV à partir de la théorie des ondelettes. Ce nouvel outil de traitement des signaux est puissant et a déjà montré sa supériorité par rapport aux outils traditionnels dans plusieurs domaines, notamment dans celui du traitement de la parole avec le débruitage de la voix. Il est donc possible que cette démarche procure de bonnes performances. La transformée en ondelettes étant très peu courante pour la détection d'activité vocale, il s'agit ici de mettre au point un algorithme presque entièrement nouveau. Pour cela, il est nécessaire d'effectuer des recherches approfondies pour déterminer les relations entre les coefficients d'ondelettes et l'activité vocale. Enfin, ces

deux approches doivent être simulées et testées à l'aide du logiciel MATLAB afin d'évaluer leurs performances, ceci dans l'optique d'assurer leur fonctionnement pour une utilisation au sein des bouchons d'oreille « intelligents ».

Plan du mémoire :

Le premier chapitre de ce mémoire permettra au lecteur de prendre connaissance des notions de base nécessaires à la bonne compréhension de ce projet de recherche. Il introduira donc la parole, le traitement numérique des signaux et la détection d'activité vocale. Le deuxième chapitre présentera les méthodes de base de la détection d'activité vocale proposées dans la littérature et leur mise en application. Le chapitre 3 sera consacré à la première approche, c'est-à-dire à la mise au point d'un DAV à partir d'une méthode existante afin qu'il soit efficace dans les milieux industriels bruités. L'algorithme initial, les modifications apportées et les résultats obtenus y seront exposés. Les trois derniers chapitres seront, quant à eux, dédiés à la deuxième approche. Ainsi, le chapitre 4 introduira la théorie des ondelettes, le chapitre 5 exposera le DAV que nous avons mis au point avec les ondelettes et le chapitre 6 présentera les résultats obtenus dans les environnements industriels bruités.

CHAPITRE 1

NOTIONS DE BASE

Ce chapitre présente brièvement les notions de parole, traitement numérique des signaux et détection d'activité vocale afin de faciliter la compréhension des sections suivantes.

1.1 La parole

1.1.1 Définition d'un son

Un son est un phénomène physique qui fait réagir notre cerveau, c'est une sensation auditive provoquée par une onde acoustique. D'un point de vue physique, il s'agit d'une vibration qui se propage dans un milieu matériel solide, liquide ou gazeux.

La perturbation associée à une onde sonore concerne la pression interne d'un milieu matériel. Ainsi plus la pression acoustique est importante, plus le volume sonore est grand. Depuis sa source, l'onde mécanique modifie la valeur de cette pression en chaque point de son trajet. Grâce à l'excitation mécanique, les molécules ayant reçu une impulsion se mettent en mouvement et entrent en collision avec les molécules voisines auxquelles elles communiquent le même mouvement. Une zone de compression est alors créée. À cause du choc, les premières reculent et dépassent leur position de repos, c'est pourquoi une détente succède toujours à une compression, tandis qu'une autre zone de compression se forme plus loin. Il s'établit ainsi des oscillations. Le mouvement des molécules voisines étant limité pour les mêmes raisons, elles oscillent à leur tour. Petit à petit, ce mouvement se propage, créant ainsi une onde sonore à l'origine du son. L'onde sonore créée par le mouvement oscillatoire des particules se disperse autour de la source émettrice selon une sphère. Plus l'onde sonore s'éloigne de la source, plus la surface de

la sphère augmente et plus l'intensité diminue. La transmission s'accompagne d'une dissipation d'énergie sous forme de chaleur, ce qui provoque l'amortissement de l'onde avec la distance. La propagation du son se fait à une vitesse dépendant des caractéristiques et des conditions de température et de pression du milieu (Matras [1]).

On peut diviser les sons en deux catégories :

- les sons purs : ils correspondent à des mouvements d'oscillations des particules pures, c'est à dire une sinusoïdale parfaite. Un son pur est en fait constitué d'une fréquence unique. Il est très peu répandu dans la nature.
- les sons complexes : ils peuvent contenir des éléments périodiques, transitoires ou aléatoires. Les éléments sonores périodiques (cas de la partie soutenue d'une note d'un instrument de musique) sont caractérisés par leur fréquence de base, dite fondamentale, et leurs harmoniques, multiples de la fréquence fondamentale, comme montré par la figure 1. Les harmoniques déterminent le timbre d'un son et selon leur nombre et leur fréquence, on peut ainsi distinguer le violon de la flûte. Les éléments transitoires (cas de l'attaque d'une note d'un instrument de musique) sont plus difficilement caractérisables. Les éléments aléatoires (cas du bruit de la turbulence générée par un écoulement) sont souvent appelés bruit en traitement des signaux et sont caractérisés par leur contenu fréquentiel et leur densité spectrale de puissance. Un cas extrême est le bruit « blanc » dont le spectre contient toutes les fréquences avec la même densité spectrale de puissance (Matras [1]).

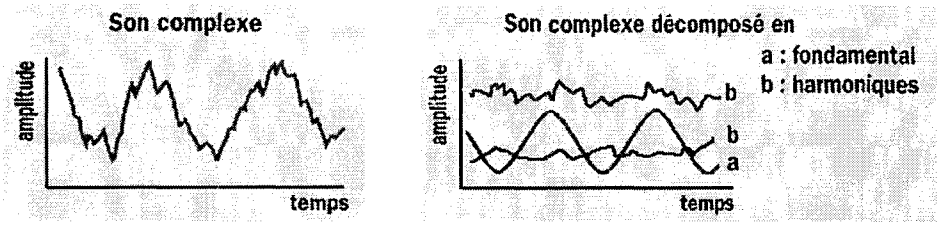


Figure 1 Les sons complexes périodiques

1.1.2 Définition de la parole

La figure 2 présente les principaux organes utilisés lors de la production de la parole :

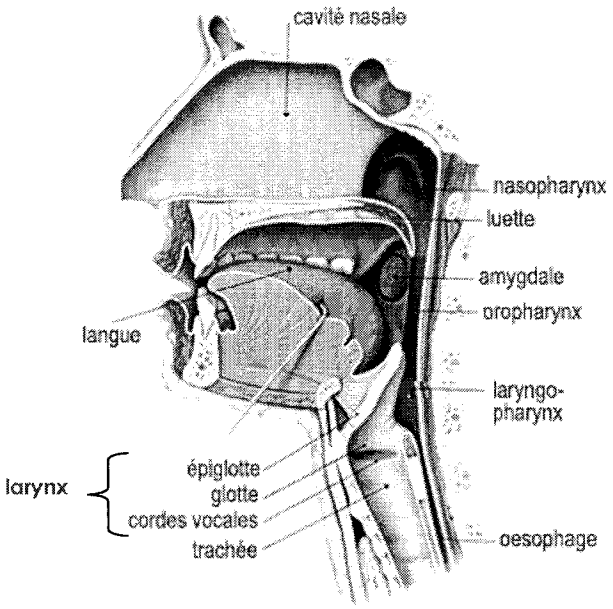


Figure 2 Les principaux organes de la production de la parole
(Adaptée de : http://www.lecerveau.mcgill.ca/flash/capsules/outil_bleu21.html, [2])

L'air contenu dans les poumons traverse la trachée et arrive dans le larynx où il rencontre les cordes vocales. Ce contact entraîne un mouvement des cordes vocales (phonation) qui lui-même provoque l'ouverture et la fermeture rapides de la glotte. Cette vibration engendre alors une onde sonore, prémisse du signal de parole. Le pharynx ainsi que les cavités buccales et nasales modifient ensuite les caractéristiques de cette onde, en atténuant ou en amplifiant certaines fréquences. Ils ont donc un rôle de résonateur. Le son laryngé devient finalement de la parole lorsqu'il a été modulé par la position de la langue, du voile du palais, des dents et des lèvres, [2].

La parole fait partie des sons complexes. Ainsi, chaque trace vocale est caractérisée par sa fréquence fondamentale, appelée aussi *pitch*, et par ses harmoniques. Le *pitch* est directement lié au nombre de fois que la glotte s'ouvre et se ferme par seconde. Pour certains sons, les cordes vocales ne vibrent pas, il n'y a alors pas de *pitch*. Les harmoniques, quant à eux, dépendent de l'endroit où a eu lieu la résonance. On appelle *formants* les zones fréquentielles entourant ces fréquences de résonance. Les traces vocales se distinguent aussi par leur forme (durée, intensité...) et sont particulières à chaque individu, idée fondamentale de la reconnaissance du locuteur (Schafer et Markel [3]).

La parole étant un son, elle se propage de la manière décrite à la section 1.1.1 et arrive finalement à l'oreille. L'être humain entend alors ce son s'il est compris entre 20Hz et 20kHz. La figure 3 schématise l'oreille humaine :

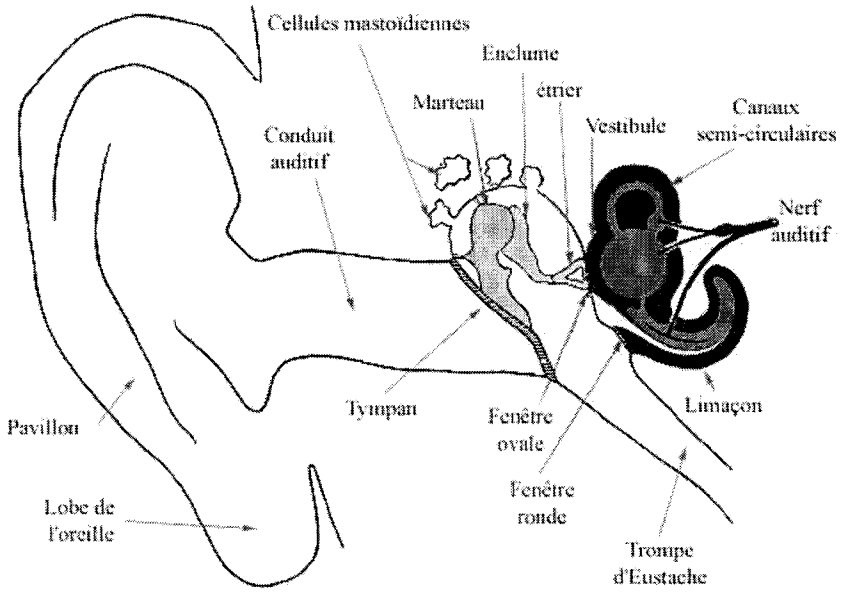


Figure 3 Anatomie de l'oreille
(Issue de http://www.medecine-et-sante.com/anatomie/anatoreille.html, [4])

Lorsqu'un son vient heurter les tympans, ils se mettent à onduler puis la vibration est transmise jusqu'à l'oreille interne dans laquelle se trouve le nerf auditif. Ce dernier transmet l'information au cerveau et une sensation sonore est alors ressentie (Matras [1]).

1.1.3 Différents types de parole

D'après Tetschner [5], les sons parlés peuvent être regroupés en deux catégories :

- les sons voisés : on dit qu'un son est voisé lorsque les cordes vocales vibrent de façon quasi-périodique, c'est-à-dire lorsqu'il y a phonation. Le signal de parole est alors caractérisé par son *pitch*.
- les sons non-voisés : on dit qu'un son est non-voisé si le phénomène de phonation est absent. Il ne possède donc pas de *pitch*.

Une autre manière de classer les sons parlés est d'utiliser leurs caractéristiques articulatoires. Il existe deux unités phonétiques (Parsons [6]) :

- les voyelles : elles sont produites en laissant passer l'air librement dans le conduit vocal et ceci sans obstruction d'aucune sorte. Elles provoquent la vibration des cordes vocales et font donc partie des sons voisés.
- les consonnes : elles sont produites par une obstruction partielle ou totale du conduit, par exemple à l'aide du palais, de la langue, des lèvres... Elles peuvent être voisées ou non.

Il existe une troisième façon de diviser les sons parlés : le modèle phonologique mais nous ne l'aborderons pas ici car cette classification est rarement utilisée dans le domaine de la détection d'activité vocale (Parsons [6]).

1.2 Le traitement numérique des signaux

Depuis plus d'une vingtaine d'années, la technologie des microprocesseurs s'est développée rapidement. Le traitement numérique des signaux est alors devenu très populaire. En effet, comparé à un système analogique, un système numérique est plus rapide, plus flexible, généralement plus simple à concevoir et surtout plus économique. On comprend donc pourquoi la majorité des traitements sur les signaux s'effectue aujourd'hui de manière numérique. C'est le cas du détecteur d'activité vocale. Ainsi pour concevoir un tel procédé, il est important de connaître le fonctionnement d'un système numérique.

Comme nous l'avons vu précédemment, les signaux de parole, et les sons en général, sont des signaux analogiques temporels. Afin de les manipuler numériquement, il faut d'abord les convertir en signaux numériques. Une fois qu'ils ont été traités, il est nécessaire de les retransformer en signaux analogiques afin que le signal de sortie soit de nouveau un son. Le schéma fonctionnel d'un tel système est décrit par la figure 4 :

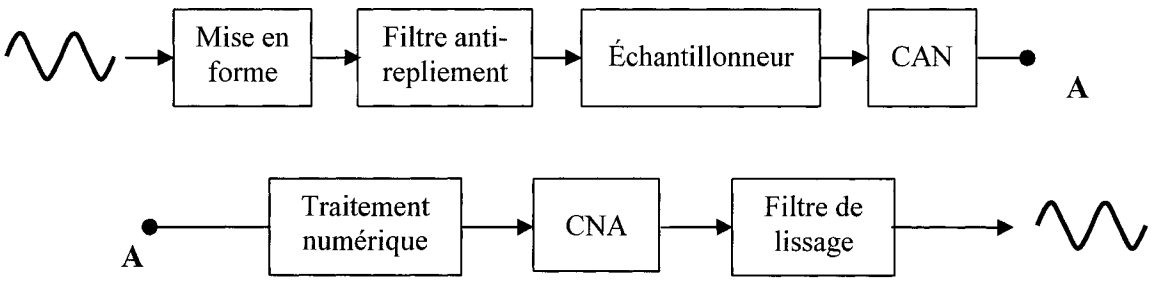


Figure 4 Schéma fonctionnel d'une chaîne de traitement numérique des signaux (Adaptée de « Discrete-Time Signal Processing », Oppenheim et Schafer [7])

D'après Oppenheim et Schafer [7], les différentes étapes mises en œuvre dans un système numérique, présenté par la figure 4, peuvent être décrites de la manière suivante :

Mise en forme : dans la plupart des cas, la donnée à traiter est une grandeur physique. Un capteur approprié est alors utilisé pour transformer cette grandeur physique en tension ou en courant. Le signal est ensuite mis en forme par des circuits spécialisés de conditionnement.

Échantillonneur : l'échantillonnage d'un signal consiste à prélever des échantillons sur le flux continu d'information analogique à des intervalles de temps discrets et constants T_e. Le choix de cette période est important pour assurer la précision du signal restitué après traitement. Il apparaît clairement que plus les échantillons sont rapprochés, plus la précision du signal sera importante, par contre un grand nombre d'échantillons par seconde conduira à un flux de données plus difficile à traiter. Le choix de la fréquence d'échantillonnage f_e doit respecter le théorème de Nyquist afin d'assurer une bonne restitution du signal et d'éviter le repliement spectral :
Pour éviter toute perte d'information et donc pour que le signal échantillonné puisse être reconstruit à partir de ses échantillons, il faut et il suffit que la fréquence d'échantillonnage f_e soit plus grande ou égale à deux fois la fréquence du signal à échantillonner :

$$f_e \geq 2 f_m = \textit{fréquence de Nyquist} \qquad (1.1)$$

Filtre anti-repliement : même en respectant le théorème de Nyquist, on peut être en présence de repliement spectral. En effet, le signal à échantillonner est le plus souvent entaché de bruit, dont le spectre est généralement plus étendu que celui du signal utile. L'échantillonnage a alors pour effet de superposer, au spectre basse fréquence du signal utile, certaines parties du spectre du bruit. Les composantes du bruit au dessus de la fréquence d'échantillonnage sont « repliées » sur le spectre du signal utile. Pour enlever

cet effet perturbateur, il faut éliminer la partie haute fréquence du bruit. Pour cela, on utilise un filtre passe-bas de fréquence de coupure égale, ou parfois inférieure, à la moitié de la fréquence d'échantillonnage. Ce filtre est appelé filtre anti-repliement et doit précéder l'échantillonneur.

Convertisseur Analogique-Numérique (CAN) : un signal numérique est un signal discret quantifié, c'est-à-dire que son amplitude est, elle aussi, discrétisée. C'est cette quantification qui rend possible la représentation d'un signal par un nombre fini de bits. Le CAN permet donc de transformer le signal discret en un signal numérique. La conversion n'étant pas instantanée, il est nécessaire d'insérer entre la sortie de l'échantillonneur et l'entrée du CAN un bloqueur d'ordre zéro qui a pour fonction de maintenir suffisamment longtemps la valeur d'un échantillon. Le temps de blocage doit bien sûr être supérieur au temps de conversion pour que le CAN puisse convertir les données correctement, le temps de conversion étant le temps qui s'écoule entre l'impulsion de demande de conversion et la stabilisation des données dans le tampon.

Traitement numérique : en pratique, les opérations sur les signaux numériques sont effectuées à l'aide d'un *Digital Signal Processor* (DSP). Ce DSP manipule les données obtenues à la sortie du CAN afin de réaliser la fonction qu'il doit assurer. À sa sortie, les données modifiées sont du même type qu'à son entrée, c'est-à-dire numériques. Il est caractérisé par plusieurs paramètres comme son architecture, sa mémoire, son jeu d'instructions, ses registres d'adressage... C'est à ce niveau que le détecteur d'activité vocale intervient. En effet, une fois celui-ci mis au point, il suffit de coder son algorithme en assembleur sur le DSP. Ce projet de recherche n'ira pas jusqu'à cette étape, il s'agit ici uniquement de développer un détecteur d'activité vocale efficace pour les milieux industriels bruités et de le simuler avec MATLAB afin de vérifier son bon fonctionnement.

Convertisseur Numérique-Analogique (CNA) : une fois le traitement effectué, il est nécessaire d'avoir recours à une conversion numérique-analogique afin que le signal de sortie de la chaîne soit encore un son. À l'entrée du CNA, il y a des échantillons dont l'amplitude est binaire. Ce convertisseur va permettre de convertir cette amplitude binaire en une amplitude analogique. Sa sortie est donc constituée d'une suite d'impulsions rectangulaires juxtaposées et de largeur égale à la période d'échantillonnage.

Filtre de lissage : le signal de sortie du CNA n'est pas encore du même type que le signal d'entrée de la chaîne de traitement. Pour enlever cet effet rectangulaire, il faut lisser ce signal. Pour cela, on utilise un filtre passe-bas de fréquence de coupure égale à la moitié de la fréquence d'échantillonnage. Ce filtre est appelé filtre de lissage et doit suivre directement le CNA. Ainsi le signal, après traitement numérique, est du même type qu'au départ, par contre ses caractéristiques ont été modifiées comme désiré.

1.3 La détection d'activité vocale

Par définition, la détection d'activité vocale consiste à déterminer les segments sonores avec et sans parole dans une conversation ou dans une simple phrase. Cette étape est cruciale dans la plupart des systèmes de traitement de la parole car elle permet d'identifier et de délimiter avec précision l'information utile, à savoir la voix. En effet, les données superflues, c'est-à-dire dépourvues de parole, n'ont pas besoin d'être traitées. Dans certaines applications, il est même primordial qu'elles ne subissent pas le traitement réservé à la parole, comme par exemple lors de l'amplification de la voix.

Lorsqu'il y a peu ou pas de bruit, cette opération est aisée. Dans le cas d'un bruit important, c'est-à-dire quand le rapport signal à bruit est faible, des parties de parole sont complètement ensevelies sous ce dernier et il devient alors difficile de détecter correctement l'activité vocale. Ceci peut être observé sur la figure 5.

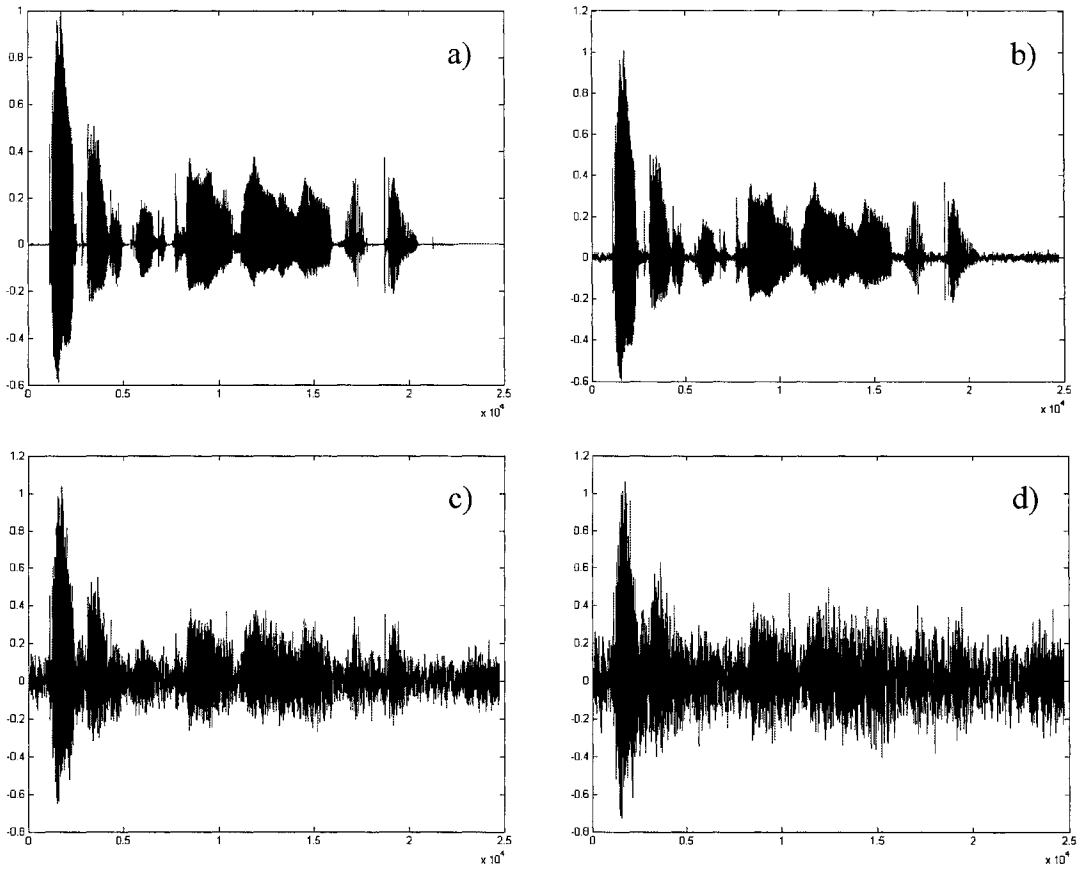


Figure 5 Exemples de signaux de parole bruitée
a) Signal de parole claire, RSB → ∞ b) Signal de parole bruitée, RSB = 20dB
c) Signal de parole bruitée, RSB = 5dB d) Signal de parole bruitée, RSB = 0dB

On rappelle que le RSB est défini par (Oppenheim et Schafer [7]) :

$$RSB = 10\log_{10}\left(\frac{P_x}{P_b}\right) \qquad (1.2)$$

Avec :

- P_x : la puissance du signal initial
- P_b : la puissance du signal de bruit

Ainsi plus le RSB est faible, plus le signal initial est enseveli sous le bruit.

Le système réalisant la délimitation de la parole s'appelle un détecteur d'activité vocale. Son principe est décrit par la figure 6 :

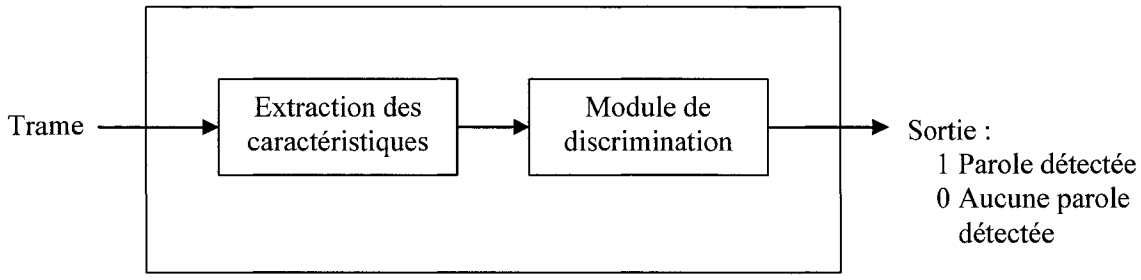


Figure 6 Schéma fonctionnel d'un DAV

Le signal à étudier est tout d'abord découpé en trames de longueur fixe. Chaque trame arrive ensuite à l'entrée du DAV. Ce dernier en extrait certaines caractéristiques. À l'aide de règles de décision portant sur les valeurs de ces paramètres, il prend alors une décision quant à l'état de la trame. La sortie du DAV ainsi obtenue est une variable logique. Si elle est unitaire, la trame contient de la parole. On dit alors que cette trame est active ou PAROLE. Si la sortie du DAV est nulle, on dit que la trame est inactive, ou BRUIT, ou encore, par abus de langage, qu'il s'agit d'une trame de silence. La figure 7 montre la sortie du DAV dans le cas du signal de parole claire vu à la figure 5 a) :

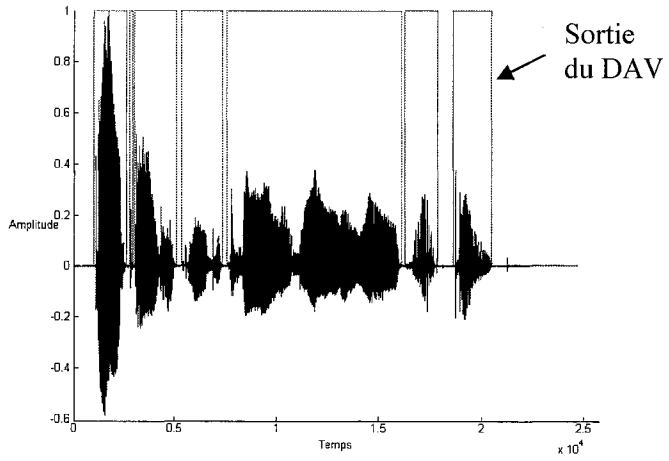


Figure 7 Exemple de sortie d'un DAV

Les DAV sont surtout utilisés dans les télécommunications, au sein des codeurs de parole pour la téléphonie fixe, mobile et multimédia. Ils permettent d'éviter le traitement de l'information inutile, bruit de fond ou silence, et ainsi de libérer le canal de transmission pour d'autres applications ou d'autres communications. En général, lorsqu'une trame est déclarée inactive par le DAV, le codeur de parole ne la transmet pas mais envoie à la place un « descripteur d'insertion de silence ». Quand le décodeur, de l'autre côté du canal de transmission, reçoit ce type d'information, il déclenche alors le « générateur de bruit de confort », simulant le bruit de fond enregistré par le codeur. Il donne ainsi l'impression aux interlocuteurs que leur conversation entière a été transmise. Si le décodeur n'utilise pas de « générateur de bruit de confort », l'interlocuteur passif entend alors le locuteur de manière hachée, un peu comme avec des talkies-walkies, à cause du caractère tout ou rien de la sortie du DAV. Le système regroupant le DAV, le « descripteur d'insertion de silence » et le « générateur de bruit de confort » permet donc à la fois une conversation agréable et une économie des transmissions, d'où l'importance d'un algorithme de détection d'activité vocale efficace. Il est à noter qu'en cas de doute le DAV devrait toujours indiquer la présence de parole afin de conserver l'information utile et ainsi la bonne qualité du message (Benyassine et Al. [8]).

Le but de ce projet de recherche est de mettre au point un DAV efficace dans les milieux industriels bruités. La plupart des méthodes proposées dans la littérature ont été développées pour les télécommunications. Le chapitre suivant en présente quelques unes.

CHAPITRE 2

LES MÉTHODES DE BASE DE LA DÉTECTION D'ACTIVITÉ VOCALE ET LEUR MISE EN APPLICATION

Comme il a été mentionné dans le chapitre précédent, la détection d'activité vocale est une étape cruciale dans la plupart des systèmes de traitement de la parole. Pour cette raison, ce domaine a été largement traité dans les années passées et il existe aujourd'hui des dizaines, voire des centaines, de méthodes différentes. Les plus anciennes sont basées sur l'énergie, le taux de passages par zéro [9], la mesure de la distance LPC [10] ou encore la périodicité [11]. Plus récemment, on retrouve des algorithmes portant sur les mesures cepstrales [12] ou fractales [13]. D'autres utilisent la séparation de sources [14] ou les modèles statistiques [15], [16]. Il y a aussi les DAV ayant recours aux réseaux de neurones ou à la logique floue [17]. À ceux-ci s'ajoutent les procédures des différents codeurs de parole [18], [19], [20]. La théorie des ondelettes commence, elle aussi, à fournir quelques résultats intéressants pour la détection d'activité vocale (ce cas particulier sera traité dans la section 4.2). En résumé, il existe de très nombreuses manières de réaliser un DAV.

Ce projet de recherche consiste à mettre au point un DAV fonctionnel pour des milieux industriels bruités. Toutefois, la majorité de ceux rapportés dans la littérature n'ont été ni développés, ni testés dans ce type d'environnement. Ce chapitre va permettre de prendre connaissance de quelques procédures utilisées en télécommunications. Ne pouvant présenter toutes les méthodes existantes, une approche historique a été choisie afin de montrer l'évolution et l'avancement dans ce domaine. Tout d'abord, les méthodes de base seront exposées car elles représentent les fondements de la détection d'activité vocale. Nous verrons ensuite leur mise en application à travers trois des plus importants

codeurs de parole, à savoir le Pan-european [18], le G729 [19], l'AMR [20]. Nous conclurons ensuite par une comparaison de ces différents algorithmes.

2.1 Les méthodes de base

Selon Tanyer et Özer [9], les méthodes de base de détection d'activité vocale sont : le seuillage direct ou adaptatif de l'énergie, le taux de passages par zéro et l'estimateur de la périodicité par les moindres carrés. Nous présenterons ici chacun de ces procédés mais nous exposerons d'abord celui basé sur la mesure de la distance LPC proposé par Rabiner et Sambur [10].

2.1.1 La mesure de la distance LPC

La première méthode que nous abordons ici est celle du DAV basé sur la mesure de la distance LPC, proposé par Rabiner et Sambur [10] en 1977. L'idée principale de cette méthode est de déterminer par entraînement une caractérisation spectrale de trois différentes classes de sons, à savoir la parole voisée, la parole non voisée et le silence. Les distances LPC et énergie par rapport à ces classes sont ensuite utilisées dans l'algorithme de détection d'activité vocale afin de prendre une décision quant à la nature de la trame étudiée. Il s'agit en fait de déterminer à quelle classe celle-ci s'apparente le plus. Il est à noter que la sortie du DAV peut prendre trois états.

Le schéma-bloc de ce DAV est donné par la figure 8 :

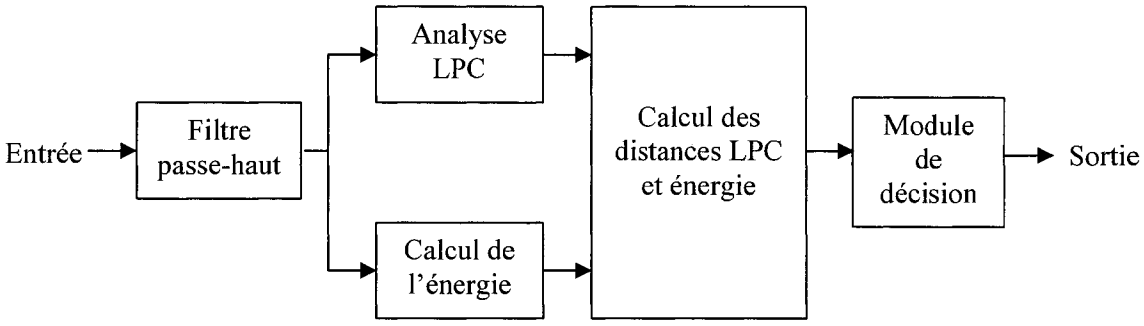


Figure 8 Schéma Bloc du DAV basé sur la distance LPC
(adaptée de l'article de Rabiner et Sambur, [10], 1977)

Nous décrivons ci-dessous les différents blocs mis en jeu.

Filtre passe-haut :

Ce filtre passe-haut est utilisé en guise de prétraitement afin d'éliminer tous les parasites basses fréquences.

Analyse LPC :

Une analyse LPC 8 pôles est effectuée sur chaque trame. On rappelle que le *Linear Predictive Coding* (LPC) est une méthode de prédiction de la valeur d'un échantillon basée sur les échantillons passés. Le prédicteur linéaire est un filtre à réponse impulsionnelle finie d'ordre p dont les coefficients sont appelés coefficients LPC. Ils sont déterminés de manière à minimiser l'erreur de prédiction par la méthode des moindres carrés (Rabiner et Sambur [10]). Pour chaque trame, on obtient donc l'ensemble des coefficients LPC :

$$a = \{a(0), a(1), a(2), ..., a(8)\} \quad \text{avec} \quad a(0) = 1 \qquad (2.\ 1)$$

Calcul de l'énergie :

En parallèle à l'analyse LPC, l'énergie en dB est calculée pour chaque trame :

$$E = 10\log_{10}\left(\sum_{n=1}^{L} x^2(n)\right) \tag{2. 2}$$

Avec : L la longueur de la trame du signal étudié $x(n)$

Calcul des distances LPC et énergie :

Connaissant les coefficients LPC et l'énergie de la trame, il est possible de calculer les distances par rapport aux trois classes.

En ce qui concerne la distance LPC, elle est calculée à l'aide de l'expression proposée par Itakura [21] :

$$D_a(k) = \frac{(a - m_k)\phi(a - m_k)^t}{(a\phi a^t)} \tag{2. 3}$$

Avec :

- a : le vecteur des coefficients LPC de la trame actuelle
- $k = 1,2,3$ représentant respectivement le silence, la parole non voisée et la parole voisée
- m_k : le vecteur moyen des coefficients LPC pour la classe k, cette valeur est déterminée lors de l'entraînement
- ϕ : la matrice de corrélation de la trame actuelle

La distance de l'énergie, quant à elle, est une distance euclidienne normalisée (Rabiner et Sambur [10]) :

$$D_E(k) = \left|\frac{E - \overline{E(k)}}{\sigma_E(k)}\right| \tag{2. 4}$$

Avec :

- E : l'énergie de la trame analysée

- $\overline{E(k)}$: l'énergie moyenne en dB pour la classe k, cette valeur est déterminée lors de l'entraînement
- $\sigma_E(k)$: la déviation standard, ou variance, de l'énergie en dB pour la classe k, elle aussi est obtenue lors de l'entraînement

À ce point, les données $D_E(k)$ et $D_a(k)$ sont connues pour chaque valeur de k : 1, 2, 3.

Module de décision :

Le module de décision est composé de 13 règles de décision comparant les différentes distances entre elles (Rabiner et Sambur [10]). Elles vont permettre de déterminer à quelle classe la trame étudiée s'apparente le plus. Certaines de ces règles prennent en considération l'état de la trame précédente. Il est intéressant de noter d'une part que l'ordre des règles est important car leur architecture est sous forme de cascade. Par exemple, si la règle i n'est pas vérifiée, la règle $i+1$ est testée. Dans le cas contraire, la sortie du DAV est trouvée et les règles suivantes ne sont pas testées. D'autre part, les cinq premières règles servent à détecter les trames de silence. Les huit suivantes permettent, quant à elles, de faire la distinction entre la parole voisée et celle non voisée, puisque si elles sont testées c'est que la possibilité d'une trame de silence a été écartée.

Les règles de décision ainsi que les grandeurs m_k, $\overline{E(k)}$ et $\sigma_E(k)$ utilisées lors des calculs de distances sont obtenues par expérimentation à l'aide d'un ensemble de trames d'entraînement, classées manuellement.

Bien que cet algorithme présente de bons résultats, c'est-à-dire de faibles erreurs de classification (Rabiner et Sambur [10]), il a été développé pour des environnements à rapports signal à bruit très élevés. Sa robustesse au bruit n'est donc pas assurée. L'approche exposée demeure toutefois très intéressante. En effet, elle met bien en évidence le fait que la détection de l'activité vocale n'est finalement qu'un problème de classification. Cette idée fondamentale est toujours utilisée aujourd'hui. Les DAV

utilisant les réseaux de neurones (Hoyt et Wechsler [22], [23]) en sont des exemples types.

2.1.2 Le seuillage de l'énergie

L'énergie est un paramètre important et très souvent utilisé dans la détection d'activité vocale. Elle est définie par :

$$E = \sum_{i=0}^{L-1} x^2(i) \tag{2. 5}$$

Avec : $x(n)$: le signal de longueur L dont il faut déterminer l'énergie.

La deuxième méthode de base est celle du seuillage de l'énergie. Elle consiste à calculer l'énergie de chaque trame du signal étudié. Cette donnée est alors comparée à un seuil dépendant du niveau de bruit. Si elle est supérieure à cette limite, la trame est dite active. Dans le cas contraire, on considère qu'elle ne contient pas de parole.

Selon Sangwan et Al. [24], la règle de classification est donc :

$$\begin{aligned} &Si\ E_j > kE_r\ alors\ la\ trame\ est\ ACTIVE \\ &Sinon \qquad\qquad la\ trame\ est\ INACTIVE \end{aligned} \tag{2. 6}$$

Avec :

- E_j : l'énergie de la $j^{ième}$ trame
- E_r : l'énergie des trames de bruit seul
- kE_r : le seuil, avec $k > 1$

La grandeur E_r est recalculée à chaque fois qu'une trame inactive est rencontrée. La règle de mise à jour utilisée est (Sangwan et Al. [24]):

$$E_{r\_new} = (1-p)E_{r\_old} + pE_{noise} \tag{2.7}$$

Avec :

- E_{r_new} : la nouvelle valeur du seuil
- E_{r_old} : l'ancienne valeur
- E_{noise} : l'énergie des plus récentes trames inactives
- p : un coefficient arbitraire compris entre 0 et 1

Selon Prasad et Al. [25], la valeur initiale du seuil est obtenue en prenant la moyenne des énergies des N premières trames du signal, considérées inactives, ou des N trames inactives pré-enregistrées et représentatives du bruit :

$$E_{r\_ini} = \frac{1}{N}\sum_{k=1}^{N} E_k \tag{2.8}$$

Cet algorithme est très simple et repose sur l'hypothèse que l'énergie en présence de parole est supérieure à celle du bruit seul. Malheureusement, ceci n'est pas toujours le cas car il existe de la parole basse énergie (Tanyer et Özer [9]). De même, plus le RSB diminue, plus la parole est ensevelie sous le bruit et moins cette hypothèse est vérifiée. Un autre inconvénient de cette procédure est que le coefficient p ne tient pas compte des statistiques du bruit. Lors d'un bruit qui varie très rapidement, cela entraîne beaucoup de coupures dans la parole et particulièrement au début et à la fin d'une bouffée (Sangwan et Al. [24]). Ce système n'est donc pas robuste.

2.1.3 Le seuillage adaptatif de l'énergie

Pour palier ce dernier problème, Prasad et Al. [25] indiquent qu'il faut utiliser les statistiques du second ordre lors de la mise à jour du seuil. La méthode proposée ici suit donc les mêmes règles de classification (2.6), d'initialisation (2.8) et de mise à jour du seuil (2.7) sauf que le coefficient p est calculé en fonction d'une nouvelle règle.

Il s'agit de stocker les m dernières trames inactives et de calculer la variance σ de leur énergie. La valeur de p est ensuite déduite de la manière suivante :

$$\begin{cases} Si \quad \frac{\sigma_{new}}{\sigma_{old}} \geq 1,25 \quad alors\ p = 0,25 \\ Si \quad 1,25 > \frac{\sigma_{new}}{\sigma_{old}} \geq 1,1 \quad alors\ p = 0,2 \\ Si \quad 1,1 > \frac{\sigma_{new}}{\sigma_{old}} \geq 1 \quad alors\ p = 0,15 \\ Si \quad 1 > \frac{\sigma_{new}}{\sigma_{old}} \quad alors\ p = 0,1 \end{cases} \tag{2. 9}$$

Avec :

- σ_{new} : la nouvelle valeur de la variance
- σ_{old} : l'ancienne

Ceci repose sur l'idée qu'un changement brusque dans le bruit est traduit par (Prasad et Al. [25]) :

$$\sigma_{new} > \sigma_{old} \tag{2. 10}$$

Cet algorithme est mieux adapté à un bruit non stationnaire. Toutefois, il reste fondé sur l'hypothèse de la supériorité de l'énergie en présence de parole, comparée à celle obtenue avec du bruit seul. Ses performances sont donc faibles lors de parole basse énergie et lors de petits RSB.

2.1.4 Le taux de passages par zéro

Pour résoudre le problème associé aux deux méthodes précédentes concernant la non-détection de la parole basse énergie, on peut rajouter un détecteur de passages par zéro (Rabiner et Sambur [26]). L'hypothèse sur laquelle repose cette méthode est que, contrairement à la parole, le bruit fluctue rapidement autour de zéro et que le nombre de fois que cela se produit est aléatoire. Ainsi, il est possible de déterminer un intervalle pour le nombre habituel de passages par zéro d'une trame de parole en fonction de sa longueur. Par exemple, pour une trame de parole de 10ms, il a été montré que ce nombre est compris entre 5 et 15. La méthode proposée ici consiste donc à faire passer les trames déclarées inactives par la méthode de l'énergie, dans le détecteur de passages par zéro. La règle de classification est :

$$\begin{aligned} &Si\ NPZ(trame_j) \in [a:b]\ alors\ la\ trame\ est\ ACTIVE \\ &Sinon \qquad\qquad\qquad\qquad\quad la\ trame\ est\ INACTIVE \end{aligned} \tag{2.11}$$

Avec :

- NPZ : le nombre de passages par zéro
- $[a:b]$: l'intervalle attendu, par exemple $[5:15]$ pour une trame de 10ms

Cet algorithme permet de retrouver les phénomènes de parole basse énergie manqués par les méthodes précédentes. Toutefois, certaines trames de bruit ont un nombre de passages par zéro trop bas et donc inclus dans l'intervalle utilisé. À l'inverse, certaines trames de parole en ont un trop élevé pour être déclarées actives. Cette méthode engendre donc des erreurs de classification. Enfin, elle ne fonctionne pas correctement lorsque le RSB est faible car dans ce cas les passages par zéro sont beaucoup trop fréquents.

2.1.5 L'estimateur de la périodicité par les moindres carrés

La dernière méthode de base que nous présentons ici est celle du DAV basé sur la mesure de la périodicité, proposé par Tucker [11]. Il s'agit non pas de déterminer les limites exactes de chaque mot mais simplement de localiser les régions contenant de la parole. Cela permet d'obtenir un fonctionnement acceptable dans des situations défavorables, c'est-à-dire à faibles RSB.

Cette méthode repose sur l'un des travaux de Irwin [27]. Il a montré que la manière optimale de mesurer la périodicité est d'appliquer directement au signal un estimateur de périodicité par les moindres carrés. Le problème de cet estimateur est sa sensibilité à tous les signaux périodiques et donc aux interférences. Compte tenu de cela, Tucker [11] a mis au point l'algorithme de détection d'activité vocale représenté par la figure 9 :

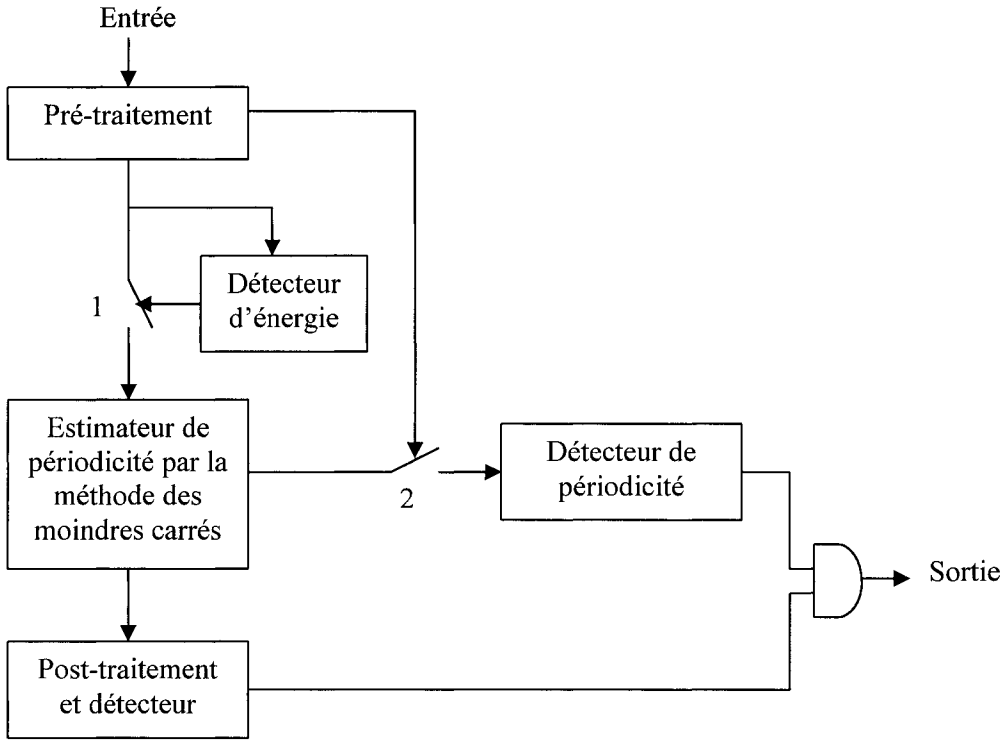


Figure 9 Schéma Bloc du DAV basé sur la mesure de la périodicité (adaptée de l'article de Tucker, 1992, [11])

Prétraitement :

Cette étape consiste à détecter et à enlever les interférences périodiques. Pour cela, la transformée de Fourier, définie ci-dessous (Oppenheim et Schafer [7]), est appliquée à plusieurs trames consécutives :

$$X(e^{j\omega}) = \sum_{n=-\infty}^{+\infty} x(n)e^{-j\omega n} \qquad (2.12)$$

Avec : $x(n)$: une trame.

Une analyse spectrale est ensuite effectuée. Elle repose sur le fait que les interférences sont caractérisées par des pics dans leur spectre. Or, il arrive que la parole ait un *pitch* constant, ce qui provoque le même phénomène fréquentiel. Afin de ne pas déclarer ce type de parole comme interférence, une recherche des harmoniques est menée (Tucker [11]). Si l'analyse indique la présence d'interférences, l'interrupteur 2 s'ouvre afin de ne pas prendre en compte la sortie de l'estimateur de périodicité dans la suite de l'algorithme.

Détecteur d'énergie :

Le détecteur d'énergie a pour but de laisser passer les signaux pouvant contenir de la parole (interrupteur 1 fermé) et de rejeter ceux qui en sont assurément dépourvus (interrupteur 1 ouvert).

Estimateur de périodicité par la méthode des moindres carrés :

La périodicité de chaque trame x est calculée. Pour cela, on part de l'idée que le signal x peut être décomposé en une composante périodique x_p et une composante non périodique x_{np} :

$$x(i) = x_p(i) + x_{np}(i) \qquad (2.13)$$

Avec :

- x_p : la composante périodique, on a donc $x_p(i) = x_p(i + kP)$
- k : un entier

- P : la période de x_p
- x_{np} : la composante non périodique

Notons $\hat{P}$ l'estimation de P et $\hat{x}_p$ celle de x_p. L'estimation de la composante périodique est donnée par :

$$\begin{cases} \hat{x}_p(i) = \sum_{k=0}^{K_0} \frac{x(i+k\hat{P})}{K_0} \quad pour \quad 1 \le i \le \hat{P} \; et \; P_{\min} \le \hat{P} \le P_{\max} \\ K_0 = \frac{L-i}{P} + 1 \end{cases} \tag{2. 14}$$

Avec :

- $P_{\min}$ et $P_{\max}$: le nombre d'échantillons minimum et maximum, respectivement, dans une période de *pitch*
- K_0 : le nombre de périodes de $\hat{x}_p$ dans la trame étudiée
- L : la longueur de la trame étudiée

Il s'agit ici d'estimer la période $\hat{P}$ du *pitch* par la méthode des moindres carrés. Cela revient à déterminer le minimum de la fonction coût : $\sum_{i=1}^{L}\left(x(i)-\hat{x}_p(i)\right)^2$ (Tucker [11]). D'après Friedman [28], cela équivaut à trouver le maximum de la fonction suivante :

$$R(\hat{P}) = \frac{I_0(\hat{P}) - I_1(\hat{P})}{\sum_{i=1}^{L} x^2(i) - I_1(\hat{P})} \tag{2. 15}$$

Avec :

$$I_0(\hat{P}) = \sum_{i=1}^{L} x_p^2(i) = \sum_{i=1}^{L}\left(\sum_{k=0}^{K_0} \frac{x(i+k\hat{P})}{K_0}\right)^2 \tag{2. 16}$$

et

$$I_1(\hat{P}) = \sum_{i=1}^{L}\left(\sum_{k=0}^{K_0} \frac{x^2(i+k\hat{P})}{K_0}\right) \tag{2. 17}$$

Ainsi pour chaque trame, les valeurs de $R(\hat{P})$ sont calculées par les équations (2.15) à (2.17), pour toutes les valeurs de $\hat{P}$ comprises entre $P_{\min}$ et $P_{\max}$. Le maximum de $R(\hat{P})$ est obtenu quand $\hat{P}$ est l'estimation correcte de la périodicité de la trame. Il est donc possible d'en déduire cette périodicité.

