الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي والبحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Projet de Fin d'Études

présenté par

M^{r} DERBAL Abdeladhim

pour l'obtention du diplôme de master en Électronique spécialité réseaux et télécommunications

Thème

Filtrage adaptatif de Wiener pour la réduction de bruit dans les systèmes Multi-microphones

Proposé par : Mr MERAOUBI Hamid

Mr GUESSOUM Abderrezak

Année Universitaire 2011-2012

Remerciements

Ce travail a été effectué à la division d'architecture des systèmes et multimédias du centre de développement des technologies avancés CDTA.

Je tiens à manifester mes sincères remerciements à mon directeur de mémoire Monsieur Meraoubi Hamid de m'a avoir donné la chance d'enrichir mesconnaissances et d'acquérir plus d'expérience en ce domaine de recherche. Ses conseils ont été très précieux pour mener à bien ce travail. L'expérience d'avoir travaillé avec lui m'a été très enrichissante.

J'exprime également mes remerciements à Monsieur Guessoum Abderrezak pour ses remarques pertinentes et son suivi, ce qui a apporté une amélioration à ce travail.

Je tiens à remercier ma famille, pour m'avoir encouragé à tracer ce chemin, et pour leur aide et leurs conseils.

Mes sincères remerciements iront aussi, à Messieurs les membres du jury d'avoir bien voulu examiner et évaluer cette modeste contribution dans ce travail.

Pour finir, un grand merci à tous ceux qui m'ont aidé et encouragés.

ملخص:

في إطار هذه الدراسة, في المرحلة الاولى, قمنا ببعض التوضيحات حول إشكالية معروفة في مجال دراسة و تحليل الصوت, وهي الكشف عن نطاق الكلام, بحيث قمنا بدراسة و تطبيق أربعة خوارزميات في هذا المجال, ومن ثم قمنا بمقارنتها و اختيار الافضل من بينها واستعمالها في النظام الذي نود تطويره, و بعد ذلك قمنا بتطوير نظام لتحسين جودة الصوت , وذلك اعتمادا على طريقتين الاولى هي(خوارزمية إزالة الفصوص الجانبية) و الثانية هي (خوارزمية ترجيح تشوه الصوت- مرشح متعدد القنوات لوينر). و قمنا بدمج هاتين الطريقتين في نظام واحد تحت مسمى خوارزمية (معالجة اولية - خوارزمية ترجيح تشوه الصوت - مرشح متعدد القنوات لوينر), أو باللغة الإنجليزية :

« the Spatially Preprocessed - Speech Distortion Weighted- Multi-channel Wiener Filter » (SP-SDW-MWF).

وتمكنا من الحصول على تحسين جيد لجودة الصوت. ولمزيد من الجودة أضفنا مصفوفة متكيفة للايقاف.

كلمات المفاتيح:

التسجيل, شبكة متعددة الميكروفونات, الصوت, ازالة الضوضاء, مرشح وينر, خوارزمية إزالة الفصوص الجانبية.

Résumé :

Dans le cadre de cette étude, nous avons établi, dans une première partie, quelques éclaircissements sur un problème très souvent rencontré dans le traitement du signal vocal qui est la détection d'activité vocale VAD. Ainsi, nous avons développé et implémenté quatre algorithmes de détection d'activité vocale, dont la comparaison a été illustrée et un choix a été effectué. Nous avons développé, par la suite, un système de rehaussement de la parole utilisant les deux techniques GSC et SDW-MWF intégrées dans un seul système. Ce dernier est nommé « Prétraitement Spatial – Distorsion Pondéré de la Parole – Filtre Multicanal de Wiener » ou en anglais « the Spatially Preprocessed - Speech Distortion Weighted- Multi-channel Wiener Filter » (SP-SDW-MWF). Il est constitué de deux parties, l'une est fixe (Prétraitement Spatial) et l'autre adaptative (SDW-MWF). Nous avons obtenu un bon rehaussement du signal de parole, avec de meilleurs résultats pour la matrice de blocage adaptative.

Mots clés :

Acquisition, Multi-microphones, Parole, Débruitage, Wiener, GSC

Abstract :

In this study, we established in the first part, some light on a problem often encountered in processing the speech signal which is the Voice Activity Detection VAD. Thus, we have developed and implemented four algorithms for voice activity detection, the comparison was illustrated and an election is made. We have developed subsequently, a system of speech enhancement using both techniques SDW-MWF and GSC integrated into one system. The latter is called "the Spatially Preprocessed - Speech-Distortion Weighted Multi-channel Wiener Filter" (SP-SDW-MWF). It consists of two parts, one is fixed (Spatially preprocessing) and the other adaptive (SDW-MWF). We got a great enhancement of the speech signal, with better results for the adaptive blocking matrix.

Keywords :

Acquisition, Multi-microphones, Speech, Denoising, Wiener, GSC

Listes des acronymes et abréviations

ANC : Adaptive Noise Canceller

BFF : Beamformer Fixe

BSD : Bark Spectral Distortion

DA: Direction d'arrivé

DSP :Digital Signal Processor

EQM : Erreur Quadratique Moyenne

ETF : Energie Temps-Fréquence

FFT: Fast Fourrier Transform

FVDS : Formation de Voie par Délais-Somme

G.729 annexe B : codec G.729 a été étendu avec l'annexe B à la compression des silences et la détection d'activité vocale

G.729 : Algorithme de compression de données qui décompose la voix en paquets de 10 millisecondes.

GS: Gradient Stochastique

GSC : Generalized Sidelobe Canceller

GSVD: Generalized Singular Value Decomposition

IFFT: Inverse Fast Fourrier Transform

ITD: Interaural Time Difference

ITU: International Telecommunication Union

LMS: Least Mean Square

MBA: Matrice de Blocage Adaptative

MMSE: Minimum Mean Square Error

MSE: Mean-Square-Error

MWF: Multichannel Wiener Filtering

NI PCI-4472B : est un module d'acquisition de données DAQ

PESQ: Perceptual Evaluation of Speech Quality

PSQM: Perceptual Speech Quality Measure

RIF : Filtre à réponse Impulsionnelle Finie

SDW: Speech Distortion Weighted

segSNR: segmental Signal to Noise Ratio

SNR: Signal to Noise Ratio

SP-SDW-MWF: Spatially Preprocessed - Speech Distortion Weighted- Multi-channel Wiener Filter

TFD : Transformée de Fourier Discrète

TFDI : Transformée de Fourrier Discrète Inverse

VAD: Voice Activity Detection

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Dans le but de réduire les bruits, dans les appareils auditifs numériques, la technique employée vise à privilégier les sons liés à la parole au détriment de ceux qui sont associés au bruit. Elle permet de différencier la parole du bruit à l'aide d'une analyse temporelle et fréquentielle du signal sonore. Le bruit étant constant dans le temps alors que la parole fluctue rapidement. Ainsi, l'amplification des fréquences apparentées au bruit est réduite comparativement à celle du signal associé à la parole. Ceci entraîne une amélioration du ratio signal/bruit de fond et, donc, une meilleure écoute.

La réduction de bruit dans les systèmes multi-microphone est très sensible aux erreurs dans le modèle de signal choisi, l'incompatibilité des microphones (mismatch), la réverbération, etc. [1], [2]. En général l'objectif est la création d'algorithmes de réduction du bruit suffisamment rapide pour réduire le gain durant les pauses de la parole sur une large bande de niveaux de sortie, tout en remplissant le double objectif de maintenir le confort d'écoute sans réduire l'intelligibilité de la parole dans un environnement bruyant. Afin d'estimer le rapport signal/bruit, le système détecte et surveille en permanence les caractéristiques spectrales et temporelles à la fois de la parole et des niveaux de bruit.

Dans un précédent travail [3], nous nous sommes intéressées au développement d'un système de réduction de bruit basé sur le « Beamforming adaptatif ». La solution proposée était à base d'annulateurs de lobe latéraux GSC «Generalized Sidelobe Canceller » [4]. Elle est composée de trois parties :

Une partie qui n'est autre qu'un formateur de voies fixe de type « delay and sum » pour la détection de l'angle d'incidence de la source sonore désirée. Elle est obtenue en recherchant le maximum de puissance dans un espace donné. La seconde partie est une matrice de blocage produisant les signaux de référence de bruit et bloquant le

signal désiré. Alors que la dernière partie utilisant une contrainte de type LMS « Least Mean Square » qui tente d'annuler le bruit dans le formateur de voies fixe de sortie, « Adaptive Noise Canceller ANC ».

Dans ce travail, pour réduire la distorsion de la parole, l'« ANC » est adapté pendant les périodes de bruit seulement [5], [6], [7], [8]. Pour mettre au point une telle solution, on a implémenté 4 algorithmes de détection d'activité vocale (VAD) afin de pouvoir détecter les périodes de silence, ainsi que les zones d'activité vocales. Ce dernier n'étant pas implémenté Norrdholm et alcités plus haut.

Théoriquement, dans le GSC, on considère que la source est bien détectée, et que les caractéristiques et les positions de microphones connus, alors que les réflexions du signal de parole seraient inconnues. Or, si ces hypothèses sont satisfaites, le GSC fournit un signal de parole sans distorsion, dans lequel la quantité de bruit résiduel est réduite au minimum. En réalité ces hypothèses sont souvent violées, entraînant une fuite de la parole et donc une distorsion de cette dernière. Lorsqu'on est en présence de réseaux de petites tailles, comme dans les appareils auditifs, une contrainte de robustesse est nécessaire pour garantir des performances, en présence des erreurs dans le modèle de signal, par exemple, à cause l'incompatibilité des microphones [9] [10], « même lorsque l'ANC est adapté pendant les périodes de bruit seulement ».