Détecteur de périodicité :
Le détecteur de périodicité prend une décision en sommant les périodicités de plusieurs trames consécutives. Si la valeur obtenue est supérieure à un seuil fixe, la dernière trame étudiée est considérée active.

Post-traitement et détecteur :
Cette partie de l'algorithme est utilisée pour prendre une deuxième décision, moins sensible mais plus robuste en présence d'interférences. Elle utilise un estimateur de qualité du signal de parole. La prise de décision est basée à la fois sur l'espacement des harmoniques de la trame étudiée et sur le déplacement des pics entre les trames (Tucker [11]).

Sortie du DAV :
La décision finale est PAROLE si les détecteurs indiquent tous deux la présence d'activité vocale.

L'algorithme présenté ici est donc basé sur la mesure de la périodicité calculée par la méthode des moindres carrés. Les modules de pré et post traitement permettent de réduire les faux déclenchements dus à la sensibilité de l'estimateur de périodicité aux interférences et aux bruits de fond périodiques. Toutefois, ce DAV ne permet pas d'obtenir une délimitation précise des bouffées de parole. En effet, le début et la fin d'une bouffée sont généralement mal identifiés car le DAV n'est pas capable de détecter certains segments de voix contenant uniquement une composante non périodique.

Les méthodes de base ont été présentées dans cette partie du chapitre. Elles sont toutes relativement simples du point de vue de la complexité de calcul. La procédure basée sur la mesure de la distance LPC ne fonctionne qu'en présence d'un bruit très faible. La méthode du taux de passages par zéro peut être vue comme une amélioration de celle du seuillage adaptatif de l'énergie qui est elle-même une amélioration du seuillage direct. Aucune de ces trois méthodes ne fonctionne correctement lorsque le RSB est faible. De plus, le seuil utilisé dans chaque algorithme n'est remis à jour que lorsque le bruit seul est rencontré. Ceci n'est pas adapté aux bruits non-stationnaires. En effet, ce type de bruit varie beaucoup et le seuil devrait prendre en considération ses variations. Pour cela, il serait nécessaire de le réajuster beaucoup plus souvent. Enfin, la méthode de l'estimateur de la périodicité par les moindres carrés donne de meilleurs résultats surtout pour de petites valeurs de RSB mais sa sensibilité aux interférences périodiques la rend délicate à utiliser.

2.2 La mise en application des méthodes de base

Dans cette deuxième partie, nous allons voir la mise en application des méthodes de base exposées précédemment. Pour cela, trois DAV utilisés dans des codeurs de parole vont être présentés. Ils utilisent tous des algorithmes qui leur sont propres mais qui reposent d'une manière ou d'une autre sur une ou plusieurs méthodes de base.

2.2.1 Le DAV du Pan-European

L'un des premiers DAV mis en œuvre dans un système complexe est celui du Pan-European, connu aussi sous le nom de « GSM Full-Rate », codeur de parole pour la téléphonie mobile proposé en 1988 par British Telecom. Selon Freeman, Southcott et Boyd [18], un bon moyen de détecter l'activité vocale, lorsque le signal de parole est noyé dans le bruit, est d'utiliser les caractéristiques spectrales.

En se basant sur cette idée, ils ont développé l'algorithme représenté par la figure 10 :

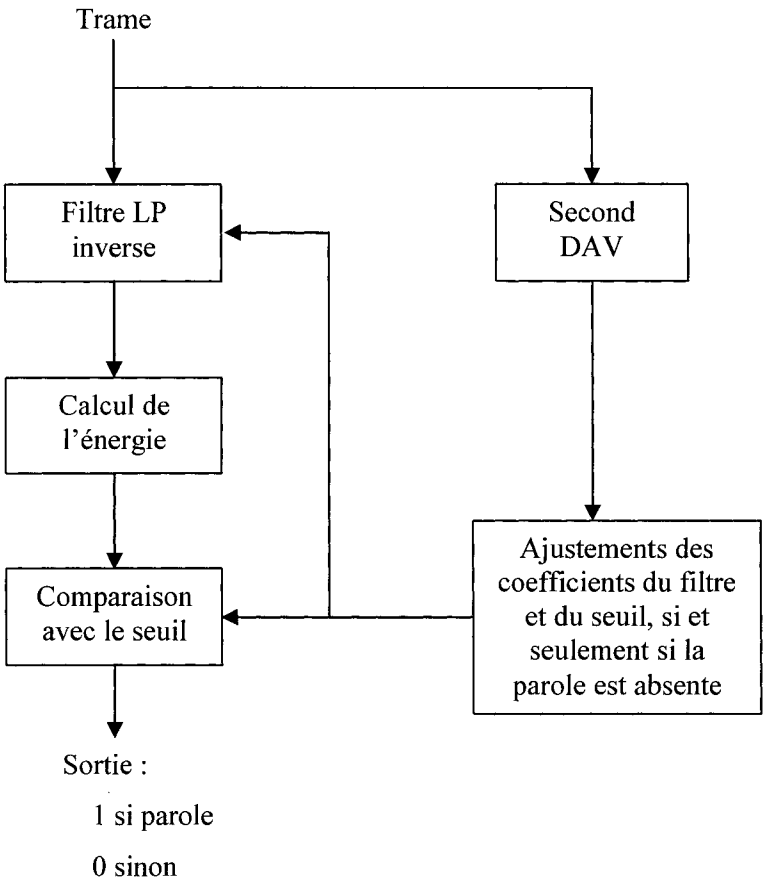


Figure 10 Schéma Bloc du DAV du Pan European
(adaptée de l'article de Barrett, 1995, [29])

Filtre LP inverse :

Chaque trame est d'abord envoyée dans un filtre adaptatif inverse. Il s'agit d'un filtre LP modélisant le bruit de fond. Il permet de réduire le bruit sans pour autant altérer sérieusement les caractéristiques de la parole.

Calcul de l'énergie et comparaison à un seuil :

L'énergie de la sortie de ce filtre est ensuite calculée par l'équation (2.5) et comparée à un seuil adaptatif. Si elle est supérieure à cette limite, la voix est détectée. Dans le cas contraire, la décision indique que la trame étudiée ne contient que du bruit ou du silence.

Second DAV et ajustements :

Le seuil et les coefficients du filtre inverse sont mis à jour uniquement lors des trames sans parole. Pour se faire, un deuxième DAV est utilisé. Il repose sur l'hypothèse que le bruit est stationnaire sur une plus longue période que la voix, c'est-à-dire que ses caractéristiques spectrales sont proches d'une trame à l'autre, contrairement à la parole (Freeman, Southcott et Boyd [18]). Il faut noter que la parole fait partie des signaux fortement non-stationnaires. Le second DAV calcule donc la distorsion spectrale entre plusieurs trames consécutives. Si cette valeur reste inférieure à un seuil fixe, cela indique que les caractéristiques spectrales n'ont pas beaucoup évoluées et donc qu'il y a de fortes chances pour que la parole soit absente. À cela, il a été ajouté un détecteur de périodicité (Barrett [29]). Si ce dernier ne détecte aucun signal périodique, l'absence de parole est confirmée et les variables qu'il faut mettre à jour sont alors ajustées en fonction des caractéristiques du bruit.

Lors de tests, un problème a été identifié : les tonalités, après un court instant, sont classées comme du bruit, alors qu'il s'agit là d'informations importantes à transmettre (Barrett [29]). Pour palier ce problème, la deuxième génération de codeur GSM, appelée « Half-Rate », a intégré un détecteur de tonalités au second DAV, afin d'empêcher l'adaptation des coefficients du filtre et du seuil en présence d'une tonalité.

Cet algorithme est intéressant de par son approche spectrale. Malheureusement, l'hypothèse concernant la stationnarité du bruit n'est pas toujours vérifiée et ce DAV présente dans de nombreux cas des performances peu satisfaisantes (El-Maleh et Kabal [30]). Toutefois, on retrouve dans la littérature plusieurs améliorations possibles permettant à la fois de résoudre ce problème et d'augmenter les performances du

système. Par exemple, Garner et Al. [31] proposent une manière de rendre le second DAV plus robuste. En effet, ils ont déterminé les paramètres les plus efficaces pour la détection des trames sans parole afin que la mise à jour des données soit par la suite optimale.

2.2.2 Le DAV du G729

Le G729 est un codeur de parole pour les communications fixes et multimédia, proposé par ITU-T en 1996. Dans son annexe B (ITU-T [19]), est présenté un DAV, que nous appellerons par la suite le G729.B, développé à la fois pour ce codeur et pour sa version basse complexité : le G729.A. Son principe est décrit par la figure 11.

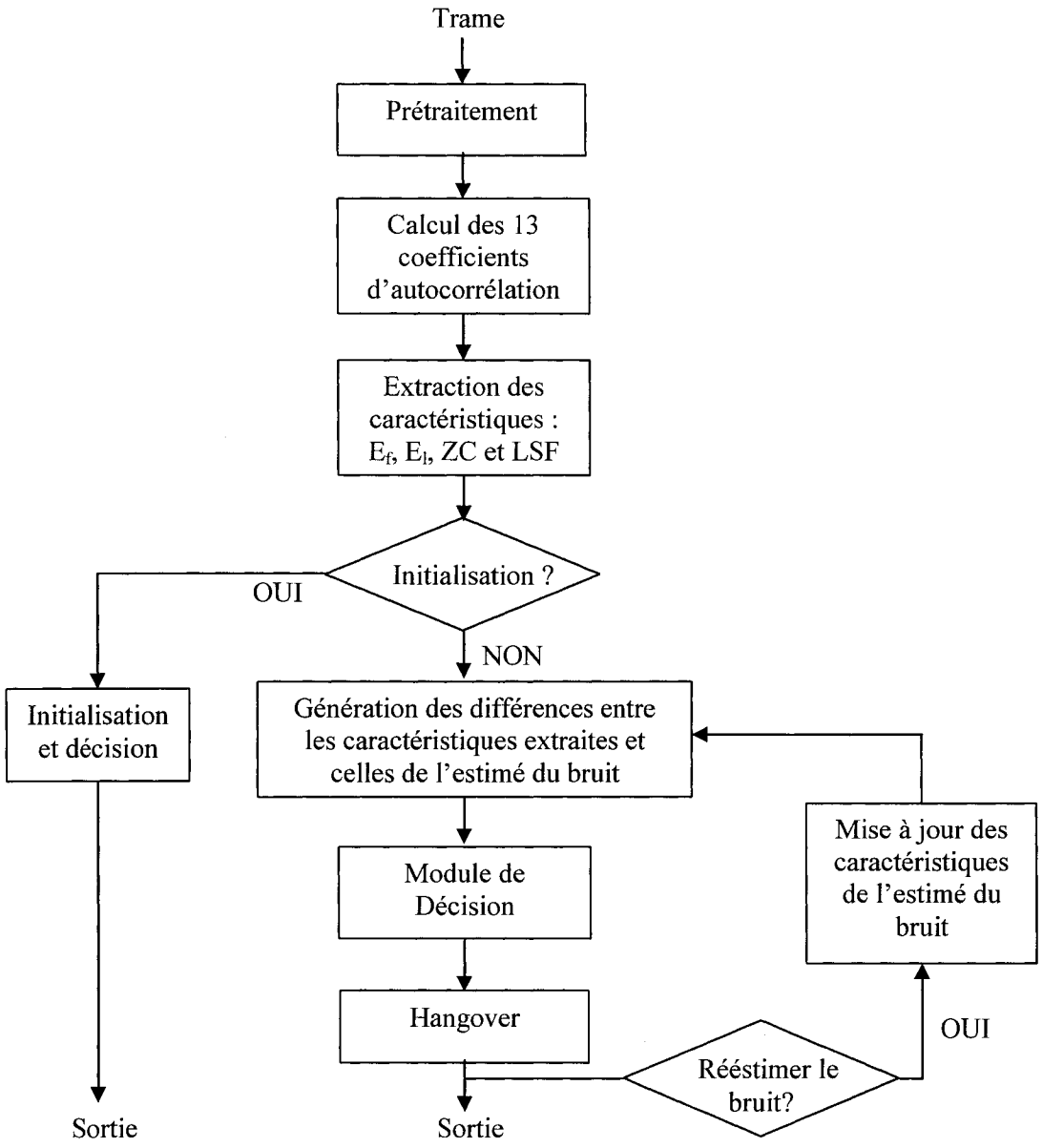


Figure 11 Schéma Bloc du DAV du G729
(Adaptée de la recommandation de ITU-T, 1996, [19])

Nous décrivons ci-après les différents blocs qui composent ce DAV.

Prétraitement :

La trame est traitée par un filtre passe-haut afin d'éliminer les parasites dans les basses fréquences (ITU-T [19]).

Calcul des 13 coefficients d'autocorrélation :

Pour pouvoir procéder au calcul des coefficients d'autocorrélation, une trame d'analyse est tout d'abord construite. Elle est constituée de la moitié de l'avant-dernière trame, de la dernière trame, de la trame actuelle et de la moitié de la trame future (ITU-T [19]), comme le montre la figure 12 :

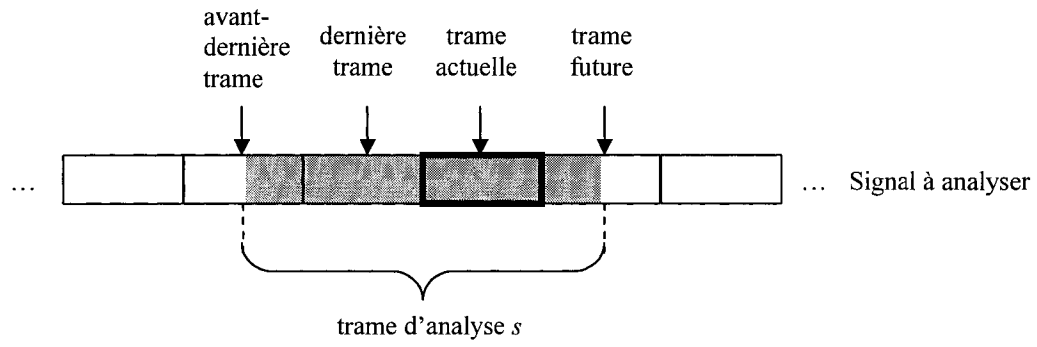


Figure 12 Formation de la trame d'analyse

L'ensemble des 13 coefficients d'autocorrélation $\{r(i)\}_{i=0}^{12}$ est ensuite déterminé à partir de cette trame d'analyse s :

$$r(k) = \sum_{n=k}^{L} s(n)s(n-k) \tag{2. 18}$$

Avec :

- $k = 0$ à 12
- L : la longueur de la trame d'analyse s

Extraction des caractéristiques :

Quatre caractéristiques sont ensuite extraites pour chaque trame à l'aide de ces coefficients d'autocorrélation :

- l'énergie dans toute la bande :

$$E_f = 10\log_{10}\left[\frac{r(0)}{L}\right] \tag{2. 19}$$

- l'énergie dans les basses fréquences [0-1kHz] :

$$E_l = 10\log_{10}\left[\frac{1}{L}h^T Rh\right] \tag{2. 20}$$

avec : h la réponse impulsionnelle finie du filtre passe-bas d'ordre 12 et de fréquence de coupure 1kHz et R la matrice de Toeplitz :

$$\begin{bmatrix} r(0) & \cdots & r(12) \\ \vdots & \ddots & \vdots \\ r(12) & \cdots & r(0) \end{bmatrix} \tag{2. 21}$$

- le taux de passages par zéro de la trame actuelle x :

$$ZC = \frac{1}{M}\sum_{i=0}^{M-1}\left|signe\left[x(i)\right] - signe\left[x(i-1)\right]\right| \tag{2. 22}$$

avec : M la longueur de la trame actuelle x

- les coefficients de raies spectrales $\{LSF_i\}_{i=1}^{10}$: les coefficients du filtre de prédiction linéaire, les coefficients LPC, sont tout d'abord générés à partir des 11 premiers coefficients d'autocorrélation, ceci grâce à la procédure de Levinson-Durbin et de manière à minimiser l'erreur de prédiction. Les 11

coefficients LPC sont ensuite convertis en un ensemble de 10 coefficients de raies spectrales $\{LSF_i\}_{i=1}^{10}$ (Benyassine et Al. [8]).

Initialisation :

Afin de prendre une décision, les quatre caractéristiques précédentes doivent être comparées à celles de l'estimé du bruit. Pour cela, il faut d'abord initialiser ces grandeurs. L'initialisation s'effectue sur les 32 premières trames et repose sur l'hypothèse qu'elles contiennent peu ou pas de parole. Cette étape revient à calculer les moyennes des quatre caractéristiques et permet d'obtenir les premiers paramètres de l'estimé du bruit : $\overline{E_f}$, $\overline{E_l}$, $\overline{ZC}$, $\overline{\{LSF_i\}_{i=1}^{10}}$. Le DAV doit fournir une décision en tout temps, donc même pendant la phase d'initialisation. Pour cela, une règle de décision basique a été mise en place (ITU-T [19]) :

$$\begin{array}{ll} Si & E_f < coefficient\ alors\ la\ trame\ est\ INACTIVE \\ Sinon & la\ trame\ est\ ACTIVE \end{array} \qquad (2.23)$$

Génération des différences des caractéristiques :

Les différences entre les caractéristiques du bruit et celles de la trame étudiée sont générées de la manière suivante :

$$\begin{cases} \Delta E_f = \overline{E_f} - E_f \\ \Delta E_l = \overline{E_l} - E_l \\ \Delta ZC = \overline{ZC} - ZC \\ DS = \sum_{i=1}^{10} \left[\overline{LSF(i)} - LSF(i) \right]^2 \end{cases} \qquad (2.24)$$

Avec : DS : la distorsion spectrale.

Module de décision :

La décision préliminaire est ensuite prise en fonction des grandeurs ΔE_f, ΔE_l, ΔZC, DS, et ceci à l'aide de 15 règles déterminées par expérimentation.

L'approche de la reconnaissance des formes a été choisie ici pour la classification. Selon les valeurs de leurs caractéristiques ΔE_f, ΔE_l, ΔZC, DS, toutes les trames de la base d'expérimentation ont été placées dans un espace euclidien à 4 dimensions, chaque grandeur correspondant à une dimension. Connaissant les trames contenant de la parole et celles qui en sont dépourvues, il a été possible d'identifier le nuage de parole et celui de bruit. Les règles du module de décision représentent la séparation entre ces deux nuages et ont été déterminées par inspection visuelle (Benyassine et Al. [8]).

Hangover :

Afin d'augmenter les performances, on utilise un module de *Hangover*, appelé aussi lissage. Celui-ci corrige la décision préliminaire afin de refléter la nature stationnaire long terme de la parole (ITU-T [19]). Il utilise quatre règles heuristiques, déterminées par expérimentation, liant les caractéristiques de la trame actuelle et des précédentes (Benyassine et Al. [8]). Il a pour but d'éviter les coupures dans une longue période de parole et de faciliter la détection des fins de bouffées de parole qui sont généralement caractérisées par de basses énergies. Son principe est le suivant : si une bouffée de parole a été détectée et que sa durée est suffisamment longue, la sortie du DAV va être bloquée à l'état PAROLE durant une période dite de *Hangover*. Ce module permet aussi de prévenir les faux déclenchements provoqués par des pics de bruit intenses et soudains dans les périodes d'inactivité vocale.

Mise à jour des caractéristiques de l'estimé du bruit :

La décision finale a été obtenue à l'aide du module de *Hangover*. Toutefois, une dernière étape peut s'avérer nécessaire : la mise à jour des caractéristiques de l'estimé du bruit. Elle est effectuée si :

$$\Delta E_f > coefficient \quad \& \quad 2^{ème}\, coefficient\ de\ réflexion < coefficient \quad \& \quad DS < coefficient$$

Ce module permet de faire face à des changements dans le bruit et utilise un système autorégressif du premier ordre pour recalculer les grandeurs $\overline{E_f}$, $\overline{E_l}$, $\overline{ZC}$, $\overline{\{LSF_i\}_{i=1}^{10}}$ (ITU-T [19]) :

$$\overline{g} = \beta_g \overline{g} + (1-\beta_g)\overline{g} \tag{2.25}$$

Avec :

- $\overline{g} = \overline{E_f}$, $\overline{E_l}$, $\overline{ZC}$ ou $\overline{LSF}$: les grandeurs à recalculer
- $\beta_g = \beta_{E_f}$, β_{E_l}, β_{ZC} ou β_{LSF} : les coefficients de régression

De plus, il contient une remise à zéro permettant d'éviter le blocage lorsque le bruit devient soudainement trop important. En effet, dans ce cas le DAV serait bloqué à l'état PAROLE sans aucune possibilité de remise à jour des caractéristiques de l'éstimé du bruit (Benyassine et Al. [8]).

L'algorithme du G729.B a l'avantage d'être rapide car même s'il est plus complexe que ceux vus précédemment, il reste simple. Il extrait quatre paramètres intéressants et les compare à ceux de l'estimé du bruit. En plus de cela, il possède un système de correction de la décision afin de réduire les erreurs de classification. Il est donc possible que ce DAV offre une bonne robustesse. Ce point sera discuté dans la conclusion de ce chapitre.

2.2.3 Les DAV de l'AMR

En 1998, ETSI [20] a proposé les standards pour les deux options de détection d'activité vocale du codeur de parole à multi-taux adaptatif (AMR). L'AMR a été développé par GSM pour les systèmes de communications mobiles de troisième génération. Ces deux DAV, que nous appellerons par la suite AMR1 et AMR2, différent à la fois par leur approche et par leur complexité. Toutefois, ils sont tous deux compatibles avec le codeur.

La première méthode, l'AMR1, peut être schématisée par la figure 13 :

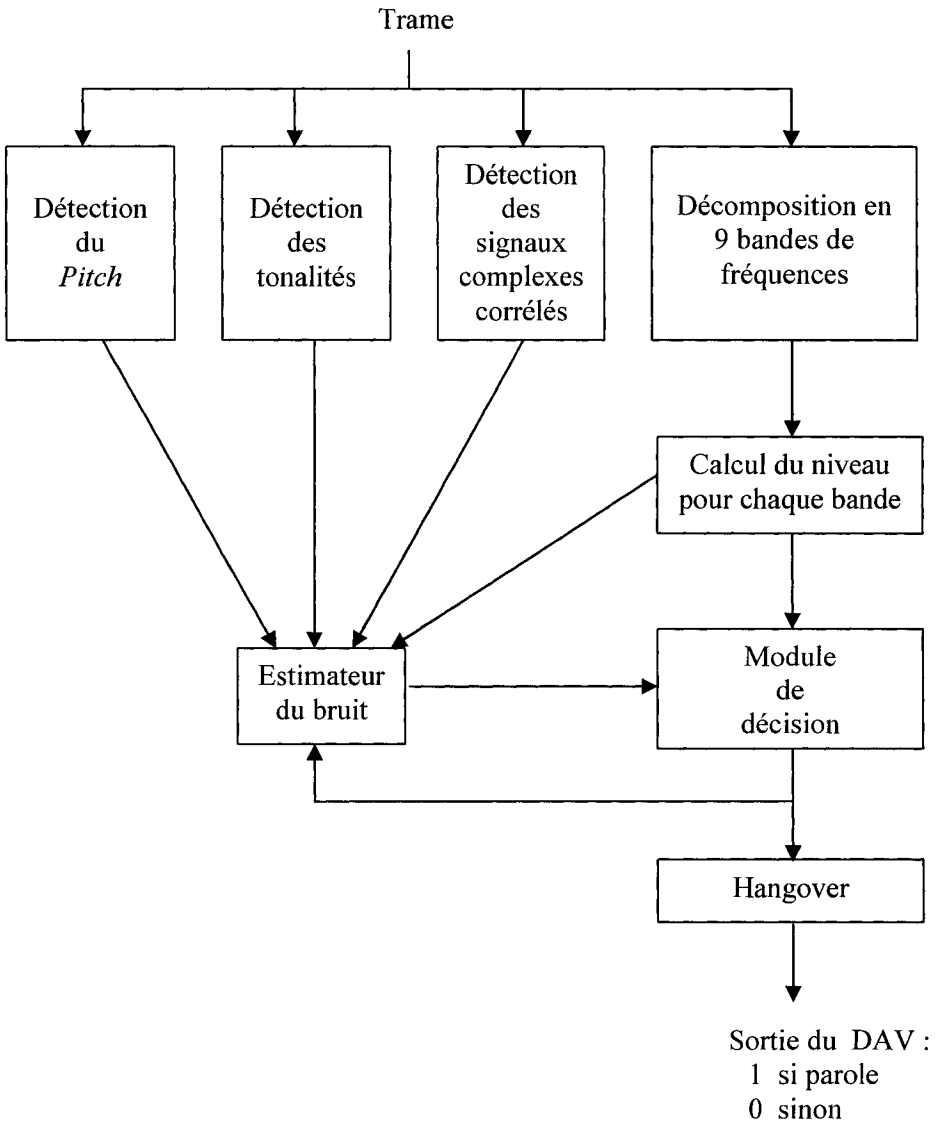


Figure 13 Schéma Bloc du DAV AMR1
(adaptée du standard de ETSI, 1998, [20])

Détection du pitch et détection des tonalités :

Dans leur article, Vähätalo et Johansson [32] expliquent l'utilité de ces modules comme suit. Ces deux détecteurs vont identifier la présence de signaux stationnaires, comme les sons voisés et les tonalités d'information. Il est important de détecter séparément ces signaux stationnaires car la mise à jour de l'estimé du bruit ne sera pas effectuée de la même manière lorsqu'ils sont présents. Il est nécessaire d'utiliser ces deux détecteurs conjointement car chacun compense les faiblesses de l'autre. En effet, le détecteur du *pitch* n'identifie pas correctement les signaux ayant plus d'une fréquence fondamentale. Celui des tonalités les détecte. À l'inverse, le détecteur de tonalités identifie les sons voisés seulement quand le RSB est élevé. Celui du *pitch* est, lui, capable de les détecter quand le RSB est plus faible.

Détection des signaux complexes corrélés :

Comme il a été mentionné dans le chapitre 1, les codeurs de parole transmettent un bruit de confort lorsque la sortie du DAV est nulle. Selon ETSI [20], en présence de signaux complexes, comme la musique, cette opération risque d'être gênante pour la conversation. Elle ne doit donc pas être effectuée. Pour cela, un détecteur de signaux complexes corrélés est nécessaire. Il consiste à filtrer le signal d'entrée par un filtre passe-haut. Si la sortie du filtre contient des hautes valeurs de corrélation alors un signal complexe corrélé est détecté. Cela repose sur l'hypothèse que les signaux musicaux possèdent des harmoniques même dans les hautes fréquences alors que le bruit, en général, n'en a que dans les basses fréquences (Vähätalo et Johansson [32]).

La description des trois étapes suivantes fait référence aux standards de ETSI [20].

Décomposition en 9 bandes de fréquences :

Afin de pouvoir extraire les caractéristiques utilisées pour la prise de décision, la trame étudiée est décomposée en 9 bandes de fréquences à l'aide du banc de filtres illustré par la figure 14.

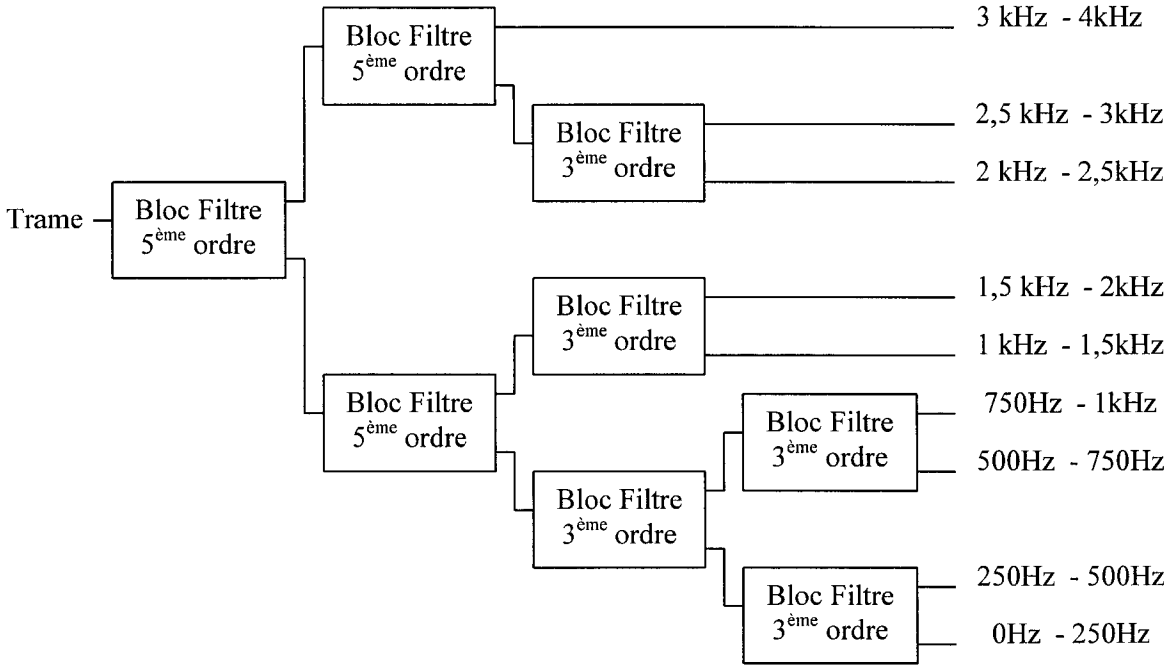


Figure 14 Banc de filtres utilisé par l'AMR1
(adaptée de l'article de Vähätalo et Johansson, 1999, [32])

Chaque bloc divise son entrée en une partie passe-haut et une passe-bas et effectue aussi une décimation par 2.

D'après Vähätalo et Johansson [32], la décomposition en plus de 9 bandes ne procure pas d'amélioration significative et augmente la complexité.

Calcul du niveau pour chaque bande :

Le niveau du signal est calculé pour chaque bande :

$$niveau(k) = \sum |x_k| \qquad (2.\ 26)$$

Avec :

- k : la bande de fréquences
- x_k : le signal de la bande k

Module de décision :

Comme pour la méthode précédente, un système d'estimation du bruit a été mis en place. À l'aide de ce dernier, la différence entre les niveaux de la trame étudiée et ceux de l'estimé du bruit est déterminée :

$$som\_rsb = \sum_{k=1}^{9} \max\left(1, \frac{niveau(k)}{estimé\_bruit(k)}\right)^2 \tag{2. 27}$$

La valeur ainsi calculée est ensuite comparée à un seuil et la décision préliminaire est obtenue:

$$\begin{aligned} &Si\ \ som\_rsb > seuil\ \ alors\ la\ trame\ est\ ACTIVE \\ &Sinon \qquad\qquad\qquad\qquad la\ trame\ est\ INACTIVE \end{aligned} \tag{2. 28}$$

Il est à noter que le seuil utilisé par la règle de décision est adaptatif et dépend du niveau de bruit dans chaque bande.

Hangover :

De même que pour le G729.B, un module de *Hangover* est utilisé pour corriger la décision préliminaire, si cela est nécessaire. Son principe est le même sauf que la longueur de la période de *Hangover* est ici variable et dépend du niveau de l'estimé du bruit (Vähätalo et Johansson [32]). De plus, ce module a un but supplémentaire : laisser passer les signaux complexes corrélés. Si le détecteur indique la présence d'un signal complexe corrélé, la sortie du DAV va être forcée à l'état PAROLE afin d'empêcher le déclenchement du générateur de bruit de confort utilisé à la suite du DAV (ETSI [20]).

Estimateur du bruit :

La mise à jour de l'estimé du bruit engendre un retard d'une trame car elle s'effectue selon les niveaux d'amplitude de la trame précédente. Ceci a pour but de faciliter la détection des débuts de bouffées de parole, ETSI [20]. Vähätalo et Johansson [32] expliquent le fonctionnement de l'estimateur du bruit de la manière suivante. Si la décision préliminaire est PAROLE ou si le *pitch* ou une tonalité est détectée, l'estimé du bruit sera revu à la baisse. Dans le cas contraire, il sera revu à la hausse. L'estimé du

bruit est réajusté dans chacune des 9 bandes de fréquences de manière indépendante et la vitesse de mise à jour α est aussi propre à chacune d'entre elles. Il est à noter que cette vitesse est plus élevée lors d'une revue à la baisse. L'estimation du bruit est régie par l'équation (2.29) :

$$bruit_{j+1}[k] = (1-\alpha) \times bruit_j[k] + \alpha \times niveau_{j-1}[k] \qquad (2.29)$$

Avec :

- k : la bande de fréquences
- j : l'indice de la trame

Comme pour le G729.B, un mécanisme pour faire face à une augmentation importante et soudaine du bruit a été mis en place. En effet, dans ce cas la sortie du DAV serait à l'état PAROLE et l'estimé du bruit serait continuellement revue à la baisse. Pour éviter cela, l'augmentation de l'estimé du bruit est permise lorsque la décision préliminaire est PAROLE pendant une période suffisamment longue et lorsque le spectre du signal est stationnaire.

La seconde méthode, celle de l'AMR2, est beaucoup plus complexe et nous ne la décrirons ici que très sommairement, en nous basant sur l'article de Cornu et Al. [33] et les standards de ETSI [20].

La trame étudiée est séparée en deux sous trames. Chacune d'entre elles subit alors la même procédure, illustrée par la figure 15.

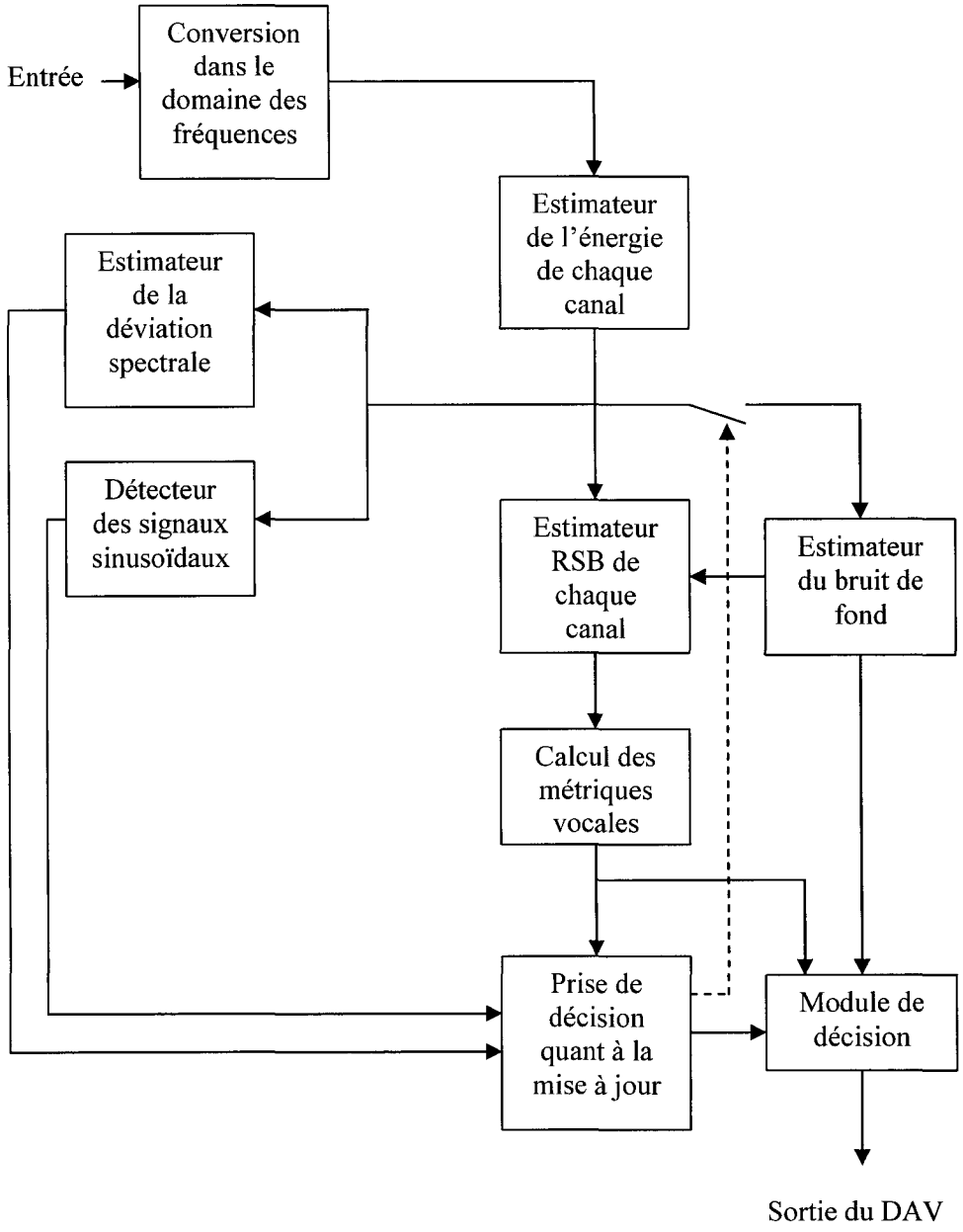


Figure 15 Schéma Bloc du DAV AMR2
(adaptée du standard de ETSI, 1998, [20])

Le signal d'entrée est d'abord converti dans le domaine des fréquences grâce à la transformée de Fourier. Les bandes de fréquences sont regroupées en N_C canaux. Les énergies de chacun d'entre eux sont déterminées. Une fois encore, un estimateur de bruit

a été mis au point et permet de calculer les rapports signal à bruit pour les N_C canaux. Les métriques vocales sont déduites de ces RSB grâce à une fonction non-linéaire. Une décision préliminaire est obtenue pour chaque sous-trame en comparant la somme des métriques vocales à un seuil adaptatif dépendant du bruit. Finalement, la trame entière est dite active si au moins l'une des deux sous-trames est elle-même active.

Les autres blocs sont utilisés pour la mise à jour du bruit. L'estimateur de la déviation spectrale et le détecteur des signaux sinusoïdaux ont pour but d'éviter les réajustements inadéquats de l'estimé du bruit. Ils génèrent des drapeaux unitaires s'il y a détection respective d'une trop grande déviation ou d'un signal sinusoïdal. Ces derniers sont utilisés pour déterminer si une mise à jour de l'estimé du bruit est nécessaire. Si c'est le cas, l'estimateur du bruit de fond réajuste les caractéristiques du bruit.

Les deux DAV présentés ici utilisent une décomposition en bandes de fréquences sur lesquelles ils extraient ensuite les caractéristiques utiles à la prise de décision. Cette approche est très intéressante car elle repose sur le fait que le bruit seul et la parole bruitée ne se répartissent pas de la même manière sur les plages de fréquences. L'AMR1 est plus complexe que le G729.B mais il a l'avantage de réestimer les caractéristiques du bruit continuellement. L'AMR2, quant à lui, est bien plus lourd d'un point de vue computationnel. La manière de déterminer si une mise à jour de l'estimé du bruit est nécessaire est beaucoup plus élaborée. Elle serait donc susceptible de rendre ce DAV plus robuste que l'AMR1.

2.3 Comparaison des méthodes de base et de leur mise en application

Les méthodes de base de la détection d'activité vocale, décrites dans la première partie de ce chapitre, présentent des performances acceptables uniquement lorsque le rapport signal à bruit est élevé. La seule capable de fonctionner en présence de bruit fort est celle utilisant un estimateur de périodicité par les moindres carrés mais sa sensibilité aux interférences périodiques la rend difficile à utiliser. Ce projet de recherche consiste à mettre au point un DAV fonctionnel dans les milieux industriels, où l'on rencontre habituellement de faibles RSB. Ces procédures de base ne sont donc pas adéquates pour ce projet. Toutefois, l'étude des DAV utilisés par trois codeurs de parole a montré que les algorithmes plus robustes reposent sur ces méthodes fondamentales. En effet, ils reprennent et améliorent un ou plusieurs de leurs concepts.

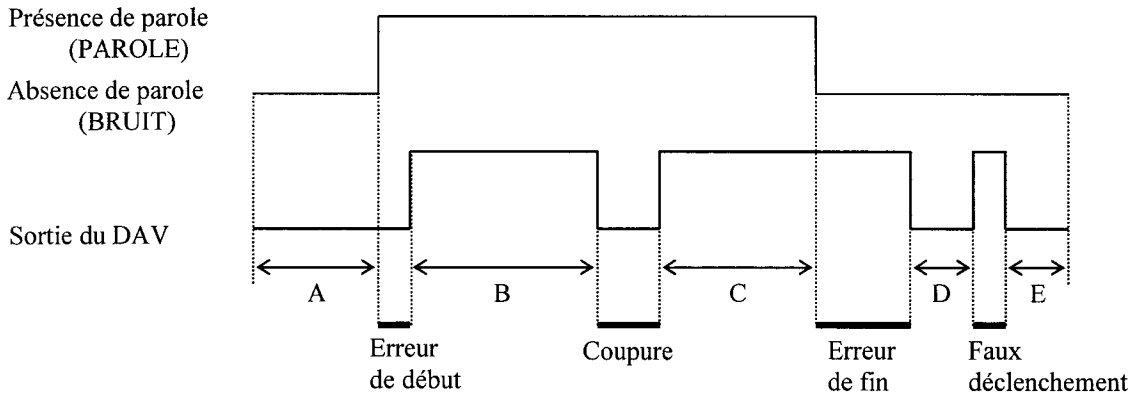
Comme il a été mentionné précédemment le DAV développé pour le Pan European a été l'un des premiers DAV à avoir été utilisé dans un système complexe. Malheureusement, il ne présente pas de résultats intéressants dans toutes les situations, surtout lorsque le bruit est non stationnaire (El-Maleh et Kabal [30]). Il reste maintenant à déterminer les avantages et les inconvénients des trois derniers DAV, soit le G729.B, l'AMR1 et l'AMR2.

Dans leur article, Beritelli et Al. [34] comparent les performances de quatre DAV : G729.B, AMR1, AMR2 et un DAV en logique floue. Nous ne retiendrons ici que les résultats obtenus pour les trois premiers. Afin d'évaluer leur efficacité, quatre critères objectifs couramment utilisés dans la littérature ont été retenus (Beritelli et Al. [35], [36]) :

- Erreur de début : erreur introduite lors du passage d'une période de bruit à une période de parole. Le DAV continue à indiquer l'absence de parole alors que l'activité vocale a déjà commencé.

- Erreur de fin : erreur introduite lors du passage d'une période de parole à une période de bruit. Le DAV continue à indiquer la présence de parole alors que celle-ci est déjà terminée.
- Coupure : erreur se produisant au cours d'une bouffée de parole, la parole est soudainement prise pour du bruit.
- Faux déclenchement : erreur se produisant au cours d'une période d'inactivité vocale, le bruit est soudainement pris pour de la parole.

La figure 16 illustre les paramètres utilisés pour l'évaluation des performances d'un DAV :



Reconnaissance de la parole = B + C
= Parole – (Erreur début + Coupure)

Reconnaissance du bruit = A + D + E
= Bruit – (Erreur fin + Faux déclenchement)

Figure 16 Paramètres objectifs pour l'évaluation des performances d'un DAV

Les résultats du G729.B, de l'AMR1 et de l'AMR2 présentés dans l'article de Beritelli et Al. [34] ne prennent toutefois en compte que deux critères objectifs:

- l'erreur « parole prise pour du bruit » : il s'agit du nombre de trames de parole qui ont été déclarées inactives par le DAV, c'est-à-dire la somme des erreurs de début et des coupures comme représentées sur la figure 16.
- l'erreur totale : c'est l'ensemble des erreurs de classification, que ce soit les trames de parole prises pour du bruit ou celles du bruit prises pour de la parole. Il s'agit donc de la somme des erreurs de début, des erreurs de fin, des coupures et des faux déclenchements, comme représentés sur la figure 16.

En plus de ça, un critère subjectif a été retenu afin d'estimer la qualité de l'intelligibilité des résultats obtenus. Il s'agit de l'*Activity Burst Corruption*, paramètre psychoacoustique issu de l'une de leurs précédentes recherches (Beritelli et Al. [37]). Ce paramètre remplace les tests habituellement effectués par un groupe de personnes qui écoutent et jugent les phrases à la sortie du DAV.

Chacun des trois DAV a été testé sur la même base de données. Cette base contient des phrases prononcées par différents locuteurs (hommes, femmes) et dans des langues différentes (Italien, Anglais, Français, Allemand). Chaque séquence contient environ 40% de parole afin de refléter le pourcentage moyen d'activité vocale dans une conversation téléphonique. En plus de cela, différents niveaux de signal (-16, -26, -36dBovl), différents bruits (auto, bureau, train, restaurant, rue) et différents rapports signal à bruit (0, 10, 20dB) ont été utilisés.

En comparant les résultats obtenus, les conclusions suivantes ont été tirées par Beritelli et Al. [34] :

- le G729.B est le plus simple mais présente les moins bonnes performances tant du point de vue objectif que subjectif. Il s'agit du DAV le moins robuste.
- l'AMR2 offre une bonne résistance aux changements de niveaux de la parole mais son comportement se dégrade très vite plus le rapport signal à bruit est faible, c'est-à-dire plus il y a de bruit. Il s'agit du DAV le plus complexe.
- l'AMR1 se comporte de la manière inverse. En effet, plus le niveau de la parole est bas, plus il a des difficultés à reconnaître la voix. Par contre, ses performances se dégradent moins brutalement avec le RSB. Sa complexité est moyenne.

Il est à noter que, quelque soit le DAV, les performances sont plus élevées lorsque la langue utilisée est l'italien ou le français car les vocalisations sont plus grandes.

Pour conclure, nous pouvons dire que, comme dans bien des cas, il faut choisir entre la facilité d'implémentation et l'efficacité. Il semble préférable pour un usage en télécommunications de privilégier l'AMR1 car il offre le meilleur compromis : complexité et robustesse. Dépendant de l'application, il est possible de choisir entre ces trois DAV.

CHAPITRE 3

LE DÉTECTEUR D'ACTIVITÉ VOCALE G729.B

Aucune des méthodes de détection d'activité vocale présentées dans le chapitre 2 n'a été mise au point pour un milieu industriel. Il est donc nécessaire de se baser sur une procédure utilisée dans les télécommunications pour pouvoir réaliser la première étape de ce projet de recherche. Nous avons présenté trois DAV intéressants à savoir le G729.B, l'AMR1 et l'AMR2. À priori c'est l'AMR1 qui offre le meilleur compromis entre complexité et efficacité. Toutefois, il n'est pas possible de prévoir le comportement de ces DAV dans un environnement industriel bruité puisque le type de bruits et les rapports signal à bruit sont différents de ceux que l'on rencontre habituellement en télécommunications. Ainsi c'est le DAV du G729 qui a été retenu comme premier sujet de notre étude car c'est la méthode la mieux documentée et aussi la plus évidente à mettre en œuvre. Le G729.B a donc tout d'abord été implémenté et simulé avec le logiciel MATLAB, en suivant les recommandations de ITU-T [19] et [38]. Par la suite, quelques tests ont été effectués dans le milieu d'étude. Compte tenu de son comportement, des modifications ont dû être apportées.

L'algorithme détaillé du G729.B comme proposé par ITU-T sera présenté dans la première partie de ce chapitre. Suite à cela, les modifications apportées à ce DAV seront exposées. Enfin, la dernière partie de ce chapitre sera consacrée aux résultats pratiques, obtenus avec l'algorithme ajusté dans les milieux industriels.

3.1 Algorithme détaillé du G729.B

Toute cette partie se base sur les recommandations de ITU-T concernant le codeur G729 [38] et le DAV en tant que tel [19].

La fréquence d'échantillonnage attendue par le G729.B est 8kHz. Chaque trame de 10ms, soit 80 échantillons, subit la procédure présentée par la figure 17 :

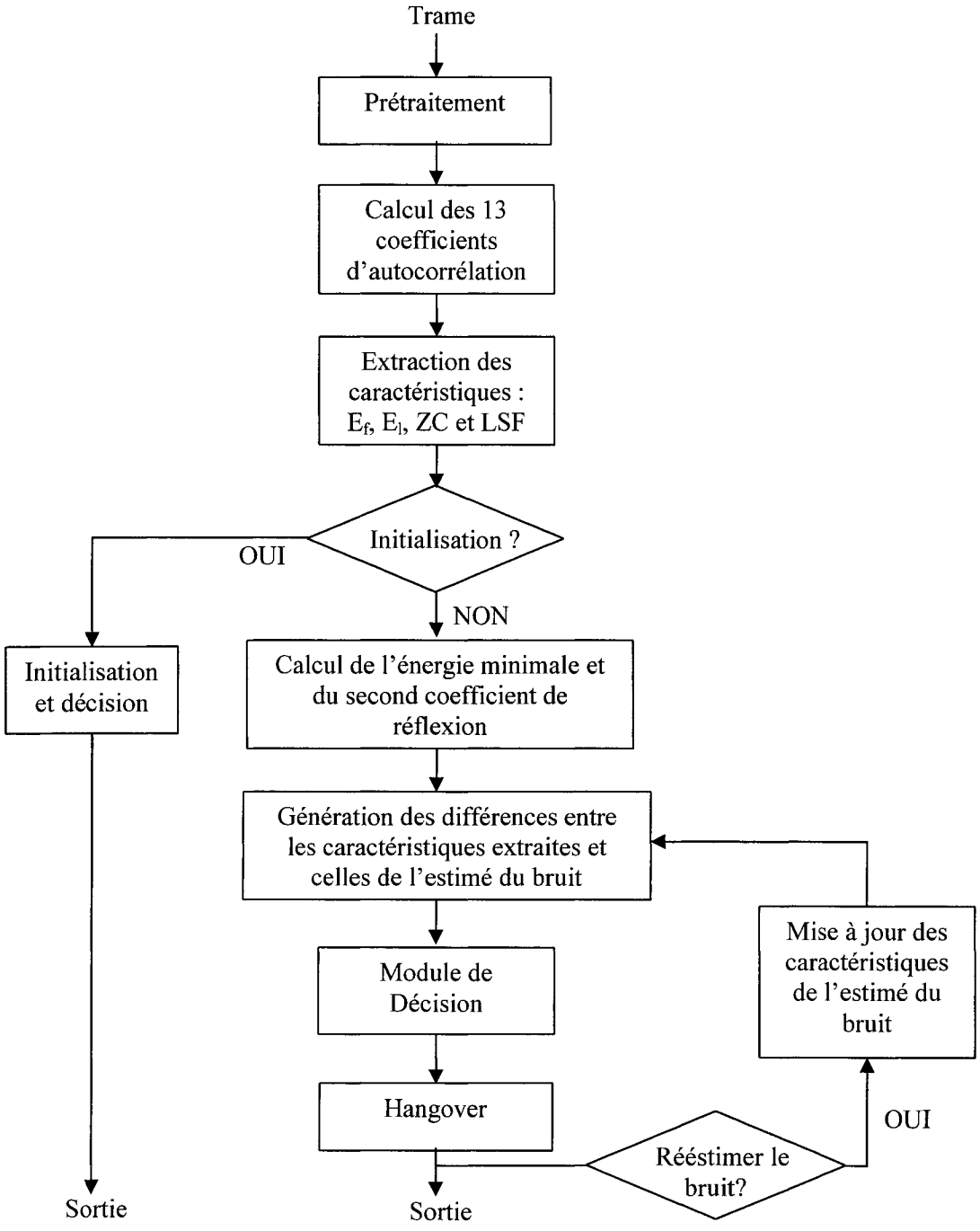


Figure 17 Schéma Bloc du G729.B
(adaptée de la recommandation de ITU-T, 1996, [19])

Prétraitement :

La trame est tout d'abord filtrée par un filtre passe-haut de fréquence de coupure 140Hz combiné à un normalisateur par 2. La fonction de transfert de ce filtre est donnée par :

$$H(z) = \frac{0.46363718 - 0.92724705z^{-1} + 0.46363718z^{-2}}{1 - 1.9059465z^{-1} + 0.9114024z^{-2}} \qquad (3.1)$$

La normalisation par deux a pour but d'éviter le dépassement de capacité lors d'une implémentation en virgule fixe. Le filtrage passe-haut, quant à lui, permet d'éliminer les parasites basses fréquences.

Calcul des 13 premiers coefficients d'autocorrélation :

Comme il a été dit dans le chapitre 2, section 2.2.2, une trame d'analyse est tout d'abord formée, comme montré par la figure 18:

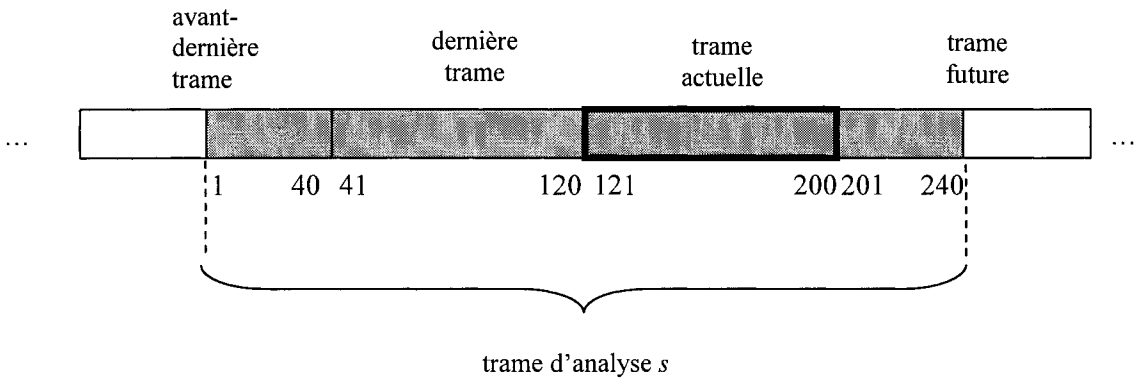


Figure 18 Formation de la trame d'analyse

Cette trame d'analyse de longueur 240 échantillons est extraite à l'aide d'une fenêtre adoucie, propre au G729.B et définie par :

$$w(i) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi(i-1)}{399}\right) & \textit{pour } i = 1 \textit{ à } 200 \\ \cos\left(\dfrac{2\pi(i-201)}{159}\right) & \textit{pour } i = 201 \textit{ à } 240 \end{cases} \qquad (3.2)$$

Cette fenêtre adoucie tient compte de l'importance de chaque trame. Comme on peut le voir sur la figure 19, les échantillons de la trame actuelle (compris entre 121 et 200) sont conservés de manière plus intacte car ils contiennent l'information la plus intéressante pour l'analyse.

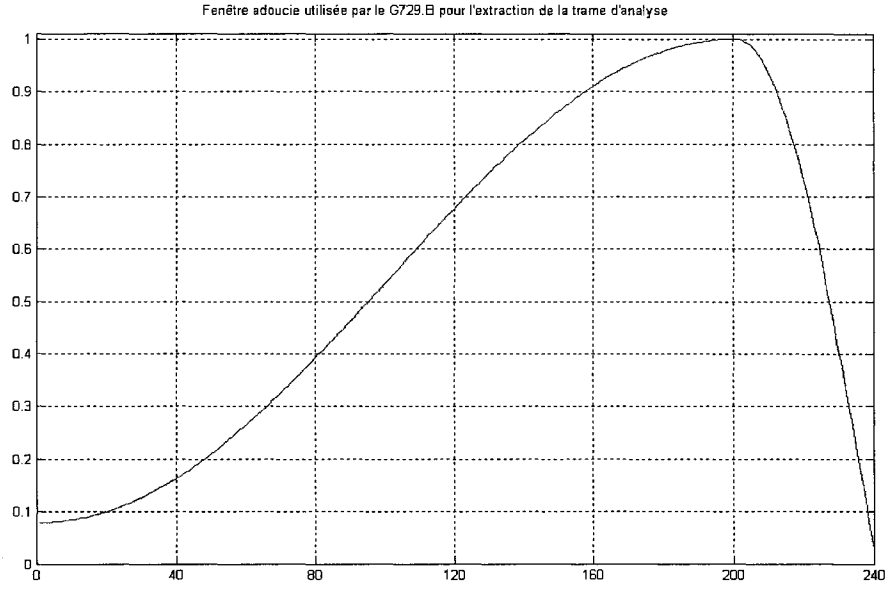


Figure 19 Fenêtre adoucie utilisée par le G729.B

Lorsqu'on utilise une fenêtre rectangulaire pour extraire des trames, cela engendre la phénomène de Gibbs : réponse fréquentielle présentant d'importantes oscillations au niveau des discontinuités. Pour éviter cela, on a recours à une fenêtre adoucie (Oppenheim et Schafer [7]). Il en existe plusieurs et le choix d'une telle fenêtre est en général arbitraire.

L'ensemble des 13 coefficients d'autocorrélation $\{r(i)\}_{i=0}^{12}$ est ensuite déterminé à partir de la trame d'analyse s :

$$r(k) = \sum_{n=k}^{240} s(n)s(n-k) \tag{3.3}$$

Avec :

- $k = 0$ à 12

Extraction des caractéristiques :

Quatre paramètres sont ensuite extraits : l'énergie dans toute la bande E_f, l'énergie dans les basses fréquences E_l, le taux de passages par zéro ZC et les 10 coefficients de raies spectrales $\{LSF_i\}_{i=1}^{10}$, voir chapitre 2 section 2.2.2.

Initialisation :

L'initialisation a pour but de déterminer les quatre premières caractéristiques de l'estimé du bruit. Elle s'effectue sur les 32 premières trames du signal et repose sur l'hypothèse qu'il y a peu ou pas de parole pendant cette période. Les moyennes de E_f, ZC et $\{LSF_i\}_{i=1}^{10}$ sur ces 32 trames permettent de déterminer les trois premiers paramètres de l'estimé du bruit :

$$\overline{E_f} = \frac{1}{32}\sum_{k=1}^{32} E_{fk} \tag{3.4}$$

$$\overline{ZC} = \frac{1}{32}\sum_{k=1}^{32} ZC_k \tag{3.5}$$

$$\overline{\{LSF_i\}_{i=1}^{10}} = \frac{1}{32}\sum_{k=1}^{32}\left(\{LSF_i\}_{i=1}^{10}\right)_k \tag{3.6}$$

La valeur de $\overline{E_f}$ est ensuite utilisée pour caractériser la présence de la parole pendant l'initialisation. Ainsi selon cette valeur, trois cas sont possibles : « pas de parole », « très peu de parole » ou « peu de parole ». Dépendant du cas, on en déduit $\overline{E_l}$ et on corrige $\overline{E_f}$ si cela est nécessaire :

- Cas où il n'y a pas de parole :

$$Si\ \ \overline{E_f} \le a\ \ alors\ \ \overline{E_l} = \overline{E_f} - c1 \tag{3.7}$$

- Cas où il y a très peu de parole :

$$Si\ \ a < \overline{E_f} \le b\ \ alors\ \ \overline{E_l} = \overline{E_f} - c2 \quad et\ \ \overline{E_f} = \overline{E_f} - c3 \tag{3.8}$$

- Cas où il a peu de parole :

$$Si \ \ b < \overline{E_f} \ \ alors \ \ \overline{E_l} = \overline{E_f} - c4 \qquad (3.9)$$
$$et \ \ \overline{E_f} = \overline{E_f} - c5$$

Avec : a, b, $c1$ à $c5$ des constantes.

Lors de l'initialisation, le DAV fournit une décision à l'aide d'une règle simple basée sur la valeur E_f de la trame actuelle :

$$\begin{array}{lll} Si & E_f < coefficient & alors \ \ la\ trame\ est\ BRUIT \\ Sinon & & la\ trame\ est\ PAROLE \end{array} \qquad (3.10)$$

Calcul de l'énergie minimale et du second coefficient de réflexion :

L'énergie minimale $E_{\min}$ est utilisée pour la remise à zéro de l'estimé du bruit si un blocage de la mise à jour survient. Il s'agit simplement de déterminer le minimum de E_f sur les 128 dernières trames.

Le second coefficient de réflexion rc est, quant à lui, utilisé afin de déterminer si une réévaluation de l'estimé du bruit est nécessaire.

Génération des différences des paramètres :

Les différences ΔE_f, ΔE_l, ΔZC, DS entre les caractéristiques de la trame et ceux de l'estimé du bruit sont obtenues comme mentionné dans le chapitre 2 section 2.2.2.

Module de décision :

Le module de décision est composé de deux parties. La première est en fait constituée d'une seule et unique règle. Elle a pour but d'identifier les trames qui sont assurément dépourvues de parole. L'idée est que si l'énergie E_f est très petite, cela implique que la trame ne peut pas contenir de la parole. La décision préliminaire, appelée aussi décision1, est alors BRUIT et on passe directement au module de *Hangover* :

$$Si \ \ E_f < coefficient \ \ alors \ \ decision1 = BRUIT \qquad (3.11)$$

Si la règle précédente n'est pas vérifiée, cela implique que la trame actuelle est susceptible de contenir de la parole. La deuxième partie du module de décision va donc déterminer la présence ou l'absence d'activité vocale. Cette prise de décision se fait en examinant les valeurs de ΔE_f, ΔE_l, ΔZC, DS. Selon la recommandation du G729.B [19] et Benyassine et Al. [8], il y a 14 règles qui ont été obtenues par expérimentation :

$$\begin{cases} Si\ \ DS > a_1 \Delta ZC + b_1 & alors\ \ decision1 = PAROLE \\ Si\ \ DS > a_2 \Delta ZC + b_2 & alors\ \ decision1 = PAROLE \\ Si\ \ \Delta E_f < a_3 \Delta ZC + b_3 & alors\ \ decision1 = PAROLE \\ Si\ \ \Delta E_f < a_4 \Delta ZC + b_4 & alors\ \ decision1 = PAROLE \\ Si\ \ \Delta E_f < b_5 & alors\ \ decision1 = PAROLE \\ Si\ \ \Delta E_f < a_6 DS + b_6 & alors\ \ decision1 = PAROLE \\ Si\ \ DS > b_7 & alors\ \ decision1 = PAROLE \\ Si\ \ \Delta E_l < a_8 \Delta ZC + b_8 & alors\ \ decision1 = PAROLE \\ Si\ \ \Delta E_l < a_9 \Delta ZC + b_9 & alors\ \ decision1 = PAROLE \\ Si\ \ \Delta E_l < b_{10} & alors\ \ decision1 = PAROLE \\ Si\ \ \Delta E_l < a_{11} DS + b_{11} & alors\ \ decision1 = PAROLE \\ Si\ \ \Delta E_l > a_{12} \Delta E_f + b_{12} & alors\ \ decision1 = PAROLE \\ Si\ \ \Delta E_l < a_{13} \Delta E_f + b_{13} & alors\ \ decision1 = PAROLE \\ Si\ \ \Delta E_l < a_{14} \Delta E_f + b_{14} & alors\ \ decision1 = PAROLE \end{cases} \qquad (3.12)$$

Ces règles sont indépendantes les unes des autres. Ainsi, si l'une ou plusieurs d'entre elles sont vérifiées, la trame est PAROLE. Si aucune de ces règles n'est vérifiée, la trame est BRUIT. Les coefficients a_i et b_i sont des constantes qui ont été déterminées par expérimentations.

Hangover :

La décision finale *vad_flag* est obtenue en corrigeant la décision préliminaire à l'aide des quatre règles du *Hangover*, ITU-T [19].

- **Règle 1 :** Si la décision préliminaire indique BRUIT mais que la trame précédente contenait de la parole et que l'énergie de la trame actuelle est bien plus grande que celle de l'estimé du bruit, il faut corriger la décision :

$$\begin{aligned} &Si\ (decision1 = BRUIT \\ &\&\ dernier\_vad\_flag = PAROLE \\ &\&\ E_f > \overline{E_f} + coefficient) \\ &Alors\ vad\_flag\ =\ PAROLE \end{aligned} \tag{3.13}$$

- **Règle 2 :** Cette étape a pour but d'assurer la détection des fins de bouffées de parole, généralement difficiles à détecter de par leur faible énergie. Il s'agit ici de forcer la sortie à 1 pendant une période de N_1 trames si les deux dernières contenaient de la parole. Une variable logique *vl* est utilisée : elle est unitaire tant que la période de N_1 trames n'est pas terminée et nulle dans le cas contraire afin d'arrêter le blocage à 1 :

$$\begin{aligned} &Si\ (vl = 1\ \&\ decision1 = BRUIT \\ &\&\ dernier\_vad\_flag = PAROLE \\ &\&\ avant\_dernier\_vad\_flag = PAROLE \\ &\&\ \left|E_f - E_{fprec}\right| \le coefficient\ ) \\ &Alors\ vad\_flag = PAROLE \quad // \ Sortie\ forcée\ à\ 1 \\ &\quad compteur++ \\ &\quad Si\ compteur > N_1 \quad // \ Période\ de\ N_1\ trames\ finie \\ &\quad\quad vl = 0 \quad // \ Arrêt\ du\ blocage \\ &\quad\quad compteur = 0 \end{aligned} \tag{3.14}$$

- **Règle 3 :** Cette règle évite les déclenchements intempestifs dans une longue période d'inactivité vocale. Si la décision est PAROLE alors que les N_2 trames d'avant étaient BRUIT et que l'énergie de la trame est plus faible que celle de la trame précédente, alors la décision doit être corrigée.

$$\begin{aligned} &Si\ (vad\_flag = PAROLE \\ &\&\ compteur\_bruit > N_2 \\ &\&\ E_f - E_{fprec} \leq coefficient\ ) \\ &Alors\ vad\_flag = BRUIT \\ &\qquad compteur\_bruit = 0 \end{aligned} \tag{3.15}$$

$$\begin{aligned} &Si\ vad\_flag = BRUIT \\ &\qquad compteur\_bruit{+}{+} \\ &Sinon \\ &\qquad compteur\_bruit = 0 \end{aligned}$$

- **Règle 4 :** Si la décision indique BRUIT mais que son énergie est largement supérieure à celle de l'estimé du bruit alors il faut corriger la décision :

$$\begin{aligned} &Si\ (vad\_flag = BRUIT \\ &\&\ numero\_trame > 128 \\ &\&\ E_f > \overline{E_f} + coefficient\ ) \\ &Alors\ vad\_flag = PAROLE \end{aligned} \tag{3.16}$$

La condition concernant le numéro de la trame permet simplement de s'assurer qu'on est assez loin de l'initialisation et donc que l'estimé du bruit est correct.

Selon Alcatel [39], cette dernière règle est la seule à ne pas tenir compte des trames précédentes et constitue une faiblesse du G729.B. C'est pourquoi ils ont mis au point une amélioration du module de *Hangover*. Ce nouveau module contient encore quatre règles. La première renforce la détection des fins de bouffées de parole. Il s'agit de forcer la sortie du DAV à 1 pendant une période d'inertie de N_3 trames.

$$
\begin{aligned}
&Si\ decision1 = PAROLE \\
&Alors\ compteur\_inertie = 0
\end{aligned}
$$

$$
\begin{aligned}
&Si\ (decision1 = BRUIT \\
&\&\ E_f > coefficient \\
&\&\ compteur\_inertie\ < N_3) \\
&Alors\ vad\_flag = PAROLE \\
&\qquad compteur\_inertie++
\end{aligned}
\tag{3.17}
$$

Les trois autres règles sont en fait les trois premières règles du module *Hangover* initial. Ainsi les règles 2 et 3 du *Hangover* Alcatel sont données respectivement par (3.13) et (3.14). La dernière règle est, quant à elle, régie par (3.15) à laquelle il faut rajouter l'assignation du *compteur_inertie* à N_3 quand les conditions de la règle sont vérifiées. La règle 4 initiale n'est plus utilisée puisqu'elle n'est pas adéquate.

Mise à jour du bruit :

Comme il a été mentionné dans le chapitre 2 section 2.2.2, la mise à jour de l'estimé du bruit s'effectue à l'aide d'un système auto-regressif du premier ordre et ceci pour chaque caractéristique : $\overline{E_f}$, $\overline{E_l}$, $\overline{ZC}$, $\overline{\{LSF_i\}_{i=1}^{10}}$.

Cette mise à jour a lieu si et seulement si la différence d'énergie, le second coefficient de réflexion et la distorsion spectrale vérifient la condition suivante :

$$\Delta E_f > coefficient\ \ \&\ \ rc < coefficient\ \ \&\ \ DS < coefficient \tag{3.18}$$

Enfin, un mécanisme de remise à zéro a été intégré afin d'éviter le blocage de la sortie du DAV sur PAROLE lors d'une montée brusque du bruit. Il est régi par :

$$Si\ E_f < E_{\min}\ alors\ remise\ à\ zéro\ nécessaire\ :\ E_f = E_{\min} \tag{3.19}$$

3.2 Ajustements du G729.B pour des milieux industriels bruités

Dans le cadre de ce projet de recherche, nous avons mis en œuvre l'algorithme détaillé du G729.B selon les règles et procédures décrites dans la partie précédente et qui proviennent des recommandations de ITU-T [19] et [38]. Nous l'avons simulé à l'aide du logiciel MATLAB. Quelques tests ont été ensuite effectués sur des signaux de parole non bruitée, issus de la base de donnée DARPA TIMIT, afin de s'assurer de son bon fonctionnement. L'inspection visuelle des sorties du DAV a montré que la distinction entre les trames de parole et celles du bruit était très bonne. La figure 20 montre un exemple des résultats obtenus.

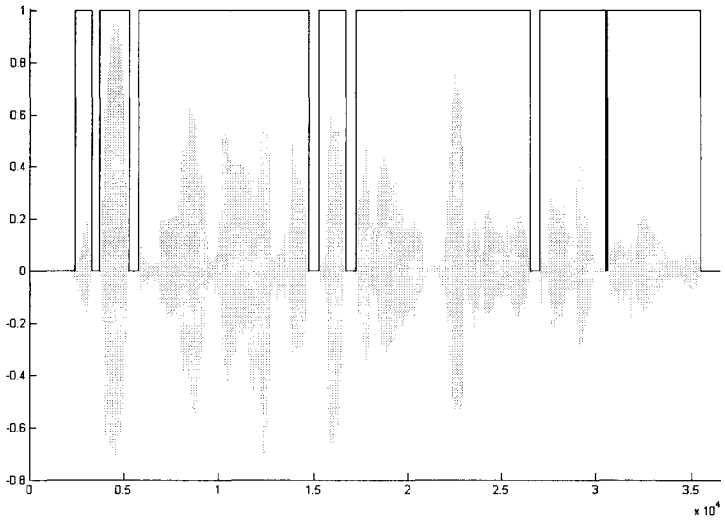


Figure 20 Sortie du DAV G729.B pour une phrase non bruitée

Suite à cela, nous avons effectué les mêmes tests mais en ajoutant cette fois-ci du bruit industriel aux signaux de parole. Le bruit utilisé est le « Factory Noise 1 » de la base de données NOISEX et son amplitude a été ajustée de manière à obtenir pour toutes les phrases testées un rapport signal à bruit de 10dB. L'inspection visuelle des sorties du DAV a permis de constater qu'il y avait beaucoup d'erreurs de classification, comme le montre la figure 21.

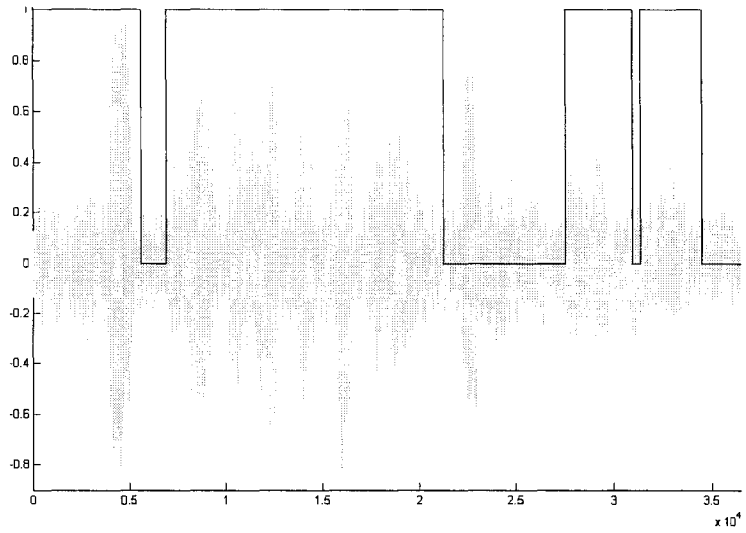


Figure 21 Sortie du DAV G729.B pour la même phrase qu'à la figure 20 mais bruitée (Factory Noise 1, RSB = 10dB)

3.2.1 Base d'expérimentation

Un ajustement de l'algorithme a donc été nécessaire. Afin de déterminer quelles parties modifier et comment les modifier, il a fallu mettre au point une base de données d'expérimentation.

Elle est composée de 32 phrases issues de la base de données en anglais DARPA TIMIT. Ces phrases sont toutes différentes et de longueur variable : 3 à 10s. De même, elles sont toutes prononcées par des locuteurs différents, originaires de trois régions des États-Unis : New York City, South Midland et Northern. Afin de respecter la parité, il y a autant de femmes que d'hommes. Bien qu'enregistrées avec une fréquence d'échantillonnage de 20kHz, ces phrases ont été traitées afin que celle-ci soit égale à 16kHz. Dans notre cas, la fréquence d'échantillonnage attendue est 8kHz. Il a donc été nécessaire de rééchantillonner ces données. Pour cela, un sous-échantillonnage par 2 a été effectué après avoir utilisé un filtre anti-repliement permettant de conserver la bonne qualité de la parole.

Pour faciliter la recherche, un seul bruit a été étudié. Il s'agit du « Factory Noise 1 » provenant de la base de données NOISEX. La fréquence d'échantillonnage utilisée par NOISEX est 19 980Hz. Une fois encore, il a fallu rééchantillonner. Un sur-échantillonnage par 2 a tout d'abord été appliqué. Le signal ainsi obtenu a ensuite été traité par un filtre permettant à la fois de ne garder qu'une image spectrale de ce signal et d'empêcher le repliement provoqué par l'étape suivante : le sous-échantillonnage par 5.

$$19980Hz \times \frac{2}{5} = 7992Hz \approx 8000Hz \tag{3.20}$$

On rappelle que le sur-échantillonnage provoque une répétition du spectre à tous les multiples de la fréquence d'échantillonnage. Un sous-échantillonnage engendre, lui, un repliement du spectre (Oppenheim et Schafer [7]).

Le but de ce projet de recherche est de mettre au point un DAV efficace avec des bruits industriels et pour des rapports signal à bruit compris entre 5dB et 15dB, voire plus faibles. Nous avons donc choisi d'ajuster le système en se basant sur les observations pour le milieu de la plage, c'est-à-dire 10dB. Ainsi, pour chaque phrase, nous avons déterminé le coefficient à appliquer à l'amplitude du bruit afin d'obtenir ce RSB et nous avons ensuite ajouté ce bruit à chacune des phrases. Finalement, un ensemble de 32 phrases bruitées avec un RSB de 10dB a été obtenu.

L'algorithme a ensuite été exécuté sur tout cet ensemble. Pour chaque trame, nous avons alors stocké les valeurs des différentes grandeurs utilisées dans les prises de décision. De plus, connaissant l'activité vocale des 32 phrases, nous avons séparé ces trames en deux ensembles : la parole bruitée d'un côté et le bruit seul de l'autre.

3.2.2 Ajustements de l'algorithme

La recherche sur cette base d'expérimentation a conduit à plusieurs modifications du système et finalement à une procédure d'ajustement. Elle peut être résumée comme suit :

Étape1 : Modification de la condition pour la mise à jour de l'estimé du bruit

Pour chaque trame de parole et de bruit, on connaît les valeurs de E_f, rc et DS, grandeurs utilisées par la condition. On rappelle que cette mise à jour devrait idéalement être effectuée sur toutes les trames de bruit et uniquement sur celles-ci. La condition de mise à jour doit donc être vérifiée par le plus de trames de bruit et le moins de trames de parole, le cas idéal ne pouvant être atteint. La recherche des coefficients à utiliser dans la condition est réalisée dans cette optique. Par essais et erreurs, on détermine les trois coefficients les plus adéquats, c'est-à-dire ceux pour lesquels la différence ci-dessous est maximale :

$$\left(\frac{nbre\ trames\ bruit\ validant\ condition}{nbre\ total\ trames\ de\ bruit} - \frac{nbre\ trames\ parole\ validant\ condition}{nbre\ total\ trames\ de\ parole} \right)$$

Étape2 : Stockage des données nécessaires pour la suite de l'ajustement

Selon les résultats de l'Étape 1, on modifie la condition de remise à jour. On doit ensuite reformer la base d'expérimentation car les valeurs de certaines grandeurs ont changé, les caractéristiques de l'estimé du bruit étant réajustées de manière plus adéquate. Pour les 32 premières trames de chaque phrase, on stocke uniquement la valeur de E_f. Pour toutes les autres, on stocke E_f, ΔE_f, ΔE_l, ΔZC et DS. Il est à noter que les trames sont encore séparées en deux classes : les trames de parole et les trames de bruit.

Étape3 : Modification de la règle de décision lors de la phase d'initialisation

La règle de décision utilisée lors de l'initialisation est :

$$\begin{aligned} &Si\ \ E_f < coefficient\ \ alors\ \ la\ trame\ est\ BRUIT \\ &Sinon \qquad\qquad\qquad\quad la\ trame\ est\ PAROLE \end{aligned} \tag{3.21}$$

Nous déterminons par dichotomie le coefficient permettant de reconnaître le plus de trames de parole tout en générant le moins d'erreur « bruit pris pour de la parole ». Le coefficient retenu est celui pour lequel la différence ci-dessous est maximale :

$$\left(\frac{nbre\ trames\ parole\ validant\ règle}{nbre\ total\ trames\ de\ parole} - \frac{nbre\ trames\ bruit\ validant\ règle}{nbre\ total\ trames\ de\ bruit} \right)$$

Étape4 : Modification des règles du module de décision

Lors de l'analyse du comportement du DAV avec un bruit industriel, nous avons constaté que sept des quinze règles du module de décision initial n'étaient jamais utilisées par les trames de notre base d'expérimentation. Aucune de ces trames ne possède en fait des caractéristiques vérifiant les conditions de ces règles. Ainsi, il est nécessaire de déterminer les règles les plus propices à notre situation ainsi que leurs coefficients. Pour cela, nous avons choisi la même approche que ITU-T [19], à savoir celle de la reconnaissance de formes dans un espace euclidien à 4 dimensions. Il s'agit donc de trouver par expérimentation les séparations des nuages de parole et de bruit. Les caractéristiques utilisées pour la prise de décision préliminaire sont : ΔE_l, ΔE_f, ΔZC, DS. Nous devons donc comparer ces caractéristiques deux à deux et trouver les droites procurant la meilleure séparation entre la parole et le bruit. La figure 22 montre les six plans à examiner.

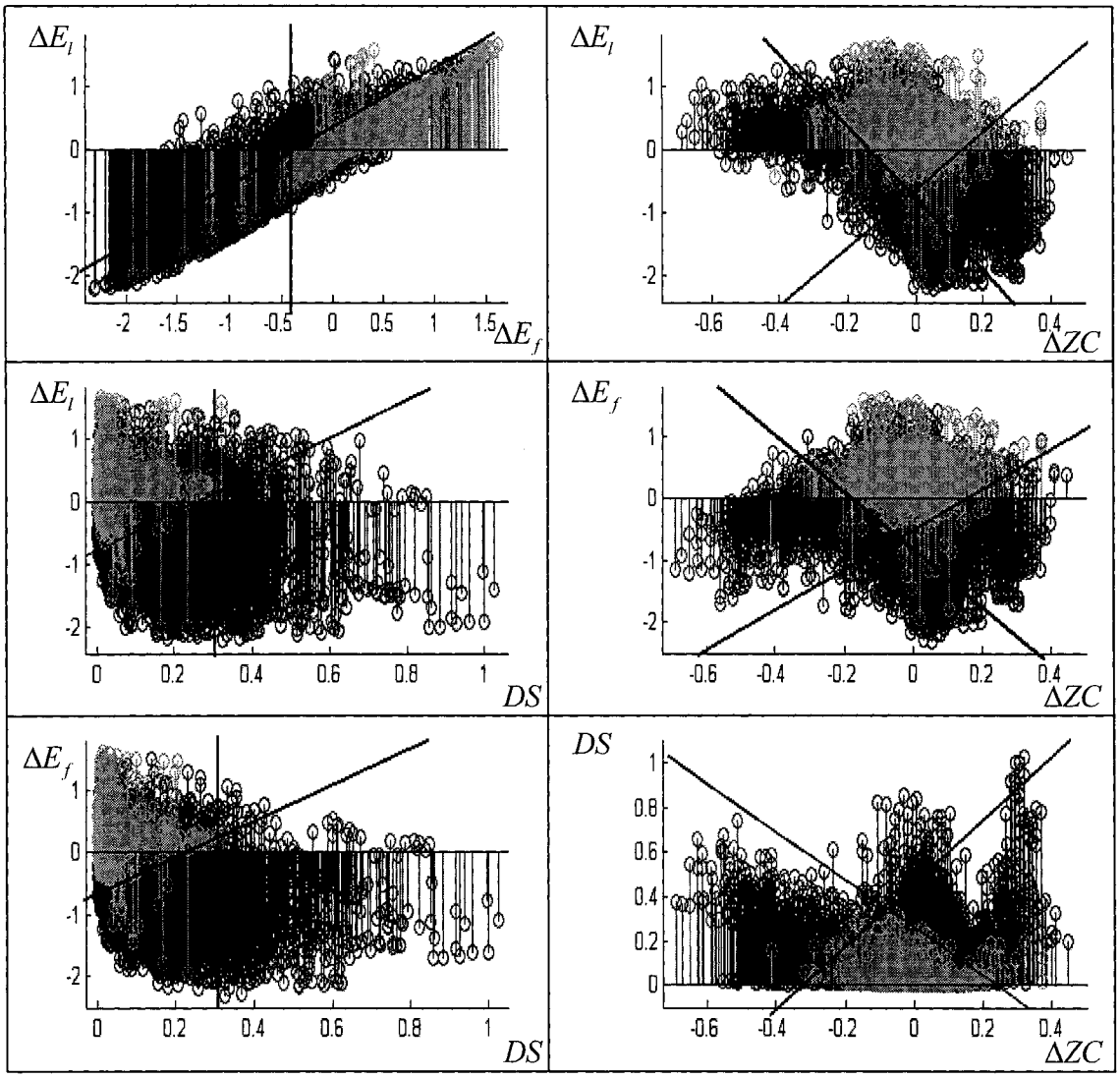


Figure 22 Les six plans à analyser pour obtenir les règles du module de décision

Sur la figure 22, les droites sont données à titre d'exemple et elles ont été dessinées arbitrairement. Elles permettent d'illustrer la notion de séparation entre les nuages de parole et de bruit. À chaque droite choisie correspond une règle. Par exemple, dans le plan $\left(\Delta E_f, \Delta E_l\right)$, en haut à gauche de la figure 22, la droite verticale montre une séparation possible entre les deux types de données. Lors de l'ajustement, si on

considère qu'elle constitue une séparation judicieuse entre les trames de parole et celles de bruit, on la choisit et la règle de décision associée à cette droite est alors :

$$Si\ \Delta E_f < -0.4\ alors\ decision1 = PAROLE \tag{3.22}$$

Finalement, pour chaque plan, il y a deux types de règles possibles. En plus de cela, pour chaque caractéristique, il y a une règle simple : la comparaison de sa valeur à un coefficient. Au total, il existe donc seize types de règles :

$$\begin{cases} \Delta E_l < a_1 \\ \Delta E_f < a_2 \\ \Delta ZC < a_3 \\ DS > a_4 \end{cases}$$

$$\begin{cases} \Delta E_l > a_5 \Delta E_f + b_5 \\ \Delta E_l < a_6 \Delta E_f + b_6 \end{cases}$$

$$\begin{cases} \Delta E_l > a_7 \Delta ZC + b_7 \\ \Delta E_l < a_8 \Delta ZC + b_8 \end{cases}$$

$$\begin{cases} \Delta E_l > a_9 DS + b_9 \\ \Delta E_l < a_{10} DS + b_{10} \end{cases}$$

$$\begin{cases} \Delta E_f > a_{11} \Delta ZC + b_{11} \\ \Delta E_f < a_{12} \Delta ZC + b_{12} \end{cases}$$

$$\begin{cases} \Delta E_f > a_{13} DS + b_{13} \\ \Delta E_f < a_{14} DS + b_{14} \end{cases}$$

$$\begin{cases} DS > a_{15} \Delta ZC + b_{15} \\ DS < a_{16} \Delta ZC + b_{16} \end{cases}$$

À l'aide de la base d'expérimentation, nous déterminons les règles intéressantes et leurs coefficients de la manière suivante :

a) Pour chaque type de règles, on fait varier les coefficients a_i et b_i entre deux bornes et avec un pas fixe. Les bornes et le pas sont déterminés par visualisation de chaque plan et de manière à étudier le maximum de droites afin de trouver par la suite celle qui sépare le mieux les trames de parole des trames de bruit. On obtient ainsi un grand nombre de règles possibles.

b) Pour chacune d'entre elles, on relève le pourcentage de trames de parole reconnues ainsi que le pourcentage de trames de bruit validées par la règle.

c) Le but du module de décision est de reconnaître le plus de parole tout en générant le moins d'erreur « bruit pris pour de la parole », la reconnaissance parfaite ne pouvant être obtenue. La règle choisie et ses coefficients sont donc ceux pour lesquels la différence ci-dessous est maximale :

$$\left(\frac{nbre\ trames\ parole\ validant\ règle}{nbre\ total\ trames\ de\ parole} - \frac{nbre\ trames\ bruit\ validant\ règle}{nbre\ total\ trames\ de\ bruit} \right)$$

d) On rappelle que les règles utilisées par le module de décision sont indépendantes les unes des autres. Il suffit qu'une règle soit validée pour que la décision préliminaire soit PAROLE. On enlève donc de la base d'expérimentation toutes les trames qui vérifient la règle sélectionnée.

Retour à a)

Condition d'arrêt :

On arrête l'algorithme lorsque aucune règle ne permet d'avoir un pourcentage de reconnaissance de parole supérieur à celui d'erreur « bruit pris pour de la parole », c'est-à-dire lorsqu'il n'est plus possible de reconnaître de parole sans engendrer beaucoup d'erreur. Toutes les règles intéressantes et leurs coefficients ont donc été trouvés.

Étape5 : Stockage des données nécessaires à l'ajustement du Hangover

Selon les résultats de l'Étape 4, on modifie le module de décision. On reforme ensuite la base d'expérimentation et on stocke pour chaque trame les valeurs des grandeurs qui vont être utilisées par le *Hangover* : E_f, ΔE_f, E_{fprec}, $decision1$, $dernier\_vad\_flag$ et $avant\_dernier\_vad\_flag$. Il est à noter que les trames sont encore séparées en deux classes : les trames de parole et les trames de bruit.

Étape6 : Modification des règles du Hangover

Le module de *Hangover* retenu est celui proposé par Alcatel [39], vu à la section 3.1. Les quatre règles sont donc connues et il s'agit ici de déterminer uniquement leurs coefficients. Pour chacune d'elles, on fait varier simultanément les paramètres utilisés et on relève les pourcentages de reconnaissance des trames de parole et de bruit.

Les trois premières règles permettent d'améliorer la reconnaissance de la parole. Les coefficients retenus sont donc ceux avec lesquels on reconnaît le plus de parole en engendrant le moins d'erreur « bruit pris pour de la parole », c'est-à-dire pour lesquels la différence ci-dessous est maximale :

$$\left(\frac{nbre\ trames\ parole\ validant\ règle}{nbre\ total\ trames\ de\ parole} - \frac{nbre\ trames\ bruit\ validant\ règle}{nbre\ total\ trames\ de\ bruit} \right)$$

La dernière règle, quant à elle, doit améliorer la reconnaissance des trames de bruit. Ses coefficients sont donc ceux avec lesquels on reconnaît le plus de bruit en engendrant le moins d'erreur « parole prise pour du bruit », c'est-à-dire pour lesquels la différence ci-dessous est maximale :

$$\left(\frac{nbre\ trames\ bruit\ validant\ règle}{nbre\ total\ trames\ de\ bruit} - \frac{nbre\ trames\ parole\ validant\ règle}{nbre\ total\ trames\ de\ parole} \right)$$

En plus de ces quatre règles, nous en avons mis au point une cinquième. Notre idée est qu'une bouffée de parole de durée égale à 10ms ne peut pas contenir une information

intéressante. De même, une pause d'une trame, soit 10ms, dans une bouffée de parole n'est pas réaliste. La cinquième règle du *Hangover* est donc :

$$\begin{cases} Si\ (vad\_flag = BRUIT \\ \&\ dernier\_vad\_flag = PAROLE \\ \&\ avant\_dernier\_vad\_flag = BRUIT) \\ Alors\ dernier\_vad\_flag = BRUIT \\ \\ Si\ (vad\_flag = PAROLE \\ \&\ dernier\_vad\_flag = BRUIT \\ \&\ avant\_dernier\_vad\_flag = PAROLE) \\ Alors\ dernier\_vad\_flag = PAROLE \end{cases} \tag{3.23}$$

Ainsi nous procédons à un lissage du type :

...1100000001000000111... → ...1100000000000000111...

...0011111110111111000... → ...0011111111111111000...

Cette règle personnelle est très intéressante car elle permet d'augmenter les pourcentages de reconnaissance de parole et de bruit. Par contre, comme elle modifie non pas la valeur de la sortie actuelle mais celle de la sortie d'avant, cette règle engendre un retard. Toutefois, étant de 10ms seulement, ce retard est tout à fait acceptable.

La procédure d'ajustement est maintenant terminée.

Le principe de l'algorithme reste donc identique à celui du détecteur d'activité vocale G729.B initial. Seules les parties suivantes ont été adaptées aux environnements industriels bruités :

- Les coefficients utilisés par la condition de la mise à jour de l'estimé du bruit.
- Le coefficient de la règle de décision, lors de l'initialisation.
- Les règles du module de décision.
- Les coefficients des règles du *Hangover*. Une règle a été ajoutée.

La sortie du DAV modifié pour la phrase vue à la figure 21 est maintenant celle présentée par la figure 23:

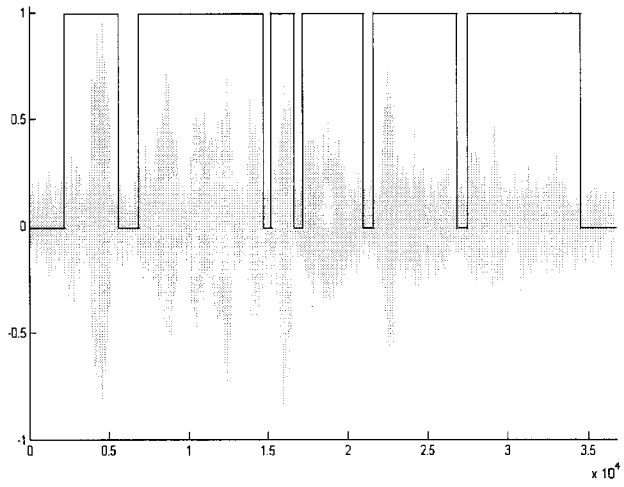


Figure 23 Sortie du DAV ajusté pour la même phrase bruitée qu'à la figure 21 (Factory Noise 1, RSB = 10dB)

3.3 Résultats pratiques obtenus avec le G729.B ajusté

La figure 23 n'est qu'un exemple de ce que nous avons obtenu sur la base d'expérimentation. Nous allons maintenant présenter les résultats pratiques obtenus avec le G729.B ajusté selon la méthode décrite dans la partie précédente. Ceci va nous permettre de vérifier son fonctionnement et de caractériser son comportement.

3.3.1 Base de validation

Afin de déterminer les performances de ce DAV, nous avons mis au point une base de validation.

Elle comporte 64 phrases issues de la base de données en anglais DARPA TIMIT. Comme pour la base d'expérimentation, elles sont toutes différentes et de longueur variable : 3 à 10s. Elles sont aussi toutes prononcées par des locuteurs différents originaires de trois régions des États-Unis (New York City, South Midland et Northern) avec autant de femmes que d'hommes. Enfin, la fréquence d'échantillonnage étant toujours de 16kHz, les phrases ont été traitées par un filtre anti-repliement puis sous-échantillonnées par 2 afin d'obtenir la fréquence d'échantillonnage attendue par le DAV, c'est-à-dire 8kHz. Il est à noter que ces 64 phrases sont différentes des 32 phrases utilisées dans la base d'expérimentation.

Pour tester le détecteur d'activité vocale, 11 bruits industriels ont été utilisés. Trois d'entre eux proviennent de la base de données NOISEX. Les huit autres nous ont été fournis gracieusement par la compagnie SONOMAX et ont été récoltés dans une affinerie de cuivre de NORANDA. Dépendant des conditions d'enregistrement, les signaux de bruit ont été rééchantillonnés, afin d'obtenir la fréquence d'échantillonnage de 8kHz, et ceci en prenant soin d'utiliser un filtre adéquat pour conserver la bonne qualité de l'information. Les caractéristiques de ces 11 bruits sont répertoriées dans le tableau I.

Tableau I

Caractéristiques des bruits industriels utilisés

Nom	Description	Niveau global à +/- 3dB	Fe
Noisex 1 (Nx1)	*Factory Noise 1* de NOISEX, salle de découpage et de soudure du métal pour des équipements automobiles	83dBA	19980Hz
Noisex 2 (Nx2)	*Factory Noise 2* de NOISEX, salle de production automobile	74dBA	19980Hz
Noisex 3 (Nx3)	*Operation Room* de NOISEX, salle d'opérations d'un *destroyer*	70dBA	19980Hz
Noranda 1 (Nor1)	Salle d'empaquetage sélénium, *Baghouse 1*	-	22050Hz
Noranda 2 (Nor2)	Salle d'empaquetage sélénium, *Baghouse 1, 2, 3* et *Vaccum Cleaner*	82dBA	22050Hz
Noranda 3 (Nor3)	Salle d'empaquetage sélénium, Micropulvérisateur de sélénium et ventilateur	79dBA	44100Hz
Noranda 4 (Nor4)	Salle d'empaquetage sélénium, Soupape de sécurité	71dBA	44100Hz
Noranda 5 (Nor5)	Salle d'empaquetage sélénium, Surpresseur à pistons rotatifs	-	44100Hz
Noranda 6 (Nor6)	Salle des transformateurs	100dBA	44100Hz
Noranda 7 (Nor7)	Brûleur propane et fournaise	102dBA	44100Hz
Noranda 8 (Nor8)	Salle hydraulique	108dBA	44100Hz

Pour tester le comportement du DAV dans chaque environnement, des parties de chaque bruit ont été choisies aléatoirement puis additionnées aux phrases. Ainsi pour chaque

milieu, nous avons 64 signaux de parole bruitée. Les niveaux sonores des phrases et des bruits sont tous différents les uns les autres, il a donc fallu déterminer dans chaque cas le coefficient à appliquer à l'amplitude du bruit afin d'obtenir les RSB voulus.

Les caractéristiques de la base de validation pour chaque bruit peuvent être résumées comme suit :

Parole :
- 64 phrases (DARPA TIMIT)
- américain
- toutes différentes
- toutes prononcées par des locuteurs différents
- 32 femmes, 32 hommes

Bruit :
- Factory Noise 1 (NOISEX)
- Factory Noise 2 (NOISEX)
- Operation Room (NOISEX)
- Noranda 1 à 8 (fournis par le compagnie SONOMAX)

Connaissant l'activité vocale de chaque phrase, il nous a été possible d'effectuer les tests pour chaque environnement. La partie suivante présente les résultats pratiques obtenus.

3.3.2 Résultats pratiques

Pour évaluer le comportement de ce DAV, nous avons relevé deux performances : le pourcentage de reconnaissance des trames de parole et le pourcentage de reconnaissance des trames de bruit. L'erreur totale, c'est-à-dire la mauvaise classification, peut être obtenue en additionnant les pourcentages des trames de parole et de bruit non reconnues.

Pour chaque bruit, nous avons testé le DAV sur la base de validation et ceci pour différents rapports signal à bruit. Les résultats sont exposés dans le tableau II :

Tableau II

Pourcentages de reconnaissance de parole et de bruit obtenus sur la base de validation avec le DAV ajusté pour « Factory Noise 1 », RSB = 10dB

RSB / Bruit	0dB		5dB		10dB		15dB		20dB	
	Rec. Parole (%)	Rec. Bruit (%)	Rec. Parole (%)	Rec. Bruit (%)	Rec. Parole (%)	Rec. Bruit (%)	Rec. Parole (%)	Rec. Bruit (%)	Rec. Parole (%)	Rec. Bruit (%)
Nx1	91,8	36,3	90,7	57,1	86,5	88,8	91,9	91,3	95,9	87,3
Nx2	95,3	44,5	89,4	75,3	87,9	90,7	93,7	89,2	96,8	86,7
Nx3	93,4	40,7	87,5	52,3	85,4	91	90	88,9	96,1	89,7
Nor1	94,3	53,5	88,8	92,4	91,5	95,5	96,3	89,4	98	83,2
Nor2	94,7	47,7	87,1	84,8	88	97,5	94,4	94	97,4	89
Nor3	98	32,6	95,3	43,8	91,9	87,3	94,7	87,8	97,3	82,4
Nor4	94,1	45,2	89,8	83,7	92	87,8	96,4	87,1	97,9	84
Nor5	84,5	58,2	83,8	63,7	82,1	96,3	87,3	96,4	94,4	95,8
Nor6	91,7	58,1	89,9	60	83,1	94,1	87,3	93,9	94,8	93,8
Nor7	99	5,8	94,6	31,4	92,2	50,7	94,8	55,4	97,4	55,7
Nor8	90,1	51,4	89,8	59,8	83,5	94,4	87,6	93,5	94,9	94,5

La première chose que nous pouvons observer dans le tableau II est que, quelque soit le bruit testé, les performances évoluent de manière similaire lors que le bruit augmente. Nous pouvons noter aussi que l'environnement ayant pour bruit Noranda 7 présente de mauvaises performances. L'ajustement déterminé précédemment n'est pas adéquat pour ce milieu. Les observations et les remarques qui suivent ne concernent donc pas ce cas particulier.

Les performances obtenues pour des RSB supérieurs ou égaux à 10dB sont satisfaisantes, voire excellentes dans certaines situations. Lors d'un RSB de 5dB, le DAV fonctionne correctement dans moins de la moitié des environnements. En dessous de 5dB, il n'est plus capable de reconnaître convenablement les trames de bruit seul. Le DAV aura tendance à laisser passer de plus en plus de signal, plus les rapports signal à bruit seront négatifs. Finalement, en dessous d'un certain RSB, sa sortie sera bloquée à l'état haut. En fait, d'après Alcatel [39], le G729.B a justement été mis au point de manière à ce que lorsqu'il est utilisé en dehors de sa plage de fonctionnement, il ne détériore pas l'information importante, c'est-à-dire la parole. Ainsi, quand le RSB est trop faible, le DAV ne sert plus à rien mais il laisse le signal utile intact. On rappelle qu'en cas de doute la sortie d'un détecteur d'activité vocale devrait toujours indiquer la présence de parole afin d'éviter des pertes lourdes et irrémédiables.