Récemment, le filtrage de Wiener multi-canal « MWF », a été proposé de telle sorte que l'estimation de l'erreur quadratique moyenne est minimale dans l'un des signaux de microphones reçus [11], [12]. Le critère d'optimisation de la minimisation de l'erreur quadratique moyenne MMSE (Minimum Mean Square Error) peut être généralisé pour permettre un compromis entre la distorsion de la parole et de réduction de bruit [13]. Nous noterons, cette généralisation comme la distorsion pondérée de la parole « SDW-MWF ».

Ainsi, après avoir développé et choisi le bon VAD pour la discrimination des zones de silence et de parole, nous avons développé un système de rehaussement de la parole utilisant les deux techniques GSC et SDW-MWF intégrées dans un seul système. Ce dernier a été proposé dans [14] et nommé « Prétraitement Spatial – Distorsion Pondéré de la Parole – Filtre Multicanal de Wiener » ou en anglais « the Spatially Preprocessed - Speech Distortion Weighted- Multi-channel Wiener Filter » (SP-SDW-

MWF). Il est constitué de deux parties, l'une est fixe (Prétraitement Spatial) et l'autre adaptative (SDW-MWF). La dernière étape est l'intégration d'une matrice de blocage adaptative pour une meilleure estimation des interférences dans le système proposé.

Pour cela, le chapitre I présente une revue de l'état de l'art dédié aux méthodes multi microphones pour la réduction du bruit, ainsi que des notions théoriques utilisées pour le développement des VADs implémentés, le filtrage de Wiener et un rappel sur le beamforming. Le chapitre II est consacré aux méthodes détection de l'activité vocale. Dans ce chapitre on présente 4 algorithmes dont les résultats ainsi que le choix du VAD utilisé dans ce travail sont illustrés dans le chapitre IV. Le chapitre III, introduit les différentes notions et techniques telles que le bruit, l'acquisition, les différents microphones utilisés ainsi que la problématique de l'utilisation des microphones dans les prothèses auditives. Ceci nous guidera vers la mise en place de l'application. Au chapitre IV, on présente nos implémentations tout en mettant en relief les résultats les plus importants. Enfin, on termine par une conclusion et des perspectives.

Chapitre 1 Etat de l'art et notions de base de traitement de parole

1.1 Introduction

Une grande partie des méthodes de réduction de bruit existantes aborde le problème de la réduction de bruit, sous son aspect fréquentiel, afin de profiter des caractéristiques de ce domaine d'étude pour appliquer de nombreuses méthodes numériques. Parallèlement aux nombreuses variantes de la soustraction spectrale, les méthodes s'inspirant d'un filtrage optimal se sont développées depuis les travaux réalisés par Wiener dès 1949 [15], [16]. Celles-ci prennent en compte souvent la statistique du signal de parole [17]. Ces méthodes appliquent des modifications sur la transformée de Fourier court terme du signal bruité, en utilisant les informations a priori disponibles sur le bruit.

1.2 Etat de l'art : Réduction de bruit multi-microphone

Pour les personnes malentendantes, les algorithmes de réduction du bruit sont essentiels pour améliorer l'intelligibilité de la parole du bruit de fond et / ou la réverbération. Les systèmes multi microphones exploitent plus l'information temporelle et spectrale du signal désiré et du bruit, donc ils sont préférés à ceux utilisant un seul microphone.

Dans [18], [19], [20], un formateur de faisceau adaptatif « Beamformer adaptatif » à large bande a été utilisé, il se compose d'un combinateur linéaire à entrées multiples et une sortie unique, et d'un algorithme adaptatif qui ajuste les poids d'une certaine manière optimale.

Griffiths & Jim [4] ont proposé une solution à base d'annulateurs de lobe latéraux (GSC) et ont reconsidéré la solution avec contraintes proposée par Fost [19], en une solution sans contraintes à base de filtres adaptatifs. Elle est composée de trois parties : Une partie qui n'est autre qu'un formateur de voies fixe de type « Delay and Sum » pour la détection de l'angle d'incidence de la source sonore désirée, basé sur la détection du maximum de puissance. La seconde partie est une matrice de blocage produisant les signaux de référence de bruit en bloquant le signal désiré. Alors que la dernière partie utilise une contrainte de type LMS qui tente d'annuler le bruit dans le formateur de voies fixe de sortie « Adaptive Noise Canceller ANC ». Pour réduire la distorsion de la parole, l'« ANC » est adapté pendant les périodes de bruit seulement [5],[6],[7],[8].

Certains auteurs ont proposé l'utilisation du GSC pour le rehaussement du signal de parole dans un milieu réverbérant. Hoshuyama et al. [21] ont utilisé une structure à 3 blocs similaire au GSC. Cependant la matrice de blocage a été modifiée pour fonctionner de manière adaptative.

Norrdholm et al. [22] ont proposé une solution à base d'une technique « GSC » dont la matrice de blocage est réalisée à l'aide de filtres passe haut, améliorant ainsi la référence du signal de bruit.

Meyer et sydow[23] ont suggéré la constitution du signal de référence du bruit en balayant les lobes d'un « Beamformer » multifaisceaux vers le bruit et la direction du signal désiré séparément. Dans [24] les auteurs ont utilisé cette technique pour l'acquisition et le perfectionnement de la parole dans les environnements bruyants et réverbérant.

Dans [25], elle a été utilisée pour l'annulation d'interférence en radioastronomie, où des réseaux de radiotélescopes sont utilisés pour des observations d'espace lointain.

Dans [26], [27], elle a été dédiée pour la conception d'antennes adaptatives dans des systèmes de communications mobiles, et enfin dans [28], [29], pour le filtrage spatial pour la réduction d'interférence.

Lorsque on est en présence est de réseaux de petite taille, comme dans les appareils auditifs, une contrainte de robustesse est nécessaire pour garantir les performances, en présence des erreurs dans le modèle de signal assumé, par exemple, à cause l'incompatibilité des microphones [9], [10], même lorsque l'ANC est adapté pendant les périodes de bruit seulement. Une méthode largement appliquée consiste à imposer une contrainte d'inégalité quadratique « QIC-GSC en anglais » [30], [6], [8], [31],[32], [33]. Cette méthode est une technique simple et efficace qui impose cette contrainte, mais elle se fait au détriment d'une faible réduction de bruit [9], [10].

Récemment, le filtrage de Wiener multi-canal « MWF », a été proposé de telle sorte que l'estimation de l'erreur quadratique moyenne est minimale dans l'un des signaux de microphones reçus [11]-[12]. Le « MWF » ne pose aucune condition sur le signal pour garantir des performances dans le cas d'un réseau de microphones de petite taille [9], [10], et d'une source de bruit plus complexe (plusieurs sources de bruit, bruit diffus).

Dans [34], un LMS basé sur le « MWF» a été développé. L'algorithme nécessite des enregistrements de signaux d'étalonnage. Dans les salles qui contiennent plusieurs personnes, les caractéristiques des microphones et la localisation de locuteur vont changer tout le temps. Le re-calibrage nécessaire rend cette approche plus lourde et plus coûteuse. Dans [35], l'algorithme du gradient stochastique SDW-MWF proposé ne nécessite pas d'étalonnage [36].

Contrairement à l'«ANC » appliqué dans le GSC, le « MWF » prend en compte les distorsions de la parole dans son critère d'optimisation. Le critère d'optimisation MMSE (Minimum Mean Square Error) peut être généralisé pour permettre un compromis entre la distorsion de la parole et de réduction de bruit [13]. Nous noterons, dans la suite, cette généralisation comme la distorsion pondérée de la parole « SDW-MWF ». La technique « SDW-MWF » est basée uniquement sur les estimations des statistiques de second ordre du signal de parole enregistrée et le signal de bruit. Une détection robuste de la parole est nécessaire, contrairement dans le cas du « GSC ».

Dans [37], [38] l'implémentation récursive de SDW-MWF a été proposée sur la base de la décomposition en valeurs singulières généralisé, « Generalized Singular Value Decomposition (GSVD) » ou en décomposition QR. L'implémentation en sous-bande [39], [12], a un niveau d'intelligibilité de la parole inférieur par rapport à une approche pleine bande.

Dans [40], le GSC et SDW-MWF sont intégrées dans un seul système généralisé, nommé « Prétraitement Spatial – Distorsion Pondéré de la Parole – Filtre Multicanal de Wiener » ou en anglais « the Spatially Preprocessed - Speech Distortion Weighted-Multi-channel Wiener Filter » (SP-SDW-MWF). Constitué de deux parties, l'une est fixe (Prétraitement Spatial) et l'autre adaptative (SDW-MWF).

Le SDR-GSC ou plus généralement le SP-SDW-MWF, ajoute de la robustesse au GSC en prenant la distorsion de la parole explicitement ainsi que le critère de conception de la partie adaptative.

1.3 Beamforming

Le « Beamforming » consiste à combiner les signaux à la sortie des microphones, qui seront convolués à des filtres pondérés optimaux (gain, retard) et additionnés pour obtenir un « faisceau » dans une direction d'intérêt spécifique. Ce faisceau rend le réseau de microphone fortement directif. La direction d'intérêt s'appellera direction de visée. Elle peut être la direction d'une source acoustique dans un environnement bruyant et/ou réverbérant par exemple.