Le DAV que nous avons ajusté précédemment avec le bruit Noisex 1 fonctionne donc correctement dans dix environnements industriels sur onze, lorsque le rapport signal sur bruit est supérieur ou égal à 10dB. Si le RSB est inférieur à cette valeur, son comportement se dégrade rapidement. Ceci peut être expliqué par le fait que l'ajustement a été effectué en fonction des données obtenues pour un RSB de 10dB. Ainsi, plus on s'éloigne de ce point, moins les modifications apportées sont adéquates. En effet, le bruit devient alors plus présent et recouvre encore davantage la voix. Ceci engendre une augmentation des caractéristiques de toutes les trames et les caractéristiques des trames de parole se rapprochent de celles du bruit. Les coefficients utilisés dans la condition de mise à jour de l'estimé du bruit, dans le module du *Hangover* et dans la règle de décision de l'initialisation sont fixes et ils ne sont donc plus adaptés à la situation. La mauvaise mise à jour des paramètres du bruit provoque alors le déplacement des nuages de parole et de bruit dans l'espace euclidien à 4 dimensions alors que les droites, dont sont issues les règles pour la prise de décision préliminaire, restent, elles, statiques. Elles ne séparent donc plus correctement les deux nuages.

Afin d'augmenter la plage de RSB sur laquelle le DAV se comporte convenablement, nous avons réajusté l'algorithme en fonction des données obtenues pour 5dB. Nous l'avons aussi réajusté pour 15dB afin de voir si les performances pour les RSB élevés pouvaient être améliorées. Les procédures d'ajustement utilisées sont identiques en tout point à celle effectuée pour 10dB, voir section 3.2.2.

Nous avons ensuite testé ces deux ajustements sur la base de validation. Les tableaux III et IV exposent les résultats obtenus pour les ajustements à 5dB et 15dB, respectivement. Seuls trois bruits, à savoir Noisex 1, Noisex 2 et Noranda 1, sont présentés ici car cela est suffisant pour tirer des conclusions.

Tableau III

Pourcentages de reconnaissance de parole et de bruit obtenus sur la base de validation avec le DAV ajusté pour « Factory Noise 1 », RSB = 5dB

RSB \ Bruit	0dB		5dB		10dB		15dB		20dB	
	Rec. Parole (%)	Rec. Bruit (%)	Rec. Parole (%)	Rec. Bruit (%)	Rec. Parole (%)	Rec. Bruit (%)	Rec. Parole (%)	Rec. Bruit (%)	Rec. Parole (%)	Rec. Bruit (%)
Nx1	86,2	42,8	79,6	74,9	87,6	78,7	93,6	80,3	96,1	79,5
Nx2	81,3	61,2	82	86,1	89,3	86,4	94,4	84,8	96,6	80,5
Nor1	77,4	87,7	85,4	92,4	92,1	92	95,8	88,1	97,1	81,2

Le tableau III montre que l'ajustement à 5dB est propice à la discrimination parole/bruit dans des milieux à RSB égal à 5dB. Certes, le pourcentage de reconnaissance de parole est moins élevé qu'avec l'ajustement à 10dB mais celui de bruit est meilleur. On peut donc considérer que le DAV fonctionne maintenant quand le RSB est de 5dB. Les performances pour un RSB de 0dB sont aussi plus intéressantes que précédemment. Par contre pour des RSB supérieurs ou égaux à 10dB, l'ajustement à 10dB offrait une

meilleure identification des trames de parole et de bruit. Il est à noter que nous avons encore une dégradation de la reconnaissance du bruit, plus on s'éloigne du RSB utilisé pour l'ajustement.

Tableau IV

Pourcentages de reconnaissance de parole et de bruit obtenus sur la base de validation avec le DAV ajusté pour « Factory Noise 1 », RSB = 15dB

RSB / Bruit	0dB		5dB		10dB		15dB		20dB	
	Rec. Parole (%)	Rec. Bruit (%)	Rec. Parole (%)	Rec. Bruit (%)	Rec. Parole (%)	Rec. Bruit (%)	Rec. Parole (%)	Rec. Bruit (%)	Rec. Parole (%)	Rec. Bruit (%)
Nx1	99,7	0,8	99,7	7,2	93,9	61,5	91,6	91,8	96	92
Nx2	99,9	1,8	98,8	21,4	90,3	81,6	92	97,8	96,6	94,9
Nor1	100	3	96	44,1	90	97,6	94,9	97,5	98	91,6

Le tableau IV montre que l'ajustement à 15dB permet d'obtenir des performances supérieures ou égales à celles obtenues avec l'ajustement à 10dB, et ceci pour les RSB de 15dB et 20dB. Par contre, pour les faibles RSB, le comportement du DAV se dégrade plus rapidement.

En conclusion, plus le RSB utilisé pour l'ajustement est petit, plus la plage de RSB sur laquelle le DAV fonctionne correctement est grande mais plus les performances sont moyennes. Le tableau V illustre ceci.

Tableau V

Influence du RSB d'ajustement sur les performances du DAV

RSB utilisé pour l'ajustement	**Rapports signal à bruit pour lesquels le DAV fonctionne bien**	**Performances**
5dB	Environ 5dB (selon les environnements) et plus	Moyennes
10dB	Environ 10dB (selon les environnements) et plus	Bonnes
15dB	Environ 15dB (selon les environnements) et plus	Excellentes

3.4 Conclusion et améliorations possibles

Grâce à nos recherches, nous avons mis au point une procédure d'ajustement du G729.B pour mieux l'adapter aux milieux industriels bruités. Pour un nombre restreint de phrases, nous avons comparé les résultats obtenus avec le G729.B initialement mis en œuvre et avec celui ajusté. Nous avons constaté une amélioration de la détection d'activité vocale, la figure 23 comparée à la figure 21 en est un exemple. Bien que nous n'ayons pas effectué de tests exhaustifs, il y a tout lieu de croire que notre procédure d'ajustement du G729.B permette d'augmenter les performances du DAV lors d'une utilisation dans les milieux industriels bruités.

La procédure d'ajustement nous a permis d'ajuster le DAV pour les RSB 5dB, 10dB et 15dB. Nous avons donc un algorithme de détection d'activité vocale et trois ensembles de règles et de coefficients. Dépendant de l'application, il est possible de choisir l'ensemble le plus intéressant. Toutefois, le tableau V montre que le meilleur compromis est offert par l'ajustement à 10dB car il permet d'avoir de bonnes performances sur une plage de RSB raisonnable, et ceci pour la quasi totalité des environnements testés.

Pour augmenter la plage de RSB de bon fonctionnement, une possibilité serait d'ajuster une seule fois l'algorithme avec différents rapports signal à bruit. Notre objectif étant la plage [5dB – 15dB], l'idée serait de créer trois bases d'expérimentation : une pour 5dB, une pour 10dB et une pour 15dB puis de les regrouper en une seule et même base sur laquelle on effectuerait ensuite la procédure d'ajustement décrite à la section 3.2.2.

Une autre possibilité pour améliorer ce DAV serait d'ajuster l'algorithme plusieurs fois. On obtiendrait donc plusieurs ensembles de règles et de coefficients, un par RSB. En rajoutant un estimateur du rapport signal à bruit dans l'étape de mise à jour de l'estimé du bruit, on pourrait alors déterminer l'ensemble le plus propice à la situation et ainsi utiliser systématiquement l'ensemble le plus intéressant. Par exemple, nous avons déterminé précédemment trois ajustements : 5dB, 10dB et 15dB. En effectuant quelques tests, nous avons pu établir que :

- l'ajustement à 5dB permet d'obtenir les meilleures performances pour [2dB – 8dB[
- l'ajustement à 10dB permet d'obtenir les meilleures performances pour [8dB – 15dB[
- l'ajustement à 15dB permet d'obtenir les meilleures performances pour des RSB supérieurs ou égaux à 15dB

Ainsi, en calculant l'estimé du RSB, on pourrait passer d'un ensemble à l'autre et ainsi obtenir les performances les plus élevées en tout temps. Plus le nombre d'ajustements est grand, meilleur devrait être le comportement du DAV.

La dernière amélioration possible serait de rendre adaptatifs les coefficients qui ont été ajustés précédemment. De la même manière que l'on remet à jour les caractéristiques du bruit, on pourrait modifier les coefficients mis en jeu dans l'algorithme. Cette solution est bien plus complexe que les deux autres mais elle pourrait fournir les meilleurs résultats.

Le détecteur d'activité vocale présenté dans ce chapitre fonctionne donc dans des environnements industriels bruités. Ses performances sont intéressantes mais elles se dégradent lors de faibles rapports signal à bruit. La transformée en ondelettes a montré sa grande efficacité dans plusieurs domaines, notamment dans le débruitage de la parole. La deuxième partie de ce projet de recherche consiste à mettre au point un DAV basé sur la théorie des ondelettes. Les chapitres suivants sont consacrés à ce sujet.

CHAPITRE 4

LES ONDELETTES

L'origine des ondelettes remonte au début du XX^{ème} siècle avec les travaux de Haar [40], mais ce n'est qu'en 1984 que ce concept a été formellement introduit par Grossman et Morlet [41]. Au cours des années suivantes, plusieurs chercheurs ont étoffé la théorie : notion de base orthogonale : 1985 par Meyer, analyse multirésolution : 1989 par Mallat, ondelettes à support compact : 1988 par Daubechies... Aujourd'hui cet outil mathématique est utilisé dans de nombreux domaines, très différents les uns des autres : détection et prédiction des crises épileptiques, rehaussement de la parole, compression d'images...

La transformée en ondelettes consiste à décomposer un signal en composantes de différentes résolutions. Elle agit comme un microscope à différents agrandissements. Elle est donc très puissante pour l'analyse des signaux et l'extraction de caractéristiques. Le principe de la théorie des ondelettes repose sur l'expression du signal d'origine à l'aide d'un ensemble de fonctions appelées fonctions d'ondelettes. Chacune de ces fonctions est en fait le résultat d'une dilatation et d'une translation d'une seule et même fonction : l'ondelette mère. Le choix de cette dernière est donc crucial et dépend de ses propriétés (type de support, régularité, nombre de moments nuls, orthogonalité...) ainsi que de l'application. La transformée en ondelettes continue permet d'évaluer la corrélation entre le signal à analyser et les fonctions d'ondelettes, c'est-à-dire le degré de similitude entre ces deux signaux.

Ce chapitre est composé de deux grandes parties : la présentation de la théorie des ondelettes et la revue de littérature concernant son usage dans le domaine de la détection de l'activité vocale.

4.1 La théorie des ondelettes

Dans cette partie, nous présenterons brièvement les avantages des ondelettes par rapport aux outils traditionnels d'analyse des signaux. Nous définirons ensuite les différents types d'analyse par ondelettes : la transformée en ondelettes continue, la transformée en ondelettes discrète, l'analyse multirésolution et enfin les paquets d'ondelettes.

4.1.1 Avantages des ondelettes

Lorsqu'un signal est analysé dans le domaine des fréquences, il est nécessaire d'utiliser un outil mathématique pour effectuer cette opération. L'un des plus connus est la transformée de Fourier, définie pour un signal $f(t)$ par :

$$F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-j\omega t} dt \qquad (4.1)$$

Le problème de cette transformée est que le spectre obtenu n'est pas localisé temporellement. Cette technique ne s'applique donc qu'aux signaux stationnaires car ils évoluent peu au cours du temps. Malheureusement, la plupart des signaux réels sont non-stationnaires. La perte de l'information temporelle est alors dramatique car il n'est plus possible d'analyser les variations temporelles. La transformée de Fourier (TF) n'est donc pas adéquate pour l'analyse des signaux non-stationnaires, tels que la parole.

L'utilisation de la transformée de Fourier à fenêtre glissante, ou *Short-Time Fourier Transform* (STFT) est une solution possible à ce problème (Oppenheim et Schafer [7]). Cette transformée consiste à découper le signal en sections à l'aide d'une fenêtre temporelle, puis à appliquer la TF à chacune d'entre elles (Vetterli et Kovacevic [42]) :

$$F(\omega, \tau) = \int_{-\infty}^{+\infty} f(t) \psi^*(t-\tau) e^{-j\omega t} dt \qquad (4.2)$$

Avec : $\psi(t-\tau)$: la fenêtre appliquée à l'instant τ.

Si $\psi(t)$ est une gaussienne, on parle alors de transformée de Gabor [43].

Il est à noter qu'on peut appliquer la TF à chacun des segments car la fenêtre de découpage étant en général relativement étroite, le signal sur cet intervalle temporel est considéré quasi-stationnaire.

La STFT procure donc une représentation temps-fréquence du signal à analyser, illustrée par la figure 24 :

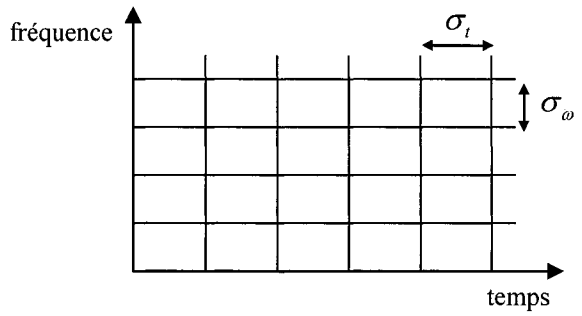


Figure 24 Pavage temps-fréquence pour la STFT

Les résolutions fréquentielle σ_ω et temporelle σ_t sont fixes. Ceci est dû à l'usage d'une seule et même fenêtre temporelle pour le découpage de tout le signal. La STFT ne permet donc pas d'étudier tous les phénomènes avec une précision adéquate. En effet, l'analyse des basses fréquences requiert des fenêtres de longueurs temporelles importantes alors que celle des hautes fréquences nécessite des fenêtres plus étroites.

La transformée en ondelettes, que nous définirons par la suite, résout ce problème. Elle fournit une représentation temps-fréquence à « résolution variable », c'est-à-dire que les fenêtres temporelles utilisées sont de largeur différentes : courtes pour mettre en évidence les hautes fréquences, longues pour révéler les basses fréquences, ce qui engendre des résolutions fréquentielles et temporelles variables.

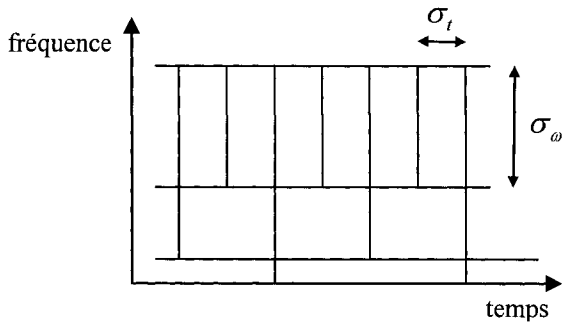


Figure 25 Pavage temps-fréquence pour la transformée en ondelettes (adaptée de « Ondelettes pour le signal numérique », Truchetet, [44])

À chaque échelle correspond une fenêtre particulière et par conséquent une résolution différente. Elle agit donc comme un microscope mathématique et permet d'identifier des phénomènes non repérés par les autres outils d'analyse habituels. La figure 25 représente un pavage typique du plan temps-fréquence obtenu par une analyse en ondelettes.

4.1.2 La transformée en ondelettes continue

Cette partie suit la présentation de la transformée en ondelettes continue (TOC) proposée par Akansu et Haddad dans leur livre « Multiresolution signal decomposition » [45].
On note $L^2(\Re)$ l'espace des fonctions réelles continues et de carré intégrable. La TOC d'une fonction $f(t) \in L^2(\Re)$ est définie par :

$$W_f(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(t)\psi\left(\frac{t-b}{a}\right)dt = \int_{-\infty}^{+\infty} f(t)\psi_{a,b}(t)dt \qquad (4.3)$$

Avec :

- $W_f(a,b)$: représentant les coefficients d'ondelettes. Plus ils sont élevés, plus la partie du signal étudiée et la fonction d'ondelette sont corrélées.
- $a \in \Re^+$: le facteur d'échelle, ou dilatation. Plus a est grand, plus l'ondelette est étirée. On a donc une vue globale lorsque ce facteur d'échelle est grand et une vue en détails lorsqu'il est petit.

- $b \in \Re$: le facteur de translation.
- $\psi(t)$: la fonction d'ondelette mère.
- $\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right)$: la fonction d'ondelette au temps b et à l'échelle a.

  Le terme $\frac{1}{\sqrt{a}}$ est utilisé pour normaliser l'énergie.

Il est important de rappeler qu'une fonction ψ ne peut être une ondelette que si elle vérifie les conditions d'admissibilité suivantes :

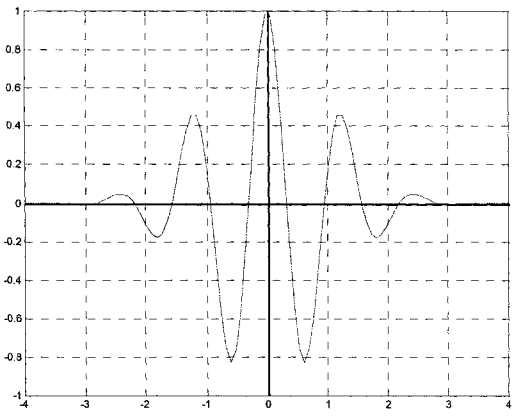
- $\psi \in L^2(\Re)$
- ψ est de moyenne nulle
- son spectre est fini

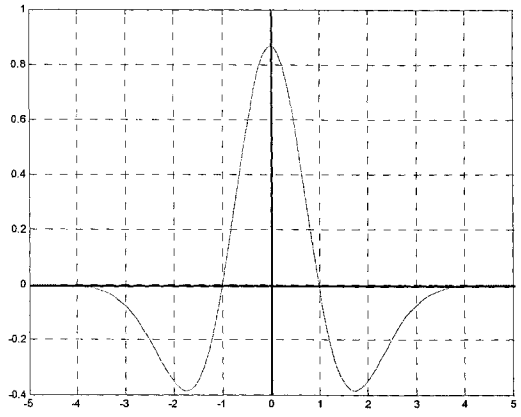
Ce qui se traduit par : $$\int_{-\infty}^{+\infty} \frac{|\psi(\omega)|^2}{|\omega|} d\omega = C_\psi < \infty \qquad (4.4)$$

Avec : $\psi(\omega)$ la transformée de Fourier de $\psi(t)$.

Il est à noter que ψ peut être vue comme la réponse impulsionnelle d'un filtre passe-bande.

Ces conditions, bien que contraignantes, sont vérifiées par plusieurs fonctions. C'est pourquoi il existe de nombreuses fonctions d'ondelettes. La figure 26 en présente deux exemples.





$$\psi_{morlet}(t) = e^{j\omega_0 t} e^{-\frac{t^2}{2}}$$

$$\psi_{chapeau\_mexicain}(t) = \frac{2}{\sqrt{3}} \pi^{-\frac{1}{4}} (1-t^2) e^{-\frac{t^2}{2}}$$

Figure 26 Exemples d'ondelettes: Morlet (à gauche), Chapeau mexicain (à droite)

Le facteur de dilatation *a* permet d'adapter la largeur de l'ondelette en fonction du type d'information à analyser. Si *a* est petit, l'ondelette est comprimée, ce qui est propice à l'étude des hautes fréquences. A contrario, si *a* est grand, l'ondelette est étirée, ce qui met en valeur les basses fréquences.

La TOC inverse existe si et seulement si $C_\psi \neq 0$ et est donnée par l'expression :

$$f(t) = \frac{1}{C_\psi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} W_f(a,b) \psi_{a,b}(t) \frac{dadb}{a^2} \qquad (4.5)$$

Avec : $W_f(a,b)$: les coefficients d'ondelettes.

4.1.3 La transformée en ondelettes discrète

Le terme « continu » dans l'expression transformée en ondelettes continue signifie que le signal est analysé pour toutes les échelles *a* et pour toutes les translations *b*, c'est-à-dire pour un nombre infini de points. Ceci n'est pas réaliste d'un point de vue computationnel et présente une grande redondance. Pour palier cette faiblesse, on

discrétise les paramètres a et b. La transformée ainsi obtenue s'appelle la transformée en ondelettes discrète (TOD) et est définie par (Mallat [46]) :

$$a = a_0^m \tag{4. 6}$$

$$b = nb_0 a = nb_0 a_0^m \tag{4. 7}$$

$$\psi_{m,n}(t) = a_0^{-\frac{m}{2}} \psi(a_0^{-m} t - nb_0) \tag{4. 8}$$

$$d_{m,n} = W_f(m,n) = a_0^{-\frac{m}{2}} \int_{-\infty}^{+\infty} f(t)\psi(a_0^{-m} t - nb_0)dt \tag{4. 9}$$

En général, on discrétise a et b de manière à obtenir une grille dyadique telle que représentée par la figure 27 :

$$a_0 = 2 \quad b_0 = 1 \tag{4. 10}$$

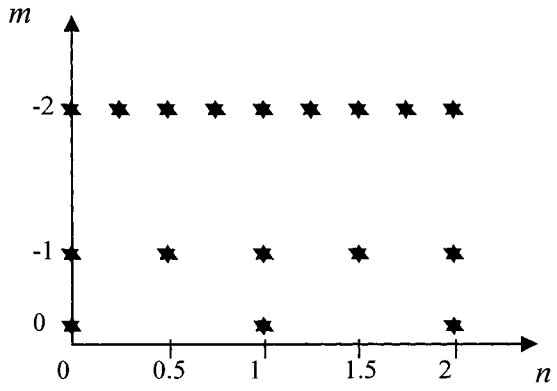


Figure 27 Grille dyadique pour la TOD
(adaptée de « Multiresolution Signal Decomposition », Akansu et Haddad [45])

On a alors : $$d_{m,n} = 2^{-\frac{m}{2}} \int_{-\infty}^{+\infty} f(t)\psi(2^{-m} t - n)dt \tag{4. 11}$$

L'inverse de la TOD est obtenue par (Akansu et Haddad [45]):

$$f(t) = \sum_m \sum_n d_{m,n} \psi_{m,n}(t) \tag{4. 12}$$

4.1.4 L'analyse multirésolution

L'analyse multirésolution permet de décomposer un signal en un ensemble d'approximation et de détails. Nous allons tout d'abord exposer les fondements mathématiques de ce concept. Nous verrons ensuite comment l'algorithme récursif de Mallat permet de procéder à une telle analyse. Nous aborderons enfin sa mise en pratique.

4.1.4.1 Les bases théoriques de l'analyse multirésolution

Partie Approximation :

On note $L^2(\Re)$ l'espace des fonctions réelles continues et de carré intégrable. L'analyse multirésolution d'une fonction f appartenant à $L^2\{\Re\}$ consiste à projeter cette fonction à différents niveaux de résolution afin de pouvoir l'examiner (Truchetet [44]):

$$A_j[f] \in V_j \tag{4. 13}$$

V_j correspond au sous-espace de L^2 associé à la résolution j et $A_j[.]$ représente la projection. C'est un opérateur linéaire.

Le passage d'un sous-espace à l'autre engendre un changement de résolution, et donc d'échelle.

Cette analyse multirésolution n'est possible que si l'ensemble des sous-espaces V_j de L^2 vérifie toutes les conditons suivantes (Mallat [47]) :

- Les sous-espaces sont inclus les uns dans les autres:

$$\ldots \subset V_{j+1} \subset V_j \subset V_{j-1} \subset \ldots \tag{4. 14}$$

- L'ensemble est complet:

$$\overline{\bigcup_{j\in Z} V_j} = L^2(\Re) \tag{4. 15}$$

$$\bigcap_{j\in Z} V_j = \{0\} \tag{4. 16}$$

- Il possède la propriété d'échelle:

$$f(x)\in V_j \Leftrightarrow f(2x)\in V_{j-1} \Leftrightarrow f\left(\frac{x}{2}\right)\in V_{j+1} \quad \forall j \in Z \tag{4. 17}$$

- Il est invariant par translation:

$$f(x)\in V_0 \Leftrightarrow f(x-k)\in V_0 \quad \forall k \in Z \tag{4. 18}$$

- Il existe une base orthonormale sur V_0 :

$$\{\varphi(x-n)/n\in Z\} \tag{4. 19}$$

Dans ces conditions, il existe une fonction permettant d'avoir une base orthonormée sur chaque sous-espace V_j. Il s'agit de la fonction d'échelle φ. Elle appartient à $L^2\{\Re\}$. Les fonctions de base sont obtenues par dilatation et translation de φ (Daubechies [48]):

$$\varphi_{j,n}(x) = 2^{-\frac{j}{2}}\varphi\left(2^{-j}x-n\right) \tag{4. 20}$$

Avec: $j \in Z$: le facteur de dilatation et $n \in Z$: celui de translation.

Cette base est orthonormée si elle vérifie (Mallat [47]) :

$$<\varphi_{j,n},\varphi_{j,k}> = \delta(n-k) \quad \forall j,n,k \in Z \tag{4. 21}$$

Puisqu'il existe une base orthonormée pour tout sous-espace V_j, la projection de f sur ces espaces est donnée par (Chui et Al. [49]) :

$$A_j[f] = \sum_n a_n^j \varphi_{j,n} \tag{4. 22}$$

Avec :

$$a_n^j = <f,\varphi_{j,n}> \tag{4. 23}$$

On parle alors d'approximation à la résolution j de la fonction f. a_n^j sont les coefficients associés à cette approximation.

Partie Détails :

Pour chaque sous-espace V_j, on peut définir son complément orthogonal: W_j dans V_{j-1}, tel que:

$$V_{j-1} = V_j \oplus W_j \tag{4. 24}$$

$$L^2(\Re) = \ldots \oplus W_{j-1} \oplus W_j \oplus W_{j+1} \oplus \ldots \tag{4. 25}$$

Les sous-espaces W_j sont tous orthogonaux entre eux et leur ensemble possède les propriétés d'échelle et d'invariance par translation (Akansu et Haddad [45]).

Dans ces conditions, il existe une fonction permettant d'avoir une base orthonormée pour chaque sous-ensemble W_j. Il s'agit de la fonction d'ondelette ψ. Elle appartient à $L^2\{\Re\}$. Les fonctions de base sont obtenues par dilatation et translation de ψ (Daubechies [48]) :

$$\psi_{j,n}(x) = 2^{\frac{-j}{2}} \psi\left(2^{-j}x - n\right) \tag{4. 26}$$

Avec: $j \in Z$: le facteur de dilatation et $n \in Z$: celui de translation.

Cette base est orthonormée si elle vérifie (Mallat [47]) :

$$< \psi_{j,n}, \psi_{i,k} >= \delta(j-i)\delta(n-k) \quad \forall j,n,i,k \in Z \tag{4. 27}$$

Puisqu'il existe une base orthonormée pour tout sous-espace W_j, la projection de f sur ces espaces est donnée par (Chui et Al. [49]) :

$$D_j[f] = \sum_n d_n^j \psi_{j,n} \tag{4. 28}$$

Avec :

$$d_n^j =< f, \psi_{j,n} > \tag{4. 29}$$

On parle alors des détails de la fonction f obtenus à la résolution j. d_n^j sont les coefficients associés à ces détails.

Conclusion :

Nous savons que:

$$V_{j-1} = V_j \oplus W_j \tag{4. 30}$$

On peut en déduire l'approximation de f à l'échelle directement inférieure :

$$A_{j-1}[f] = A_j[f] + D_j[f] \tag{4. 31}$$

Compte tenu des expressions (4. 20) et (4. 26), ceci est équivalent à :

$$A_{j-1}[f] = \sum_n a_n^j 2^{-\frac{j}{2}} \varphi(2^{-j}x - n) + \sum_n d_n^j 2^{-\frac{j}{2}} \psi(2^{-j}x - n) \tag{4. 32}$$

La décomposition pour L niveaux de résolution est représentée par la figure 28 :

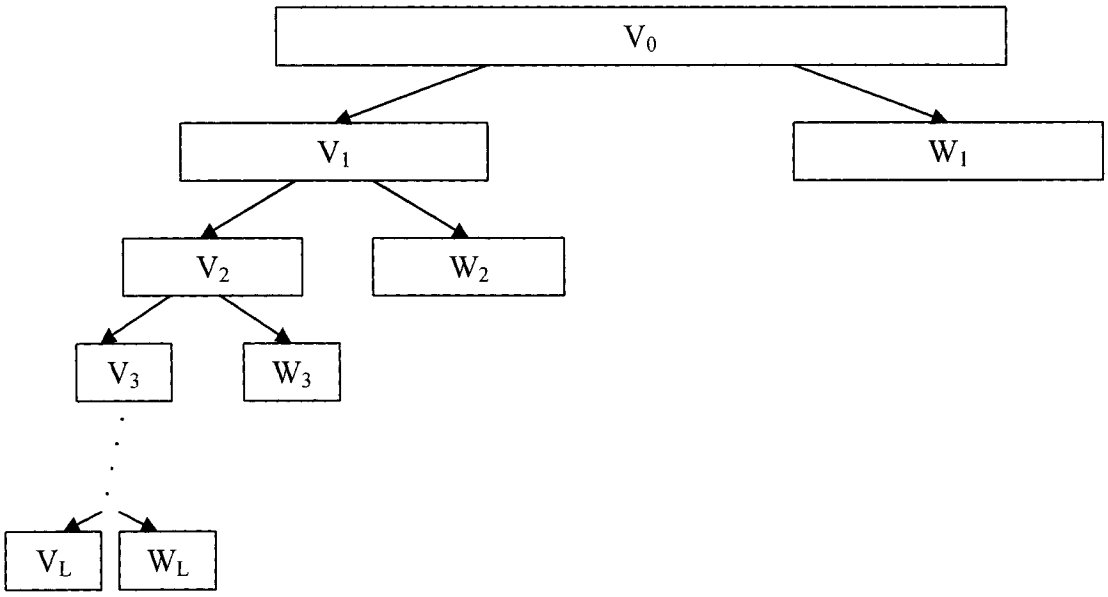


Figure 28 Schéma de l'analyse multirésolution pour L niveaux
(adaptée de « Ondelettes pour le signal », Truchetet [44])

On a donc :

$$V_0 = V_L \oplus W_1 \oplus W_2 \oplus ... \oplus W_L \tag{4. 33}$$

En utilisant l'expression (4.32), on obtient l'analyse multirésolution de $f \in V_0$:

$$f(x) = \sum_n a_n^L 2^{-\frac{L}{2}} \varphi(2^{-L}x - n) + \sum_{j=1}^{L} \sum_n d_n^j 2^{-\frac{j}{2}} \psi(2^{-j}x - n) \qquad (4.34)$$

Avec :

- a_n^j : coefficients d'échelles
- φ : fonction d'échelle
- d_n^j : coefficients d'ondelettes
- ψ : fonction d'ondelette

Finalement, $f(x)$ peut être représentée sous la forme d'une somme composée de (Akansu et Haddad [45]):

- une approximation au niveau L, obtenue à l'aide de la fonction d'échelle $\varphi_{L,n}$ (terme de gauche)
- L niveaux de détails, obtenus à l'aide de la fonction d'ondelette $\psi_{j,n}$ (terme de droite)

Ceci constitue les bases théoriques de l'analyse multirésolution.

4.1.4.2 De la théorie à la pratique : Algorithme récursif de Mallat

Afin d'effectuer une analyse multirésolution, il faut trouver un moyen pratique de passer d'une résolution à l'autre. Pour cela, il existe deux procédures : l'algorithme à trous (Dutilleux [50]), pour les analyses non-orthogonales, et l'algorithme de Mallat [47], pour celles orthogonales et biorthogonales. Ici, seul celui de Mallat sera présenté.

Passage à la résolution suivante :

Il s'agit ici de déterminer a_n^j et d_n^j à partir de a_n^{j-1}.

Partie Approximation :

Considérons le fait que φ, la fonction d'échelle, appartient à V_0. Comme $V_0 \subset V_{-1}$ et qu'il existe une base orthonormée dans V_{-1}, on peut exprimer φ dans cette base:

$$\varphi(x) = \sum_n h[n]\varphi_{-1,n}(x) \tag{4. 35}$$

Avec : $$h[n] = <\varphi, \varphi_{-1,n}> \tag{4. 36}$$

$h[n]$ peut être vu comme la réponse impulsionnelle d'un filtre (Akansu et Haddad [45]).

Il est important de noter que φ est par construction normée (en énergie), c'est-à-dire que:

$$\left\|\varphi(x)\right\|^2 = <\varphi, \varphi> = 1 \tag{4. 37}$$

Il est possible de montrer que cette norme se conserve à travers les différentes échelles (Truchetet [44]) :

$$\left\|\varphi_j(x)\right\|^2 = <\varphi_j, \varphi_j> = 1 \quad \forall j \in Z \tag{4. 38}$$

Il en découle que :

$$\sum_n h^2[n] = 1 \tag{4. 39}$$

D'après Mallat [46], la même décomposition est possible pour des échelles quelconques. En effet, de l'expression (4.20), on a :

$$\varphi_{j,n}(x) = 2^{\frac{-j}{2}}\varphi(2^{-j}x - n) \tag{4. 40}$$

Compte tenu de l'expression (4.35), on peut écrire :

$$\varphi(2^{-j}x - n) = \sum_k h[k]\varphi_{-1,k}(2^{-j}x - n) \tag{4. 41}$$

Or :

$$\varphi_{-1,k}(x) = 2^{\frac{1}{2}} \varphi(2x-k) \tag{4. 42}$$

Donc :

$$\varphi_{-1,k}(2^{-j}x-n) = 2^{\frac{1}{2}} \varphi(2^{-j+1}x-2n-k) \tag{4. 43}$$

En combinant les expressions (4.43), (4.41) et (4.40), on obtient :

$$\varphi_{j,n}(x) = 2^{-\frac{j}{2}} \sum_k h[k] 2^{\frac{1}{2}} \varphi(2^{-j+1}x-2n-k) \tag{4. 44}$$

$$\Leftrightarrow \varphi_{j,n}(x) = \sum_k h[k] 2^{-\frac{j-1}{2}} \varphi(2^{-(j-1)}x-(k+2n)) \tag{4. 45}$$

La décomposition est donc possible pour des échelles quelconques :

$$\varphi_{j,n}(x) = \sum_k h[k] \varphi_{j-1,k+2n}(x) \tag{4. 46}$$

Or nous savons que :

$$a_n^j = < f, \varphi_{j,n} > \tag{4. 47}$$

Donc :

$$a_n^j = \sum_k h[k] < f, \varphi_{j-1,k+2n} > \tag{4. 48}$$

En posant : $l = 2n + k$, on obtient :

$$a_n^j = \sum_l h[l-2n] < f, \varphi_{j-1,l} > \tag{4. 49}$$

Or :

$$< f, \varphi_{j-1,l} > = a_l^{j-1} \tag{4. 50}$$

D'où :

$$a_n^j = \sum_l h[l-2n] a_l^{j-1} \tag{4. 51}$$

Finalement, cette expression peut être considérée comme la convolution entre h et a^{j-1} pour un indice sur deux (Mallat [46]). Ainsi, pour obtenir a^j, il faut filtrer a^{j-1} par le filtre de réponse impulsionnelle h puis sous-échantillonner par 2, comme montré par la figure 29 :

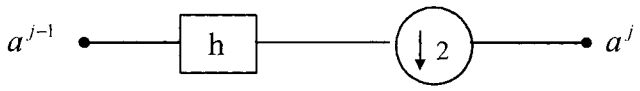


Figure 29 Schéma d'analyse pour l'obtention des coefficients d'approximation

Partie Détails :

Il s'agit ici de déterminer d^j à partir de a^{j-1}.

La fonction d'ondelette ψ de W_0 peut être exprimée dans la base orthonormée de V_{-1} :

$$\psi(x) = \sum_n g[n]\varphi_{-1,n}(x) \tag{4. 52}$$

Avec :
$$g[n] = <\psi, \varphi_{-1,n}> \tag{4. 53}$$

$g[n]$ peut être vu comme la réponse impulsionnelle d'un filtre (Akansu et Haddad [45]).

En utilisant la même démarche que pour la partie approximation, on obtient d'après Mallat [46]:

$$d_n^j = \sum_k g[k] < f, \varphi_{j-1,k+2n} > \tag{4. 54}$$

En posant : $l = 2n + k$, on obtient :

$$d_n^j = \sum_l g[l-2n] < f, \varphi_{j-1,l} > \tag{4. 55}$$

Or :

$$< f, \varphi_{j-1,l} >= a_l^{j-1} \tag{4. 56}$$

D'où :

$$d_n^j = \sum_l g[l-2n] a_l^{j-1} \qquad (4.57)$$

Finalement, cette expression peut être considérée comme la convolution entre g et a^{j-1} pour un indice sur deux (Mallat [46]). Ainsi, pour obtenir d^j, il faut filtrer a^{j-1} par le filtre de réponse impulsionnelle g puis sous-échantillonner par 2, comme montré par la figure 30 :

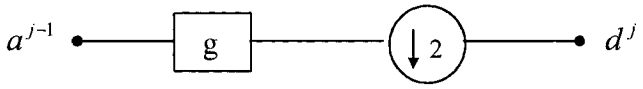


Figure 30 Schéma d'analyse pour l'obtention des coefficients de détails

Conclusion :

Finalement pour passer à la résolution suivante, c'est-à-dire obtenir a_n^j et d_n^j à partir de a_n^{j-1}, il faut utiliser l'algorithme d'analyse représenté par la figure 31 :

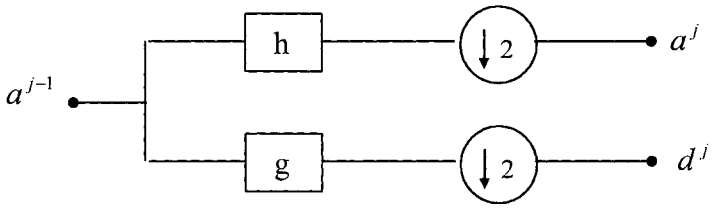


Figure 31 Algorithme récursif d'analyse de Mallat

Retour à la résolution précédente :

Il s'agit ici de déterminer a_n^{j-1} à partir de a_n^j et d_n^j.

Nous savons que :

$$A_{j-1}[f] = A_j[f] + D_j[f] = \sum_n a_n^j \varphi_{j,n} + \sum_n d_n^j \psi_{j,n} \tag{4.58}$$

Avec :

$$a_n^j = < f, \varphi_{j,n} > \tag{4.59}$$

$$d_n^j = < f, \psi_{j,n} > \tag{4.60}$$

Or (Vetterli et Kovacevic [42]) :

$$A_{j-1}\left[A_{j-1}[f]\right] = A_{j-1}[f] \tag{4.61}$$

Donc :

$$A_{j-1}\left[A_{j-1}[f]\right] = \sum_n < A_{j-1}[f], \varphi_{j-1,n} > \varphi_{j-1,n} = A_{j-1}[f] \tag{4.62}$$

Des expressions (4.58) et (4.62), on déduit que (Truchetet [44]) :

$$a_n^{j-1} = \sum_k a_k^j < \varphi_{j,k}, \varphi_{j-1,n} > + \sum_k d_k^j < \psi_{j,k}, \varphi_{j-1,n} > \tag{4.63}$$

Or de l'expression (4.46), on a :

$$\varphi_{j,k}(x) = \sum_l h[l] \varphi_{j-1,l+2k}(x) \tag{4.64}$$

Donc:

$$< \varphi_{j,k}, \varphi_{j-1,n} > = \sum_l h[l] < \varphi_{j-1,l+2k}, \varphi_{j-1,n} > \tag{4.65}$$

De par la propriété d'orthonormalité de la base d'échelle, on a:

$$< \varphi_{j-1,l+2k}, \varphi_{j-1,n} > = \delta(n - l - 2k) \tag{4.66}$$

En combinant (4.65) et (4.66), on obtient :

$$< \varphi_{j,k}, \varphi_{j-1,n} > = h[n - 2k] \tag{4.67}$$

Le même raisonnement permet d'obtenir (Mallat [46]) :

$$< \psi_{j,k}, \varphi_{j-1,n} > = g[n - 2k] \tag{4.68}$$

Enfin, de ces deux dernières expressions ainsi que de (4.63), on déduit l'équation permettant la reconstruction, c'est-à-dire l'obtention de a_n^j et d_n^j à partir de a_n^{j-1} :

$$a_n^{j-1} = \sum_k a_k^j h[n-2k] + \sum_k d_k^j h[n-2k] \qquad (4.69)$$

La signification de cette expression peut être représentée par la figure 32 :

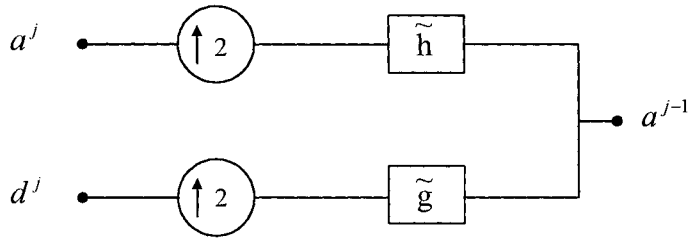


Figure 32 Algorithme récursif de reconstruction de Mallat

Avec : $$\tilde{h}[n] = h[-n] \text{ et } \tilde{g}[n] = g[-n] \qquad (4.70)$$

En conclusion, dans le cas d'une analyse orthogonale, on peut décomposer un signal en différents niveaux de résolution, à l'aide du schéma d'analyse de Mallat [47] (figure 31). On parle de transformée en ondelettes rapide. Pour reconstruire le signal, il faut utiliser le schéma de synthèse de Mallat [47] (figure 32).

4.1.4.3 La réalisation pratique de l'analyse multirésolution

L'algorithme récursif de Mallat, représenté par la figure 33 permet donc de passer d'une résolution à l'autre :

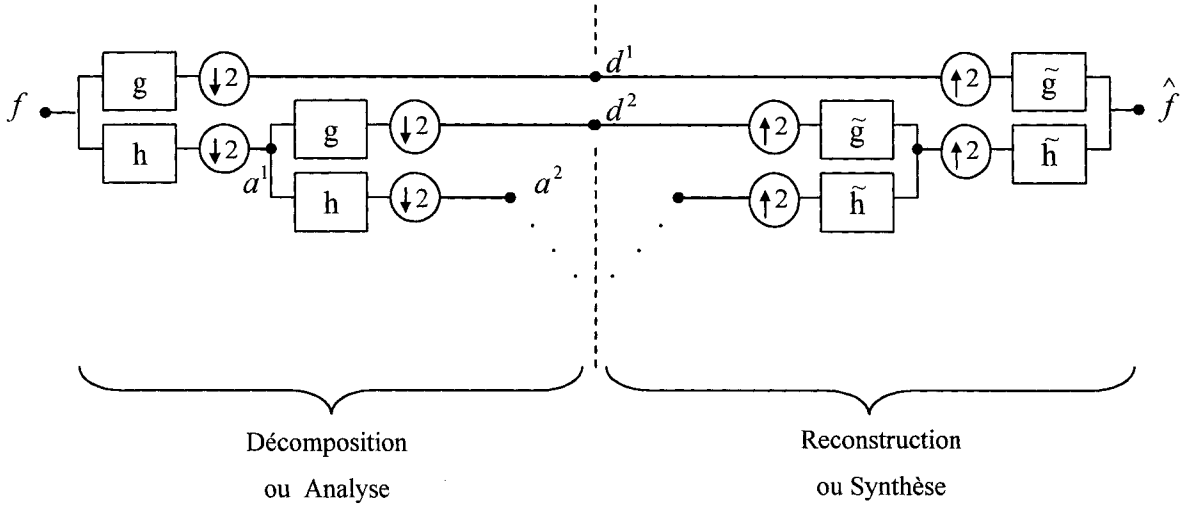


Figure 33 Schéma pour l'analyse et la recomposition à plusieurs niveaux de résolution

La réalisation pratique de l'analyse multirésolution est donc possible à l'aide d'un banc de filtres. La question est : quels types de filtres utiliser ?

Nous avons vu dans la section 4.1.4.1 que la décomposition de f en L niveaux de résolution est donnée par :

$$f(x) = \sum_n a_n^L 2^{-\frac{L}{2}} \varphi(2^{-L}x - n) + \sum_{j=1}^{L} \sum_n d_n^j 2^{-\frac{j}{2}} \psi(2^{-j}x - n) \qquad (4.71)$$

Avec :

- a_n^j : coefficients d'échelles
- φ : fonction d'échelle
- d_n^j : coefficients d'ondelettes
- ψ : fonction d'ondelette

Selon Akansu et Haddad [45], il peut être démontré que l'approximation au niveau L (terme de droite) est obtenue par des filtres h passe-bas et les L niveaux de détails (terme de gauche) par des filtres g passe-haut. On rappelle que les détails sont généralement formés de hautes fréquences et que l'approximation est la partie plus monotone du signal, il s'agit donc des basses fréquences. Ce sont les choix de g et h qui permettent de définir les fonctions d'échelles et d'ondelettes.

Prenons le cas simple d'un seul niveau de résolution, figure 34, et plaçons nous dans le domaine Z:

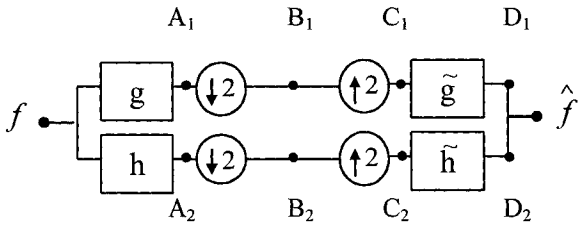


Figure 34 Schémas d'analyse et de synthèse pour un niveau de résolution

On a donc:

$$A_1(z) = F(z)G(z) \tag{4. 72}$$

$$A_2(z) = F(z)H(z) \tag{4. 73}$$

Le sous-échantillonnage par 2, qui revient à sélectionner un échantillon sur 2, se traduit dans le domaine Z par :

$$X_{\downarrow 2}(z) = \frac{1}{2}\left[X\left(z^{\frac{1}{2}}\right) + X\left(-z^{\frac{1}{2}}\right)\right] \tag{4. 74}$$

On obtient donc :

$$B_1(z) = \frac{1}{2}\left[A_1\left(z^{\frac{1}{2}}\right) + A_1\left(-z^{\frac{1}{2}}\right)\right] = \frac{1}{2}\left[F\left(z^{\frac{1}{2}}\right)G\left(z^{\frac{1}{2}}\right) + F\left(-z^{\frac{1}{2}}\right)G\left(-z^{\frac{1}{2}}\right)\right] \tag{4. 75}$$

$$B_2(z) = \frac{1}{2}\left[A_2\left(z^{\frac{1}{2}}\right) + A_2\left(-z^{\frac{1}{2}}\right)\right] = \frac{1}{2}\left[F\left(z^{\frac{1}{2}}\right)H\left(z^{\frac{1}{2}}\right) + F\left(-z^{\frac{1}{2}}\right)H\left(-z^{\frac{1}{2}}\right)\right] \tag{4. 76}$$

Le sur-échantillonnage par 2, qui revient à rajouter un échantillon nul entre chaque échantillon, se traduit dans le domaine Z par :

$$X_{\uparrow 2}(z) = X(z^2) \tag{4. 77}$$

D'où :

$$C_1(z) = B_1(z^2) = \frac{1}{2}\left[F(z)G(z) + F(-z)G(-z)\right] \tag{4. 78}$$

$$C_2(z) = B_2(z^2) = \frac{1}{2}\left[F(z)H(z) + F(-z)H(-z)\right] \tag{4. 79}$$

Après filtrage par $\tilde{h}$ et $\tilde{g}$, on a :

$$D_1(z) = C_1(z)\tilde{G}(z) = \frac{1}{2}\left[F(z)G(z) + F(-z)G(-z)\right]\tilde{G}(z) \tag{4. 80}$$

$$D_2(z) = C_2(z)\tilde{H}(z) = \frac{1}{2}\left[F(z)H(z) + F(-z)H(-z)\right]\tilde{H}(z) \tag{4. 81}$$

Enfin, en additionnant ces deux dernières expressions, on obtient :

$$\hat{F}(z) = \frac{1}{2}\left[F(z)G(z) + F(-z)G(-z)\right]\tilde{G}(z) + \frac{1}{2}\left[F(z)H(z) + F(-z)H(-z)\right]\tilde{H}(z) \tag{4. 82}$$

En regroupant les termes associés à $F(z)$ et à $F(-z)$, ceci équivaut à :

$$\hat{F}(z) = \frac{1}{2}\left[\tilde{G}(z)G(z) + \tilde{H}(z)H(z)\right]F(z) + \frac{1}{2}\left[\tilde{G}(z)G(-z) + \tilde{H}(z)H(-z)\right]F(-z) \tag{4. 83}$$

Or nous souhaitons que le signal reconstruit soit identique au signal initial, c'est-à-dire que la reconstruction soit parfaite :

$$\hat{F}(z) = F(z) \tag{4. 84}$$

De l'expression (4.83), on déduit donc les deux conditions de reconstruction parfaite :

$$\left[\tilde{G}(z)G(z)+\tilde{H}(z)H(z)\right]=2z^{-l} \qquad l\in Z \tag{4.85}$$

$$\left[\tilde{G}(z)G(-z)+\tilde{H}(z)H(-z)\right]=0 \tag{4.86}$$

Il existe plusieurs méthodes qui permettent d'obtenir la reconnaissance parfaite tout en satisfaisant d'autres conditions comme par exemple l'orthogonalité des filtres (Mallat [46], Croisier, Esteban et Galand [51], Chan [52]...). Elles font en général usage des filtres miroirs en quadrature. Ainsi *h* est un filtre passe-bas et *g* est le passe-haut miroir de *h* par rapport à $\frac{\pi}{2}$. La figure 35 représente leur réponse fréquentielle respective :

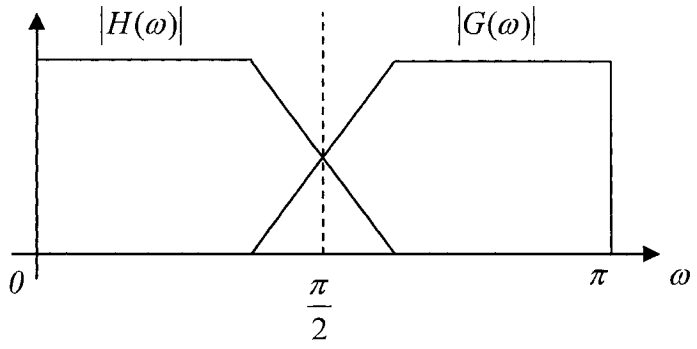


Figure 35 Filtres miroirs en quadrature

En respectant ces procédures, il est donc possible de déterminer les filtres *h* et *g* utilisés par l'algorithme de Mallat et ainsi réaliser une analyse multirésolution.

Selon Chan [52], les filtres à réponse impulsionnelle finie miroirs en quadrature ne sont pas à phase linéaire, sauf dans le cas des ondelettes de Haar, ce qui ne convient pas à toutes les applications, comme par exemple le traitement d'images.

Pour obtenir la propriété de linéarité de la phase, il faut utiliser des filtres biorthogonaux, c'est-à-dire vérifiant le système suivant, (Chan [52]) :

$$\begin{cases} \sum_{m=0}^{p-1} g(m)\left[\tilde{h}(m+k)+\tilde{h}(m-k)\right]=0 \quad k \text{ impair} & (4.87) \\ \sum_{m=0}^{p-1} \tilde{g}(m)\left[h(m+k)+h(m-k)\right]=0 \quad k \text{ impair} & (4.88) \\ \sum_{m=0}^{p-1} g(m)\tilde{g}(k-m)=\delta(k-l) & (4.89) \\ \sum_{m=0}^{q-1} h(m)\tilde{h}(k-m)=\delta(k-l) & (4.90) \end{cases}$$

Avec : p : l'ordre de g et q : celui de h.

La biorthogonalité entraîne des filtres d'ordre différents.

4.1.4.4 Exemples de fonctions d'ondelettes et d'échelles

La figure 36 montre deux types de fonctions d'échelles et d'ondelettes couramment utilisées.

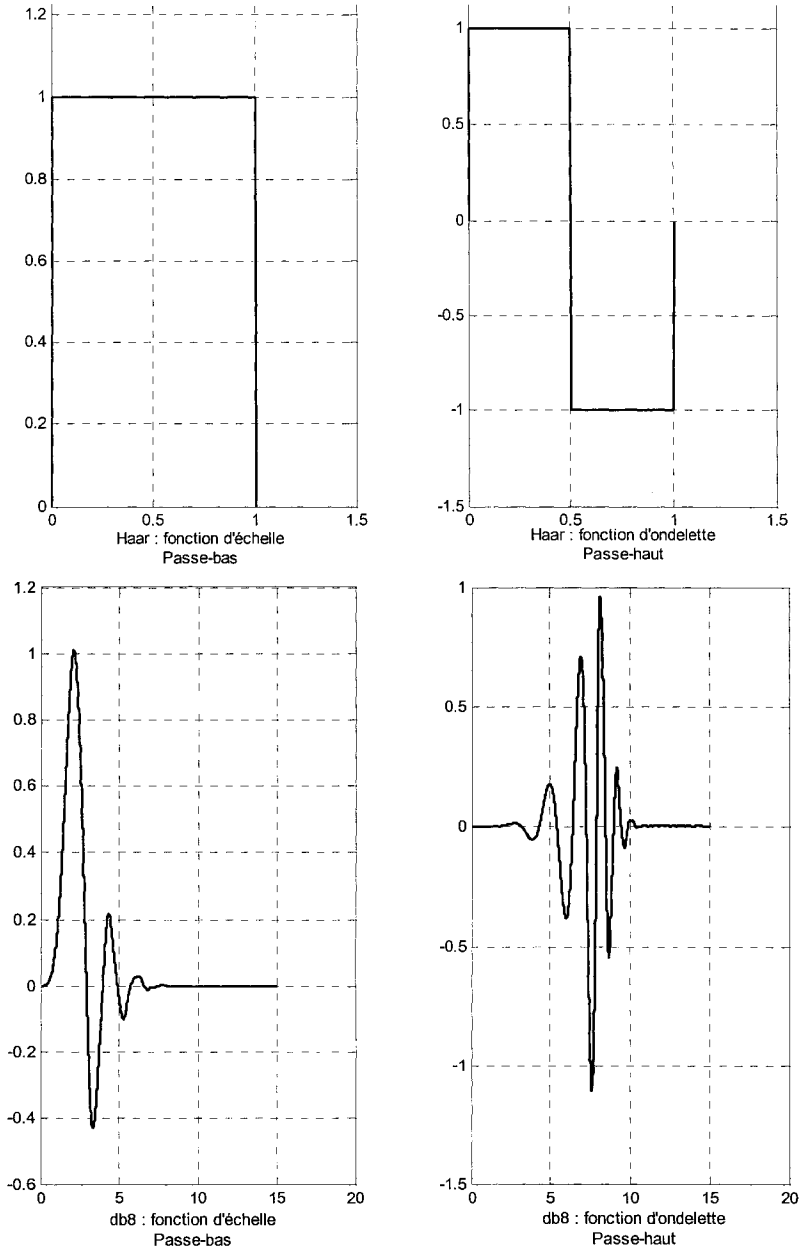
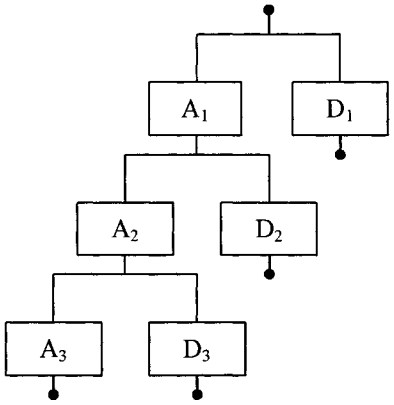


Figure 36 Exemples de fonctions d'échelles et d'ondelettes : Haar (en haut), Daubechies d'ordre 8 (en bas)

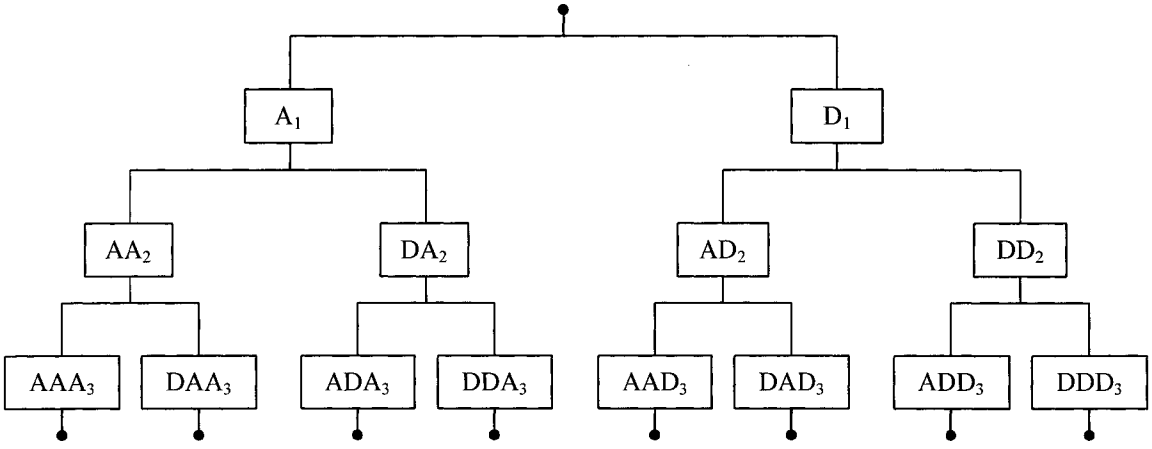
4.1.5 Les paquets d'ondelettes

Les paquets d'ondelettes sont une extension de l'analyse multirésolution vue dans la partie précédente. En effet, les détails peuvent être ici décomposés de la même manière que les approximations. La figure 37 met en évidence ceci :



Analyse multirésolution

Pour n niveaux, il y a $n+1$ décompositions possibles du signal.



Paquets d'ondelettes

Pour n niveaux, il y a $2^{2^{n+1}}$ décompositions possibles du signal.

Figure 37 Comparaison Analyse multirésolution / Paquets d'ondelettes

Les paquets d'ondelettes décomposent le signal en différentes bandes de fréquences : les basses fréquences sont complètement à gauche et les hautes fréquences complètement à droite (sur la figure 37). La largeur des bandes dépend du choix de l'arbre. Cet outil est donc très flexible et utile car il offre une résolution fréquentielle variable, pour pouvoir s'adapter au mieux au signal à analyser.

Il est à noter que la reconstruction est parfaite car la base utilisée pour la synthèse est la même que celle pour l'analyse (Vetterli et Kovacevic [42]). Ici encore, la base d'ondelettes est orthogonale.

Comme il est indiqué à la figure 37, pour n niveaux, il y a $2^{2^{n+1}}$ possibilités de décomposer le signal. La question est : quelle décomposition choisir ? D'après Mallat [53], la meilleure décomposition est celle pour laquelle l'entropie est la plus faible. Ainsi seuls quelques coefficients auront une valeur d'entropie beaucoup plus élevée, ce qui facilitera par la suite l'analyse de la décomposition en paquets d'ondelettes.

L'entropie du signal peut être exprimée en fonction des coefficients d'ondelettes :

$$E(f) = \sum_i E(f_i) \tag{5.91}$$

Avec : f_i : les coefficients d'ondelettes de f.

Le but est donc de minimiser cette fonction coût qui doit d'ailleurs être nulle en 0. Il existe plusieurs manières de calculer $E(f_i)$, comme par exemple :

- l'entropie par Shannon :

$$E(f_i) = -f_i^2 \log(f_i^2) \tag{5.92}$$

- l'entropie « logarithme de l'énergie » :

$$E(f_i) = \log(f_i^2) \tag{5.93}$$

L'algorithme de recherche du meilleur arbre se déroule de la manière suivante :

- Calcul de l'entropie du signal d'origine
- Décomposition du nœud en deux

- Calcul de l'entropie à chacun des nœuds : si la somme de ces deux entropies est inférieure à l'entropie du nœud en amont, la décomposition de ce dernier est retenue. Dans le cas contraire, le nœud en amont n'aura pas de fils.

La même méthode est appliquée à chaque nouveau nœud de l'arbre jusqu'à ce que l'entropie atteigne un minimum. Il est à noter que si l'entropie d'un nœud est nulle, il ne sert à rien de le décomposer.
Une fois le meilleur arbre trouvé, l'analyse en paquets d'ondelettes du signal initial peut être effectuée.

Nous venons de voir l'essentiel de la théorie des ondelettes. Nous allons maintenant aborder la détection d'activité vocale basée sur les ondelettes en présentant quelques algorithmes proposés dans la littérature.

4.2 Les DAV ondelettes proposés dans la littérature

Comme il a été mentionné précédemment, les ondelettes permettent dans de nombreux cas d'atteindre des performances bien supérieures à celles obtenues avec les outils traditionnels. Dans le domaine du traitement de la voix, elles sont très puissantes pour le débruitage de la parole (cette notion sera abordée par la suite, section 5.1.2). Cependant, les recherches bibliographiques ont révélé que cet outil est pour le moment très peu courant pour la détection d'activité vocale. Seuls quelques DAV basés sur les ondelettes ont été proposés dans la littérature au cours de la dernière décennie. Nous allons ici présenter les deux algorithmes qui semblent être les plus pertinents.

4.2.1 L'algorithme de Stegmann et Schröder

En 1997, Stegmann et Schröder [54] ont suggéré un DAV basé sur les ondelettes. La procédure utilisée est relativement simple et repose sur l'extraction de quatre caractéristiques d'énergie, obtenues à l'aide des coefficients d'ondelettes. La majorité des DAV détecte la présence de parole à l'aide d'un ensemble de règles. Ainsi si aucune d'entre elles n'est vérifiée, la trame est considérée inactive. Celui-ci raisonne dans le sens inverse, c'est-à-dire que si aucune des conditions établies n'est vérifiée, le segment étudié contient de la voix. Il détecte donc non pas la présence mais l'absence de parole.
Afin de prendre une décision quant à l'activité vocale, chaque trame du signal initial est analysée de la manière suivante :

Transformée en ondelettes :
Le passage dans le domaine temps-fréquence s'effectue à l'aide de la TOD Daubechies d'ordre 4. Il en résulte alors un ensemble de coefficients d'approximation : A_L et L ensembles de coefficients de détails : D_1 à D_L.

Extraction des 4 caractéristiques d'énergie :
Les quatre caractéristiques extraites sont :

- l'énergie totale : il s'agit de la somme des énergies des $L+1$ ensembles de coefficients : E_{totale}
- la différence des énergies des L niveaux de détails entre la trame actuelle et la précédente : Δ^k avec k : numéro de la trame actuelle
- l'estimation du bruit pour les parties de détails D_2 et D_L : B_2 et B_L respectivement

- les P sous-énergies du niveau de détails D_2 : D_2 est en fait décomposé en P intervalles et l'énergie est calculée pour chacun d'entre eux : $E_{2,i}$ avec $i = 1$ à P

Génération de 4 drapeaux : silence, stationnarité, bruit en D_2, bruit en D_L :

Les caractéristiques sont ensuite comparées à des seuils afin de générer quatre drapeaux :

- si $E_{totale} < T_1$ alors le drapeau *silence* est unitaire
- si Δ^k et $\Delta^{k-1} < T_2$ alors le drapeau *stationnarité* est unitaire
- si $E_{2,i} - B_2 < T_3 \quad \forall i$, c'est-à-dire si toutes les sous-énergies de D_2 sont inférieures à l'estimation du bruit B_2, alors le drapeau *bruit en D_2* est unitaire
- si $E_L - B_L < T_4$, c'est-à-dire si l'énergie en D_L est inférieure à l'estimation du bruit B_L, alors le drapeau *bruit en D_L* est unitaire

Prise de décision :

La décision préliminaire quant à l'état de la trame étudiée est obtenue à l'aide des états des quatre drapeaux précédents :

- si du silence a été détecté, c'est-à-dire : $silence = 1$, alors la trame est dite inactive
- si du bruit a été détecté en D_2 et en D_L et si la stationnarité a été vérifiée, c'est-à-dire : *bruit en* $D_2 = 1$ & *bruit en* $D_L = 1$ & *stationnarité* $= 1$, alors la trame est dite inactive
- sinon la trame contient de la parole et elle est déclarée active

Révision de la décision :

Comme dans la plupart des DAV, la décision finale est obtenue après un algorithme de *Hangover*. Ce type de procédure vise à réduire les coupures dans les longues bouffées de parole et à faciliter les fins de bouffées, généralement basses énergies. De même, ce procédé évite les faux déclenchements dus à des pics de bruit soudains et intenses dans une longue période d'inactivité vocale. Les auteurs ont opté pour celui du codeur de parole « GSM Enhanced Full Rate » [55]. Le principe est le même que pour le *Hangover* du G729, vu dans le chapitre 3, mais les règles utilisées sont différentes.

L'algorithme du DAV peut être schématisé par la figure 38 :

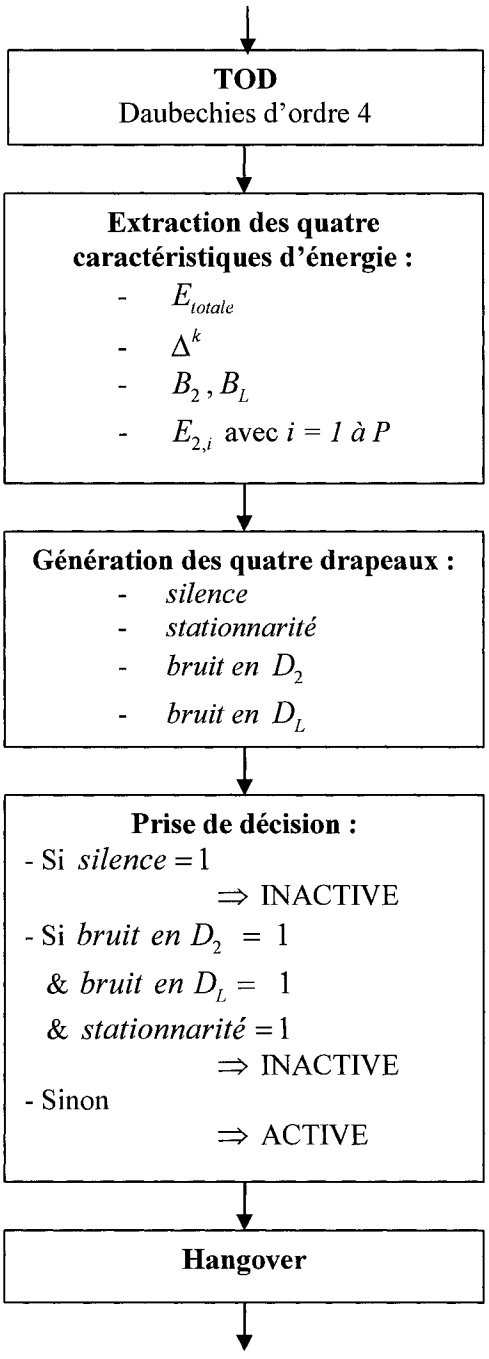


Figure 38 Schéma Bloc du DAV de Stegmann et Schröder

Afin de déterminer la robustesse de ce DAV, Stegmann et Schröder ont procédé à une série de tests : 8 phrases en allemand (environ 47% d'activité vocale), 4 locuteurs : 2 hommes, 2 femmes, 4 bruits : voiture, *babble*, harmoniques, bruit blanc et une plage de rapports signal à bruit de 50 à 10dB. Ils ont ensuite comparé les performances obtenues avec celles du G729.B et du « GSM Enhanced Full Rate ». Il semblerait que le DAV proposé représente un compromis entre les deux systèmes. En effet, les résultats montrent que les faux déclenchements sont beaucoup moins nombreux que pour le GSM alors que les coupures dans la parole sont moins fréquentes qu'avec le G729.B et ceci pour tous les bruits étudiés et toute la plage de RSB. D'après les auteurs, le comportement de leur DAV se dégrade de manière conséquente pour les RSB inférieurs à 10dB.

4.2.2 L'algorithme de Chen et Wang

En 2002, lors de son doctorat sur l'application des ondelettes à la parole [56], [57], Shi-Huang Chen, en collaboration avec son professeur Jhing-Fa Wang, a mis au point un DAV basé sur les ondelettes. L'idée maîtresse de cet algorithme est de générer une unique caractéristique : la Forme d'Activité Vocale (FAV). Une simple comparaison à un seuil adaptatif suffit ensuite à obtenir une décision quant à la présence de parole.
La procédure de détection est la suivante :

Décomposition en paquets d'ondelettes :
Le signal est tout d'abord décomposé en paquets d'ondelettes selon l'échelle de Bark. Cette dernière modélise l'oreille humaine de manière psychoacoustique, c'est-à-dire qu'elle reflète le fait que l'homme ne distingue pas toutes les plages de fréquences de la même manière. Elle est définie par (Zwicker et Terhardt [58]):

$$z(f) = 13\arctan(7.6\times10^{-4} f) + 3,5\arctan(1.33\times10^{-4} f)^2 \qquad (4.94)$$

Elle peut être représentée par la figure 39 :

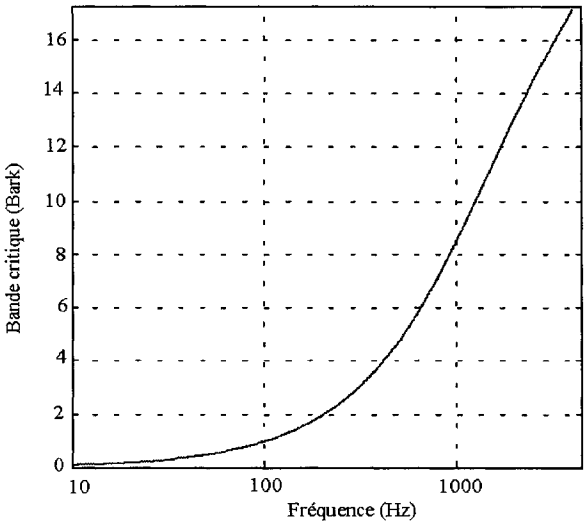


Figure 39 Échelle de Bark
(adaptée de la thèse de Shi-Huang Chen, 2002, [56])

Shi-Huang Chen, [56], [57], a déterminé l'arbre de décomposition permettant d'approximer cette échelle et qui est représenté par la figure 40 :

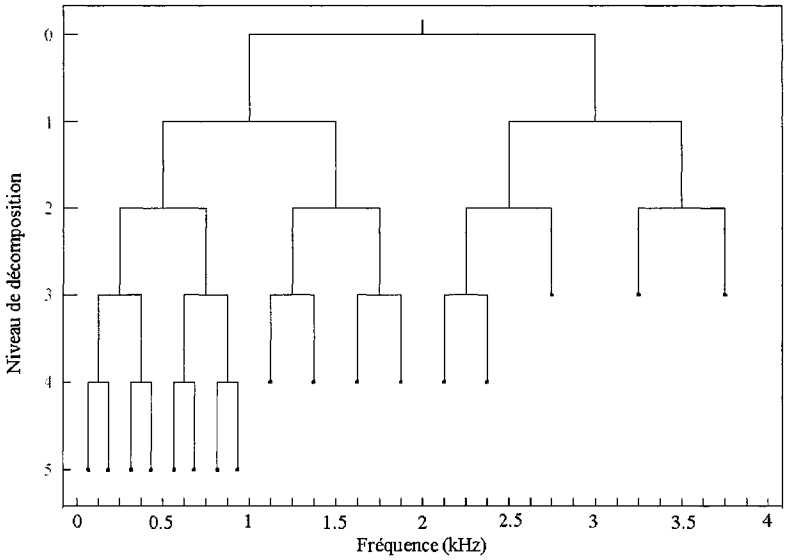


Figure 40 Arbre de décomposition selon l'échelle de Bark
(adaptée de la thèse de Shi-Huang Chen, 2002, [56])

De plus, pour cet algorithme, l'auteur a choisi d'utiliser une ondelette Daubechies d'ordre 10.

Traitement :

De l'étape précédente, il résulte une décomposition du signal initial en 17 sous-bandes de fréquences. Le *Teager Energy Operator* (TEO) (Kaiser [59]) est alors appliqué à chacune d'entre elles. Cet opérateur non linéaire reflète l'énergie du signal et permet d'augmenter les différences ente les composantes de parole et celles du bruit [56], [57]. Il peut être défini par :

$$TEO[y(n)] = y^2(n) - y(n+1)y(n-1) \qquad (4.95)$$

Sélection des bandes :

Dans le but d'éliminer les bandes de fréquences composées uniquement de bruit, la variance de chaque nœud est comparée à un seuil. Si elle est supérieure à ce seuil, les coefficients restent intacts, dans le cas contraire ils sont mis à zéro. Cette sélection s'apparente en fait au seuillage dur utilisé dans le débruitage de la parole. Cette notion sera vue plus en profondeur à la section 5.1.2.

Masquage :

Les bandes de fréquences sélectionnées lors de l'étape précédente sont convoluées avec une fenêtre de Hamming.

Calcul de la FAV :

La FAV, caractéristique utilisée pour la prise de décision, est calculée. D'après Chen, [56], il s'agit de la somme des 17 transformées inverses des bandes obtenues après masquage.

Prise de décision :

La FAV est une fonction du temps et selon l'auteur, elle présente une amplitude plus forte lors de parole que lors de bruit seul. Une comparaison de son amplitude à un seuil adaptatif suffit donc à prendre une décision.

L'algorithme du DAV peut être schématisé par la figure 41 :

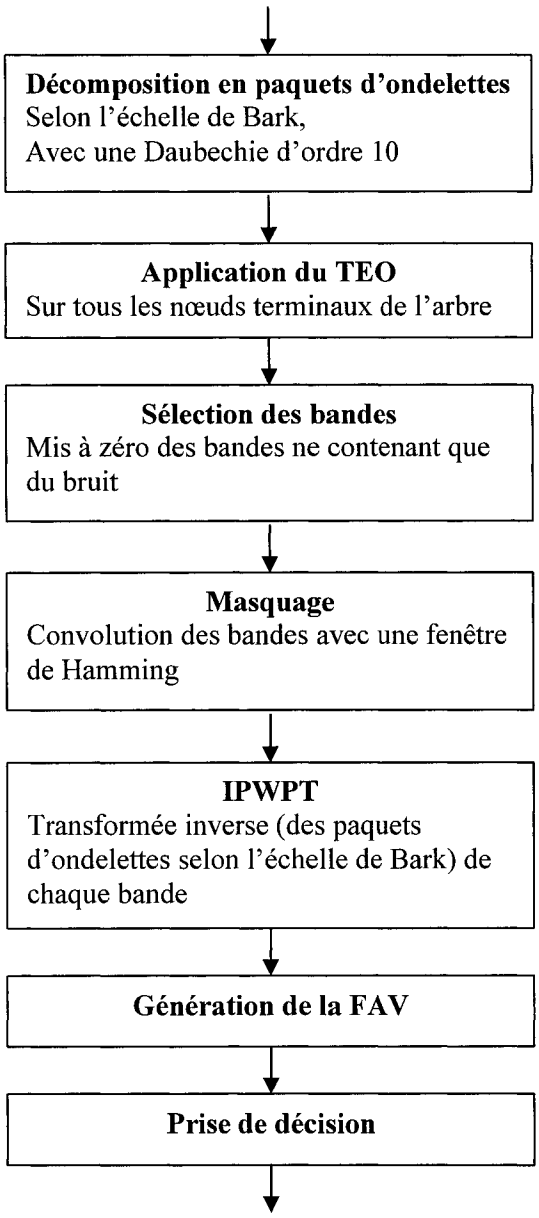


Figure 41 Schéma Bloc du DAV de Chen et Wang

Le système ainsi développé a été testé pour 3 bruits : blanc, rue, voiture et pour des rapports signal à bruit de 25dB à 0dB. De même que précédemment, les performances ont été comparées à celles du G729.B. Les résultats concernant la reconnaissance des régions de parole sont très bons et ceci dans toutes les situations. En effet, la plus basse performance est obtenue pour le bruit blanc à un RSB de 0dB et s'élève tout de même à 84%. De même, le pourcentage de faux déclenchements, bruit pris pour de la parole, est très acceptable, atteignant un maximum de 14% à 0dB pour le bruit de la rue. Comparé au G729.B, le DAV proposé ici semble bien plus robuste.

Compte tenu du petit nombre d'articles proposant des DAV basés sur les ondelettes, il semble qu'il est encore difficile aujourd'hui d'exploiter correctement l'information fournie par un tel outil. Pour palier à ce problème, des chercheurs ont proposé d'utiliser un réseau de neurones afin d'identifier les différences entre les coefficients d'ondelettes des classes « parole » et « bruit » et de pouvoir par la suite réaliser la détection d'activité vocale, Hoyt et Wechsler [22], [23]. En parallèle, d'autres chercheurs tentent de déterminer le type de l'ondelette mère le plus approprié à ce domaine, Kacur et Al. [60]. De l'étude des deux algorithmes présentés plus haut, il ressort que les ondelettes sont susceptibles d'être très efficaces pour la détection d'activité vocale. Dans le cadre de ce projet, nos recherches ont abouti à un nouvel algorithme de DAV basé sur les ondelettes permettant d'obtenir des performances intéressantes. Ceci est abordé dans les deux prochains chapitres.

CHAPITRE 5

LE DÉTECTEUR D'ACTIVITÉ VOCALE BASÉ SUR LES ONDELETTES

Dans le cadre de ce projet de recherche, l'approche des ondelettes appliquées à la détection d'activité vocale a été étudiée et nos recherches ont abouti à un nouvel algorithme, appelé DAV INNES (**I**ndustriel o**N**delettes joh**N**stone silv**E**rman aju**S**tement). Il repose sur deux notions fondamentales : la décomposition en paquets d'ondelettes selon l'échelle de Mel et la prise de décision en fonction des valeurs du Paramètre du Seuil de Johnstone et Silverman (PSJS) et des énergies. La première partie de ce chapitre va permettre de définir les notions théoriques utilisées par ce DAV. Suite à cela, l'algorithme sera présenté en détails et les raisons qui nous ont poussé à développer un tel système seront exposées. Une procédure d'ajustement du DAV à un environnement particulier a été mise au point afin d'obtenir systématiquement des performances accrues. Pour faciliter son utilisation, nous l'avons entièrement automatisée. La dernière partie du chapitre sera consacrée à l'explication de cette procédure.

5.1 Notions théoriques

Nous allons dans un premier temps aborder la décomposition en paquets d'ondelettes selon l'échelle de Mel, puis définir les deux caractéristiques utilisées dans la prise de décision, à savoir le PSJS et l'énergie. Enfin deux concepts mathématiques seront définis : la moyenne et la variance.

5.1.1 Décomposition en paquets d'ondelettes selon l'échelle de Mel

La plage de fréquences audibles par l'homme est de l'ordre de 20 à 20 000Hz. Toutefois, il ne possède pas la même capacité à discriminer les sons sur toute cette plage. En effet,

pour que l'oreille humaine entende deux sons distincts, leur différence de fréquences doit être plus ou moins grande selon qu'ils se trouvent dans les basses ou dans les hautes fréquences. Plusieurs études psychoacoustiques ont montré ce phénomène et ont abouti à des modèles perceptifs de l'ouïe. L'un d'entre eux est l'échelle de Bark qui a été introduite à la section 4.2.2. Un autre modèle très utilisé en reconnaissance de la parole est l'échelle de Mel (Umesh et Al. [61]). Elle a été proposée en 1937 par Stevens, Volkman et Newman.

L'homme a la faculté de discerner très facilement différents types de sons et par conséquent de distinguer la parole du bruit. L'utilisation d'un modèle perceptif de l'audition humaine dans l'algorithme de détection d'activité vocale semble donc très intéressante. C'est pourquoi le DAV INNES proposé ici repose sur l'échelle de Mel. Avant d'expliquer comment utiliser conjointement cette échelle et les ondelettes, il est important de présenter ses caractéristiques.

La figure 42 montre la correspondance entre le domaine des fréquences et celui des Mels :

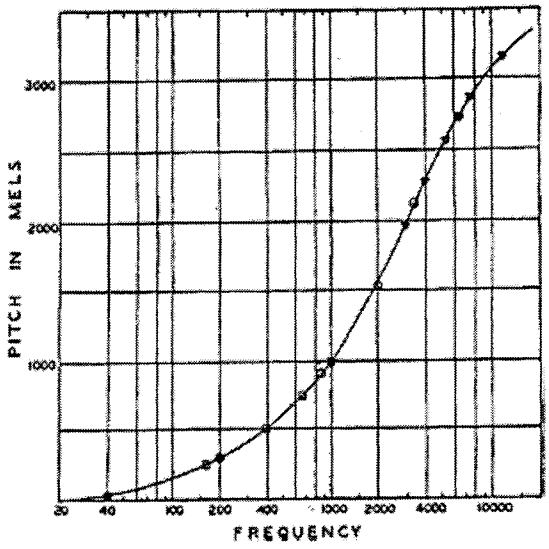


Figure 42 Échelle de Mel
(adaptée de Umesh, Cohen et Nelson, 2002 [61])

La relation liant ces deux grandeurs est la suivante :

$$m(f) = x \log(1 + \frac{f}{y}) \tag{5. 1}$$

Avec : f : la fréquence, en Hertz.

Plusieurs types de logarithmes et plusieurs valeurs de x et y ont été utilisés dans le passé. Aujourd'hui, comme par exemple dans l'article de Wu et Lin [62], il est très courant de trouver la formule suivante :

$$m(f) = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{5. 2}$$

À l'instar de tout modèle perceptif de l'ouïe, celui-ci peut être vu comme une décomposition en bandes de fréquences de différentes largeurs. Ainsi, l'utilisation de l'échelle de Mel avec les ondelettes ne peut se faire que si l'on a recours aux paquets d'ondelettes. Malheureusement, compte tenu de leurs caractéristiques, il n'est pas possible de respecter exactement cette séparation de l'espace fréquentiel. Il s'agit donc d'une approximation.

Farooq et Datta, dans leur article de 2001 [63], ont justement établi le lien entre l'échelle de Mel et la décomposition en paquets d'ondelettes, comme le montre le tableau VI. C'est cette correspondance qui a permis d'obtenir l'arbre de décomposition en ondelettes utilisé dans l'algorithme du DAV INNES.

Tableau VI

Bandes de fréquences pour les paquets d'ondelettes et l'échelle de Mel
(adapté de Farooq et Datta, 2001 [63])

Filtre Ondelettes				Filtre Mel	
Numéro du filtre	Fréquence de coupure basse (Hz)	Fréquence de coupure haute (Hz)	Bande passante (Hz)	Fréquence centrale (Hz)	Bande passante (Hz)
1	0	125	125	100	100
2	125	250	125	200	100
3	250	375	125	300	100
4	375	500	125	400	100
5	500	625	125	500	100
6	625	750	125	600	100
7	750	875	125	700	100
8	875	1000	125	800	100
9	1000	1125	125	900	100
10	1125	1250	125	1000	124
11	1250	1375	125	1149	160
12	1375	1500	125	1320	184
13	1500	1750	250	1516	211
14	1750	2000	250	1741	242
15	2000	2250	250	2000	278
16	2250	2500	250	2297	320
17	2500	2750	250	2639	367
18	2750	3000	250	3031	422
19	3000	3500	500	3482	484
20	3500	4000	500	4000	556
21	4000	5000	1000	4595	639
22	5000	6000	1000	5278	734
23	6000	7000	1000	6063	843
24	7000	8000	1000	6954	969

La fréquence d'échantillonnage des signaux de parole attendue par le DAV INNES est 8kHz. Compte tenu du théorème de Nyquist, introduit à la section 1.2, nous devons avoir :

$$f_e \geq (2 \times f_m) = \textit{fréquence de Nyquist} \qquad (5.3)$$

Les signaux qui vont être examinés par le DAV INNES ont donc une bande passante de 4kHz. Ainsi, seuls les 20 premiers filtres du tableau VI nous intéressent.

On peut à présent en déduire la décomposition en paquets d'ondelettes selon l'échelle de Mel :

les trois premiers niveaux de l'arbre sont générés entièrement. Ceci permet d'obtenir huit bandes de fréquences. Les trois plus basses, à savoir 0-0.5kHz, 0.5-1kHz et 1-1.5kHz, sont ensuite redécomposées à l'aide de deux niveaux supplémentaires, d'où l'obtention de 12 bandes de largeur 125Hz. Les trois bandes moyennes du niveau 3, c'est-à-dire 1.5-2kHz, 2-2.5kHz et 2.5-3kHz, sont, quant à elle, redécomposées à l'aide d'un seul niveau, ce qui engendre 6 bandes de fréquences de largeur 250Hz. Enfin, les deux dernières, soit les plus hautes, sont laissées telles quelles et correspondent aux 2 bandes de largeur 500Hz.

La figure 43, qui suit, représente cet arbre de décomposition en paquets d'ondelettes.

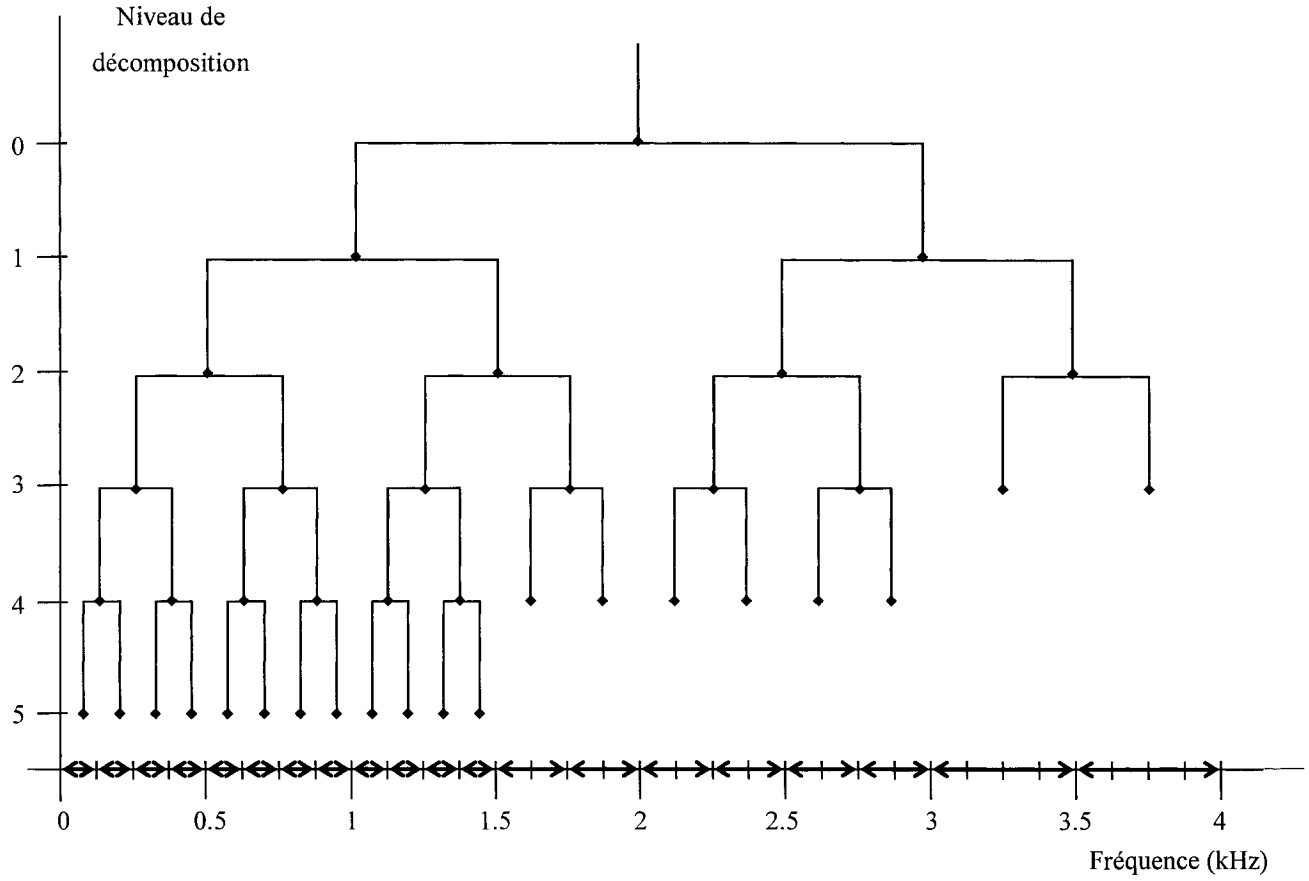


Figure 43 Arbre de décomposition en paquets d'ondelettes selon l'échelle de Mel

5.1.2 Paramètre du Seuil de Johnstone et Silverman

L'un des domaines privilégiés des ondelettes est le débruitage de la parole. En effet, ce type d'algorithmes est relativement simple et fournit de très bons résultats. Il s'agit d'effectuer une transformée en ondelettes du signal à débruiter puis d'appliquer un seuillage sur les coefficients obtenus et enfin de revenir au domaine temporel.

Il existe aujourd'hui une multitude de méthodes différentes puisqu'à chaque nouvelle forme de seuillage correspond une nouvelle méthode. Toutefois, elles reposent toutes sur les deux premières procédures proposées, à savoir les seuillages dur et doux.

Selon Chang, Kwon, Yang et Kim [64], on peut les définir de la manière suivante :
Soit x le signal de parole p corrompu par du bruit b. On a donc :

$$x = p + b \tag{5.4}$$

À l'aide de la transformée en ondelettes, on passe dans le domaine temps-fréquence et on obtient :

$$X = P + B \tag{5.5}$$

Si l'on veut utiliser un seuillage dur, il faut effectuer l'opération suivante :

$$X = \begin{cases} X & |X| > \lambda \\ 0 & |X| \leq \lambda \end{cases} \tag{5.6}$$

Pour le seuillage doux, il faut appliquer :

$$X = \begin{cases} signe(X)(|X| - \lambda) & |X| \geq \lambda \\ 0 & |X| < \lambda \end{cases} \tag{5.7}$$

λ représente, dans les deux cas, la valeur du seuil utilisé.

Le seuil λ utilisé dans les algorithmes de débruitage par ondelettes peut différer d'une méthode à l'autre. Le premier seuil a avoir été proposé est le seuil de Donoho [65], du nom de son créateur, souvent appelé seuil universel. Il est défini par :

$$\lambda = \sigma\sqrt{2\log(N)} \tag{5. 8}$$

Avec :

- $\sigma = \frac{MAD}{0.6745}$ qui représente l'estimation du niveau de bruit (*MAD* étant la « Median Absolute Deviation »)
- N : la longueur du signal

Dans le cas de l'utilisation des paquets d'ondelettes, le seuil de Donoho devient :

$$\lambda = \sigma\sqrt{2\log(N\log_2(N))} \tag{5. 9}$$

Ces deux expressions ont été trouvées lors de recherches utilisant le bruit blanc Gaussien comme bruit additif. Malheureusement, ce type de bruit n'est pas celui que l'on retrouve dans les situations réelles et de ce fait les systèmes de débruitage conçus pour le réduire ou l'éliminer présentent plusieurs problèmes lorsqu'ils sont utilisés avec d'autres types de bruit.

Pour palier ce problème, Johnstone et Silverman [66] ont mis au point un autre seuil, efficace avec les bruits corrélés :

$$\lambda_j = \sigma_j\sqrt{2\log(N)} \tag{5. 10}$$

Avec :

- $\sigma_j = \frac{MAD_j}{0.6745}$ qui représente l'estimation du niveau de bruit à l'échelle j
- N : la longueur du signal considéré

Lorsque les paquets d'ondelettes sont utilisés, le niveau de bruit est différent à chaque niveau et à chaque sous-bande de fréquences. Ainsi Chang, Kwon, Yang et Kim [64] ont eu l'idée d'appliquer ce seuil à chacun des nœuds. La formule devient donc :

$$\boxed{\lambda_{j,k} = \sigma_{j,k}\sqrt{2\log(N)}} \qquad (5.11)$$

Avec :

- $\sigma_{j,k} = \dfrac{MAD_{j,k}}{0.6745}$ qui représente cette fois-ci l'estimation du niveau de bruit à l'échelle j et dans la sous-bande de fréquences k
- N : le nombre de coefficients d'ondelettes du nœud j, k

Finalement, ce seuil correspond au seuil universel mais appliqué à chaque échelle, ou à chaque nœud s'il s'agit des paquets d'ondelettes.

C'est cette expression (5.11) qui correspond au « Paramètre du Seuil » de Johnstone et Silverman, PSJS. Cette formule a été initialement créée pour être utilisée comme seuil. Toutefois, dans l'algorithme du DAV INNES, elle est employée comme critère de décision, c'est-à-dire qu'il s'agit d'une caractéristique qui va être comparée à un seuil afin de déterminer la présence ou l'absence d'activité vocale. C'est pourquoi on lui a donné le nom de « Paramètre du Seuil » de Johnstone et Silverman.

5.1.3 Énergie

La seconde caractéristique extraite par le DAV INNES, basé sur les ondelettes, est l'énergie à chaque nœud terminal de l'arbre.

Par définition, l'énergie d'un signal *x(n)* est obtenue par :

$$E[x(n)] = \sum_{k=0}^{k=N-1} |x(k)|^2 \qquad (5.12)$$

Avec : N : longueur du signal *x(n)*

5.1.4 Moyenne et variance

Une fois les caractéristiques extraites, le DAV INNES utilise des règles afin de prendre une décision quant à l'activité vocale de la trame étudiée. Certaines d'entre elles s'appliquent aux moyennes et variances des caractéristiques. Ainsi, il est nécessaire de présenter ces deux notions.

La moyenne d'un signal *x(n)* est définie par :

$$\mu[x(n)] = \frac{\sum_{k=0}^{N-1} x(k)}{N} \tag{5. 13}$$

Avec : N : longueur du signal *x(n)*

La variance d'un signal *x(n)* est, quant à elle, définie par :

$$\sigma^2[x(n)] = \frac{\sum_{k=0}^{N-1} |x(k) - \mu|^2}{N} = \mu\left[x^2(n)\right] - \mu[x(n)]^2 \tag{5. 14}$$

Avec : N : longueur du signal *x(n)*

5.2 Description détaillée de l'algorithme du détecteur d'activité vocale basé sur les ondelettes

Cette partie va permettre de décrire en détails l'algorithme du DAV basé sur les ondelettes qui a été développé dans le cadre de ce projet. Toutefois, les raisons qui ont poussé à développer un tel système seront exposées dans la section suivante 5.3.

Toutes les 10ms, le DAV INNES extrait une trame du signal à analyser. La fréquence d'échantillonnage étant de 8kHz, cela correspond à une trame de 80 échantillons. Afin de prendre une décision quant à son état, c'est-à-dire indiquer la présence ou l'absence de parole, le DAV procède au traitement et à l'analyse présentés par la figure 44.

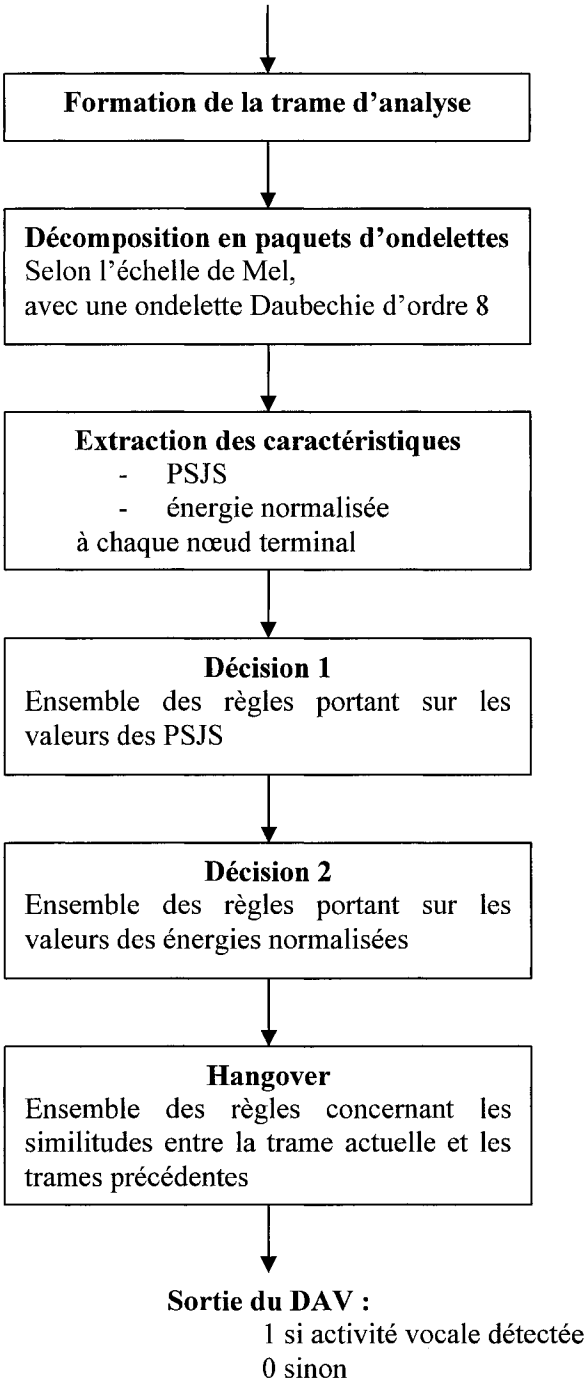


Figure 44 Schéma Bloc de l'algorithme du DAV INNES

ÉTAPE 1 : Formation de la trame d'analyse

Comme il a été mentionné précédemment, le DAV fournit une décision tous les 80 échantillons. Toutefois, l'analyse permettant d'obtenir cette décision s'effectue sur une trame beaucoup plus large : la trame d'analyse. Elle est composée de 240 échantillons provenant des trames :

- future (80 échantillons)
- actuelle (80 échantillons)
- précédente (80 échantillons)

La figure 45 permet de montrer la différence entre la trame actuelle et celle d'analyse :

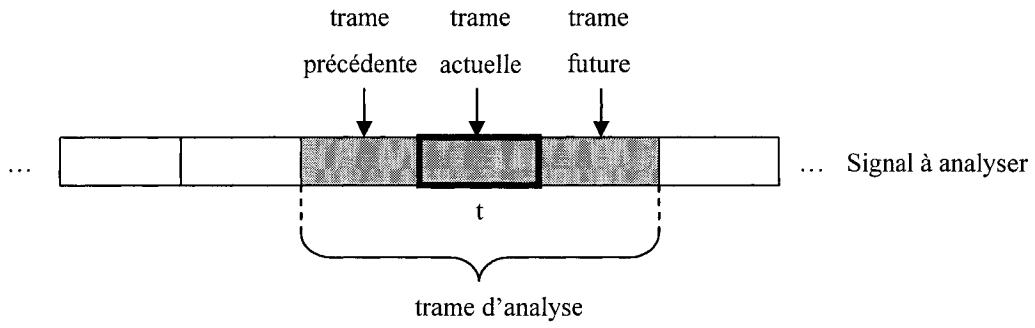


Figure 45 Trame actuelle et trame d'analyse à l'instant t

La trame d'analyse est extraite à l'aide d'une fenêtre de Hamming :

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{M}\right) & 0 \le n \le M-1 \\ 0 & \text{sinon} \end{cases} \qquad (5.15)$$

M étant la longueur de la fenêtre, elle est égale à 240.