Le « beamforming» peut être fixe ou adaptatif. Les approches de « beamforming» conventionnelles ont pour but de trouver un filtre w complexe linéaire et invariant dans le temps TI (Time Invariant), tel que sa sortie est de la forme:

$$\boldsymbol{y}(t) = \frac{1}{N}\boldsymbol{w}^H\boldsymbol{x}(t) \tag{1.1}$$

Et optimise un critère au second ordre sous d'éventuelles contraintes, où **x**(t) désigne le vecteur des signaux observés en sortie des microphones. Cette sortie correspond ainsi à une estimée au second ordre du signal utile venant d'une direction particulière et potentiellement corrompu par des interférences et d'un bruit de fond.

1.4 Filtrage de Wiener

Le filtrage de Wiener est adéquat pour les situations dans lesquelles le signal ou le bruit sont stationnaires figure 1.1.

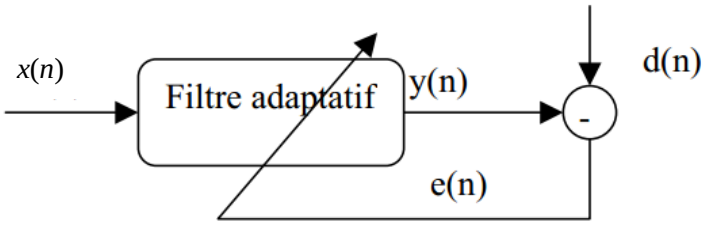


Figure 1. 1. Le schéma principal du filtrage adaptatif

Les applications sont diverses. On dispose d'une entrée ***u*** ainsi que de la réponse désirée ***d*** et l'erreur ***e*** entre la sortie ***y*** et ***d*** sert à contrôler (adapter) les valeurs des coefficients du filtre ***w***. Ce qui différencie essentiellement les applications provient de la façon de définir la réponse désirée ***d***. On peut distinguer quatre grandes classes d'applications : L'identification de systèmes, la prédiction, la modélisation inverse, et l'annulation d'interférences.

• On a un ensemble d'échantillons d'un signal d'entrée {$x(0)$, $x(1)$, $x(2)$, ⋯ } et un ensemble d'échantillons d'une réponse désirée {$d(0)$,$d(1)$, $d(2)$, ⋯}

• Dans la famille des filtres calculant leur sortie selon:

$$y(n) = \sum_{l=0}^{L-1} h_l\, x(n-l), \quad n = 0,1,2, \quad \cdots \tag{1.2}$$

• Trouver les paramètres {h_0, h_1 ,h_2 , ⋯} de telle manière de minimiser l'erreur quadratique moyenne(EQM) ou critère.

$$\boldsymbol{J} = \mathrm{E}\{e^2(n)\} \tag{1.3}$$

Où le signal d'erreur est:

$$e(n) = d(n) - y(n) = d(n) - \sum_{l=0}^{L-1} h_l x(n-l), \tag{1.4}$$

La famille des filtres (1.1) est la famille des filtres linéaires RIF.

C'est plus pratique d'utiliser une notation matricielle pour la sortie du filtre:

$$\mathbf{y}(n) = \sum_{l=0}^{L-1} h_l x(n-l)$$

$$y(n) \quad = \boldsymbol{h}^T \boldsymbol{x}(n) \; = \boldsymbol{x}^T(n)\, \mathbf{h}\,, \tag{1.5}$$

Où

$$\mathbf{h}=[h_0 h_1 \ldots . h_{L-1}]^T$$

est un vecteur de longueur L contenant les coefficients du filtre RIF et

$$\boldsymbol{x}(n) = [\, x(n) \;\; x(n\text{-}1) \;\; \ldots \;\; \ldots x(n\text{-}L\text{+}1)\,]^T$$

est le vecteur des L données d'entrée les plus récentes.

1.4.1 Principe d'orthogonalité

Le vecteur optimum $\mathbf{h_{opt}}$ est celui qui annule le gradient du critère:

$$\frac{\partial J}{\partial h} = 0_{L\times 1} \tag{1.6}$$

On a :

$$\frac{\partial J}{\partial h} = 2\, E\, \{\, e(n)\, \frac{\partial e(n)}{\partial h} \;\}$$

$$\frac{\partial J}{\partial h} = -\,2\, E\, \{\mathbf{e}(n)\, \boldsymbol{x}(n)\}\,. \tag{1.7}$$

Par conséquent, à l'optimum, on a:

$$E\, \{\mathbf{e}_{min}\,(n)\, \mathbf{x}(n)\,\} = 0_{\,L\times 1}, \tag{1.8}$$

Où $\mathbf{e}_{min}(n)$ est l'erreur pour laquelle J est minimisée (filtre optimal).

C'est le principe d'orthogonalité signifiant que toutes les entrées $x(n-l)$, $0 \le l \le L-1$, sont décorréllées de l'erreur $\mathbf{e}_{min}(n)$.

En d'autres termes, le critère *J* atteint son minimum si et seulement si l'erreur $e(n)$ est orthogonale aux échantillons du signal d'entrée $x(n-l)$.

A l'optimum, on a aussi:

$$E\, \{\boldsymbol{e}_{min}(n)\boldsymbol{y}(n)\} = E\, [\, \boldsymbol{e}_{min}(n) \sum_{l=0}^{L-1} \boldsymbol{h}_{opt,l} x(n\text{-}l)\,]$$

$$E\, \{\mathbf{e}_{min}(n) y(n)\} = \sum_{l=0}^{L-1} \boldsymbol{h}_{opt,l}\, E\, [\, \mathbf{e}_{min}(n)\mathbf{x}(n\text{-}l)\,] = 0\,, \tag{1.9}$$

C'est le corollaire du principe d'orthogonalité. $\boldsymbol{h}_{opt,l}$ sont les coefficients du filtre optimal $\boldsymbol{h}_{opt}$:

$$\boldsymbol{h}_{opt} = [\, h_{opt,0} h_{opt,1} \;\; \cdots \; h_{opt,L-1}\,]^T\,.$$

En d'autres termes, quand le critère J atteint son minimum alors l'erreur $\boldsymbol{e}_{\min}(n)$ est orthogonale à la sortie du filtre$\boldsymbol{y}(n)$.

1.4.2 Equation de Wiener-Hopf

Nous savons que pour le filtre optimum $\boldsymbol{h}_{\text{opt}}$, nous avons E $\{\boldsymbol{e}_{\min}(n)\, \boldsymbol{x}(\text{n})\} = 0_{\text{L}\times 1}$. En développant cette équation, nous obtenons:

$$\text{E}\ \{\ \boldsymbol{x}(n)\ [\boldsymbol{d}(n) - \boldsymbol{x}^{\text{T}}(n)\, \boldsymbol{h}_{\text{opt}}]\} = 0_{\text{L}\times 1}\ ,$$

Soit

$$\text{E}\ \{\ \boldsymbol{x}(n)\ \boldsymbol{x}^{\text{T}}(n)\}\ \boldsymbol{h}_{\text{opt}} = \text{E}\ \{\boldsymbol{x}(n)\, \boldsymbol{d}(n)\ \}\ . \qquad (1.10)$$

Ou encore

$\mathbf{R}\boldsymbol{h}_{\text{opt}} = \mathbf{P}$avec solution

$$\boldsymbol{h}_{\text{opt}} = \mathbf{R}^{-1}\mathbf{P}\ . \qquad (1.11)$$

R= E $\{\ \boldsymbol{x}(n)\ \ \boldsymbol{x}^{\text{T}}(n)\}$ est la matrice d'autocorrélation du signal d'entrée **x**(n). Cette matrice est définie positive, de Toeplitz et symétrique. **P**= E $\{\ \boldsymbol{x}(n)\ \boldsymbol{d}(n)\ \}$ est le vecteur d'intercorrélation entre la sortie désirée $\boldsymbol{d}(n)$ et l'entrée **x**(n).

L'équation (1.11) est appelée l'équation de Wiener-Hopf

1.4.3 Surface de l'erreur quadratique moyenne (EQM)

Rappelons que le signal d'erreur est:

$\boldsymbol{e}(n) = \boldsymbol{d}(n) - \boldsymbol{h}^{\text{T}}\boldsymbol{x}(n)$, donc la fonction coût peut encore s'écrire:

$$\boldsymbol{J}\ (\boldsymbol{h}) = \boldsymbol{J} = \text{E}\{\mathbf{e}^2(n)\}$$

$$= \text{E}\{\boldsymbol{d}^2(n)\} - 2\text{E}\{\boldsymbol{d}(n)\boldsymbol{x}^{\text{T}}(\text{n})\}\boldsymbol{h} + \boldsymbol{h}^{\text{T}}\ \text{E}\{\boldsymbol{x}(n)\boldsymbol{x}^{\text{T}}(n)\}\boldsymbol{h}$$

$$\mathbf{J}\ (\text{h}) = \sigma^2\boldsymbol{d} - 2\mathbf{P}^{\text{T}}\boldsymbol{h} + \boldsymbol{h}^{\text{T}}\ \mathbf{R}\boldsymbol{h}, \qquad (1.12)$$

où $\sigma^2\boldsymbol{d}$ = E$\{\mathbf{d}^2(n)\}$ est la variance du signal désiré.

$\boldsymbol{J}\ (\boldsymbol{h})$ est une fonction quadratique des paramètres $\{\ h_0, h_1, \cdots, h_{\text{L}-1}\}$.

$\boldsymbol{J}$ ($\boldsymbol{h}$) est un paraboloïde de dimension L qui a un minimum unique obtenu en prenant le gradient égal à zéro : ∂J (h) /∂h = $0_{L\times1}$.

L'équation (1.12) représente la surface de l'EQM.