Comme il a été mentionné dans le chapitre 3, section 3.1, l'utilisation d'une fenêtre rectangulaire entraîne l'apparition du phénomène de Gibbs. Pour éviter cela, il suffit d'utiliser une fenêtre plus adoucie telle que la fenêtre de Hamming (Oppenheim et Schafer [7]). Ce type de fenêtre a malheureusement pour effet de diminuer fortement l'amplitude au niveau des extrémités. Il est donc nécessaire de choisir une longueur

suffisamment grande pour que l'information importante reste la plus intacte possible. Finalement, plus la fenêtre adoucie est large et plus les données sont préservées, mais ceci entraîne aussi l'extraction des informations voisines ainsi que des retards de plus en plus importants.

Dans notre cas, le fait d'utiliser une fenêtre trois fois plus grande que la trame étudiée, soit 240 échantillons, semble être un bon compromis : l'information contenue dans la trame actuelle est bien conservée et le retard engendré par cette opération n'est que de 10ms, ce qui est tout à fait acceptable. Comme le montre la figure 46, ceci provoque alors un chevauchement des fenêtres :

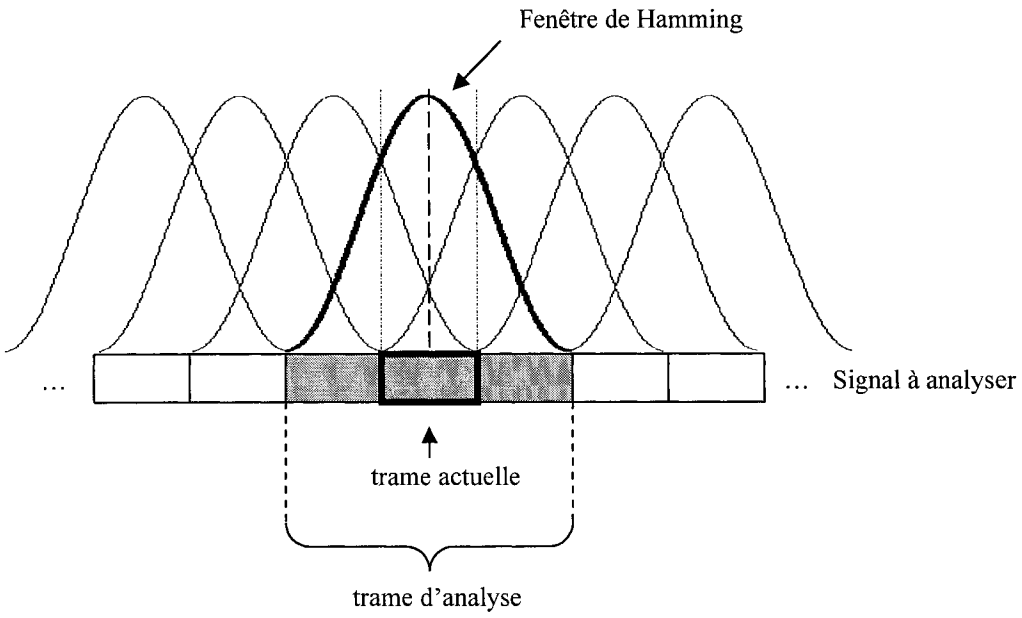


Figure 46 Chevauchement des fenêtres de Hamming

ÉTAPE 2 : Décomposition en paquets d'ondelettes

Une fois la trame d'analyse extraite, cette dernière est décomposée en paquets d'ondelettes à l'aide d'un arbre de décomposition approximant l'échelle de Mel. Comme

il a été vu dans la section 5.1.1, cet arbre de décomposition peut être représenté par la figure 47 :

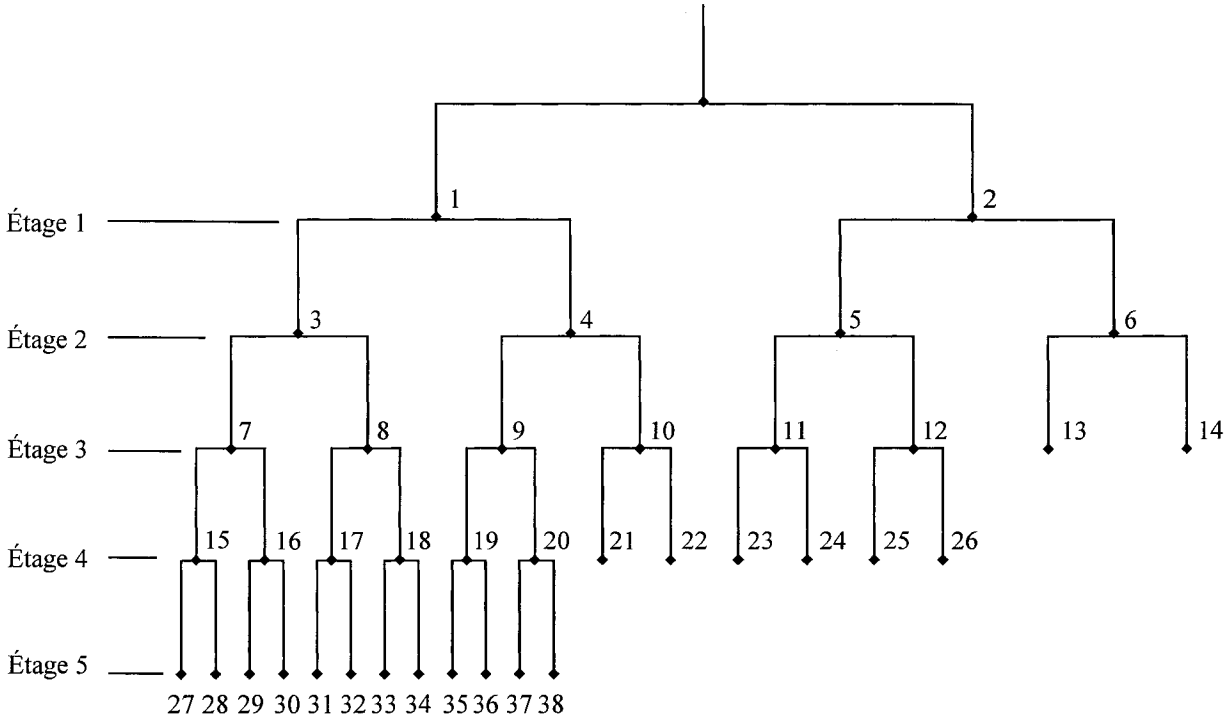


Figure 47 Décomposition en paquets d'ondelettes selon l'échelle de Mel

L'ondelette utilisée pour cette décomposition est l'ondelette Daubechie d'ordre 8. La raison de ce choix sera exposée plus loin, à la section 5.3.3.

Il est à noter que seuls les coefficients des 20 nœuds terminaux (ou externes), c'est-à-dire les nœuds : 13, 14, 21 à 26 et 27 à 38, seront utilisés dans la suite de l'algorithme. En effet, dans la mesure où l'échelle de Mel est utilisée, il est nécessaire de les analyser, sinon celle-ci n'a plus d'intérêt. De plus, ces nœuds contiennent toute l'information située en amont, simplement représentée d'une manière différente. Il est donc inutile d'analyser les nœuds internes de l'arbre.

ÉTAPE 3 : Extraction des caractéristiques

Les étapes 1 et 2 constituent la partie traitement de notre algorithme. Elles nous permettent de passer dans le domaine temps-fréquence, domaine dans lequel nous allons pouvoir effectuer notre analyse pour déterminer la présence ou l'absence de parole. Comme tout détecteur d'activité vocale, le DAV INNES a besoin d'examiner certaines caractéristiques afin de pouvoir prendre une décision quant à l'état de la trame étudiée. L'étape 3 consiste donc à calculer les valeurs du PSJS et de l'énergie pour chacun des nœuds terminaux (voir les sections 5.1.2 et 5.1.3 pour leur expression analytique respective).

Il est important de noter que l'énergie déterminée aux nœuds terminaux est normalisée par l'énergie de la trame. Ceci permet de prévenir les changements de niveaux de parole et de bruit qui ne modifient pas le rapport signal à bruit. Dans la suite du chapitre, lorsque nous utilisons le terme « énergie », cela sous-entend l'énergie normalisée.

Les raisons de l'utilisation de ces deux grandeurs (le Paramètre du seuil de Johnstone et Silverman et l'énergie) comme caractéristiques seront exposées dans la section 5.3.2.

Prise de décision :

Les modules Décision 1 et Décision 2 vont permettre d'obtenir une décision préliminaire quant à l'état de la trame. Cette dernière sera ensuite révisée par le module *Hangover* et la décision finale sera obtenue.

La décision préliminaire est unitaire, c'est-à-dire indiquant la présence d'activité vocale, si au moins une des règes du module Décision 1 ou Décision 2 est validée. L'ordre des règles n'a donc aucune importance. Ces deux modules pourraient être réunis en une seule et même étape puisque toutes les règles sont indépendantes les unes des autres. Toutefois, pour faciliter la compréhension, toutes les règles concernant le PSJS ont été regroupées dans le module Décision 1 et toutes celles concernant l'énergie normalisée dans le module Décision 2.

ÉTAPE 4 : Décision 1

À la sortie de l'étape 3, la valeur du PSJS de chacun des 20 nœuds terminaux est connue. Le module de Décision 1 est constitué de toutes les règles appliquées à ces valeurs et peut être divisé en deux parties.

Un premier ensemble de règles concerne la variance et la moyenne des PSJS à différents étages, illustrés par la figure 48 :

- À l'étage 3 : sur les nœuds 13 et 14
- À l'étage 4 : sur les noeuds 21 à 26
- À l'étage 5 partie droite : sur les nœuds 33 à 38
- À l'étage 5 partie gauche : sur les nœuds 27 à 32
- Aux étages 3, 4, 5 partie droite réunis : nœuds 13 et14, 21 à 26, 33 à 38

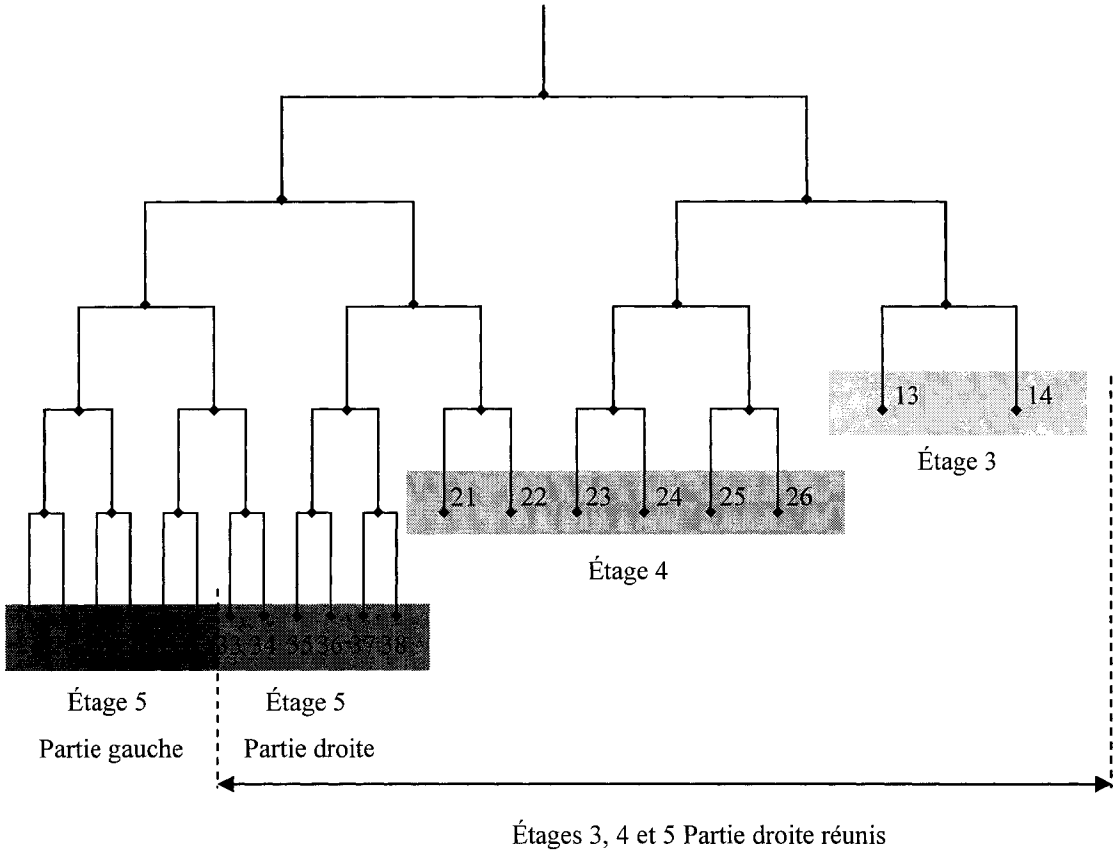


Figure 48 Les 5 étages pour lesquels la variance et la moyenne sont utilisées

Comme il a été vu dans la définition des paquets d'ondelettes (chapitre 4, section 4.1.5), la partie la plus à gauche de l'arbre de décomposition, figure 48, représente l'approximation du signal. L'information est donc plus difficile à exploiter que dans la partie complètement à droite. C'est pourquoi l'étage 5 a été décomposé en deux parties afin de manipuler la partie très basses fréquences avec plus de précaution. Le fait de partager l'étage 5 en deux parties égales (six noeuds pour chacune d'elles) a été choisi arbitrairement.

Les premières règles du module Décision 1 sont énoncées de la façon suivante :

$$
\begin{cases}
\text{Si } \textit{variance des PSJS à l'étage 3} > a_1 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\text{Si } \textit{moyenne des PSJS à l'étage 3} > a_2 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\\
\text{Si } \textit{variance des PSJS à l'étage 4} > a_3 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\text{Si } \textit{moyenne des PSJS à l'étage 4} > a_4 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\\
\text{Si } \textit{variance des PSJS à l'étage 5 droite} > a_5 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\text{Si } \textit{moyenne des PSJS à l'étage 5 droite} > a_6 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\\
\text{Si } \textit{variance des PSJS à l'étage 5 gauche} > a_7 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\text{Si } \textit{moyenne des PSJS à l'étage 5 gauche} > a_8 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\\
\text{Si } \textit{variance des PSJS étages 3, 4 et 5 droite} > a_9 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\text{Si } \textit{moyenne des PSJS étages 3, 4 et 5 droite} > a_{10} & \text{alors } \textit{décision} = \textit{PAROLE}
\end{cases}
\qquad (5.16)
$$

La manière de déterminer les seuils a_k utilisés par ces règles, avec $k = 1$ à 10, sera présentée en détails dans la partie « Entraînement du DAV INNES » (section 5.4).

Le second ensemble de règles du module Décision 1 concerne la valeur du PSJS en chacun des nœuds terminaux de l'arbre. Ainsi, il y a 20 règles du type :

$$\text{Si } PSJS(i) > \alpha_i \text{ alors } décision = PAROLE \qquad (5.17)$$

Avec :

- $PSJS(i)$: la valeur du PSJS au nœud i
- α_i : le seuil utilisé au nœud i
- i : le numéro du nœud, i = 13, 14, 21 à 38

La manière de déterminer les seuils α_i utilisés par ce deuxième ensemble de règles sera, elle aussi, présentée en détails dans la partie « Entraînement du DAV INNES » (section 5.4).

Les règles utilisées dans ce module Décision 1 sont, à notre connaissance, complètement originales. Elles ont été établies par expérimentations. Comme pour le G729.B, c'est l'approche de la reconnaissance des formes qui a été retenue ici. Les raisons qui nous ont poussées à les choisir seront exposées dans la partie 5.3.2.

ÉTAPE 5 : Décision 2

Le module de décision 2 fonctionne exactement de la même manière que le module précédent, à l'exception du fait que les règles s'appliquent non pas aux PSJS mais aux valeurs des énergies, normalisées par l'énergie de la trame actuelle.

Ainsi, les premières règles du module Décision 2 sont :

$$
\begin{cases}
\text{Si } \textit{variance des énergies à l'étage 3} > b_1 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\text{Si } \textit{moyenne des énergies à l'étage 3} > b_2 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\text{Si } \textit{variance des énergies à l'étage 4} > b_3 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\text{Si } \textit{moyenne des énergies à l'étage 4} > b_4 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\text{Si } \textit{variance des énergies à l'étage 5 droite} > b_5 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\text{Si } \textit{moyenne des énergies à l'étage 5 droite} > b_6 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\text{Si } \textit{variance des énergies à l'étage 5 gauche} > b_7 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\text{Si } \textit{moyenne des énergies à l'étage 5 gauche} > b_8 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\text{Si } \textit{variance des énergies étages 3, 4 et 5 droite} > b_9 & \text{alors } \textit{décision} = \textit{PAROLE} \\
\text{Si } \textit{moyenne des énergies étages 3, 4 et 5 droite} > b_{10} & \text{alors } \textit{décision} = \textit{PAROLE}
\end{cases}
\tag{5. 18}
$$

La manière de déterminer les seuils b_k utilisés par ces règles, avec $k = 1$ à 10, sera présentée en détails dans la partie « Entraînement du DAV INNES » (section 5.4).

Le second ensemble de règles du module Décision 2 concerne la valeur de l'énergie normalisée en chacun des nœuds terminaux de l'arbre. Ainsi, il y a 20 règles du type :

$$\text{Si } \textit{énergie(i)} > \beta_i \text{ alors } \textit{décision} = \textit{PAROLE} \tag{5. 19}$$

Avec :

- *énergie(i)* : la valeur de l'énergie au nœud i normalisée par l'énergie de la trame
- β_i : le seuil utilisé au nœud i
- i : le numéro du nœud, i = 13, 14, 21 à 38

La manière de déterminer les seuils β_i utilisés par ce deuxième ensemble de règles sera, elle aussi, présentée en détails dans la partie « Entraînement du DAV INNES » (section 5.4).

Tout comme pour le module Décision 1, les règles utilisées dans ce module Décision 2 sont originales. Elles ont été établies par expérimentations grâce à l'approche de la reconnaissance des formes. Les raisons qui nous ont poussées à les choisir seront exposées dans la partie 5.3.2.

ÉTAPE 6 : Hangover

Ce module est similaire au *Hangover* utilisé par le G729.B, présenté dans le chapitre 3, section 3.1. Il va réviser la décision préliminaire en fonction des caractéristiques des trames précédentes. On rappelle qu'il vise à empêcher les coupures dans les bouffées de parole mais aussi les déclenchements intempestifs dans les longues périodes d'inactivité vocale. Le *Hangover* utilisé ici est composé de six règles.

On note par la suite E l'énergie de la trame actuelle et E_{prec} celle de la trame précédente.

Les trois premières règles permettent d'améliorer la détection des trames de parole :

- **Règle 1 :** Si la décision préliminaire indique BRUIT mais que la trame précédente contenait de la parole et que l'énergie de la trame actuelle est du même ordre de grandeur que celle de la trame précédente, il faut corriger la décision :

$$\begin{aligned} &Si\ (décision\ préliminaire\ =\ BRUIT \\ &\&\ dernier\_vad\_flag\ =\ PAROLE \\ &\&\ \left|E\ -\ E_{prec}\right|\ \leq\ coefficient) \\ &Alors\ vad\_flag\ =\ PAROLE \end{aligned} \tag{5.20}$$

- **Règle 2 :** Si la décision préliminaire indique BRUIT mais que les deux dernières trames contenaient de la parole et que l'énergie de la trame actuelle est du même ordre de grandeur que celle de la trame précédente, il faut corriger la décision :

$$\begin{aligned} &Si\ (décision\ préliminaire\ =\ BRUIT \\ &\&\ dernier\_vad\_flag\ =\ PAROLE \\ &\&\ avant\_dernier\_vad\_flag\ =\ PAROLE \\ &\&\ \left|E\ -\ E_{prec}\right|\ \leq\ coefficient) \\ &Alors\ vad\_flag\ =\ PAROLE \end{aligned} \tag{5.21}$$

- **Règle 3 :** Si la décision préliminaire indique BRUIT mais que les trois dernières trames contenaient de la parole et que l'énergie de la trame actuelle est du même ordre de grandeur que celle de la trame précédente, il faut corriger la décision :

$$\begin{aligned} &Si\ (décision\ préliminaire\ =\ BRUIT \\ &\&\ dernier\_vad\_flag\ =\ PAROLE \\ &\&\ avant\_dernier\_vad\_flag\ =\ PAROLE \\ &\&\ avant\_avant\_dernier\_vad\_flag\ =\ PAROLE \\ &\&\ \left|E\ -\ E_{prec}\right|\ \leq\ coefficient) \\ &Alors\ vad\_flag\ =\ PAROLE \end{aligned} \tag{5.22}$$

La recherche a montré qu'au-delà des 3 trames précédentes, ce type de règles n'a plus d'utilité.

La quatrième règle évite les faux déclenchements dus à des changements brusques du bruit dans une période sans parole :

- **Règle 4 :** Si la décision est PAROLE alors que les N_1 trames d'avant étaient BRUIT et que l'énergie de la trame est plus faible que celle de la trame précédente, alors la décision doit être corrigée.

$$\begin{aligned} &Si\ (vad\_flag = PAROLE \\ &\&\ compteur\_bruit > N_1 \\ &\&\ \left|E - E_{prec}\right| \le coefficient\ ) \\ &Alors\ vad\_flag = BRUIT \end{aligned} \tag{5. 23}$$

Enfin, les deux dernières règles du *Hangover* préviennent aussi bien des coupures dans la parole que des déclenchements intempestifs :

- **Règle 5 :** Si la trame actuelle n'a pas le même état que la précédente et que le rapport de leur énergie est inférieur à un certain seuil alors il faut modifier la décision pour que la trame actuelle ait le même état que la trame précédente.

$$\begin{aligned} &Si\ (vad\_flag \neq dernier\_vad\_flag \\ &\&\ \frac{E}{E_{prec}} < coefficient) \\ &Alors\ vad\_flag = dernier\_vad\_flag \end{aligned} \tag{5. 24}$$

- **Règle 6 :** Il s'agit de la règle de lissage que nous avons mis au point pour la version du G729.B ajustée aux bruits industriels, voir chapitre 3, section 3.2.2. Nous rappelons son énoncé ci-dessous :

$$\begin{cases} Si\ (vad\_flag = BRUIT \\ \&\ dernier\_vad\_flag = PAROLE \\ \&\ avant\_dernier\_vad\_flag = BRUIT) \\ Alors\ dernier\_vad\_flag = BRUIT \\ \\ Si\ (vad\_flag = PAROLE \\ \&\ dernier\_vad\_flag = BRUIT \\ \&\ avant\_dernier\_vad\_flag = PAROLE) \\ Alors\ dernier\_vad\_flag = PAROLE \end{cases} \quad (5.25)$$

Elle permet de procéder au lissage du type :

...11000000010000000111... → ...11000000000000000111...

...00111111101111111000... → ...00111111111111111000...

Les six règles de ce module de *Hangover* ont été obtenues par expérimentations. La manière de déterminer les coefficients utilisés dans leurs conditions sera présentée en détails dans la partie « Entraînement du DAV INNES » (section 5.4). Il est à noter que les quatre premières règles ont été inspirées de celles proposées par Alcatel [39] pour le G729.B, voir chapitre 3 section 3.1. Cependant, elles diffèrent toutes d'une manière ou d'une autre de celles de Alcatel. Enfin, les deux dernières règles sont, à notre connaissance, complètement originales.

Sortie du DAV :

Grâce au module de *Hangover*, la décision finale pour la trame actuelle est obtenue. Si elle indique la présence d'activité vocale, la sortie du DAV est 1. Dans le cas contraire, la sortie est 0.

5.3 Raisons d'un tel algorithme

Comme il a été mentionné dans la section 5.1.1, la décomposition en paquets d'ondelettes selon l'échelle de Mel a été retenue afin d'intégrer un modèle perceptif de l'oreille humaine dans l'algorithme de détection d'activité vocale. Cette partie va maintenant expliquée comment la recherche a permis d'aboutir à notre algorithme. La base d'expérimentation sur laquelle a été menée cette recherche va tout d'abord être présentée. Les raisons qui nous ont poussé à utiliser le PSJS et l'énergie comme caractéristiques seront ensuite exposées. Enfin, nous verrons pourquoi l'ondelette Daubechie d'ordre 8 a été choisie.

5.3.1 Base d'expérimentation

Afin de procéder à des tests permettant la conception du DAV INNES, reposant sur la théorie des ondelettes, une base d'expérimentation a été mise au point.

Elle est similaire à celle utilisée pour l'adaptation du G729.B aux milieux industriels, présentée dans le chapitre 3, section 3.2.1. Il y a donc 32 phrases issues de la base de données DARPA TIMIT et un bruit, le « Factory Noise 1 » de la base de données NOISEX.

Nous avons mentionné dans le chapitre 3 qu'une amélioration possible du G729.B modifié serait d'effectuer un ajustement sur une base de données comportant plusieurs rapports signal à bruit. Nous avons pris cela en compte lors de la mise au point du DAV INNES et nous avons choisi d'effectuer nos recherches sur une base d'expérimentation contenant trois RSB différents : 5dB, 10dB et 15dB, ceci dans le but de couvrir au mieux la plage de RSB visée : [5dB – 15dB] et d'obtenir ainsi un unique ensemble de coefficients pour les règles de décision.

Le bruit a donc été additionné à la totalité des phrases avec trois amplitudes différentes afin d'obtenir trois ensembles de données:

- 32 phrases bruitées, RSB = 5dB
- 32 phrases bruitées, RSB = 10dB
- 32 phrases bruitées, RSB = 15dB

Pour chacun de ces ensembles, nous avons ensuite exécuté les deux premières étapes de l'algorithme, à savoir la formation de la trame d'analyse et la décomposition en paquets d'ondelettes selon l'échelle de Mel. Nous avons ainsi obtenu les trames d'analyse de toutes les phrases, et ceci pour les trois RSB. Pour chacune d'entre elles, nous avons stocké l'énergie de la trame actuelle et les coefficients d'ondelettes aux nœuds terminaux de l'arbre.

Le problème abordé ici est la distinction entre la parole bruitée et le bruit sans parole. Connaissant l'activité vocale de chacune des 32 phrases, nous avons pu séparé les trames en deux ensembles : parole bruitée et bruit seul.

Finalement, la base d'expérimentation est constituée de deux grands groupes : les trames avec activité vocale, appelées trames de parole, et les trames qui en sont dépourvues, appelées trames de bruit. À chacune d'elles sont associés les coefficients d'ondelettes aux nœuds terminaux ainsi que leur énergie. La figure 49 illustre la formation de cette base d'expérimentation.

Remarque : tous les résultats présentés dans cette section ont été obtenus à partir de cette base d'expérimentation.

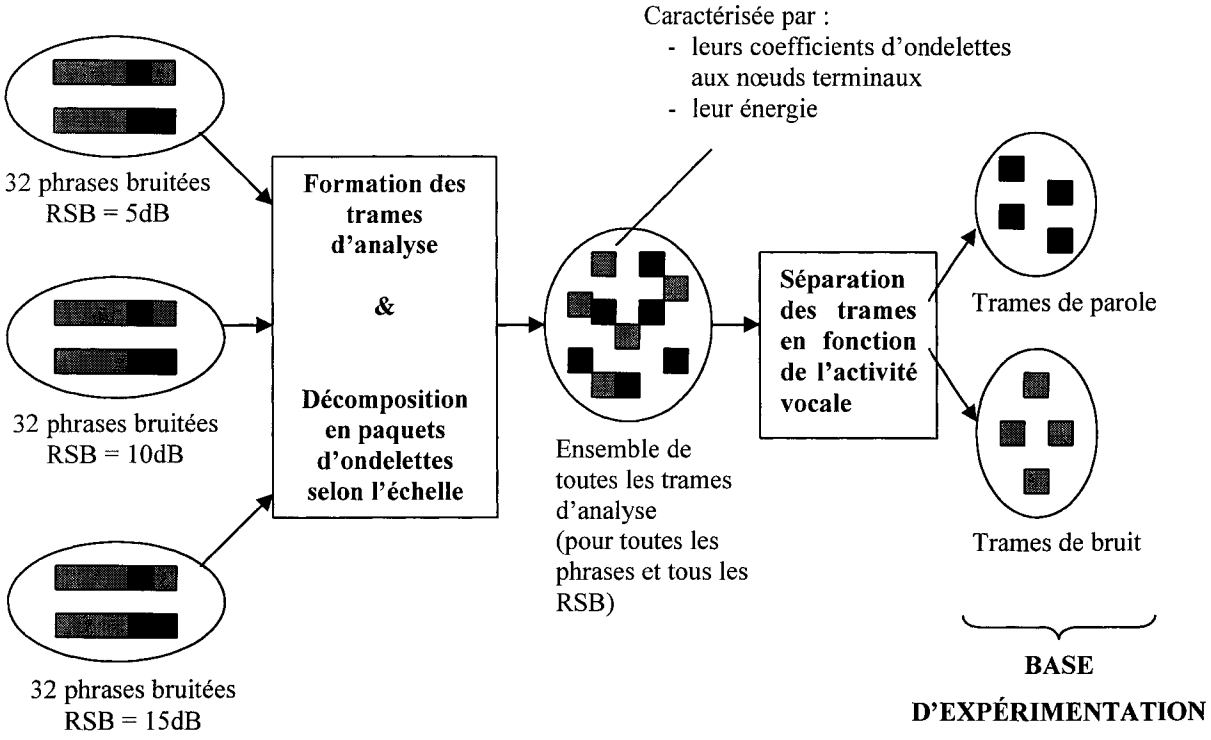


Figure 49 Formation de la base d'expérimentation

5.3.2 Pourquoi utiliser le PSJS et l'énergie comme caractéristiques?

Comme il a été mentionné dans le chapitre 2, l'énergie a été l'une des toutes premières caractéristiques utilisées pour la détection d'activité vocale. Son utilisation seule ne permet pas d'obtenir de bonnes performances en présence de bruit fort. Cependant, combinée avec d'autres caractéristiques et appliquée après certains traitements, elle peut être d'une très grande utilité. C'est pourquoi on la retrouve dans la plupart des DAV.

Ainsi, lors de la recherche de l'algorithme du DAV INNES, nous avons commencé par examiner les valeurs des énergies normalisées à chaque nœud terminal. Nous avons donc visualisé et comparé les distributions des énergies des trames de parole et des trames de

bruit, à chacun des 20 nœuds terminaux. Sur la figure 50, qui est un exemple des résultats obtenus, nous avons représenté les valeurs des énergies normalisées en fonction d'elles-mêmes pour justement pouvoir observer ces distributions. En haut, il s'agit des trames de parole et en bas celles de bruit. Il est à noter que toutes les trames de la base d'expérimentation (trames de parole et de bruit pour les RSB 5, 10 et 15dB) sont représentées sur cette figure.

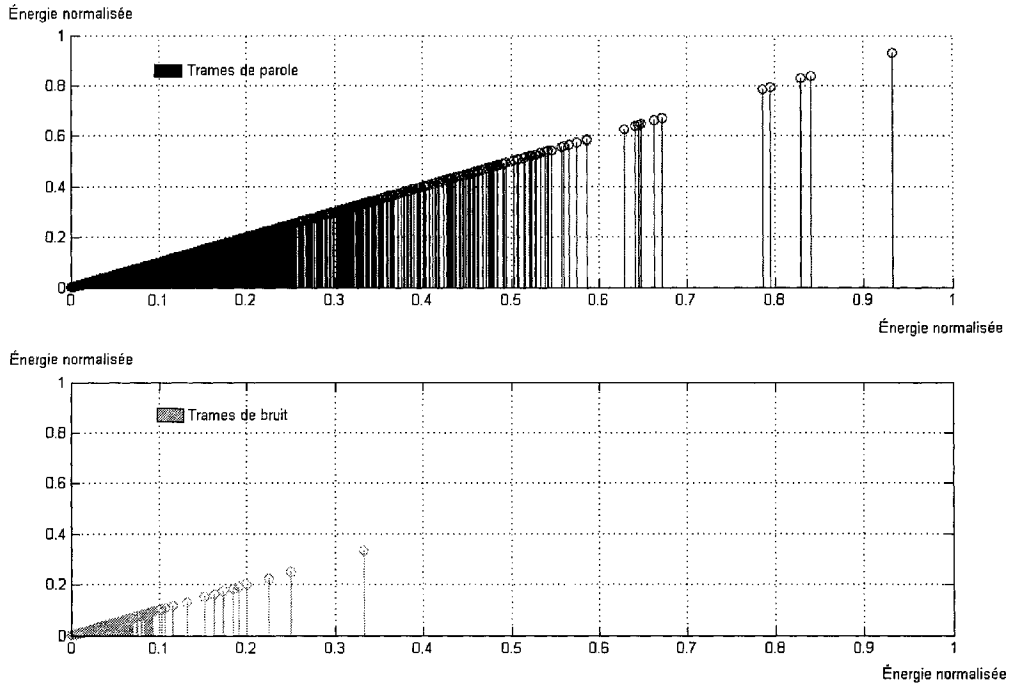


Figure 50 Observation de l'énergie normalisée au nœud terminal 26

Remarque : le nœud 26 a été choisi arbitrairement pour illustrer toute notre démarche. Toutefois, d'autres nœuds auraient pu être retenus.

Comme nous pouvons le voir sur cette figure 50, les trames de parole s'étendent sur l'intervalle d'énergie [0 - 0,93], alors que celles du bruit couvrent une plage d'énergie beaucoup plus étroite : [0 - 0,33]. Il paraît donc possible de déterminer un seuil délimitant la parole du bruit et d'utiliser une règle du type :

Si l'énergie normalisée au nœud 26 > seuil alors la trame contient de la parole

Après visualisation des distributions des énergies à chaque nœud terminal de notre arbre de décomposition, nous avons constaté que ce type de figures est obtenu à chaque fois. Nous avons effectué la même étude pour les variances et les moyennes des énergies normalisées aux différents étages de l'arbre de décomposition (voir figure 48) et nous avons encore obtenu ce type de figures. En plus de cela, nos recherches ont montré que les trames de parole dont les grandeurs sont très élevées (c'est-à-dire celles dont l'énergie est supérieure à 0,33 dans le cas de la figure 50) ne sont pas systématiquement les mêmes d'un nœud à l'autre, ou d'un étage à l'autre. Ainsi, un module de décision utilisant des règles qui portent sur les énergies normalisées de tous les nœuds terminaux permet d'identifier une bonne partie des trames de parole. Ceci explique pourquoi nous avons choisi d'utiliser les règles sur l'énergie, présentées précédemment section 5.2.

Compte tenu de ces observations, le module de décision, basé sur l'énergie normalisée, a été mis au point à l'aide du logiciel MATLAB, puis testé. Les résultats ont montré que la reconnaissance des trames de parole est bien moins bonne pour le RSB de 5dB que pour les RSB de 10dB et 15dB. Ceci s'explique par le fait que le choix du seuil de chaque règle s'effectue en fonction de la base d'expérimentation toute entière, donc en fonction des données pour les trois RSB. Or, de nos observations, nous avons constaté que plus le RSB est faible et plus la distribution des énergies s'étend mais plus celle des énergies normalisées se rétrécit. Illustrons ceci à l'aide du cas du nœud 26 vu à la figure 50. Les énergies normalisées se répartissent de la manière suivante:

- [0 – 0,4] pour la parole, [0 – 0,09] pour le bruit, dans le cas d'un RSB de 5dB
- [0 – 0,585] pour la parole, [0 – 0,18] pour le bruit, dans le cas d'un RSB de 10dB
- [0 – 0,93] pour la parole, [0 – 0,33] pour le bruit, dans le cas d'un RSB de 15dB

La partie basse de la figure 51 montre ces intervalles :

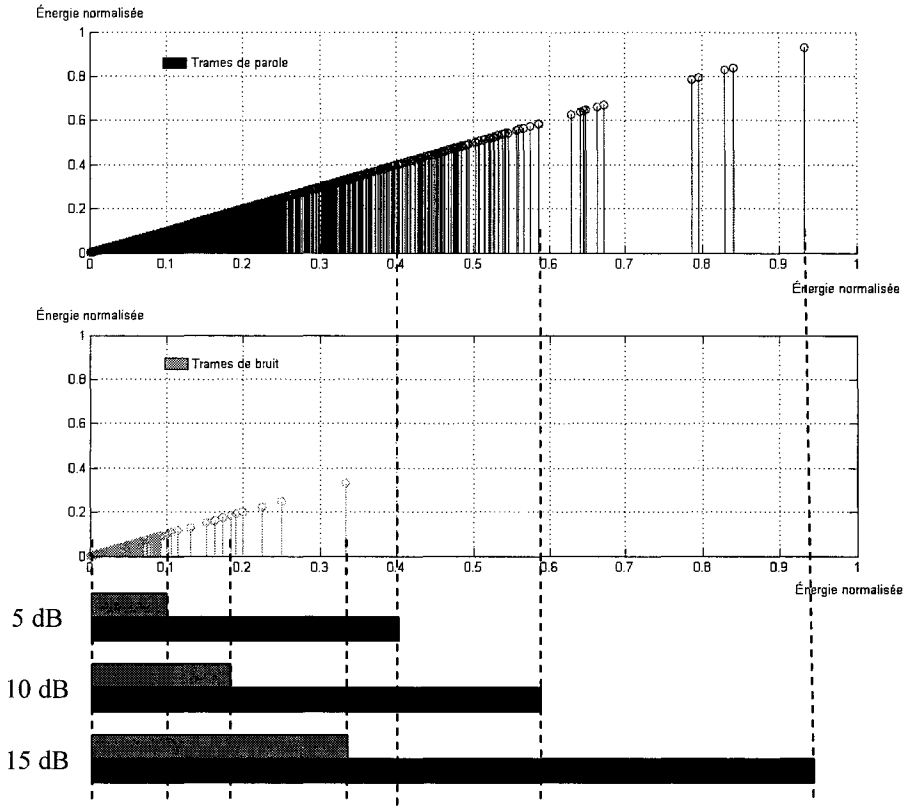


Figure 51 Répartition des énergies normalisées au nœud 26 en fonction du RSB

Attention : les bandes noire et grise indiquent seulement les largeurs des intervalles. La distribution des valeurs à l'intérieur de ceux-ci est similaire à celle représentée sur la figure, c'est à dire très concentrée vers 0 et beaucoup moins dense à l'approche de la borne supérieure.

Choisissons par exemple un seuil autour de 0,1 afin de reconnaître beaucoup de trames de parole sans générer trop d'erreur « bruit pris pour la parole ». Nous détectons alors beaucoup moins de trames de parole pour un RSB de 5dB que pour un de 15dB. Si nous

abaissons le seuil, le taux d'erreur « bruit pris pour de la parole » augmente très rapidement.

Le module de décision sur l'énergie normalisée est donc propice à la détection de la parole jusqu'à une certaine limite de RSB. Pour des RSB plus faibles, il n'est pas possible d'obtenir de performances élevées.

Pour résoudre ce problème, l'idée d'effectuer un débruitage de la parole par ondelettes a été retenue. En effet, ceci permettrait de diminuer le bruit et donc de relever le RSB. Le module de décision sur l'énergie serait alors capable de fournir de meilleurs résultats. Lors des tests de quelques méthodes de débruitage usuelles utilisant le PSJS comme seuil (voir section 5.1.2), une question a été soulevée : pourquoi ne pas utiliser le seuil de Johnstone et Silverman comme caractéristique dans l'algorithme du DAV INNES?

Les valeurs de ce seuil, c'est-à-dire les PSJS, à chaque nœud terminal ont donc été stockées pour toutes les trames de bruit et pour celles de parole qui n'avaient pas été reconnues par le module de décision basé sur l'énergie normalisée. Comme pour l'énergie, nous avons ensuite visualisé et comparé les distributions des PSJS des trames de parole et des trames de bruit, à chacun des 20 nœuds terminaux. Sur la figure 52, qui est un exemple des résultats obtenus, nous avons représenté les valeurs PSJS en fonction d'elles-mêmes pour justement pouvoir observer ces distributions. En haut, il s'agit des trames de parole, en bas celles de bruit.

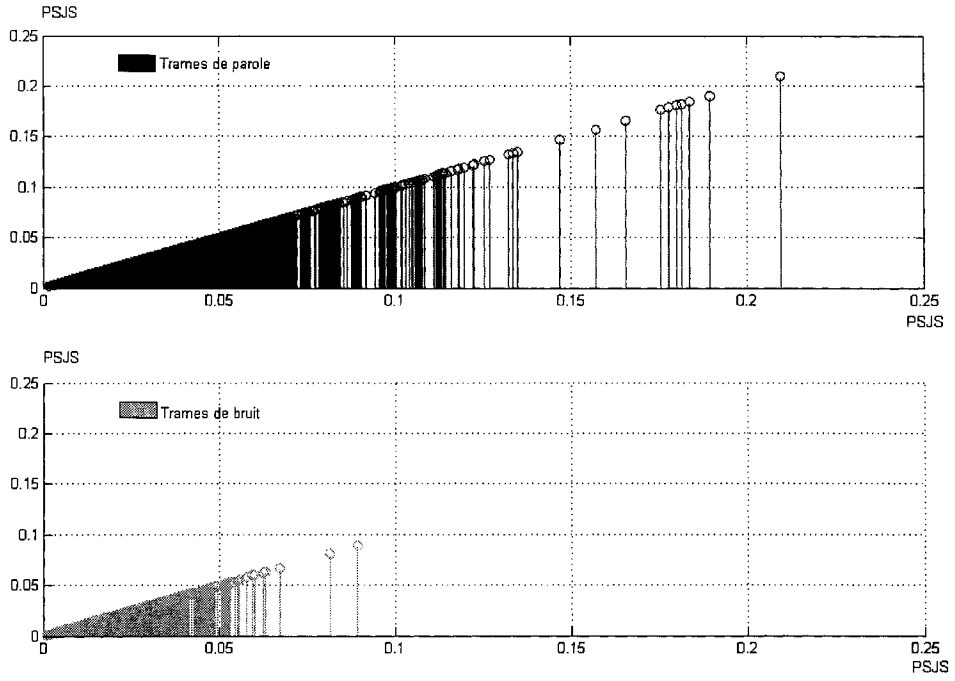


Figure 52 Observation du PSJS au noeud terminal 26

Comme nous pouvons le voir sur cette figure 52, nous avons le même phénomène qu'avec l'énergie normalisée. Les valeurs des PSJS des trames de parole s'étendent sur l'intervalle [0 – 0,21], alors que celles du bruit couvrent la plage [0 – 0,09]. Il paraît donc possible de déterminer un seuil délimitant la parole du bruit et d'utiliser une règle du type :

Si le PSJS au nœud 26 > seuil alors la trame contient de la parole

Comme pour l'énergie normalisée, il apparaît que ce type de figures est obtenu pour chacun des nœuds terminaux ainsi que pour la variance et la moyenne aux différents étages (voir figure 48). Il s'avère aussi que les trames de parole au dessus des seuils ne sont pas systématiquement les mêmes d'un nœud à l'autre, ou d'un étage à l'autre. En plus de cela, les trames de parole dont nous discutons ici sont celles qui n'ont pas été reconnues par le module de décision basée sur l'énergie normalisée.

L'idée du débruitage a donc été remplacée par celle d'utiliser le PSJS comme caractéristique puisqu'un module de décision sur les valeurs du PSJS permet d'identifier une autre partie des trames de parole. Il permet donc de compléter la reconnaissance obtenue avec le module basé sur l'énergie normalisée. Ceci explique pourquoi nous avons choisi d'utiliser les règles sur le PSJS, présentées précédemment section 5.2.

Le module de décision basé sur les valeurs des PSJS aux nœuds terminaux a été mis au point à l'aide du logiciel MATLAB puis évalué. Les résultats montrent un phénomène contraire à celui obtenu avec l'énergie, c'est-à-dire que la reconnaissance des trames de parole est meilleure pour le RSB de 5dB que pour les RSB de 10dB et 15dB. Ceci s'explique de la même manière. Le choix du seuil de chaque règle s'effectue en fonction des données pour les trois RSB. Or, de nos observations nous avons constaté que plus le RSB est faible et plus la distribution des PSJS s'étend, à l'inverse de celle de l'énergie normalisée. Par exemple, dans le cas du nœud 26 vu à la figure 52, les PSJS se répartissent de la manière suivante:

- [0 – 0,21] pour la parole, [0 – 0,09] pour le bruit, dans le cas d'un RSB de 5dB
- [0 – 0,13] pour la parole, [0 – 0,06] pour le bruit, dans le cas d'un RSB de 10dB
- [0 – 0,11] pour la parole, [0 – 0,03] pour le bruit, dans le cas d'un RSB de 15dB

La partie basse de la figure 53 ci-dessous montre ces intervalles :

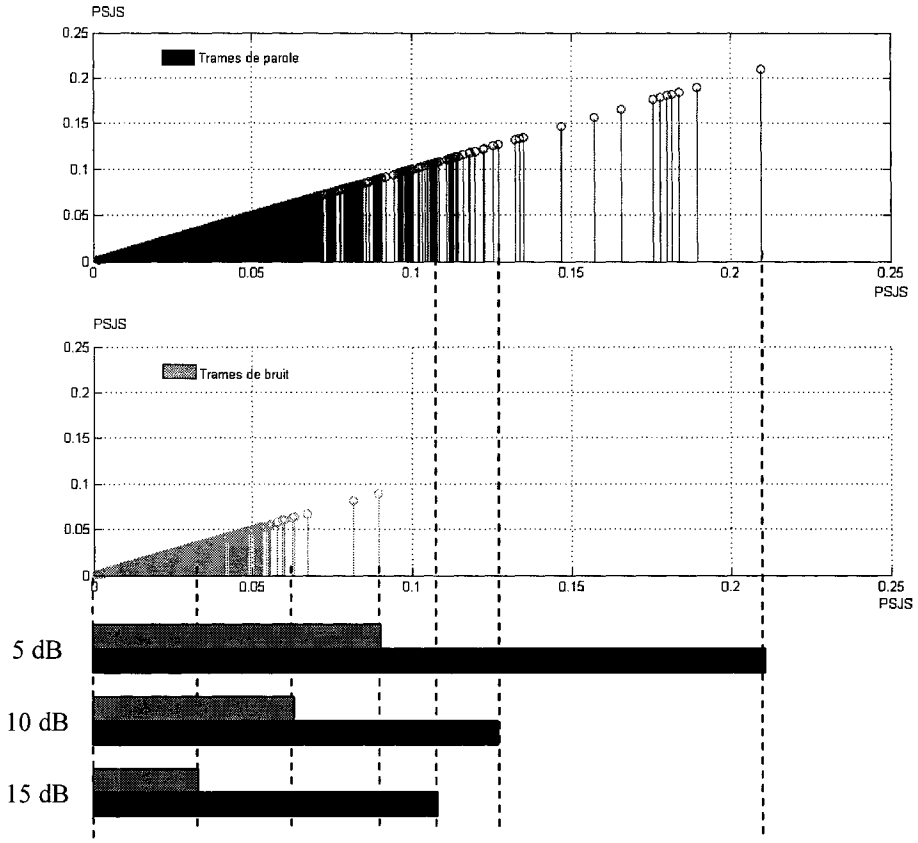


Figure 53 Répartition des PSJS au nœud 26 en fonction du RSB

Attention : les bandes noire et grise indiquent seulement les largeurs des intervalles. La distribution des valeurs à l'intérieur de ceux-ci est similaire à celle représentée sur la figure, c'est à dire très concentrée vers 0 et beaucoup moins dense à l'approche de la borne supérieure.

Ainsi en choisissant un seuil autour de 0,05, pour ne pas engendrer trop d'erreur du type « bruit pris pour de la parole », on détecte beaucoup moins de trames de parole pour 15dB que pour 5dB.

Le module de décision basé sur l'énergie normalisée permet donc une meilleure détection des trames de parole lors d'un RSB de 15dB tandis que celui basé sur le PSJS détecte mieux celles pour 5dB. L'utilisation conjointe de ces modules procure une bonne reconnaissance des trames de parole sur l'ensemble de la plage de RSB visée, soit [5dB - 15dB]. C'est pourquoi nous les avons utilisé dans notre algorithme. Il est à noter que toutes les règles sont, à notre connaissance, originales et que la principale nouveauté du DAV INNES, basé sur la théorie des ondelettes, est l'utilisation du Seuil de Johnstone et Silverman comme critère de décision.

5.3.3 Pourquoi utiliser une ondelette Daubechies d'ordre 8?

L'ondelette la plus couramment utilisée dans le domaine de traitement de la parole est celle de Daubechies. C'est pourquoi elle a été choisie ici. Notre premier système de détection d'activité vocale basé sur les ondelettes et utilisant les deux modules de décision précédents a été développé avec une ondelette Daubechies d'ordre 9. Une question s'est alors posée : est-ce que le fait d'utiliser un ordre différent permettrait d'augmenter les performances? Pour répondre à ceci, les ordres de 6 à 12 ont été testés. Notre DAV INNES a donc été ajusté pour chacun des ordres, c'est-à-dire que les seuils utilisés dans les règles de décision ont été déterminés. Il est à noter que nous avons opté ici pour un ajustement rapide plutôt qu'un ajustement permettant d'obtenir des performances élevées. Ainsi chaque seuil correspond à la valeur maximale des trames de bruit. Par exemple, pour le nœud 26 vu aux figures 50 et 52, le seuil de l'énergie est 0,33 et celui du PSJS est 0,07. Ce type d'ajustement consiste donc à déterminer les seuils de manière à ce que le pourcentage d'erreur « bruit pris pour la parole » soit nul.

Les résultats obtenus dans ces conditions et sur la base d'expérimentation sont présentés dans le tableau VII :

Tableau VII

Reconnaissance de la parole en fonction de l'ordre de l'ondelette Daubechies

Ordre de l'ondelette Daubechies	**Pourcentage de reconnaissance de la parole**
6	59,02%
7	60,76%
8	62,65%
9	62,38%
10	62,62%
11	62,34%
12	61,64%

Ce sont donc les ordres 8 et 10 qui permettent d'obtenir le meilleur pourcentage de reconnaissance de parole. Sachant que plus l'ordre de l'ondelette est élevé, plus la longueur des filtres est grande et plus le temps de calcul est long, nous avons opté pour l'ordre le plus petit. C'est ainsi que l'ondelette Daubechies d'ordre 8 a été retenue.

La description et la justification de l'algorithme du DAV INNES, reposant sur les ondelettes, viennent d'être présentées. Il est maintenant possible d'aborder la partie « Entraînement du DAV INNES ». Nous allons voir comment déterminer les coefficients utilisés par les règles de décision.

5.4 Entraînement du DAV INNES

Notre détecteur d'activité vocale basé sur les ondelettes, comme décrit précédemment, a été développé avec le logiciel MATLAB. Lors de nos recherches, les coefficients des règles de décision ont été déterminés manuellement à l'aide de la base d'expérimentation présentée section 5.3.1, et ceci afin que notre DAV soit efficace sur la plage de RSB visée, à savoir [5dB – 15dB]. Toutefois, pour s'assurer que le DAV INNES proposé dans ce rapport soit performant dans la majorité des milieux industriels, nous avons mis au point une procédure d'ajustement des seuils que nous avons par la suite entièrement automatisée. Ainsi, si un ajustement des coefficients des règles de décision s'avère nécessaire, il suffit de lancer cet algorithme pour les déterminer. Ceci rend notre système très facile à utiliser. Cette phase est appelée l'entraînement et c'est ce que nous allons présenter dans cette partie. Il est à noter qu'une fois entraîné, le DAV INNES, basé sur les ondelettes, peut alors fonctionner dans l'environnement désiré sans aucune modification supplémentaire.

Pour réaliser l'entraînement, il faut tout d'abord enregistrer un échantillon représentatif du bruit rencontré. De même, il est nécessaire de définir la plage de rapport signal à bruit dans laquelle le DAV va fonctionner: [RSB_{min} - RSB_{max}]. La procédure d'ajustement des coefficients se déroule ensuite comme illustrée par la figure 54.

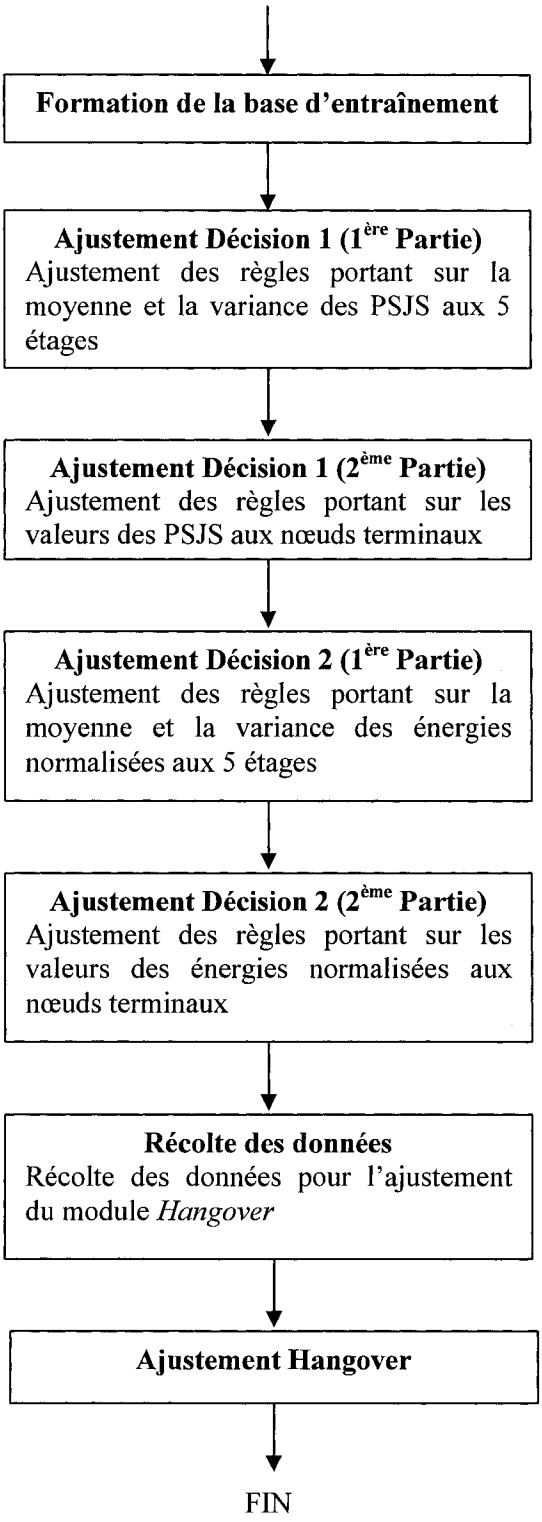


Figure 54 Schéma Bloc de l'ajustement des seuils du DAV INNES

ÉTAPE 1 : Formation de la base d'entraînement

La base de données d'entraînement est très similaire à celle d'expérimentation présentée à la section 5.3.1. Elle est composée des mêmes 32 phrases issues de DARPA TIMIT. Afin d'obtenir la fréquence d'échantillonnage attendue par le DAV INNES, à savoir 8kHz, ces phrases sont rééchantillonnées, et ceci en prenant soin d'utiliser un filtre adéquat pour conserver la bonne qualité de l'information.

Selon les conditions d'enregistrement du bruit spécifié par l'utilisateur, ce dernier est lui aussi rééchantillonné.

Des parties du bruit sont alors choisies aléatoirement et additionnées aux phrases afin d'obtenir 32 signaux de parole bruitée. Le signal de bruit fourni par l'utilisateur est sensé être représentatif du milieu industriel et donc avoir une longueur suffisante pour que les 32 fractions de bruit soient différentes les unes des autres.

Afin d'obtenir un ajustement adapté à la plage de RSB : [RSB_{min} – RSB_{max}], trois ensembles vont être construits :

- 32 phrases bruitées à un rapport signal sur bruit de RSB_{min}
- 32 phrases bruitées à un rapport signal sur bruit de $\frac{RSB_{min} + RSB_{max}}{2}$ (milieu de la plage)
- 32 phrases bruitées à un rapport signal sur bruit de RSB_{max}

Pour chaque phrase et chaque RSB, les coefficients à appliquer à l'amplitude du bruit sont calculés par dichotomie.

Rappel : dans le cadre de notre projet de recherche, la plage de RSB visée est [5dB – 15dB], ainsi les rapports signal à bruit des trois ensembles sont : 5dB, 10dB et 15dB.

Les signaux de parole bruitée permettant l'entraînement sont maintenant disponibles. Pour chacun d'entre eux, on exécute alors les trois premières étapes de l'algorithme du DAV INNES, à savoir la formation de la trame d'analyse, la décomposition en paquets

d'ondelettes selon l'échelle de Mel et l'extraction des caractéristiques. On obtient ainsi les trames d'analyse de toutes les phrases et ceci pour les trois RSB. Chacune d'entre elles est caractérisée par ses valeurs d'énergie normalisée et de PSJS aux 20 nœuds terminaux.

Le problème abordé ici est toujours la distinction entre la parole bruitée et le bruit sans parole. Connaissant l'activité vocale des 32 phrases, on sépare les trames en deux groupes : parole bruitée et bruit seul.

La base d'entraînement est à présent formée. La figure 55 récapitule sa formation.

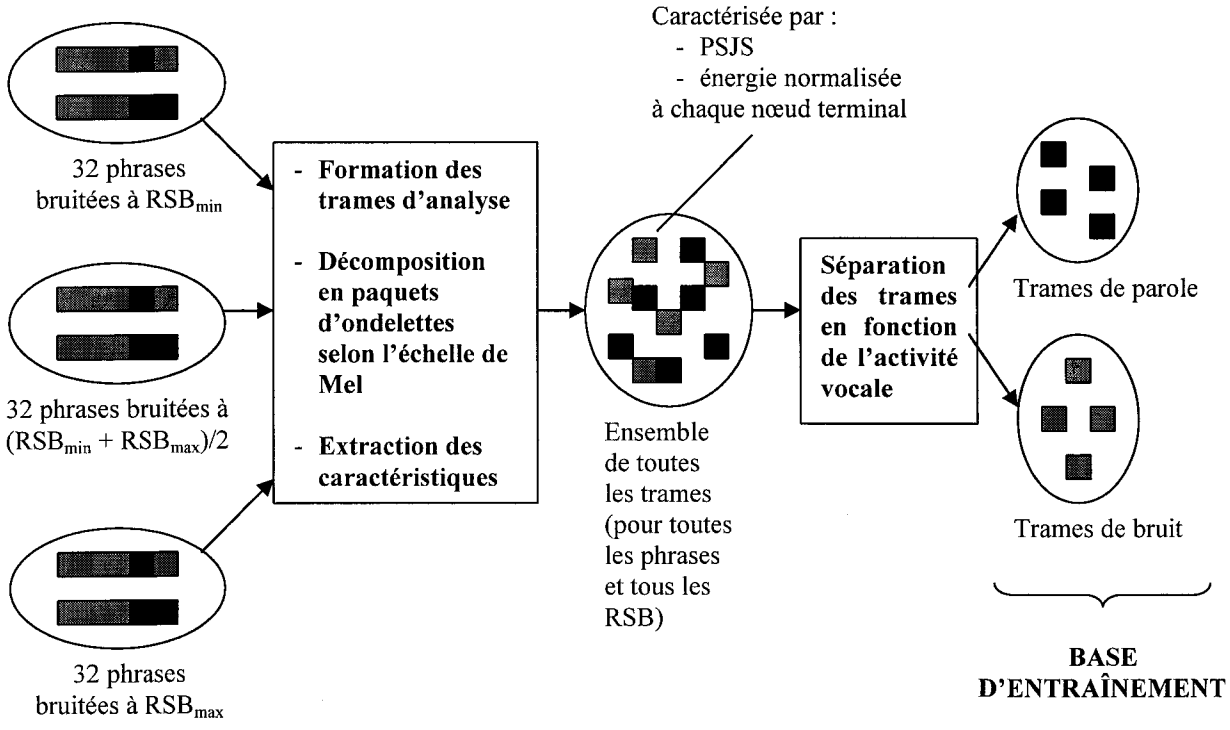


Figure 55 Formation de la base d'entraînement

Rappel : les « trames de parole » sont les trames avec activité vocale, les « trames de bruit » sont celles qui en sont dépourvues.

Ajustement des modules de Décision 1 et 2

Une détection parfaite de la parole n'est pas réaliste. L'ajustement des règles doit donc être un compromis entre une bonne reconnaissance de l'activité vocale et un pourcentage d'erreur « bruit pris pour de la parole » raisonnable. Le problème est de quantifier ces notions de « bonne reconnaissance » de parole et de pourcentage d'erreur « raisonnable ». Nous devons déterminer des seuils. Si nous les choisissons trop élevés, nous allons passer à côté de beaucoup de trames de parole, pourtant facilement identifiables. Si nous les choisissons trop petits, nous allons engendrer beaucoup d'erreur « bruit pris pour de la parole ». Avant de chercher à déterminer les seuils, il faut donc nous imposer un pourcentage d'erreur « bruit pris pour de la parole » à ne pas dépasser lors de l'ajustement des deux modules de décision. L'expérience acquise au cours de la recherche du DAV INNES a montré que pour 5dB la génération de 20% d'erreur sur le bruit permet d'obtenir un bon compromis entre reconnaissance de parole satisfaisante et erreur « bruit pris pour de la parole » raisonnable. Pour 10dB, ce pourcentage s'élève à 15% et pour 15dB à 10%. Une formule liant ce pourcentage d'erreur et le rapport signal sur bruit a été déduite de ces observations :

$$\textit{Pourcentage acceptable d'erreur sur le bruit} = 25 - RSB \qquad (5.26)$$

Cette équation, utilisée dans la procédure d'ajustement, donne des résultats intéressants. Toutefois, elle a été choisie arbitrairement et n'est peut être pas optimale.

Comme il a été vu dans la section 5.3.2, le module Décision 1, basé sur les PSJS, permet une meilleure reconnaissance des trames de parole pour de faibles RSB donc pour la borne inférieure RSB_{min}. Le module Décision 2, basé sur les énergies normalisées, identifie mieux celles pour les RSB plus élevés donc pour la borne supérieure RSB_{max}. Ainsi l'ajustement des coefficients utilisés par les règles sur les PSJS s'effectue en fonction du pourcentage acceptable d'erreur sur le bruit à RSB_{min}, tandis que celui du module Décision 2 se fait par rapport au pourcentage obtenu à RSB_{max}.

ÉTAPE 2 : Ajustement du module Décision 1

1ère Partie :

Il s'agit ici de déterminer les seuils utilisés par les règles portant sur la moyenne et la variance des PSJS aux 5 étages (voir figure 48). Comme il a été vu précédemment, pour cela, il faut examiner les distributions des grandeurs pour les ensembles parole et bruit. Ainsi, on visualise chacune de ces dix grandeurs en fonction d'elle-même. La figure 56 est un exemple des résultats obtenus. Elle représente les distributions des variances du PSJS à l'étage 2. En haut, il s'agit des trames de parole, en bas celles de bruit.

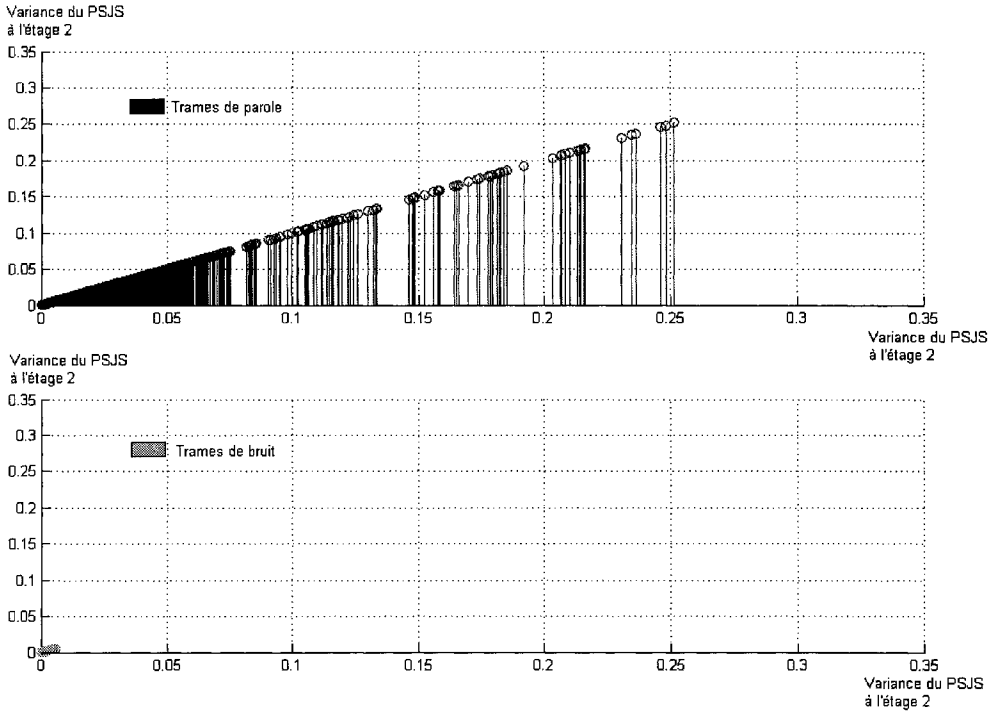


Figure 56 Variance du PSJS à l'étage 2 (= aux noeuds 21 à 26)

Comme nous pouvons le voir, les trames de bruit sont très concentrées. Pour déterminer le seuil adéquat, il est nécessaire d'agrandir cette zone. La figure 57, page suivante, présente cet agrandissement. Il est à noter que les trames de bruit y sont superposées aux trames de parole afin de comprendre plus facilement la manière de déterminer le seuil.

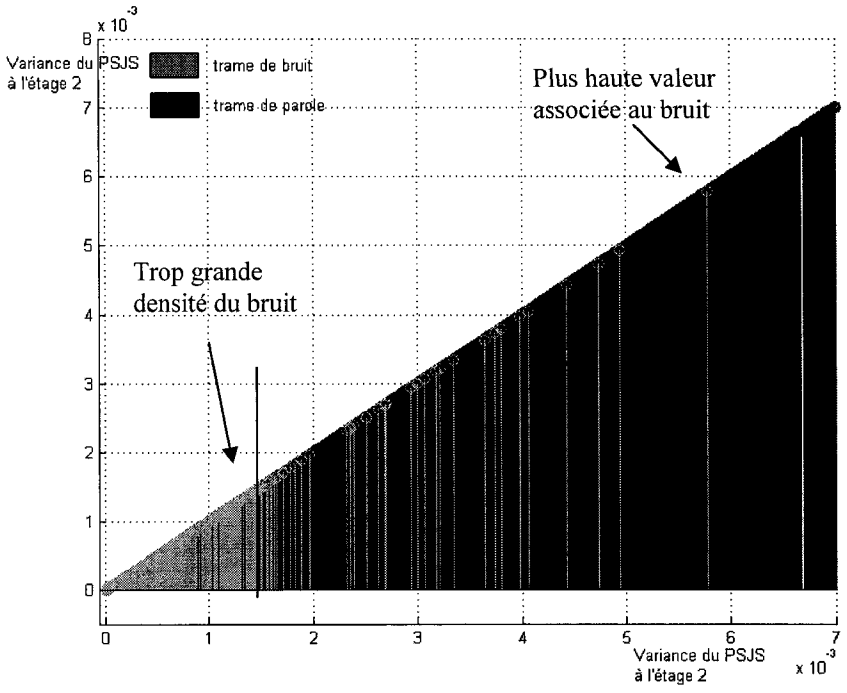


Figure 57 Variance du PSJS à l'étage 2 (agrandissement)

Lors d'un ajustement manuel, la simple visualisation de la figure 57 suffit à déterminer le seuil offrant un bon compromis : $1{,}5\times10^{-3}$. En effet, si l'on abaisse encore plus ce seuil, l'erreur sur le bruit va augmenter très rapidement alors que la reconnaissance de la parole ne sera que légèrement améliorée. A contrario, si on monte le seuil, beaucoup de trames de parole seront manquées sans pour autant diminuer grandement le pourcentage d'erreur du bruit.

Lors d'un ajustement automatique, la visualisation des distributions des données n'est pas envisageable. Il a donc été nécessaire de concevoir un « détecteur de trop forte densité de bruit ». Celui-ci part de la plus haute valeur associée au bruit. Il diminue ensuite le seuil jusqu'à rencontrer une trop grande densité de trames de bruit. Bien sûr, il tient compte du pourcentage d'erreur autorisé.

L'ajustement automatique des règles sur la moyenne et la variance des PSJS s'effectue de la manière suivante :

a) Le pourcentage d'erreur du bruit acceptable pour le RSB_{min} est évalué à l'aide de la formule (5.26)
b) Les valeurs des moyennes et des variances des PSJS sont calculées pour toutes les trames de la base d'entraînement
c) Le détecteur de trop forte densité du bruit détermine le seuil pour chaque règle ainsi que le nombre de trames de parole et de bruit validées par celle-ci
d) La règle qui reconnaît le plus de parole est sélectionnée et son coefficient est enregistré
e) Les trames validées sont retirées de la base d'entraînement
f) Le pourcentage d'erreur restant est réévalué en fonction du nombre de trames de bruit mal identifiées par cette règle (**Attention** : ici seulement, il s'agit des trames de bruit pour le RSB_{min})
   Retour à c)

Conditions d'arrêt : Lorsque toutes les règles sont ajustées, OU
Lorsqu'aucune des règles restantes ne permet d'identifier de parole supplémentaire, OU
Lorsque le pourcentage d'erreur restant est nul.

Une trame est déclarée PAROLE lorsqu'elle satisfait au moins une règle de décision, Décision 1 et Décision 2 confondus. Lors de l'ajustement d'une nouvelle règle, il n'est donc pas nécessaire de prendre en compte les trames de parole déjà validées. De même, le fait de retirer les trames du bruit mal classées par les règles précédentes offre la possibilité de descendre encore plus le seuil de la nouvelle règle et donc de reconnaître davantage de parole. La réduction de la base d'entraînement après l'ajustement de chaque seuil est donc très importante.

Rappel : pour les deux modules de décision, il n'y a qu'un seuil à déterminer par règle.

2ème Partie :

Il s'agit maintenant de déterminer les seuils associés aux règles sur les PSJS des 20 nœuds terminaux :

$$\text{Si} \quad PSJS(i) > \alpha_i \quad \text{alors } décision = PAROLE \tag{5.27}$$

Avec :

- $PSJS(i)$: la valeur du PSJS au nœud i
- α_i : le seuil utilisé au nœud i
- i : le numéro du nœud, i = 13, 14, 21 à 38

La procédure d'ajustement est semblable :

a) Le pourcentage restant (d'erreur du bruit pour le RSB_{min}) est évalué
b) Le détecteur de trop forte densité du bruit détermine le seuil pour chaque règle ainsi que le nombre de trames de parole et de bruit validées par celle-ci
c) La règle qui reconnaît le plus de parole est sélectionnée et son coefficient est enregistré
d) Les trames validées sont retirées de la base d'entraînement
e) Le pourcentage d'erreur restant est réévalué en fonction du nombre de trames de bruit mal identifiées par cette règle (**Attention** : ici seulement, il s'agit des trames de bruit pour le RSB_{min})
   Retour à b)

Conditions d'arrêt : Lorsque toutes les règles sont ajustées, OU
Lorsqu'aucune des règles restantes ne permet d'identifier de parole supplémentaire, OU
Lorsque le pourcentage d'erreur restant est nul.

La séparation entre les trames de parole et de bruit n'est pas la même sur chaque nœud ni sur chaque étage. Certains nœuds terminaux et certains étages sont plus propices à la reconnaissance de l'activité vocale. Les deux ajustements que nous venons de présenter tiennent compte de cela et déterminent systématiquement le meilleur nœud ou le

meilleur étage. C'est pourquoi les seuils des règles restantes sont réévalués à chaque itération.

ÉTAPE 3 : Ajustement du module Décision 2

L'ajustement des coefficients des règles portant sur les énergies normalisées est identique. La seule différence est le pourcentage d'erreur. En effet, il concerne ici les trames de bruit pour le plus grand rapport signal à bruit de la plage : RSB_{max}, au lieu de RSB_{min}. Il est à noter que la première fois où ce pourcentage est déterminé, les quelques trames de bruit à RSB_{max} mal classées par Décision 1 sont prises en compte dans le calcul.

ÉTAPE 4 : Récolte des données

Le module *Hangover* permet de réviser la décision préliminaire en fonction des caractéristiques des trames précédentes. Il est donc nécessaire de récolter des données supplémentaires pour pouvoir l'ajuster correctement. Pour cela, on doit reformer la base d'entraînement. On reprend donc les trois ensembles de 32 phrases bruitées à différents RSB et on exécute les cinq premières étapes de l'algorithme du DAV INNES pour chacune d'entre elles. On rappelle que les cinq premières étapes sont : la formation des trames d'analyse, la décomposition en paquets d'ondelettes selon l'échelle de Mel, l'extraction des caractéristiques, le module Décision 1 et le module Décision 2. Il est à noter que les coefficients utilisés par les règles de décision sont ceux que nous venons de déterminer aux étapes 2 et 3. Finalement, on obtient encore toutes les trames d'analyse mais cette fois-ci elles sont caractérisées par :

- la décision préliminaire concernant leur activité vocale
- les états des 3 dernières trames
- leur énergie ainsi que celle de la trame précédente
- le nombre de trames de bruit qui les précédent

On les sépare ensuite en deux groupes : les trames qui devraient être déclarée PAROLE et celles qui devraient être déclarées BRUIT.

ÉTAPE 5 : Ajustement du Hangover

L'ajustement du module Décision 1 se fait en fonction des trames à RSB_{min}, celui de Décision 2 en fonction des trames à RSB_{max}. Afin d'adapter au mieux le DAV INNES sur toute la plage de RSB, le *Hangover* est, quant à lui, configuré en fonction du milieu de cette plage : $(RSB_{min} + RSB_{max})/2$.

Les trois premières règles du *Hangover* sont :

Si (*décision préliminaire* = *BRUIT*
& *les* 1,2 *ou* 3 *dernières trames* = *PAROLE*
& $|E - E_{prec}| \leq coefficient$)
Alors vad_flag = *PAROLE*

Il s'agit donc de déterminer le *coefficient* le plus adéquat pour chacune d'elles. Pour cela, on effectue leur ajustement de la manière suivante :

a) Sélection des trames qui vérifient les deux premières conditions, dans le groupe « parole » de la base d'entraînement
b) Sélection des trames qui vérifient les deux premières conditions, dans le groupe « bruit » de la base d'entraînement
c) Calcul de la différence entre l'énergie normalisée de la trame actuelle et celle précédente pour toutes les trames sélectionnées
d) Détermination, par dichotomie, du seuil qui améliore la reconnaissance de la parole pour le RSB du milieu de la plage, sans trop augmenter le pourcentage d'erreur du bruit

Après l'ajustement de chaque règle, deux des paramètres caractérisant les trames, à savoir les états des trois dernières trames et le nombre de trames de bruit qui les précédent, sont réévaluées puisque les décisions concernant certaines trames ont été modifiées.

La quatrième règle est :

$$Si\ (vad\_flag = PAROLE$$
$$\&\ compteur\_bruit > N_1$$
$$\&\ \left|E - E_{prec}\right| \leq coefficient\ )$$
$$Alors\ vad\_flag = BRUIT$$

La manière de déterminer les deux seuils, à savoir N_1 et *coefficient*, est :

a) Création de deux groupes : « trames de parole reconnues » et « trames de bruit prises pour de la parole » à partir de la base d'entraînement
b) Calcul de la différence entre l'énergie normalisée de la trame actuelle et celle précédente pour chaque trame de ces deux groupes
c) Détermination simultanée des deux seuils. Cette fois-ci, on part des plus hautes valeurs pour groupe « bruit pris pour de la parole » et on descend ces seuils par petits pas jusqu'à trouver ceux qui permettent de réduire au mieux le pourcentage d'erreur sans trop détériorer la bonne reconnaissance de parole.

La cinquième règle permet à la fois de corriger les erreurs sur la parole et sur le bruit :

$$Si\ (vad\_flag \neq dernier\_vad\_flag$$
$$\&\ \frac{E}{E_{prec}} < coefficient)$$
$$Alors\ vad\_flag = dernier\_vad\_flag$$

Pour effectuer l'ajustement de ce dernier coefficient, il faut faire :

a) Création de deux groupes : « trames qui devraient avoir le même état que la précédente mais qui ne l'ont pas » et « trames qui ont à juste titre le même état que la précédente » à partir de la base d'entraînement
b) Calcul du rapport des énergies normalisées entre la trame actuelle et la précédente pour chaque trame de ces deux groupes
c) Détermination du seuil pour que les performances obtenues pour le RSB du milieu de la plage soient améliorées. Ceci se fait par dichotomie et en fonction des deux groupes.

Remarque : pour les cinq ajustements du *Hangover*, s'il n'y a pas de seuil permettant d'augmenter les performances, la règle est inhibée.

La dernière règle de lissage, n'utilisant aucun coefficient, n'a pas besoin d'être ajustée.

La procédure automatique d'ajustement des règles est maintenant terminée. Le DAV INNES ainsi conçu est prêt à être utilisé.

Dans ce chapitre, l'algorithme de détection d'activité vocale basé sur les ondelettes que nous avons développé a été vu en détails. Il présente de nombreuses nouveautés. La plus importante est certainement l'utilisation du seuil de Johnstone et Silverman comme critère de décision. Comme nous l'avons mentionné, il est très utile car il permet de renforcer la robustesse du DAV surtout dans les cas de RSB faibles. L'ensemble des règles, que ce soit pour la décision préliminaire ou la décision finale, constitue aussi une originalité puisqu'elles ont été mises au point spécialement pour des environnements industriels. Enfin, la procédure d'ajustement du DAV est, elle aussi, à notre connaissance complètement unique. En effet, elle s'effectue sur des bruits industriels et utilise une recherche du meilleur nœud et du meilleur coefficient pour chaque nouvel environnement. Manuellement, cette procédure est relativement simple mais très fastidieuse. C'est pourquoi nous avons décidé de l'automatiser. La visualisation des distributions des données permettant de déterminer les seuils n'étant alors plus envisageable, nous avons mis au point un « détecteur de trop forte densité ». Ainsi notre procédure automatique d'ajustement ne nécessite aucune intervention humaine, ce qui rend notre système très facile à utiliser. Le chapitre suivant présente les résultats pratiques obtenus pour notre DAV INNES, c'est-à-dire ses performances et son comportement dans différents milieux industriels bruités.

CHAPITRE 6

RÉSULTATS PRATIQUES DU DÉTECTEUR D'ACTIVITÉ VOCALE BASÉ SUR LES ONDELETTES

Nous allons ici analyser le comportement du DAV INNES, proposé dans le chapitre 5, afin de déterminer sa robustesse dans des milieux industriels bruités. Comme il a été mentionné précédemment, il faut entraîner le DAV sur chaque nouvel environnement pour pouvoir déterminer les règles de décision les plus adéquates et obtenir des performances accrues. Ainsi, les bases d'entraînement et de validation permettant d'obtenir les résultats pratiques seront tout d'abord présentées. Les performances du DAV pour différents bruits et différents rapports signal à bruit seront ensuite exposées et son comportement pourra être caractérisé. Enfin, la dernière partie sera consacrée à la généralisation du système dans l'optique d'aboutir à un unique DAV efficace dans tous les environnements industriels.

6.1 Bases d'entraînement et de validation

Comme pour le G729.B adapté aux bruits industriels vu au chapitre 3, nous allons établir les performances du DAV INNES dans onze environnements différents. Trois bruits sont issus de la base de données NOISEX. Les huit autres nous ont été fournis par la compagnie SONOMAX et ont été récoltés dans une raffinerie de cuivre de NORANDA. Le tableau VIII rappelle leurs caractéristiques.

Tableau VIII

Caractéristiques des bruits industriels utilisés

Nom	Description	Niveau global à +/- 3dB	Fe
Noisex 1 (Nx1)	*Factory Noise 1* de NOISEX, salle de découpage et de soudure du métal pour des équipements automobiles	83dBA	19980Hz
Noisex 2 (Nx2)	*Factory Noise 2* de NOISEX, salle de production automobile	74dBA	19980Hz
Noisex 3 (Nx3)	*Operation Room* de NOISEX, salle d'opérations d'un *destroyer*	70dBA	19980Hz
Noranda 1 (Nor1)	Salle d'empaquetage sélénium, *Baghouse 1*	-	22050Hz
Noranda 2 (Nor2)	Salle d'empaquetage sélénium, *Baghouse 1, 2, 3* et *Vaccum Cleaner*	82dBA	22050Hz
Noranda 3 (Nor3)	Salle d'empaquetage sélénium, Micropulvérisateur de sélénium et ventilateur	79dBA	44100Hz
Noranda 4 (Nor4)	Salle d'empaquetage sélénium, Soupape de sécurité	71dBA	44100Hz
Noranda 5 (Nor5)	Salle d'empaquetage sélénium, Surpresseur à pistons rotatifs	-	44100Hz
Noranda 6 (Nor6)	Salle des transformateurs	100dBA	44100Hz
Noranda 7 (Nor7)	Brûleur propane et fournaise	102dBA	44100Hz
Noranda 8 (Nor8)	Salle hydraulique	108dBA	44100Hz

Les phrases utilisées par les bases d'entraînement et de validation sont toutes issues de la base de données DARPA TIMIT. La base d'entraînement contient les 32 signaux de parole qui ont été utilisés pour la recherche des deux algorithmes de détection d'activité vocale proposés dans ce mémoire (voir section 3.2.1 et section 5.3.1). Celle de validation se compose des 64 signaux de parole utilisés pour tester le G729.B modifié (voir section 3.3.1).

Dépendant des conditions d'enregistrement, tous les signaux, que ce soit de parole ou de bruit, ont été rééchantillonnés à 8kHz en prenant soin d'utiliser un filtre adéquat pour conserver la bonne qualité de l'information. Suite à cela, des parties de chaque bruit ont été choisies aléatoirement puis additionnées aux phrases afin d'obtenir nos signaux de parole bruitée. La plage de rapports signal à bruit utilisée pour l'entraînement est [5dB - 15dB]. Les niveaux sonores des phrases et des bruits sont tous différents les uns les autres, il a donc fallu déterminé dans chaque cas le coefficient à appliquer à l'amplitude du bruit afin d'obtenir les RSB voulus.

Le tableau IX résume les caractéristiques des bases d'entraînement et de validation pour chaque bruit.

Tableau IX

Caractéristiques des bases d'entraînement et de validation

Base d'entraînement	Base de validation
Parole : - 32 phrases (DARPA TIMIT) - américain - toutes différentes - toutes prononcées par des locuteurs différents - 16 femmes, 16 hommes	Parole : - 64 phrases (DARPA TIMIT) - américain - toutes différentes et différentes de celles d'entraînement - toutes prononcées par des locuteurs différents - 32 femmes, 32 hommes
Bruit : - Factory Noise 1 (NOISEX) - Factory Noise 2 (NOISEX) - Operation Room (NOISEX) - Noranda 1 à 8 (fournis par SONOMAX)	Bruit : - Factory Noise 1 (NOISEX) - Factory Noise 2 (NOISEX) - Operation Room (NOISEX) - Noranda 1 à 8 (fournis par SONOMAX)
SNR : - 5dB - 10dB - 15dB	SNR : - 5dB - 10dB - 15dB - [0dB – 20dB] par pas de 1dB pour une vue en détails des performances

vue globale

Connaissant l'activité vocale de chaque phrase, il nous a été possible d'effectuer l'entraînement et la validation pour chaque environnement. La partie suivante présente les résultats pratiques obtenus.

6.2 Résultats pratiques

6.2.1 Performances

Afin de caractériser le comportement du DAV INNES, les quatre critères objectifs présentés à la section 2.3 ont été retenus. Pour faciliter la compréhension des résultats qui suivent, on rappelle ici leur définition (Beritelli et Al. [35], [36]):

- Erreur de début : erreur introduite lors du passage d'une période de bruit à une période de parole. Le DAV continue à indiquer l'absence de parole alors que l'activité vocale a déjà commencé.
- Erreur de fin : erreur introduite lors du passage d'une période de parole à une période de bruit. Le DAV continue à indiquer la présence de parole alors que celle-ci est déjà terminée.
- Coupure : erreur se produisant au cours d'une bouffée de parole, la parole est soudainement prise pour du bruit.
- Faux déclenchement : erreur se produisant au cours d'une période d'inactivité vocale, le bruit est soudainement pris pour de la parole.

À ceux-ci ont été ajoutés deux autres critères:

- Reconnaissance de la parole : il s'agit du pourcentage de parole correctement reconnue par le DAV.
- Reconnaissance du bruit : il s'agit du pourcentage d'absence de parole correctement reconnue par le DAV.

La figure 58 illustre les six paramètres utilisés pour l'évaluation des performances du DAV :

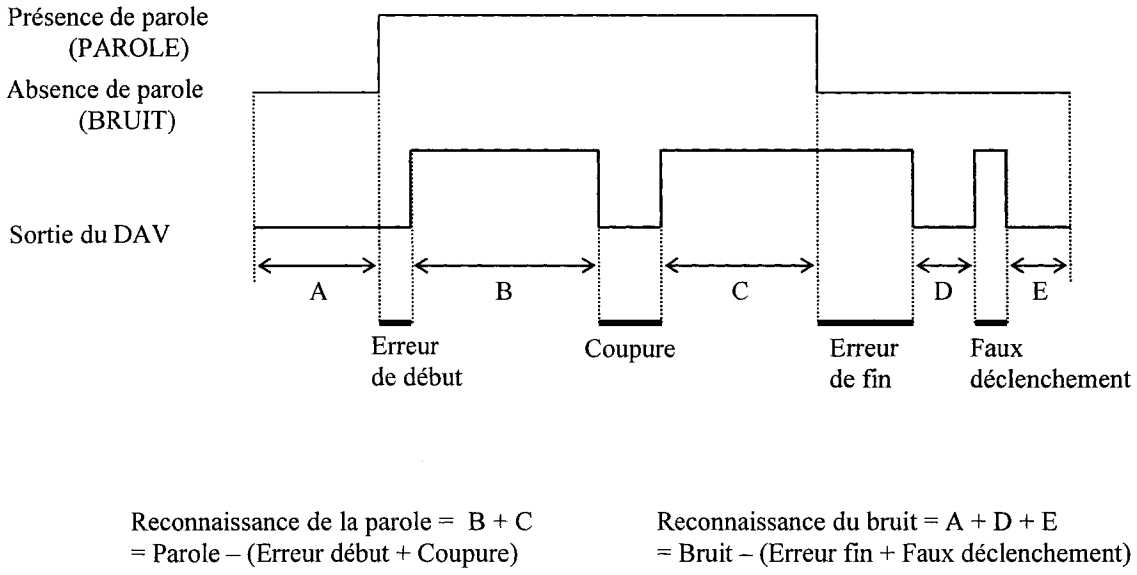


Figure 58 Paramètres objectifs pour l'évaluation des performances du DAV

Pour chacun des 11 bruits industriels, le DAV INNES, développé avec le logiciel MATLAB, a été entraîné sur la base d'entraînement correspondante, ceci dans le but d'obtenir l'ensemble des coefficients, utilisés par les règles de décision, le plus propice à chaque environnement. La procédure automatique d'ajustement a été présentée à la section 5.4. Les DAV ainsi obtenus ont ensuite été testés sur leur base de validation. Pour faciliter la compréhension des résultats pratiques, les erreurs, c'est-à-dire les quatre premiers critères objectifs, sont illustrées sous forme d'histogrammes, figure 59. Les pourcentages de reconnaissance de parole et de bruit sont, quant à eux, présentés dans les tableaux X et XI.

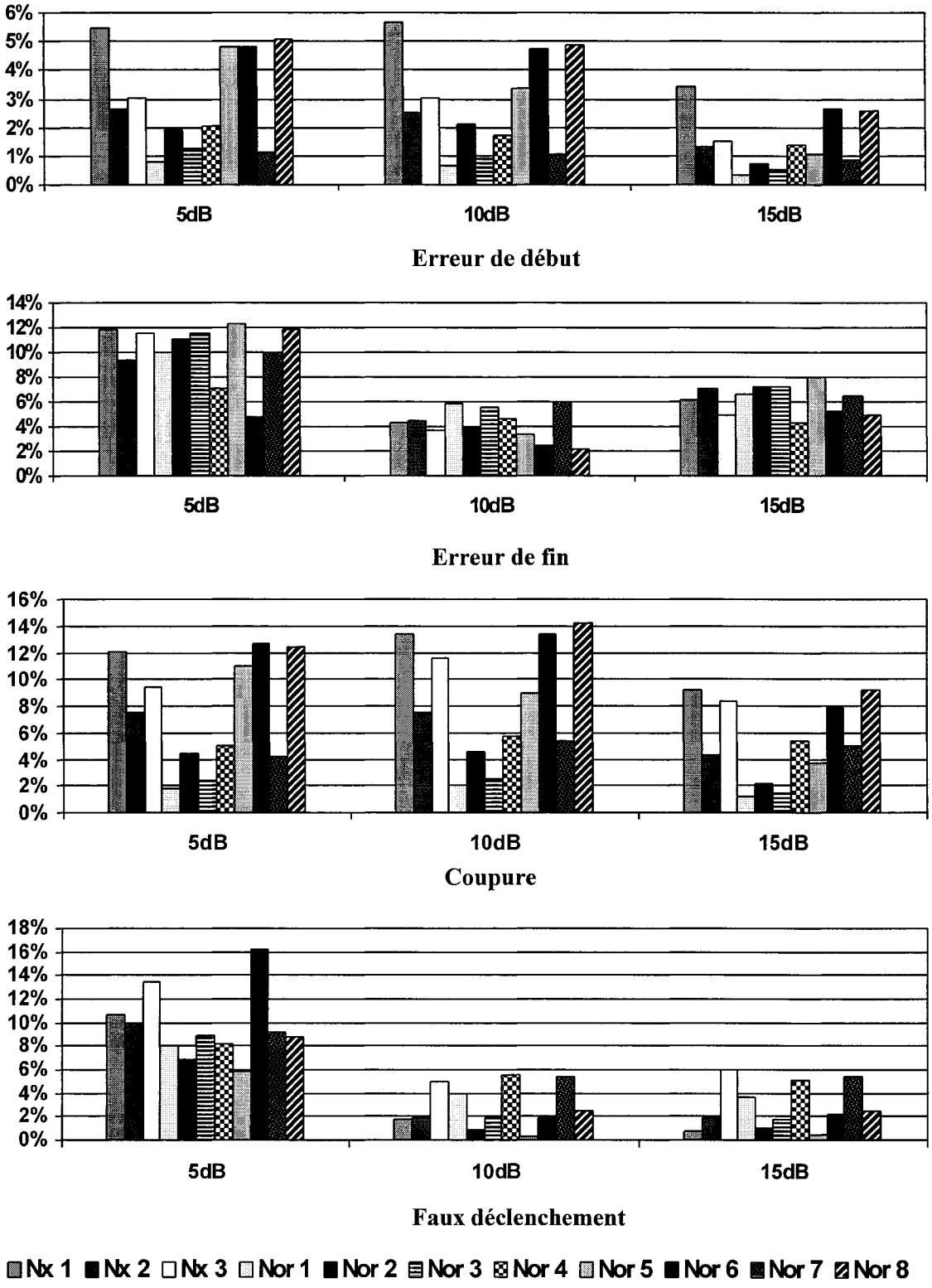


Figure 59 Taux d'erreur de début, de fin, coupure et faux déclenchement en validation

Sur la figure 59, il est observable que les pourcentages d'erreur sont relativement faibles quelque soit la situation. Le DAV INNES présenté dans ce mémoire possède donc une bonne robustesse dans les milieux industriels même pour de petits rapports signal à bruit de l'ordre de 5dB. Sur les premier et dernier histogrammes, on peut remarquer que les erreurs de début et les faux déclenchements sont de moins en moins fréquents, plus le RSB augmente. Cette tendance ne se retrouve pas pour les deux autres critères. En effet, pour 10dB il y a moins d'erreurs de fin que pour 15dB alors que le bruit est plus fort. De même, les coupures dans la parole sont plus fréquentes pour 10dB que pour 5dB. On rappelle que plus le RSB est grand, moins le bruit est présent et plus les erreurs devraient être petites. Ce phénomène, qui sera expliqué dans la partie suivante, est aussi visible sur les deux tableaux X et XI, ci-après, qui présentent les pourcentages de reconnaissance de parole et de bruit obtenus lors de l'entraînement et de la validation.

Tableau X

Pourcentages de reconnaissance de parole et de bruit obtenus en entraînement

RSB / Perf. / Bruit	5dB		10dB		15dB	
	Rec. Parole (%)	**Rec. Bruit (%)**	**Rec. Parole (%)**	**Rec. Bruit (%)**	**Rec. Parole (%)**	**Rec. Bruit (%)**
Noisex 1	82,2	81,4	81,6	93,8	87,6	92,6
Noisex 2	90,3	80,6	90,7	91,5	94,6	90,4
Noisex 3	87,2	80	85,8	93,6	90,6	91,2
Noranda 1	97,9	84,4	97,4	90,8	98,5	90,2
Noranda 2	94,4	82	94,2	94,3	97,1	90,8
Noranda 3	97	82,9	96,5	92,6	98,1	90,8
Noranda 4	93,4	83,6	92,9	87,6	93,4	88,6
Noranda 5	87,1	81,7	88,8	95,1	95,5	91,4
Noranda 6	84,2	80,4	82,9	96,2	89,7	93,3
Noranda 7	95	84	93,7	90,1	94,2	89,5
Noranda 8	83,9	80,6	82,1	95	88,6	92,8

Tableau XI

Pourcentages de reconnaissance de parole et de bruit obtenus en validation

RSB / Perf. / Bruit	5dB		10dB		15dB	
	Rec. Parole (%)	Rec. Bruit (%)	Rec. Parole (%)	Rec. Bruit (%)	Rec. Parole (%)	Rec. Bruit (%)
Noisex 1	82,4	77,5	81	93,8	87,4	93,1
Noisex 2	89,8	80,6	90	93,7	94,3	91
Noisex 3	87,6	75,9	85,4	91,3	90,1	89
Noranda 1	97,4	82	97,3	90,3	98,5	89,7
Noranda 2	93,6	82,1	93,4	95,2	97,2	91,8
Noranda 3	96,4	79,5	96,5	92,6	98,1	91
Noranda 4	93	84,8	92,5	89,9	93,3	90,5
Noranda 5	84,2	81,8	87,7	96,4	95,2	91,6
Noranda 6	82,6	78,9	81,9	95,7	89,5	92,7
Noranda 7	94,7	80,7	93,6	88,6	94,1	88,2
Noranda 8	82,5	79,4	81	95,3	88,2	92,7

L'analyse de ces deux tableaux X et XI montre que les performances obtenues en validation sont proches de celles de l'entraînement. Elles sont quasi-identiques dans certains cas. Les pourcentages de reconnaissance de parole et de bruit sont élevés quelque soit l'environnement, et ceci même lorsque le rapport signal à bruit est petit. Cela confirme la bonne robustesse de notre DAV basé sur les ondelettes développé dans le cadre de ce projet de recherche. Le DAV INNES présente donc des performances très satisfaisantes pour tous les environnements industriels testés et pour la plage complète de RSB visée, à savoir [5dB – 15dB].

Que ce soit en entraînement ou en validation, la comparaison des performances entre les trois RSB montre le phénomène énoncé précédemment : les pourcentages de reconnaissance de parole pour 10dB sont légèrement plus faibles que pour 5dB, et ceux de reconnaissance du bruit pour 10dB sont légèrement plus élevés que pour 15dB. Les deux parties qui suivent vont expliquer les raisons de ce phénomène ainsi que la manière de l'influencer.

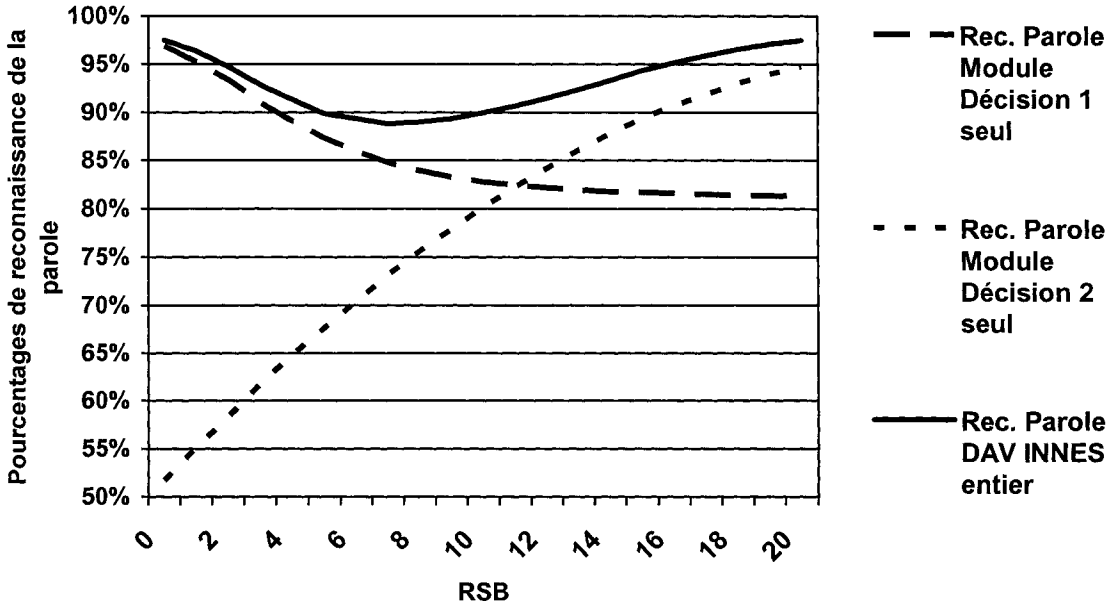
6.2.2 Explication du phénomène

Les DAV ne peuvent pas toujours fonctionner correctement. Passé un certain rapport signal à bruit, leurs performances chutent rapidement avec le RSB. Les reconnaissances des régions de parole et de bruit seul deviennent alors antagonistes. En effet, si l'on se place dans le cas extrême de la reconnaissance nulle des parties dépourvues de voix, il paraît évident que l'identification des régions de parole sera excellente, voire parfaite. La sortie du DAV sera en fait bloquée sur PAROLE. Nous avons déjà discuté de cela pour le G729.B au chapitre 3 section 3.3.2. Ce dernier a été mis au point de manière à ce que lorsqu'il est utilisé en dehors de sa plage de fonctionnement, il ne détériore pas l'information importante, à savoir la parole. Ainsi, quand le RSB est trop faible, le G729.B n'est plus capable de faire la distinction entre la parole et le bruit. Il bloque alors sa sortie sur PAROLE afin de laisser le signal utile intact (Alcatel [39]).