1.4.4 Erreur quadratique moyenne (EQM) minimale

La fonction coût s'écrit:

$$\boldsymbol{J}\ (\boldsymbol{h}) = \sigma^2\mathbf{d} - 2\mathbf{P}^{\mathrm{T}}\boldsymbol{h} + \boldsymbol{h}^{\mathrm{T}}\mathbf{R}\boldsymbol{h}. \tag{1.13}$$

A l'optimum, sachant que $\boldsymbol{h}_{\mathrm{opt}} = \mathbf{R}^{-1}\mathbf{P}$, nous avons:

$$\begin{aligned} \boldsymbol{J}_{\min} &= \boldsymbol{J}\ (\boldsymbol{h}_{\mathrm{opt}}) = \sigma^2\boldsymbol{d} - 2\mathbf{P}^{\mathrm{T}}\boldsymbol{h}_{\mathrm{opt}} + \boldsymbol{h}^{\mathrm{T}}{}_{\mathrm{opt}}\mathbf{R}\boldsymbol{h}_{\mathrm{opt}} \\ &= \sigma^2\boldsymbol{d} - \boldsymbol{h}^{\mathrm{T}}{}_{\mathrm{opt}}\mathbf{R}\boldsymbol{h}_{\mathrm{opt}} = \sigma^2\boldsymbol{d} - \mathbf{P}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{P} \\ \boldsymbol{J}_{\min} &= \sigma^2\boldsymbol{d} - \sigma^2\boldsymbol{d}_{\mathrm{opt}}, \end{aligned} \tag{1.14}$$

où $\mathbf{d}_{\mathrm{opt}}(n) = \boldsymbol{h}^{\mathrm{T}}{}_{\mathrm{opt}}\boldsymbol{x}(n)$ est le signal filtré optimal et $\sigma^2\boldsymbol{d}_{\mathrm{opt}} =$ E { $\boldsymbol{d}^2{}_{\mathrm{opt}}$ (n) } la variance de ce signal. Cette relation montre que pour le filtre optimal, l'EQM est la déférence entre la variance du signal désiré et celle de l'estimée de ce signal produite par le filtre. Ainsi, la valeur de l'EQM minimale pour le filtre optimal de Wiener est :

$$\boldsymbol{J}_{\min} = \min \boldsymbol{J}\ (\boldsymbol{h}) = \boldsymbol{J}\ (\boldsymbol{h}_{\mathrm{opt}})\ . \tag{1.15}$$

On a définit l'EQM minimale normalisée comme suit:

$$\widehat{\mathbf{J}}_{\min} = \boldsymbol{J}_{\min} / \sigma^2\boldsymbol{d} \tag{1.16}$$

$$= 1 - \frac{\sigma^2_{\mathrm{d\,opt}}}{\sigma^2_{\mathrm{d}}}\ .$$

L'EQM minimale normalisée satisfait $0 \leq \tilde{\mathrm{J}}\min \leq 1$.

1.4.5 Forme canonique de l'EQM

La fonction coût s'écrit:

$$\boldsymbol{J}(\boldsymbol{h}) = \sigma^2 \boldsymbol{d} - 2\mathbf{P}^T\boldsymbol{h} + \boldsymbol{h}^T\mathbf{R}\boldsymbol{h}. \tag{1.17}$$

En ajoutant et retranchant le terme $\mathbf{P}^T\mathbf{R}^{-1}\mathbf{P}$ dans l'expression précédente, nous avons:

$$\boldsymbol{J}(\boldsymbol{h}) = \sigma^2 \boldsymbol{d} - \mathbf{P}^T\mathbf{R}^{-1}\mathbf{P} + (\boldsymbol{h} - \mathbf{R}^{-1}\mathbf{P})^T\mathbf{R}(\boldsymbol{h} - \mathbf{R}^{-1}\mathbf{P})$$

$$\boldsymbol{J}(\boldsymbol{h}) = \boldsymbol{J}_{\min} + (\mathbf{h} - \boldsymbol{h}_{\text{opt}})^T\mathbf{R}(\boldsymbol{h} - \boldsymbol{h}_{\text{opt}}). \tag{1.18}$$

Soient $\lambda_0, \lambda_1, \cdots, \lambda_{L-1}$ les valeurs propres de la matrice d'autocorrélation **R** et $q_0, q_1, \cdots, q_{L-1}$ les vecteurs propres correspondants, qui satisfont:

$$\mathbf{R}q_l = \lambda l q_l, \qquad l = 0, 1, \cdots, L-1, \tag{1.19}$$

alors la matrice $\mathbf{Q} = [q_0\ q_1 \cdots q_{L-1}]$ peut diagonaliser **R** comme suit:

$$\mathbf{R} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T, \tag{1.20}$$

Où **Λ** est une matrice diagonale.
Utilisant la décomposition précédente dans l'EQM, on a:

$$\boldsymbol{J}(\boldsymbol{h}) = \boldsymbol{J}_{\min} + (\boldsymbol{h} - \boldsymbol{h}_{\text{opt}})^T\mathbf{R}(\boldsymbol{h} - \mathbf{h}_{\text{opt}})$$

$$= \boldsymbol{J}_{\min} + \mathbf{v}^T\boldsymbol{\Lambda}\mathbf{v}$$

$$\boldsymbol{J}(\boldsymbol{h}) = \text{Jmin} + \sum_{l=0}^{L-1} \lambda l v 2 l, \tag{1.21}$$

Où le vecteur

$$\mathbf{v} = \mathbf{Q}^T(\boldsymbol{h}_{\text{opt}} - \boldsymbol{h})$$

$$= [v_0\ v_1 \cdots v_{L-1}]^T$$

L'équation (1.21) est la forme canonique de l'EQM. Cette expression ne contient plus de termes croisés. Comme les valeurs propres $\lambda_l$sont non négatives, il apparait clairement que la surface d'erreur est du type paraboloïde elliptique dans un hyperespace.

1.5 Traitement court-terme du signal vocal

Une simple inspection visuelle de la forme d'onde du signal vocal met en évidence la nature non stationnaire du celui-ci. On peut facilement observer les variations en amplitude ou dans la fréquence. Etant donné la nature non stationnaire du signal vocal, une analyse globale ou à long terme est dans la majorité des cas inefficace.

D'autre part, on possède des moyens très puissants pour l'étude des systèmes linéaires et invariants dans le temps. Dans ces conditions l'hypothèse le plus utilisée dans le traitement de la parole est le fait que les propriétés du signal vocal changent lentement dans le temps [41]. Cette hypothèse conduit vers un traitement à court terme. Les segments du signal vocal sont isolés et traités comme s'ils étaient des fragments composant des sons soutenus et invariants. Pour ces segments, qui d'habitude se chevauchent, on peut faire usage des mêmes outils que dans le cas d'un système linéaires et invariant dans le temps SLIT [42-43].

Du point de vue statistique, le signal vocal pour des segments courts de temps est considéré la réalisation d'un processus aléatoire stationnaire et ergodique. Ces deux propriétés permettent respectivement d'être indépendant d'un décalage temporel et d'identifier les moyennes d'ensembles avec les moyennes temporelles [42].

Le résultat du traitement d'un tel segment peut être un seul numéro ou un set de numéros qui devient une nouvelle représentation du signal vocal. La représentation mathématique de ce processus est décrite par la relation [41] :

$$Q_n = \sum_{-\infty}^{\infty} T\big(x(m)\big)\, w(n-m) \tag{1.22}$$

Le signal vocal est transformé par l'opérateur *T*( ) et le résultat est multiplié par une fenêtre alignée à l'échantillon *n*. D'habitude cette fenêtre contient un nombre fini d'échantillons mais ce n'est pas toujours le cas. La relation (1.22) est le produit de convolution de la fenêtre *w*(*n*) et de la séquence *T*(*x*(*n*)). Donc Q_n peut être vu comme étant la sortie d'un SLIT qui pourrait être un filtre caractérisé par la réponse

impulsionnelle *h*(*n*) = *w*(*n*). Cette interprétation est représentée graphiquement dans la figure 1.2.

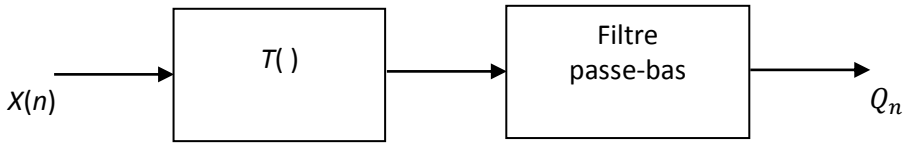


Figure 1. 2. Représentation graphique du principed'analyse court-terme

Pour les choix de la fenêtre on regarde deux aspects : la dimension en nombre d'échantillons et la forme, chaque aspect ayant des répercussions différentes sur les analyses ultérieures. L'effet du fenêtrage peut être mis en évidence par l'étude des propriétés de deux fenêtres représentatives : la fenêtre rectangulaire et la fenêtre de Hamming.

La fenêtre rectangulaire est donnée par [44] :

$$h(n) = \begin{cases} 1 \leq n \leq N-1 \\ 0 \; ailleurs \end{cases} \tag{1.23}$$

Le module de la réponse fréquentielle de la fenêtre rectangulaire est [8] :

$$\left|H(e^{jw})\right| = \frac{\left|\sin\left(\omega\frac{N}{2}\right)\right|}{\left|\sin\left(\omega\frac{1}{2}\right)\right|} \tag{1.24}$$

Le premier zéro dans la relation (1.24) se réalise pour *F*=F_s/N où F_s est la fréquence d'échantillonnage. Cette fréquence *F* délimite la largeur du premier lobe qui caractérise le module de la réponse fréquentielle de la fenêtre rectangulaire, figure 3et qui diminue quand N augmente.

La fenêtre de Hamming est décrite par [8] :

$$h(n) = \begin{cases} 0.54 - 0.46\cos\big(2\pi n(\mathrm{N}-1)\big) & 0 \leq n \leq N-1 \\ 0 & ailleurs \end{cases} \tag{1.25}$$

La largeur du premier lobe du module de la réponse fréquentielle de la fenêtre de Hamming est le double du premier lobe correspondant à la fenêtre rectangulaire de même longueur. D'autre part la fenêtre de Hamming a une plus grande atténuation à l'extérieur du premier lobe par rapport à la fenêtre rectangulaire.

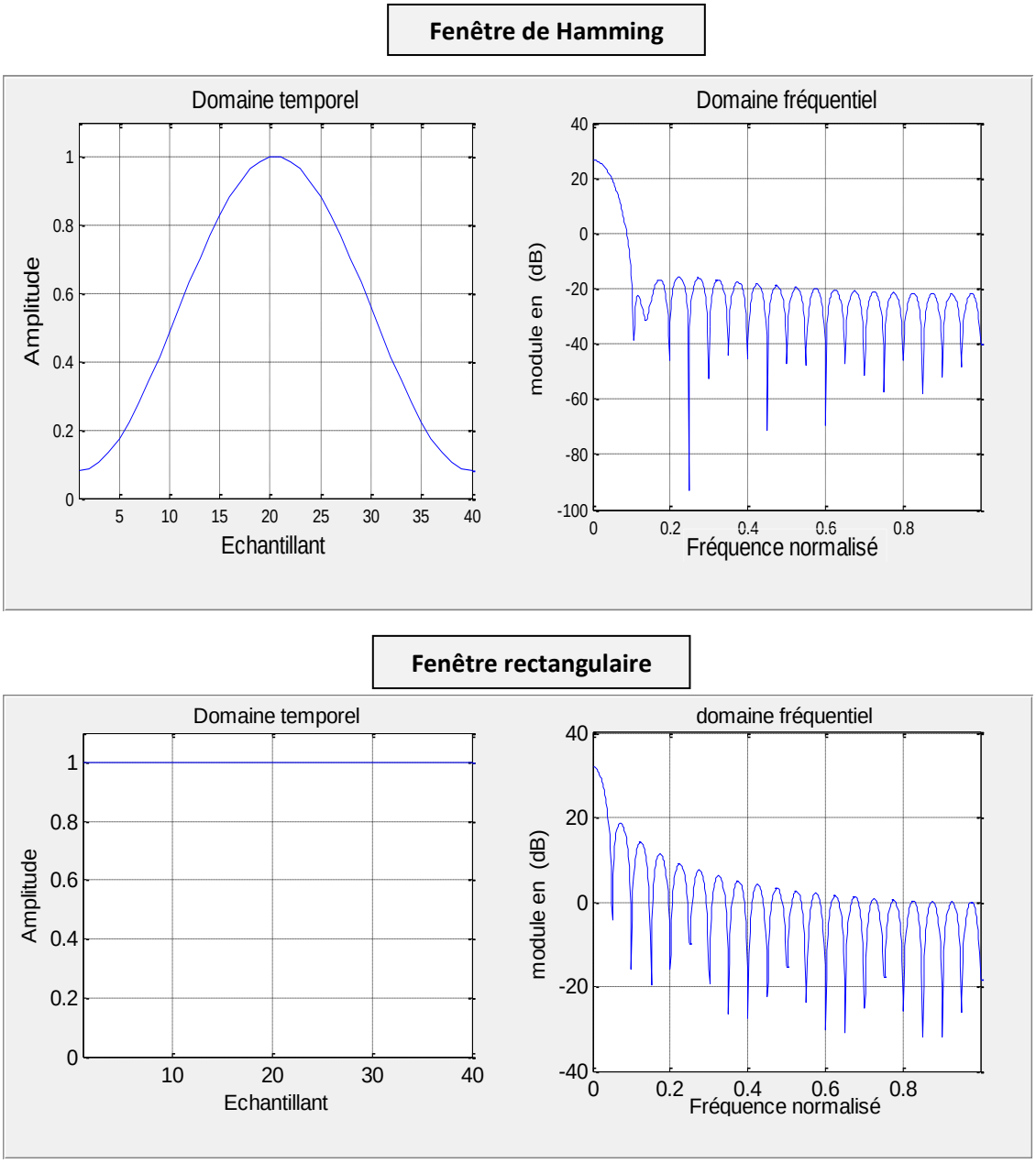


Figure 1. 3. Comportement temporel et fréquentiel pour la fenêtre rectangulaire et la fenêtre de Hamming N = 40

Pour les deux fenêtres l'augmentation de N conduit vers une diminution de la largeur du lobe principal. Donc une meilleure résolution spectrale est obtenue avec une fenêtre plus longue.

Parfois dans la pratique on est obligé d'utiliser une fenêtre assez longue pour surprendre certaines aspects qui caractérisent le signal vocal, par exemples la fréquence fondamentale de la voix d'un homme qui peut nécessiter jusqu'à 200 échantillons à une fréquence d'échantillonnage Fe = 8 kHz. D'autre part une valeur

trop grande pour N se concrétise par une fenêtre trop longue pour laquelle l'hypothèse d'invariance temporaire est violée.

Compte tenu de ces aspects, le choix du N se fait en fonction du problème concret à résoudre. Il est généralement dans la plage de 80 à 160 échantillons pour une fréquence d'échantillonnage Fe= 8 kHz.

1.6 Energie court-terme

On a déjà vu que l'amplitude du signal vocal varie d'une façon importante dans le temps. L'amplitude des régions voisées du signal vocal est généralement plus grande que celle des régions non voisées. L'énergie court-terme est un paramètre qui reflète ces variations d'amplitude dans le signal vocal et elle a été un de premiers paramètres utilisés dans la détection d'activité vocale. La définition de ce paramètre est [41] :

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \tag{1.26}$$

Ou encore

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)\ h(n-m) \tag{1.27}$$

Où

$$h(n) = w^2(n) \tag{1.28}$$

L'équation (1.27) voit l'énergie court-terme comme étant la sortie d'un filtre défini par la réponse impulsionnelle $h(n)$, relation (1.28), excité par le signal d'entrée $x^2(n)$. La figure 1.4représente l'énergie court-terme pour la même phrase utilisant deux fenêtres rectangulaires de longueurs différentes. La longueur de la deuxième fenêtre N = 160 échantillons est le double de la première pour laquelle N = 80 échantillons.

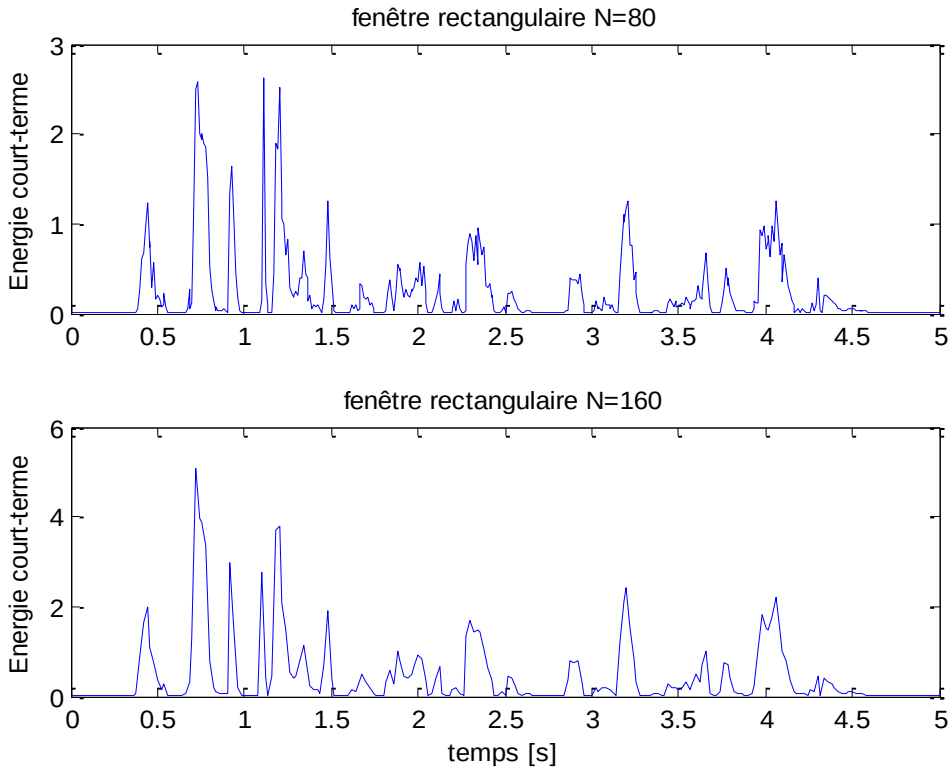


Figure 1. 4. Énergie court-terme pour deux fenêtres rectangulaires de dimensions différentes

On peut voir clairement l'effet de lissage induit par la deuxième fenêtre par rapport à la première.

1.7 Taux de passage par zéro

Le taux de passage par zéro est une estime grossière du contenu fréquentiel du signal analysé. Il est aussi un des premiers paramètres utilisés dans le VAD car la structure spectrale du bruit est différente de celle de la parole.