Nous avons le même phénomène pour notre DAV INNES. Passé un certain RSB, dépendant du milieu industriel étudié, les performances obtenues deviennent antagonistes. La reconnaissance de la parole suit une parabole : elle diminue avec le RSB jusqu'à atteindre un minimum. À partir de ce point, elle remonte au lieu de continuer à diminuer. La même chose se produit avec l'identification du bruit seul qui suit elle aussi une parabole mais inversée : elle augmente jusqu'à un maximum puis diminue avec le RSB, au lieu de poursuivre sa montée.

Ce phénomène peut être expliqué par le choix de nos critères de décision. En effet, comme il a été dit dans la section 5.3.2, la caractéristique de l'énergie est plus propice à la détection de l'activité vocale lorsque les RSB sont élevés alors que le Paramètre du Seuil de Johnstone et Silverman est plus adéquat pour la reconnaissance de la parole pour les RSB plus faibles. Ainsi les performances obtenues avec uniquement le module de décision basé sur l'énergie sont : la diminution du pourcentage d'identification de la voix et l'augmentation de celui du bruit seul, plus le RSB est petit. Par contre, si l'on utilise uniquement le module concernant les valeurs du PSJS, les évolutions des performances sont inverses. La combinaison de ces deux modules engendre donc le phénomène discuté ici, puisque leur comportement influence tous deux les résultats finaux. Ceci est mis en évidence par les figures 60 et 61 qui suivent.

Influences de Décision 1 et 2 sur la reconnaissance de la parole avec Noisex 2 :



Influences de Décision 1 et 2 sur la reconnaissance du bruit seul avec Noisex 2 :

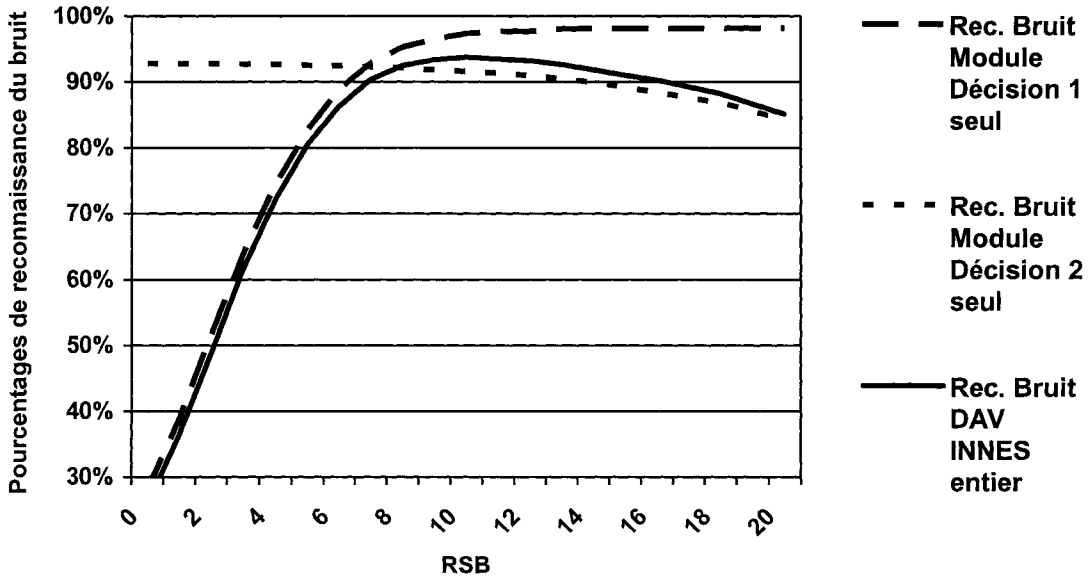
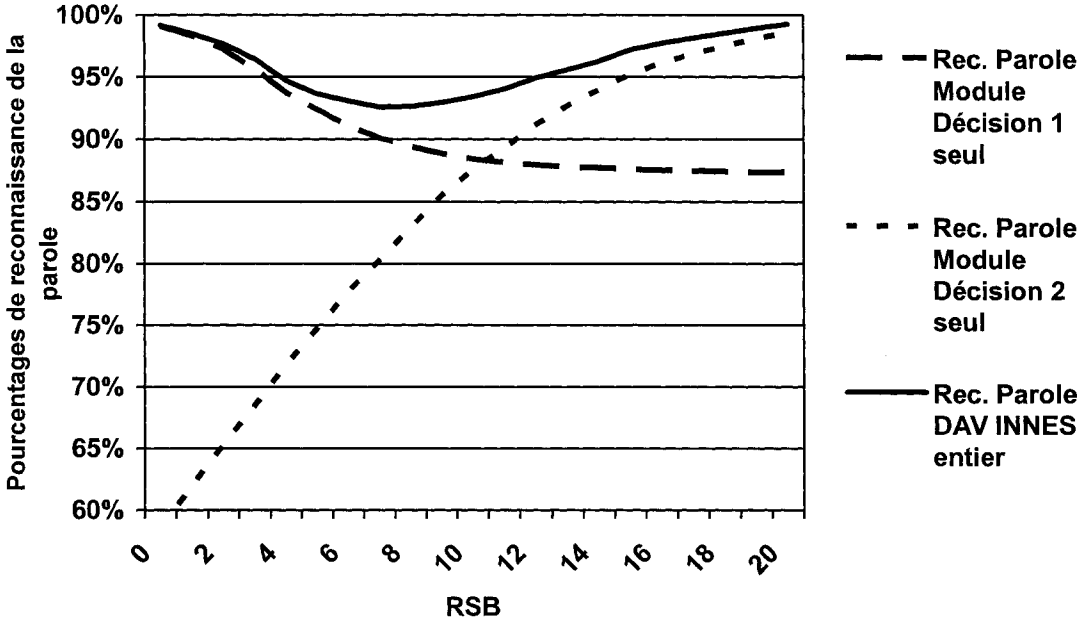


Figure 60 Mise en évidence du phénomène pour le DAV entraîné avec le bruit Noisex 2

Influences de Décision 1 et 2 sur la reconnaissance de la parole avec Noranda 2 :



Influences de Décision 1 et 2 sur la reconnaissance du bruit seul avec Noranda 2 :

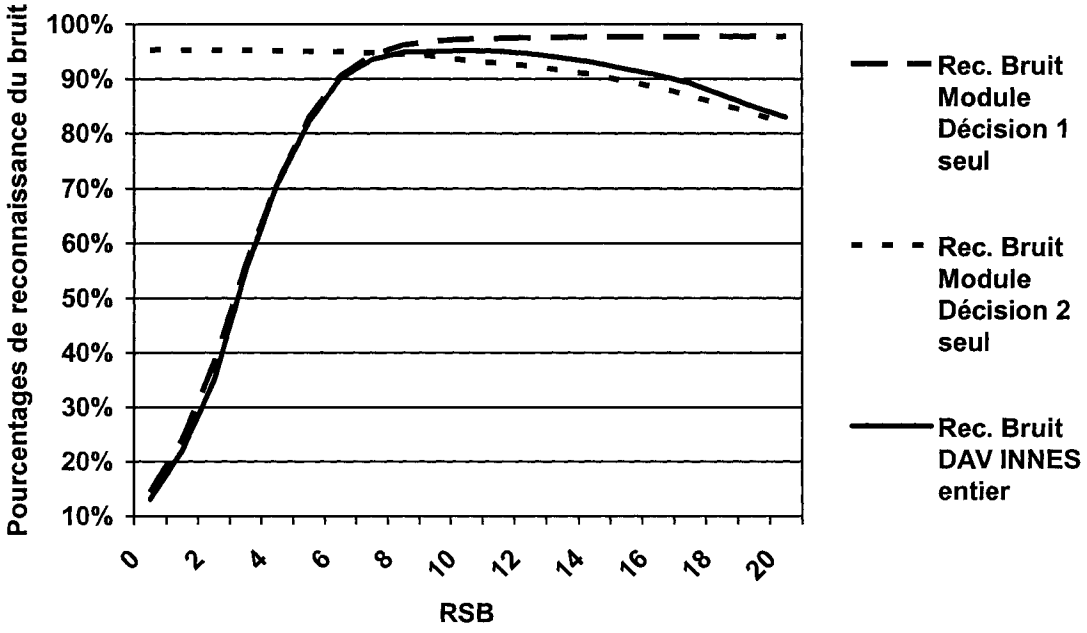


Figure 61 Mise en évidence du phénomène pour le DAV entraîné avec le bruit Noranda 2

Ces deux figures, 60 et 61, mettent clairement en évidence les raisons du phénomène, c'est-à-dire les influences des modules Décision 1 et Décision 2 sur la sortie finale du DAV, que ce soit pour la reconnaissance de la parole ou celle du bruit seul. Ces deux modules ont des comportements opposés. Leur utilisation conjointe engendre donc des performances qui évoluent paraboliquement en fonction du RSB.

Remarque : afin de ne pas surcharger inutilement ce mémoire, le phénomène n'a été présenté que pour deux bruits : Noisex 2 et Noranda 2. Toutefois, il est important de noter que les expériences pratiques ont montré que ceci est obtenu quelque soit le bruit étudié.

Notre DAV INNES présente des performances satisfaisantes sur la plage de RSB [5dB – 15dB] pour les milieux industriels. En dehors de cet intervalle, plus le RSB diminue, plus notre DAV laisse passer de signal. Finalement, quand il n'est plus capable de fournir une discrimination correcte entre la parole et le bruit, il indique continuellement la présence d'activité vocale afin d'éviter les pertes lourdes et irrémédiables dans le message de parole à transmettre.

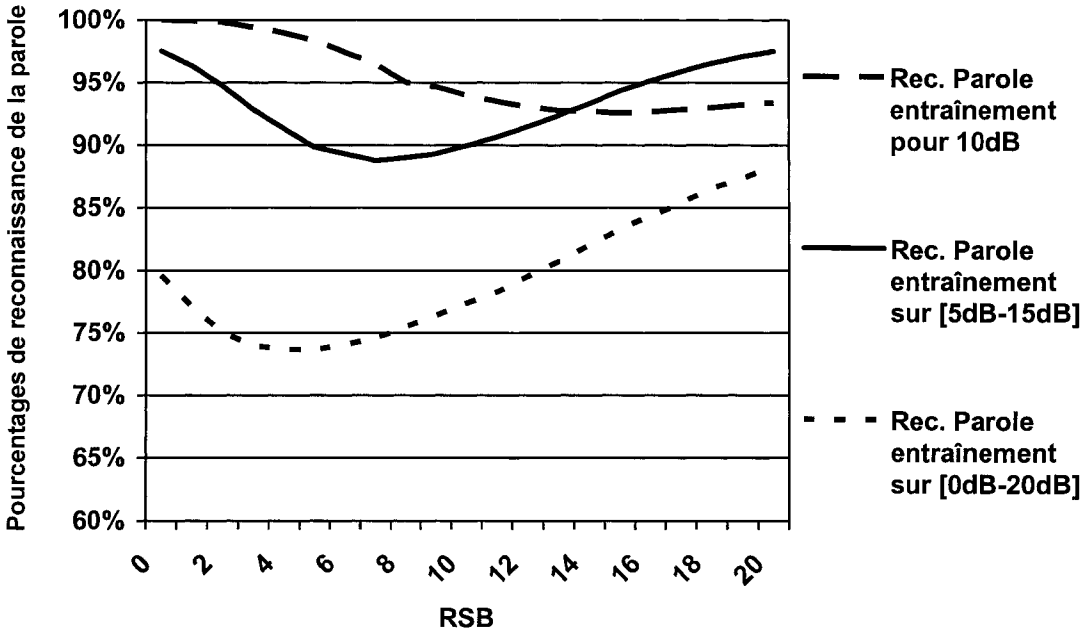
6.2.3 Influences sur le phénomène

Comme il a été vu à la section 5.4, les coefficients du module Décision 1, basé sur les valeurs du PSJS, sont déterminés en fonction des trames de parole à RSB_{min} (c'est-à-dire 5dB dans le cadre de notre projet de recherche) puisque ce module est plus propice à la détection d'activité vocale lors de faibles RSB. Ceux du module Décision 2, basé sur les valeurs des énergies, sont déterminés en fonction des trames de parole à RSB_{max} (c'est-à-dire 15dB dans notre projet) puisque ce module est, lui, plus propice à la détection d'activité vocale pour les RSB élevés. Les maximum et minimum des pourcentages de reconnaissance du bruit seul et de la parole, respectivement, sont atteints pour des rapports signal à bruit se trouvant entre ces deux valeurs. Les résultats présentés

précédemment ont été obtenus dans le cadre d'un entraînement sur [RSB_{min} = 5dB - RSB_{max} = 15dB]. Les deux figures suivantes, 62 et 63, montrent ce qu'il se passe lorsque l'entraînement du DAV INNES est effectué sur d'autres plages de RSB, à savoir : [RSB_{min} = 0dB - RSB_{max} = 20dB] et RSB_{min} = RSB_{max} = 10dB.

Remarque : les deux bruits présentés sont encore : Noisex 2 et Noranda 2. Toutefois, il est important de noter que les expériences pratiques ont montré des résultats similaires quelque soit le bruit étudié.

Effets de la plage de RSB utilisée pour l'entraînement sur les pourcentages de reconnaissance de la parole, avec Noisex 2 :



Effets de la plage de RSB utilisée pour l'entraînement sur les pourcentages de reconnaissance du bruit seul, avec Noisex 2 :

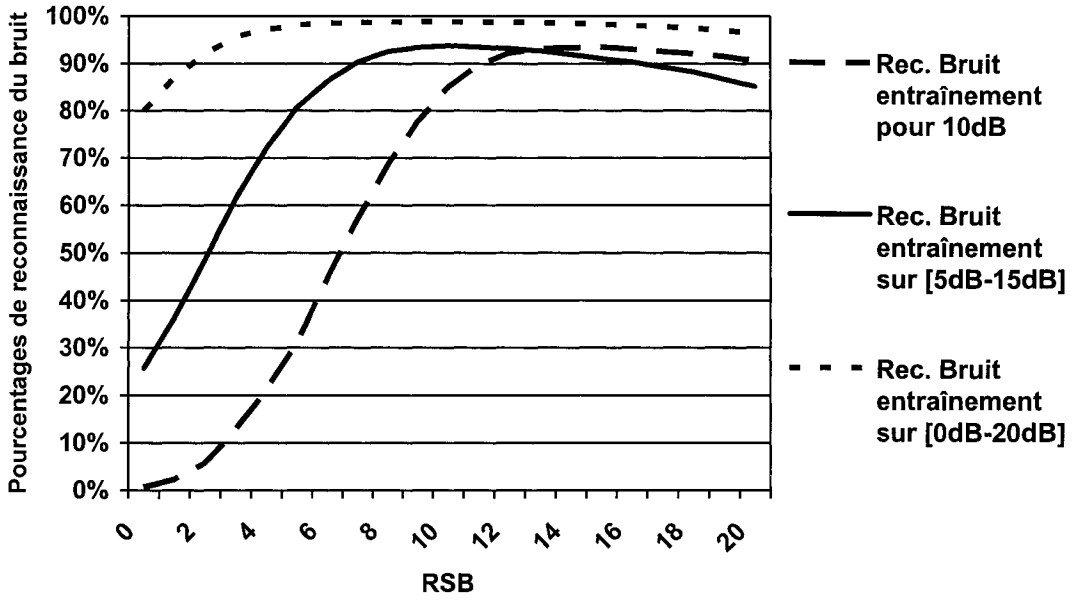
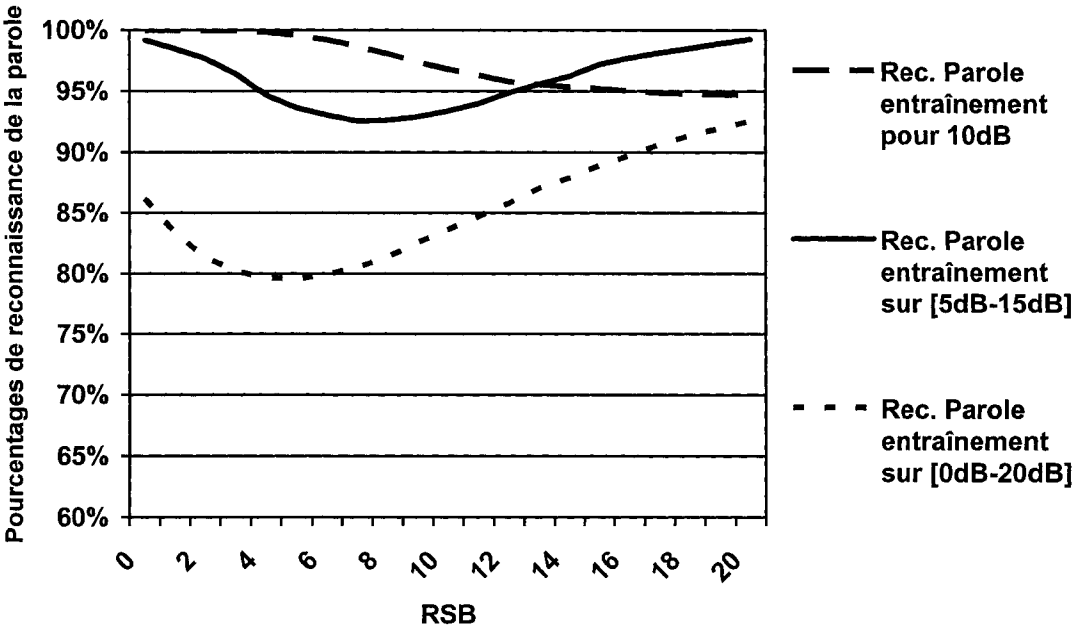


Figure 62 Performances du DAV pour Noisex 2 avec différentes plages de RSB utilisées pour l'entraînement

Effets de la plage de RSB utilisée pour l'entraînement sur les pourcentages de reconnaissance de la parole, avec Noranda 2 :



Effets de la plage de RSB utilisée pour l'entraînement sur les pourcentages de reconnaissance du bruit seul, avec Noranda 2 :

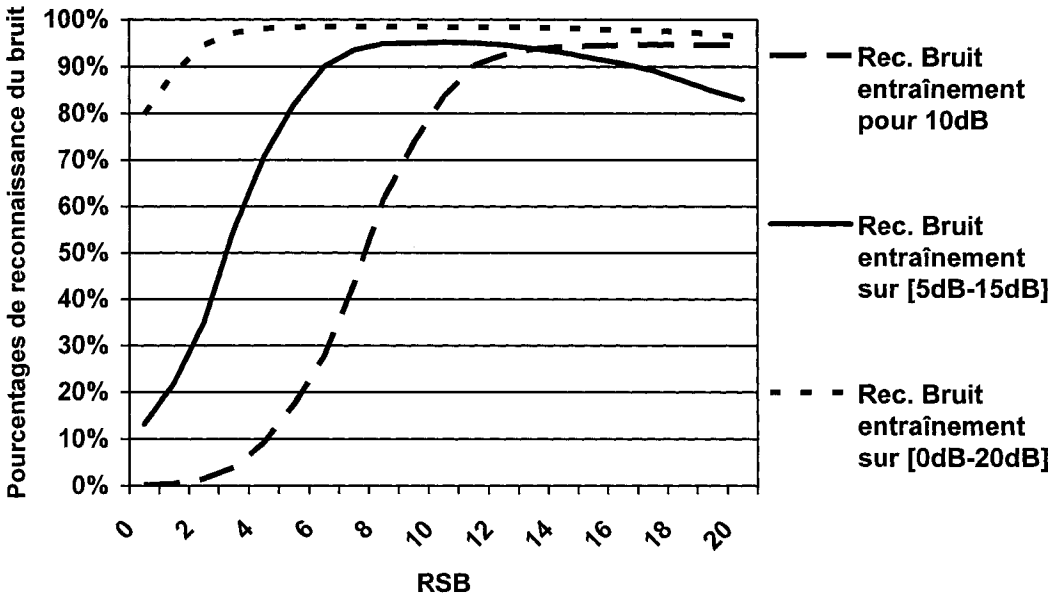


Figure 63 Performances du DAV pour Noranda 2 avec différentes plages de RSB utilisées pour l'entraînement

Comme le montrent ces figures, 62 et 63, plus la plage de RSB utilisée pour l'entraînement est large, plus le phénomène est prononcé. En effet, le maximum du pourcentage de reconnaissance du bruit est alors de plus en plus élevé alors que le minimum du pourcentage de reconnaissance de la parole est, lui, de plus en plus bas. Par contre, le fonctionnement du DAV, moins bon mais toujours acceptable, s'étend alors sur une plage de RSB plus large. Ceci est illustré par le tableau XII :

Tableau XII

Influence de la plage de RSB utilisée pour l'entraînement sur les performances du DAV INNES

Plage de RSB utilisée pour l'entraînement	**Plage de RSB sur laquelle le DAV INNES fonctionne correctement (± quelques dB selon les environnements)**	**Performances**
[0dB – 20dB]	[0dB – 20dB]	Moyennes
[5dB – 15dB]	[5dB – 15dB]	Bonnes
10dB	10dB	Excellentes

Ainsi si l'on souhaite avoir des performances élevées, il faut choisir une plage de RSB étroite pour l'entraînement. A contrario, si l'on souhaite que le DAV INNES fonctionne dans un grand nombre de situations, il faut choisir une plage de RSB large pour l'entraînement, par contre les performances obtenues seront moins bonnes. Dépendant du milieu dans lequel fonctionnera le DAV, il est possible de choisir la plage d'entraînement la plus adéquate. Toutefois, les environnements industriels bruités présentent en général un rapport signal à bruit relativement constant. Ainsi, pour obtenir des performances accrues, il sera plus intéressant de choisir un intervalle de RSB étroit, centré sur le rapport signal à bruit moyen mesuré dans l'environnement étudié.

6.3 Généralisation du système

Jusqu'ici nous n'avons présenté que les résultats obtenus lorsque notre DAV INNES fonctionne avec le même bruit que celui utilisé pour l'entraînement. Ainsi, pour chaque nouvel environnement, il faut réajuster les coefficients des règles de décision afin d'obtenir par la suite des performances accrues, ceci grâce à la procédure automatique d'ajustement décrite à la section 5.4. Ceci est tout à fait acceptable pour notre application dans les milieux industriels. Toutefois, nous avons tenté de généraliser ce système afin d'avoir un seul et unique ensemble de coefficients fonctionnel dans toutes les situations. Pour cela, nous avons effectué quelques tests.

Le premier a été d'entraîner le DAV INNES avec le bruit Noisex 1 et de faire la validation dans les environnements Noisex 2 et Noranda 2. Les résultats obtenus sont exposés dans le tableau XIII.

Tableau XIII

Performances pour Noisex 2 et Noranda 2 avec les seuils utilisés pour Noisex 1

		5dB		10dB		15dB	
Seuils utilisés	**Environnement de validation**	**Rec. de la parole**	**Rec. du bruit**	**Rec. de la parole**	**Rec. du bruit**	**Rec. de la parole**	**Rec. du bruit**
Ceux de Noisex 1	Noisex 2	74,5%	96,5%	78,6%	97,2%	86,3%	95,5%
Ceux de Noisex 2	Noisex 2	89,8%	80,6%	90%	93,7%	94,3%	91%
Ceux de Noisex 1	Noranda 2	90,9%	93,4%	88,3%	95,3%	90,3%	96,8%
Ceux de Noranda 2	Noranda 2	93,6%	82,1%	93,4%	95,2%	97,2%	91,8%

Les pourcentages de reconnaissance du bruit sont meilleurs lorsque les seuils utilisés sont ceux de Noisex 1. Les pourcentages de reconnaissance de la parole restent, eux, acceptables mais ils sont moins bons qu'avec les coefficients ajustés spécialement pour les environnements étudiés. Ceci s'explique par le fait que lors de l'entraînement, il y a une recherche des meilleurs nœuds et des meilleurs étages. D'un bruit à l'autre, ces nœuds et ces étages sont généralement différents.

La recherche des meilleurs nœuds a été mise au point car tous les nœuds terminaux de la décomposition en paquets d'ondelettes ne possèdent pas le même pouvoir de discrimination entre les trames de parole et celles de bruit seul. Certains procurent une très bonne séparation entre ces deux types de trames. A contrario, un seuillage, même très faible, sur les mauvais nœuds engendre une énorme erreur « bruit pris pour de la

parole ». Il en est de même pour les étages. Les meilleurs nœuds et les meilleurs étages diffèrent d'un bruit à l'autre.

En tenant compte de cela, nous avons généré un nouvel ensemble de seuils à partir des 11 ensembles trouvés précédemment. Nous rappelons qu'à chacun des 11 environnements industriels testés correspond un ensemble de seuils. Le nouvel ensemble est obtenu en choisissant pour chaque seuil la plus haute valeur des 11 seuils correspondants. Soit $seuil_k$ l'ensemble des M seuils ajustés pour l'environnement industriel k (avec $k = 1$ à 11), le nouvel ensemble de seuils est défini par :

$$seuil\_maxi(j) = \max_k \{seuil_k(j)\} \quad avec\ j \in [1:M]\ \ et\ \ k \in [1:11] \tag{6. 1}$$

Nous avons alors effectué quelques tests avec les bruits industriels Noisex 1 (Nx1), Noisex 2 (Nx2) et Noranda 2 (Nor2). Le tableau XIV présente les résultats obtenus.

Tableau XIV

Performances pour Noisex 1, Noisex 2 et Noranda 2 avec les maxima des seuils

		5dB		10dB		15dB	
Seuils utilisés	**Bruit**	**Rec. de la parole**	**Rec. du bruit**	**Rec. de la parole**	**Rec. du bruit**	**Rec. de la parole**	**Rec. du bruit**
Ceux ajustés pour chaque environnement $seuil_k$	Nx1	82,4%	77,5%	81%	93,8%	87,4%	93,1%
	Nx2	89,8%	80,6%	90%	93,7%	94,3%	91%
	Nor2	93,6%	82,1%	93,4%	95,2%	97,2%	91,8%
$seuil\_maxi(j) = \max_k \{seuil_k(j)\}$ avec $j \in [1:M]$ et $k \in [1:11]$	Nx1	70,5%	88,9%	68,4%	97,9%	73,4%	98%
	Nx2	64,5%	99,4%	67,7%	99,4%	73,2%	98,3%
	Nor2	79,1%	96,7%	76,2%	98,1%	77,5%	99,8%
90%× $seuil\_maxi$	Nx1	75,3%	83,1%	71,6%	96,8%	76,5%	97,1%
	Nx2	67,9%	98,9%	70,5%	99,3%	76,2%	97,9%
	Nor2	85,2%	93,6%	81,6%	96,8%	82%	98,5%
80%× $seuil\_maxi$	Nx1	81,3%	73%	76,1%	94,5%	80,2%	95,1%
	Nx2	72,2%	97%	74%	98,8%	79,7%	97,2%
	Nor2	91%	89,5%	86,8%	95,8%	86,3%	87,9%
70%× $seuil\_maxi$	Nx1	87,6%	59,6%	80,9%	90%	84,6%	91,5%
	Nx2	77,7%	91,6%	78,1%	98,1%	83,5%	96,1%
	Nor2	95,1%	82,9%	92,1%	92,8%	90,4%	94,9%

Comme nous pouvons le voir sur le tableau XIV, le nouvel ensemble correspondant aux maxima des seuils ne permet pas d'obtenir une bonne reconnaissance de l'activité vocale. Ceci est dû au fait que les seuils utilisés pour les nœuds et les étages les plus intéressants propres à chaque environnement sont trop élevés. A contrario, la reconnaissance des régions de bruit est très bonne car les seuils utilisés pour les nœuds et les étages à très faibles pouvoir discriminatoire parole/bruit sont suffisamment élevés pour n'engendrer qu'une très faible erreur « bruit pris pour de la parole ». L'ensemble des maxima des seuils n'est donc pas vraiment adéquat car les seuils sont trop élevés. La partie basse du tableau XIV montre ce qu'il se passe lorsque l'on abaisse ces seuils de manière proportionnelle : la reconnaissance de la parole augmente mais l'identification du bruit seuil diminue (de manière importante dans le cas de Noisex 1). Ceci s'explique encore par le fait que les meilleurs nœuds et étages ainsi que les nœuds et étages très sensibles ne sont, en général, pas les mêmes d'un bruit à l'autre.

Le DAV INNES que nous avons développé dans le cadre de ce projet présente des performances relativement élevées lorsqu'il est amené à fonctionner dans le même environnement que celui dans lequel il a été entraîné. Les tentatives de généralisation de ce système montrent qu'il paraît difficile d'égaler ces performances, lorsqu'un même ensemble de seuils est utilisé pour tous les milieux industriels. Toutefois, une possibilité d'amélioration serait de réunir les 11 bases d'entraînement, voire plus, et d'ajuster les coefficients du DAV sur cette immense base d'entraînement. Ceci permettrait de déterminer les nœuds et les étages à fort et faible pouvoir discriminatoire parole/bruit communs à la plupart des environnements et ainsi de choisir les seuils de manière plus adéquate.

À travers ce chapitre, nous avons pu observer le comportement de notre DAV basé sur les ondelettes. Ses performances sont élevées dans tous les environnements industriels testés et pour toute la plage de RSB visée [5dB – 15dB]. Pour certaines situations, elles sont même excellentes. L'utilisation d'un modèle perceptif de l'ouïe, échelle de Mel, et de la théorie des ondelettes, paquets d'ondelettes, semble donc être intéressante pour la détection d'activité vocale dans les milieux industriels bruités. Nous avons aussi identifié un phénomène dans l'évolution des performances en fonction du RSB. Il procure un avantage important à notre DAV puisqu'il empêche la perte irrémédiable de l'information quand le RSB est trop faible, sa sortie étant alors bloquée sur PAROLE. Nous avons vu que la caractéristique d'énergie est plus propice à la discrimination entre la parole et le bruit lorsque le RSB est élevé alors que celle du Paramètre du Seuil de Johnstone et Silverman est, elle, plus utile quand le RSB est plus petit. Le seuil de Johnstone et Silverman utilisé comme critère de décision permet donc d'augmenter la robustesse de notre DAV INNES. Enfin, l'analyse des résultats pratiques a permis de constater que pour obtenir un fonctionnement accru, il faut entraîner le DAV dans le même environnement que celui dans lequel il sera utilisé et choisir une plage de RSB étroite centrée sur le RSB moyen de ce milieu.

En résumé, le DAV basé sur la théorie des ondelettes développé dans le cadre de ce projet de recherche : le DAV INNES est à notre connaissance original. Son algorithme est tout à fait satisfaisant puisqu'il est efficace dans les milieux industriels bruités. La procédure d'ajustement des règles de décision permet d'adapter au mieux le système à chaque situation, ceci grâce aux recherches des meilleurs nœuds, des meilleurs étages et des meilleurs coefficients correspondants. Le fait qu'elle soit entièrement automatisée rend notre système encore plus intéressant car très facile à utiliser.

Enfin, la comparaison des performances du DAV INNES avec celles du G729.B adapté aux bruits industriels montre selon les tableaux II (p.74) et XI (p.174) que c'est en général le DAV INNES qui présente le meilleur comportement. En plus de cela, ce

dernier fonctionne avec tous les environnements industriels testés et pour toute la plage de RSB [5dB – 15dB]. Dans le cadre de notre étude, il semblerait donc que le DAV INNES soit à privilégier lors d'une utilisation dans les milieux industriels. Pour confirmer cela, il serait toutefois adéquat de faire des tests rigoureux prenant en compte différents ajustements possibles à la fois pour le DAV INNES et le G729.B modifié et de les comparer.

CONCLUSION

Au cours de ce mémoire, nous avons étudié la détection d'activité vocale dans les milieux industriels bruités. Le but de notre travail était de mettre au point un DAV efficace dans ce type d'environnements et pour des rapports signal à bruit compris entre 5dB et 15dB, afin que, par la suite, il puisse être utilisé au sein des bouchons d'oreille « intelligents », système développé par la compagnie SONOMAX. Pour cela, deux approches ont été abordées : l'adaptation d'une méthode existante et la conception d'un DAV à partir de la théorie des ondelettes. Nos recherches nous ont permis de tirer certaines conclusions et finalement d'aboutir à deux algorithmes intéressants.

Première approche

Pour cette première approche, nous avons tout d'abord étudié quelques-unes des méthodes de détection d'activité vocale proposées dans la littérature. Nous avons pris connaissance des méthodes de base et avons vu leur mise en application à travers quatre DAV utilisés dans des codeurs de parole : le Pan-European, le G729.B et l'AMR option 1 et 2. De cette étude, nous avons pu constater que la majorité des procédures existantes ont été développées pour les télécommunications. D'autre part, la comparaison de ces quatre DAV a montré que pour un tel usage le DAV de l'AMR option 1 offre le meilleur compromis entre complexité et robustesse.

Nous n'avons pas pu utiliser cette conclusion pour déterminer notre premier sujet d'étude car il n'est pas possible de prévoir le comportement de ces DAV dans les environnements industriels dont le type de bruits et les rapports signal à bruit sont différents de ceux que l'on rencontre dans les télécommunications. Nous avons donc choisi de nous baser sur le G729.B pour réaliser la première étape de ce projet car c'est

la méthode la mieux documentée et aussi la plus évidente à mettre en œuvre. Compte tenu de son comportement dans notre milieu d'étude, nous avons dû apporter des modifications. À l'aide de nos recherches, nous avons mis au point un ensemble de règles et de coefficients pour trois RSB : 5dB, 10dB et 15dB.

Nous avons obtenu des résultats satisfaisants pour la quasi-totalité des environnements testés. Nous avons pu constater que plus le RSB utilisé pour l'ajustement du DAV est grand, meilleures sont les performances mais plus la plage de RSB de fonctionnement est étroite. Dépendant de l'application, il est donc possible de choisir entre les trois ensembles de règles et de coefficients.

Pour éviter ce choix, une amélioration possible serait d'insérer un estimateur du RSB dans l'algorithme afin de basculer d'un ensemble à l'autre quand cela est nécessaire et ainsi d'utiliser systématiquement le plus adapté à la situation.

Deuxième approche

La deuxième approche a consisté à concevoir un DAV à partir de la théorie des ondelettes. L'algorithme de détection d'activité vocale que nous avons mis au point et qui repose sur ce puissant outil de traitement des signaux a été nommé le DAV INNES. Il intègre un modèle perceptif de l'oreille humaine grâce à l'utilisation de la décomposition en paquets d'ondelettes selon l'échelle de Mel. La prise de décision s'effectue en fonction de l'énergie, caractéristique classique utilisée par les DAV, et du paramètre du seuil de Jonhstone et Silverman. Ce seuil est normalement utilisé dans le débruitage de la parole pour indiquer quand et comment débruiter. Il s'agit, à notre connaissance, de la première fois qu'il est utilisé comme critère de décision. Les premiers résultats pratiques ont montré que ce paramètre est intéressant pour la détection d'activité vocale puisqu'il permet de distinguer la parole bruitée du bruit seul lorsque le

RSB est faible. L'utilisation conjointe de ces deux caractéristiques permet à notre DAV d'être robuste sur [5dB – 15dB] car l'énergie est, elle, plus propice à la discrimination parole/bruit lorsque le RSB est plus grand. Il est à noter que, comme pour le G729.B de la première approche, le DAV INNES a l'avantage de laisser passer tout le signal lorsque le RSB est trop faible pour obtenir un fonctionnement correct. Ceci évite des pertes conséquentes dans le signal de parole.

Afin d'obtenir des performances élevées pour chaque environnement industriel, nous avons mis au point une procédure automatique d'ajustement du DAV. À l'aide d'une recherche des nœuds et des étages à fort pouvoir discriminatoire parole/bruit, elle détermine les règles de décision les plus adaptées à la situation. L'automatisation de cette procédure rend notre système très facile à utiliser. Les résultats expérimentaux ont montré que lorsque l'ajustement s'effectue sur [5dB – 15dB], les performances du DAV INNES, basé sur les ondelettes, sont très satisfaisantes pour tous les milieux industriels testés et pour toute la plage de RSB visée. Nos objectifs ont donc été atteints. Nous avons aussi constaté que la plage de RSB de fonctionnement de ce DAV est proche de celle utilisée pour l'entraînement et que plus cette dernière est étroite, meilleures sont les performances.

Du point de vue de la mise en œuvre pratique, nous recommandons donc d'entraîner le DAV INNES dans le même environnement que celui dans lequel il fonctionnera et pour un intervalle de RSB étroit centré sur le rapport signal à bruit moyen rencontré, ceci dans le but d'atteindre des performances accrues

Comparaison des deux approches

À la fin du chapitre 6, nous avons fait une comparaison succincte des résultats obtenus avec nos deux approches. Bien qu'elle ne soit pas rigoureuse, elle semble indiquer que le DAV INNES est à privilégier dans le cas d'une utilisation dans les milieux industriels.

Travaux futurs

Compte tenu de cette comparaison, les travaux futurs concernant la détection d'activité vocale dans des environnements industriels bruités devraient porter sur l'approche des ondelettes. Nous pourrions améliorer notre système en l'ajustant sur plusieurs environnements différents afin d'obtenir un ensemble unique de règles et de coefficients efficace dans toutes les situations. Nous pourrions aussi chercher un autre algorithme. Par exemple, nous pourrions conserver la décomposition en paquets d'ondelettes selon l'échelle de Mel ainsi que les deux caractéristiques : énergie et PSJS et chercher une signature significative de la parole ou du bruit. Il s'agirait donc d'examiner plusieurs nœuds à la fois au lieu d'un seul. En gardant l'approche de la reconnaissance de formes, cela reviendrait à visualiser les nuages de parole et de bruit dans un espace à plusieurs dimensions. Enfin, nous pourrions penser à utiliser un autre type d'analyse en ondelettes ou encore d'autres caractéristiques pour la prise de décision.

BIBLIOGRAPHIE

[1] Matras, J. J. (1948). *Le son*. Paris: Que sais-je?

[2] Mc Gill. http://www.lecerveau.mcgill.ca/flash/capsules/outil_bleu21.html (Consulté le 17 Octobre 2005).

[3] Schafer, R. W., & Markel, J. D. (1979). *Speech analysis*. New York: IEEE Press.

[4] http://www.medecine-et-sante.com/anatomie/anatoreille.html (Consulté le 17 Octobre 2005).

[5] Tetschner, W. (1993). *Voice processing*. Boston: Artech House.

[6] Parsons, T. (1986). *Voice and speech processing*. New York: Mc Graw-Hill.

[7] Oppenheim, A. V., & Schafer, R. W. (1989). *Discrete-time signal processing*. New Jersey: Prentice Hall.

[8] Benyassine, A., Shlomot, E., Su, H.-Y., Massaloux, D., Lamblin, C., & Petit, J.-P. (1997). 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. *Communications Magazine, IEEE, 35*(9), pp. 64-73.

[9] Tanyer, S. G., & Ozer, H. (1998). Voice activity detection in nonstationary gaussian noise. *Signal Processing Proceedings, 1998. ICSP '98 Fourth International Conference on*, Vol. 2, pp. 1620-1623.

[10] Rabiner, L. R., & Sambur, M. R. (1977). Voiced-unvoiced-silence detection using the Itakura LPC distance measure. *Proc. Intl. Conf. Acoust., Sp. and Sig. Proc.*, pp. 323-326.

[11] Tucker, R. (1992). Voice activity detection using a periodicity measure. *Communications, Speech and Vision, IEE Proceedings I, 139*(4), pp. 377-380.

[12] Haigh, J. A., & Mason, J. S. (1993). Robust voice activity detection using cepstral features. *TENCON '93. Proceedings. Computer, Communication, Control and Power Engineering.1993 IEEE Region 10 Conference on*, Vol. 3, pp. 321-324.

[13] Yang, S., Li, Z.-G., & Chen, Y.-Q. (2003). A fractal based voice activity detector for Internet telephone. *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on,* Vol. 1, pp. I-808- I-811.

[14] Doukas, N., Stathaki, T., & Naylor, P. (1997). Stability of a voice activity detector based on source separation. *Digital Signal Processing Proceedings, 1997. DSP 97, 1997 13th International Conference on*, Vol. 2, pp. 749-752.

[15] Sohn, J., Kim, N. S., & Sung, W. (1999). A statistical model-based voice activity detection. *Signal Processing Letters, IEEE, 6*(1), pp. 1-3.

[16] Cho, Y. D., Al-Naimi, K., & Kondoz, A. (2001). Improved voice activity detection based on a smoothed statistical likelihood ratio. *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, Vol. 2, pp. 737-740.

[17] Tian, Y., Wu, J., Wang, Z., & Lu, D. (2003). Fuzzy clustering and Bayesian information criterion based threshold estimation for robust voice activity detection. *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, Vol. 1, pp. I-444- I-447.

[18] Freeman, D. K., Cosier, G., Southcott, C. B., & Boyd, I. (1989). The voice activity detector for the Pan-European digital cellular mobile telephone service. *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pp. 369-372.

[19] ITU-T. (1996). *Annex B: A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70* ,Recommendation G.729B : ITU-T.

[20] ETSI. (1998). *Digital cellular telecommunications system (Phase 2+); Voice Activity Detector (VAD) for adaptative Multi-Rate (AMR) speech traffic channels; General description*, GSM, 06.94 version 7.1.1 release 1998 : ETSI.

[21] Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *Acoustics, Speech, and Signal Processing, IEEE Transactions on, 23*(1), pp. 67-72.

[22] Hoyt, J. D., & Wechsler, H. (1994). Detection of human speech using hybrid recognition models. *Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on*, Vol. 2, pp. 330-333.

[23] Hoyt, J. D., & Wechsler, H. (1994). RBF Models for detection of human speech in structured noise. *IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on*, Vol. 7, pp. 4493-4496.

[24] Sangwan, A., Chiranth, M. C., Jamadagni, H. S., Sah, R., Venkatesha Prasad, R., & Gaurav, V. (2002). VAD techniques for real-time speech transmission on the Internet. *High Speed Networks and Multimedia Communications 5th IEEE International Conference on*, pp. 46-50.

[25] Venkatesha Prasad, R., Sangwan, A., Jamadagni, H. S., Chiranth, M. C., Sah, R., & Gaurav, V. (2002). Comparison of voice activity detection algorithms for VoIP. *Computers and Communications, 2002. Proceedings. ISCC 2002. Seventh International Symposium on*, pp. 530-535.

[26] Rabiner, L. R., & Sambur, M. R. (1975). An algorithm for determining the endpoints of isolated utterances. *Bell System Tech. Jour., 54*(2), pp. 297-315.

[27] Irwin, M. J. (1980). Periodicity estimation in the presence of noise. *Inst. Acoust. Conf.*, Vol. 3, pp. 213-221.

[28] Friedman, D. (1977). Pseudo-maximum-likelihood speech pitch extraction. *Acoustics, Speech, and Signal Processing, IEEE Transactions on, 25*(3), pp. 213-221.

[29] Barrett, P. A. (1995). Information tone handling in the half-rate GSM voice activity detector. *Communications, 1995. ICC 95 Seattle, Gateway to Globalization, 1995 IEEE International Conference on,* Vol. 1, pp. 72-76.

[30] El-Maleh, K., & Kabal, P. (1997). Comparison of voice activity detection algorithms for wireless personal communications systems. *CCECE '97. Canadian Conference on Electrical and Computer Engineering. Engineering Innovation: Voyage of Discovery. Conference Proceedings, 25-28 May 1997*, vol.2, pp. 470-473.

[31] Garner, N. R., Barrett, P. A., Howard, D. M., & Tyrrell, A. M. (1997). Robust noise detection for speech detection and enhancement. *Electronics Letters, 33*(4), pp. 270-271.

[32] Vähätalo, A., & Johansson, I. (1999). Voice activity detection for GSM adaptive multi-rate codec. *Speech Coding Proceedings, 1999 IEEE Workshop on*, pp. 55-57.

[33] Cornu, E., Sheikhzadeh, H., Brennan, R. L., Abutalebi, H. R., Tam, E. C. Y., Iles, P., et al. (2003). ETSI AMR-2 VAD: evaluation and ultra low-resource implementation. M*ultimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on,* Vol. 2, pp. II-841-II-844.

[34] Beritelli, F., Casale, S., & Ruggeri, G. (2001). Performance evaluation and comparison of ITU-T/ETSI voice activity detectors. *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, Vol. 3, pp. 1425-1428.

[35] Beritelli, F., Casale, S., & Cavallaero, A. (1998). A robust voice activity detector for wireless communications using soft computing. *Selected Areas in Communications, IEEE Journal on, 16*(9), pp. 1818-1829.

[36] Beritelli, F., Casale, S., Rugerri, G., & Serrano, S. (2002). Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors. *Signal Processing Letters, IEEE, 9*(3), pp. 88-85.

[37] Beritelli, F., Casale, S., & Rugerri, G. (2000). A psychoacoustic auditory model to evaluate the performance of a voice activity detector. *Signal Processing Proceedings, 2000. WCCC-ICSP 2000. 5th International Conference on*, Vol. 2, pp. 807-810.

[38] ITU-T. (1996). *Coding of speech at 8kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)*, Recommendation G.729 : ITU-T.

[39] Alcatel. (2002). *Brevet Européen EP 1 267 325 A1 Procédé pour détecter l'activité vocale dans un signal, et codeur de signal vocal comportant un dispositif pour la mise en oeuvre de ce procédé.*

[40] Haar, A. (1910). Zur theorie der orthogonalen funktionen-systeme. *Math. Annal., 69*, pp. 331-371.

[41] Grossman, A., & Morlet, J. (1984). Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM J. Math. Anal., 15*, pp. 723-736.

[42] Vetterli, M., & Kovacevic, J. (1995). *Wavelets and subband coding*. New Jersey : Prentice hall.

[43] Gabor, D. (1946). Theory of communication. *J. of the IEE, 93*, pp. 429-457.

[44] Truchetet, F. (1998). *Ondelettes pour le signal numérique*. Paris : Hermes.

[45] Akansu, A. N., & Haddad, R. A. (2001). *Multiresolution signal decomposition*. San diego : Academic press.

[46] Mallat, S. (2000). *Une exploration des signaux en ondelettes*. Palaiseau : Les éditions de l'école polythechnique.

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

[48] Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *Information Theory, IEEE Transactions on, 36*(5), pp. 961-1005.

[49] Chui, C. K., L., M., & L., P. (1994). *Wavelets: theory, algorithms and applications*. San diego : Academic press.

[50] Dutilleux, P. (1989). An implementation of the algorithme à trous to compute the wavelet transform. *Wavelets: Time-Frequency Methods and Phase Space, Springer IPTI (Berlin)*, pp. 286-297.

[51] Croisier, A., Esteban, D., & Galand, C. (1976). Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques. *International Conference on Information Science and Systems (Patras)*, pp. 443-446.

[52] Chan, Y. T. (1995). *Wavelet basics*. Boston : Klumer Academic Publishers.

[53] Mallat, S. (1997). *Traitement du signal : des ondes planes aux ondelettes*. Paris : Diderot Éditeur.

[54] Stegmann, J., & Schroder, G. (1997). Robust voice-activity detection based on the wavelet transform. *Speech Coding For Telecommunications Proceeding 1997, 1997 IEEE Workshop on,* pp. 99-100.

[55] ETSI. (1996). *GSM enhanced full rate (EFR) speech codec.*

[56] Chen, S.-H. (2002). *A study on speech signal processing using wavelet transforms.* National Cheng Kung University, Tainan.

[57] Chen, S.-H., & Wang, J.-F. (2002). A wavelet-based voice activity detection algorithm in noisy environments. *Electronics, Circuits and Systems, 2002. 9th International Conference on*, Vol. 3, pp. 995-998.

[58] Zwicker, E., & Terhardt, E. (1980). Analytical espressions for citical-band rate and critical bandwith as a function of frequency. *JASA, 68*, pp. 1523-1525.

[59] Kaiser, J. F. (1990). On a simple algorithm to calculate the "energy" of a signal. *Proc. ICASSP'90*, pp. 381-384.

[60] Kacur, J., Frank, J., & Rozinaj, G. (2003). Speech detection in the noisy environment using wavelet transform. *Video/Image Processing and Multimedia Communications, 2003. 4th EURASIP Conference focused on*, Vol. 2, pp. 661-666.

[61] Umesh, S., Cohen, L., & Nelson, D. (2002). Frequency warping and the Mel scale. *Signal Processing Letters, IEEE, 9*(3), pp. 104-107.

[62] Wu, G.-D., & Lin, C.-T. (2000). Word boundary detection with mel-scale frequency bank in noisy environment. *Speech and Audio Processing, IEEE Transactions on, 8*(5), pp. 541-554.

[63] Farooq, O., & Datta, S. (2001). Mel filter-like admissible wavelet packet structure for speech recognition. *Signal Processing Letters, IEEE, 8*(7), pp. 196-198.

[64] Chang, S., Kwon, Y., Yang, S.-i., & Kim, I.-j. (2002). Speech enhancement for non-stationary noise environment by adaptive wavelet packet. *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, Vol. 1, pp. I-561-I-564.

[65] Donoho, D. L. (1995). De-noising by soft-thresholding. *Information Theory, IEEE Transactions on, 41*(3), pp. 613-627.

[66] Jonhstone, D. L., & Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *J. ROY. Statist. Soc. B, 59*, pp. 319-351.