Pour un signal discret il y un passage par zéro quand deux échantillons successifs ont le signe différent. Ce paramètre est estimé par l'équation :

$$Z_n = \sum_{m=-\infty}^{\infty} \left|\mathrm{sgn}\left[\left(x(m)\right)\right] - \mathrm{sgn}\left[\left(x(m-1)\right)\right] w(n-m)\ \right| \tag{1.29}$$

Où

$$\mathrm{sgn}[x(n)] = \begin{cases} 1 & s(n) \geq 0 \\ -1 & s(n) \leq 0 \end{cases} \tag{1.30}$$

Et

$$w(n) = \begin{cases} \frac{1}{2N} & 0 \leq n \leq N-1 \\ 0 & ailleurs \end{cases} \tag{1.31}$$

On sait que les sons voisés sont caractérisés par une forte composante de basse fréquence, due à l'excitation, et que pour les sons non voisés plus d'énergie est concentrée dans la région de haute fréquence [42]. On s'attend donc d'avoir un taux de passage par zéro moins élevé pour les régions voisées du signal vocal que pour les régions non voisées. En observant des segments de 10 ms, on trouve des distributions gaussiennes avec une moyenne de 14 pour les sons voisés et une moyenne de 49 pour les sons non voisés, ces deux répartitions se recouvrent partiellement [45].

Un exemple pour ce paramètre est représenté dans la figure 1.5 où l'on a utilisé une fenêtre de 10 ms. Dans le cas d'une utilisation pratique de ce paramètre il faut diminuer au maximum certains types de bruit qui ont un impact très important sur le résultat. Par exemple si le convertisseur analogue numérique introduit un décalage par rapport à la valeur zéro, comparable avec l'amplitude du signal, le taux de passage par zéro devient nul ou très petit. Cet effet est plus important pour les régions de silence ou non voisées du signal vocal. Il est préférable d'utiliser un filtre passe-bande capable d'éliminer les bruits de base fréquence au lieu du filtre passe-bas anti-recouvrement.

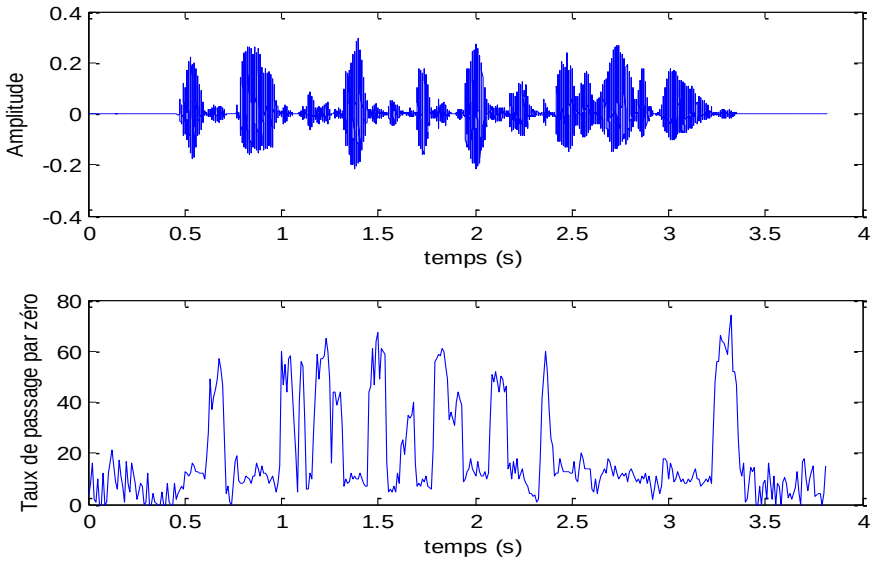


Figure 1. 5. Taux de passage par zéro en utilisant une fenêtre de 10 ms

1.8 Lissage médian et filtrage linéaire

Une technique largement utilisée pour éliminer le bruit dans un signal est le filtrage linéaire. Dépendant du type de donnée utilisée, cette technique ne donne pas toujours les meilleurs résultats. Un exemple est le cas de signaux caractérisés par des points de discontinuité qu'il faut préserver mais aussi des points largement erronés qu'il faut éliminer. Bien qu'un algorithme idéal pour résoudre ce problème n'existe pas, certaines techniques de lisage qui utilisent une combinaison des moyennes temporaires et filtrage linéaire donne de bons résultats.

On utilise cette technique pour réduire la variance des paramètres employés dans l'analyse court-terme pour décrire le signal vocal. Dans le cas spécifique de la détection d'activité vocale cette approche est justifiée car les régions de parole et de silence sont des régions compactes formées des plusieurs trames.

Dans le cas de filtrage linéaire le signal est vu comme une somme pondérée de sinusoïdes et l'effet du filtrage est de modifier les amplitudes de ces sinusoïdes. Dans le cas du lisage non linéaire il est plus utile de voir le signal comme étant formé d'une composante lisse et d'une autre bruyante. Une description analytique pour ce signal est de la forme [41] :

$$x(n) = S[x(n)] + R\,[x(n)] \tag{1.32}$$

Où $S[x(n)]$ est la partie lisse est $R[x(n)]$ est la partie bruyante du signal. Une technique non linéaire qui est capable de séparer le deux composantes S et R est le lisage médian. Ceci est la moyenne des valeurs comprises dans la fenêtre alignée au point n. Le lisage médian préserve les discontinuités qui se manifestent dans le signal sur une durée plus grande que la longueur de la fenêtre L et suit la tendance générale du signal mais n'est pas assez efficace en terme d'élimination de la composante bruyante du signal.

Un bon compromis est assuré par une combinaison de lisage médian et filtrage linéaire comme dans la figure 1.6(a). Le rôle du filtrage linéaire est d'éliminer la partie bruyante qui reste après le lisage médian. On utilise d'habitude un filtre de type FIR à coefficients symétriques.

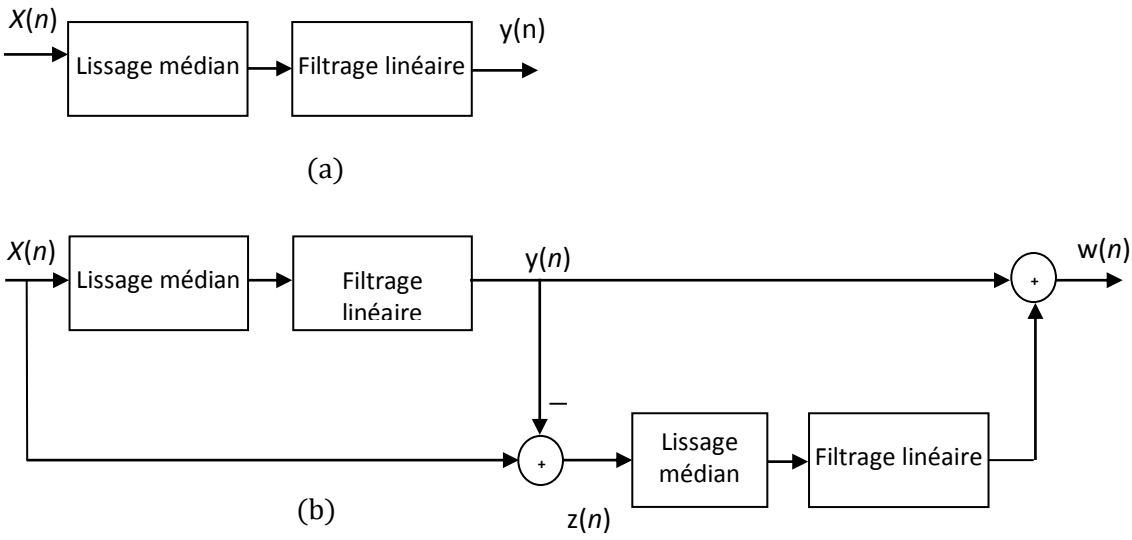


Figure 1. 6. Schéma bloc d'un système de lisage non linéaire

Dans la figure 1.6.(a) le signal y(*n*) est une approximation de *S*[*x*(*n*)]. Une réalisation pratique est suggérée dans la figure 1.6.(b) ou :

$$y(n) = S[x(n)] \tag{1.33}$$

et alors

$$z(n) = x(n) - y(n) = R[x(n)] \tag{1.34}$$

Dans la figure 1.6 (b) le signal *z*(*n*) est à son tour lissé et additionné au *y*(*n*) pour obtenir une meilleure approximation de *S*[*x*(*n*)]. Le signal *w*(*n*) satisfait la relation :

$$w(n) = S[x(n)] + S\big[R[x(n)]\big] \tag{1.35}$$

Dans le cas d'un lissage idéal la quantité z(n) serait juste la partie bruyante du signal pour laquelle la valeur *S*[*R*[*x*(*n*)]] est nulle et donc la correction inutile.

Il faut souligner que le lisage linéaire introduit un retard de (*L*-1)/2 échantillons ou *L* est la longueur de la fenêtre utilisée.

1.9 Transformée de Fourier court-terme

La représentation des signaux par une somme d'exponentiels complexes ou des sinusoïdes est connue sous le nom de transformée de Fourier. Du à ses propriétés, cette transformée est un outil largement utilisé dans le traitement de signal [46-47].

Puisque le signal vocal peut être considéré stationnaire pour des intervalles courts de temps, il est utile d'introduire la notion de transformée de Fourier court-terme. Cette

extension logique de la transformée de Fourier classique d'un signal discret introduit pour surprendre les variations temporelles du spectre du signal vocal nouvelle notion est une et elle est définie par [42,43,48] :

$$X_n\left(e^{j\omega}\right) = \sum_{m=-\infty}^{\infty} w(n-m)x(m)e^{-i\omega m} \tag{1.36}$$

Cette représentation du signal vocal est une fonction de deux variables : l'indice du temps *n* qui prend des valeurs discrètes et la pulsation ω qui est une variable continue. Cette équation peut être interprétée de deux façons différentes. Premièrement pour *n* fixe $X_n(e^{j\omega})$est la transformée de Fourier classique de la séquence *w*(*n*-*m*)*x*(*m*) et donc en possède toutes les propriétés. L'existence de cette transformée est assurée car la quantité *w*(*n*-*m*)*x*(*m*) est toujours absolument sommable dans le cas d'une fenêtre w finie en temps. La fenêtre w est utilisée pour délimiter les segments du signal vocal. De plus on peut retrouver la valeur x(n) par l'intermédiaire de la transformée de Fourier inverse [44-47] :

$$x(n) = \frac{1}{2\pi w(0)} \int_{-\pi}^{\pi} X_n(e^{j\omega})e^{j\omega n}\, d\omega \tag{1.37}$$

À condition que w(0) ne soit pas nul.

Deuxièmement pour ω fixe$X_n(e^{j\omega})$ est écrit sous la forme d'une convolution et donc on peut interpréter la relation (1.36) en terme de filtrage linéaire.

Un facteur important dans le calcul de la transformée de Fourier court-terme est la fenêtre utilisée. On suppose que les transformées de Fourier suivantes existent :

$$X_n\left(e^{j\omega}\right) = \sum_{m=-\infty}^{\infty} x(m)e^{-i\omega m} \tag{1.38}$$

et

$$W_n\left(e^{j\omega}\right) = \sum_{m=-\infty}^{\infty} w(m)e^{-i\omega m} \tag{1.39}$$

alors la transformé de Fourier de *w*(*n*-*m*)*x*(*m*) pour *n* fixe est la convolution entre les transformées de *w*(*n*-*m*) et *x*(*m*). La transformée de *w*(*n*-*m*) est $w(m)e^{-i\omega m}$et donc [41]

$$X_n(e^{j\omega}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} w(e^{j\theta})e^{-i\theta n}X(e^{j(\omega+\theta)})d\theta \quad (1.40)$$

L'équation précédente est correcte dans le cas où on supposerait que $X_n(e^{j\omega})$est la transformée de Fourier d'un signal qui a le même spectre ou qui est nul à l'extérieur de la fenêtre étudiée. Dans ces conditions la transformée de Fourier court-terme peut être vue comme une version lisse du spectre idéal du signal. Le lissage induit dans le spectre du signal est la conséquence de l'utilisation du filtre médian résultant de la transformée de Fourier de la fenêtre temporelle *w*. L'effet de la fenêtre devient clair maintenant, la meilleure fenêtre temporaire du point de vue de la résolution est celle qui s'approche le plus possible d'une impulsion dans le domaine fréquentiel.

Comme on l'a déjà vu au début de ce chapitre, la largeur du lobe central de $W(e^{j\omega})$ est en rapport inverse avec la largeur de la fenêtre lorsque l'amplitude des lobes latéraux est indépendante de la dimension de la fenêtre. On peut diminuer l'amplitude des lobes latéraux au détriment de l'épaisseur du lobe central par un choix judicieux de la forme de la fenêtre. De ce point de vue la figure 1.3 montre un meilleur comportement de la fenêtre de Hamming par rapport à la fenêtre rectangulaire.

1.10 Conclusion

Dans ce chapitre, nous nous sommes intéressés à l'état de l'art concernant le débruitage multi microphones en général et celui adopter récemment par la communauté scientifique dans le cas de prothèses auditives. Aussi, nous avons introduit des notions théoriques, concernant les détecteurs d'activités vocales, le filtrage de « wiener », ainsi que le « beamforming ». Toutes ces notions sont utilisées dans les chapitres suivants et sont utiles à la compréhension des différents concepts utilisés.

Chapitre 2 Etude théorique des détecteurs d'activité vocales

2.1 Introduction

Un problème très souvent rencontré dans le traitement du signal vocal est la détection d'activité vocale VAD. Autrement dit, il faut discriminer entre les régions où la parole est présente et les régions où la parole est absente dans le signal analysé. Un algorithme de détection d'activité vocale fonctionne d'après une logique binaire. Il produit la valeur logique 1 ou 0 pour chaque segment de signal analysé, indiquant respectivement la présence ou l'absence de la parole.

Le VAD est un module important dans une large gamme d'applications concernant le traitement de la parole soit la reconnaissance, la transmission ou le rehaussement de la parole.

Dans le domaine de reconnaissance de la parole le VAD est utilisé pour localiser le début et la fin des régions à reconnaître. La précision du VAD utilisé se matérialise dans une amélioration du taux de reconnaissance.

Pour les systèmes de transmission de la parole telle que la téléphonie cellulaire, le taux d'activité vocale est en moyenne de 40%, donc 60% du temps le système serait inutilisé [49]. Le VAD est utilisé pour contrôler la transmission discontinue qui active la transmission uniquement pendant les périodes d'activité vocale. La transmission discontinue permet d'augmenter la capacité du système pour l'opérateur et pour l'abonné prolonge l'autonomie du mobile [49].

2.2 VAD basé sur l'énergie courte-terme et le taux de passage par zéro

L'idée de base de cet algorithme[50]est d'utiliser l'estimée de l'énergie court terme Mn comme un paramètre robuste pour découvrir les régions voisées. La décision est raffinée par la suite pour inclure les régions non voisées à l'aide d'un deuxième paramètre, le taux de passage par zéro Z_n.

Pour cela l'analyse est basée sur une trame de 10 ms. Les 100 premières ms sont considérées justes du bruit et sont utilisées pour calculer la moyenne $\overline{\text{IZC}}$et l'écart type δ_{IZC}du taux de passage par zéro, et la moyenne de l'estimateur de l'énergie court-terme IMN. Le seuil minime de passages par zéro pour les régions non voisées est choisi d'après la relation :

$$\text{IZTC} = \text{MIN}\ (\text{IF}, \overline{\text{IZC}} + 2\delta_{\text{IZC}}) \qquad (2.1)$$

$$\text{IF} = 0.25 \text{ longueur de w}$$

IF est une constante, dépendante de la longueur de la fenêtre d'analyse en nombre d'échantillons. On définit encore deux seuils d'énergie ITL et ITU :

$$\begin{aligned} I1 &= 0.03(\text{IMX} - \text{IMN}) + \text{IMN} \\ I2 &= 4.\text{IMN} \\ \text{ITL} &= \text{MIN}(I_1, I_2) \\ \text{ITU} &= 5\text{ITL} \end{aligned} \qquad (2.2)$$

Ou IMX est le maximum de l'énergie court terme pour le signal entier.

L'algorithme cherche tout d'abord les régions pour lesquelles l'énergie dépasse la valeur ITL et après ITU sans retomber au-dessous de ITL. Cela donne un premier indice N_1 du debout de l'activité vocale. Un premier indice de la fin de l'activité vocale N_2 est quand l'énergie court terme devient inférieure à ITL après avoir dépassé ITU.

A partir de N_1 on se déplace 250 ms vers le début du signal et on regarde cette fois le taux de passage par zéro. Si Z_n dépasse trois fois ou plus le seuil IZTC le point N_d où IZTC a été dépassé pour la dernière fois est le début de l'activité vocale si non la décision initiale reste inchangée.

La même procédure s'exécute à partir de N_2 vers la fin du signal pour trouver l'indice final $N_f$de la fin de l'activité vocale.

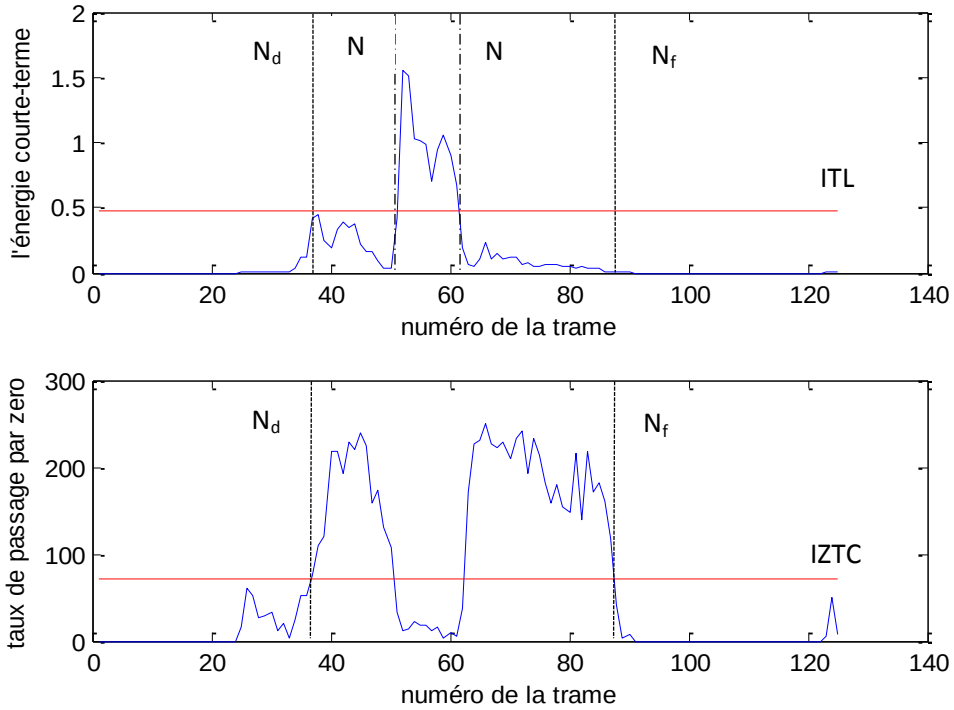


Figure 2. 1. VAD basé sur l'énergie court-terme et le taux de passage par zéro

La figure 2.1est un exemple obtenu avec cet algorithme pour le mot six qui est un exemple de mot qui commence et se termine avec un son non voisé. L'intervalle [N_1, N_2] correspondant à la voyelle / i / est découvert par l'algorithme grâce au paramètre M_n. La décision finale [N_d, N_f] qui inclut les deux consonnes / s / au début et à la fin du mot est prise en se basant sur le deuxième paramètre Z_n. Cet algorithme se comporte bien pour un RBS supérieur à 30 dB quand le paramètre Z_n est fiable, pour des RSB plus petits, la valeur de Z_n est vite corrompue par le bruit et conduit vers des conclusions erronées. De plus, l'algorithme ne s'adapte pas à l'évolution du bruit, les valeurs des paramètres qui caractérisent le bruit déterminé au début de la période restent inchangées pour tout l'intervalle d'analyse.

2.3 VAD basé sur un filtrage optimal de l'énergie court-terme

2.3.1 Conception du filtre optimal

Cet algorithme [51]est inspiré par une des techniques utilisées pour la détection du contour dans le traitement des images [52]. La procédure est similaire au lissage médian mais cette fois les caractéristiques du filtre f(y) utilisé sont optimisées pour :

- éliminer les effets du bruit
- être capable de détecter les débuts et les fins des régions d'intérêt
- avoir une longueur finie
- avoir le niveau de la réponse finie
- avoir une réponse maximale et précise dans le cas d'un changement dans
- l'évolution du paramètre utilisé
- minimiser la probabilité de fausse alarme

Suite à ces conditions on obtient un filtre antisymétrique de longueur finie qui décroît vers zéro aux extrémités. On utilise la méthode de multiplicateurs de Lagrange pour trouver les paramètres du filtre optimal. La solution est donnée par la relation (2.3) et développée dans [53] :

$$f(y)= e^{Ay}[\, K_1\sin(Ay) + K_2\cos(Ay)] + e^{-Ay}\,[\, K_3\sin(Ay) + K_4\cos(Ay)] + K_5 + K_6\, e^{sy} \qquad (2.3)$$

Où A = 0.41s.

Pour une longueur du filtre W = 7 et s = 1 on trouve K = [1.583, 1.468, -0.078,-0.036, -0.872, -0.56] [30]. Les coefficients du filtre sont obtenus en utilisant l'équation (2.4) :

$$h\,(i) = \{- f\,(\,-W \le i \le 0),\, f\,(\,1 \le i \le W\,)\} \qquad (2.4)$$

Pour la détection d'activité vocale une valeur W = 13 donne de bons résultats, le filtre optimal utilisé est présenté dans la figure 2.2.

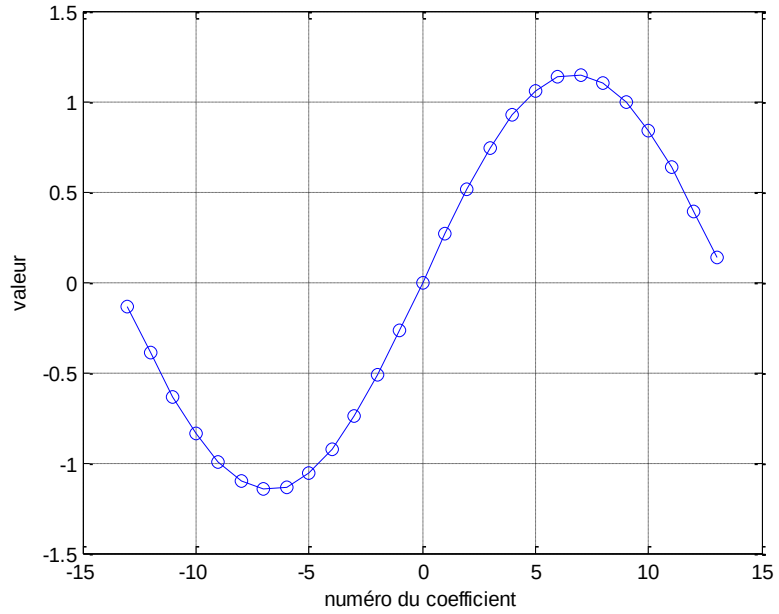


Figure 2. 2. Filtre optimal W= 13

Le paramètre utilisé est le logarithme de l'énergie court terme et le détecteur d'activité vocale fonctionne comme dans le cas du filtrage médian, pour chaque trame on a :

$$F(n) = \sum_{i=0}^{2W} h(i - W)\log_{10} E\ (n + i) \qquad (2.5)$$

2.3.2 Algorithme de décision

La valeur *F*(*n*) doit être comparée à des seuils prédéterminés et évalués par un diagramme à trois états pour obtenir la décision finale. Le diagramme de décision emploie trois états silence, parole et quitter parole et il est représenté dans la figure 2.3. Au début on se trouve dans l'état silence pour *F*(1). Les entrées dans le diagramme sont les valeurs de *F*(*n*) et les sorties sont les points de début et de fin de l'activité vocale. Le compteur est un compteur de trames, *Tu* et *TL* sont deux seuils *Tu*>*Tl*, et l'écart est un nombre entier qui indique le nombre de trames nécessaires pour passer dans l'état silence après la détection d'un point de fin de l'activité vocale. Les conditions de transition sont marquées sur le diagramme à côté des flèches indiquant la transition et les actions sont entre parenthèses. Le fonctionnement du diagramme est expliqué à l'aide d'un exemple. La figure 2.4 partie (A) présente l'énergie court terme du mot six et la partie (B) la sortie *F*(*n*) du filtre optimal.

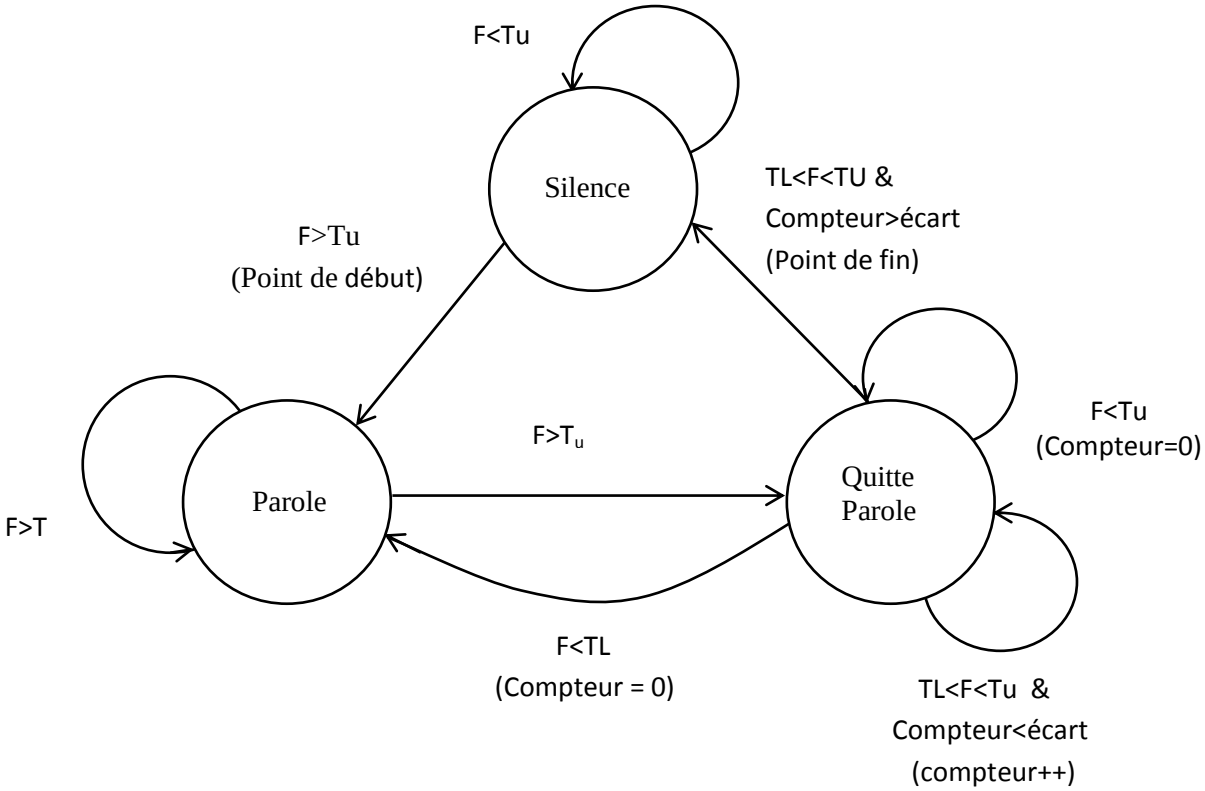


Figure 2. 3. Diagramme de décision à trois états

Le diagramme des états reste dans l'état silence jusqu'au point A où $F(n)>T_u$ et un point de début d'activité vocale est détecté, l'état courant devient parole. Quand $F(n)<T_L$ au point B, l'état courant devient quitte parole et le compteur est tenu à zéro tant que $F(n)$ ne dépasse pas T_L. Au point C, $F(n)>T_L$, et le compteur est incrémenté tant que $T_u>F(n)>T_L$. Si l'écart est dépassé par la valeur du compteur, un point de fin d'activité vocale est détecté est le diagramme des états revient à l'état de départ. La valeur de l'écart est choisie égale à 30 et correspond à la période de descendant de l'énergie avant d'arriver à un point de fin d'activité vocale. Les valeurs de T_L et T_u sont choisies empiriquement, à partir de quelques exemples, l'algorithme est assez stable par rapport aux valeurs de T_L et T_u.

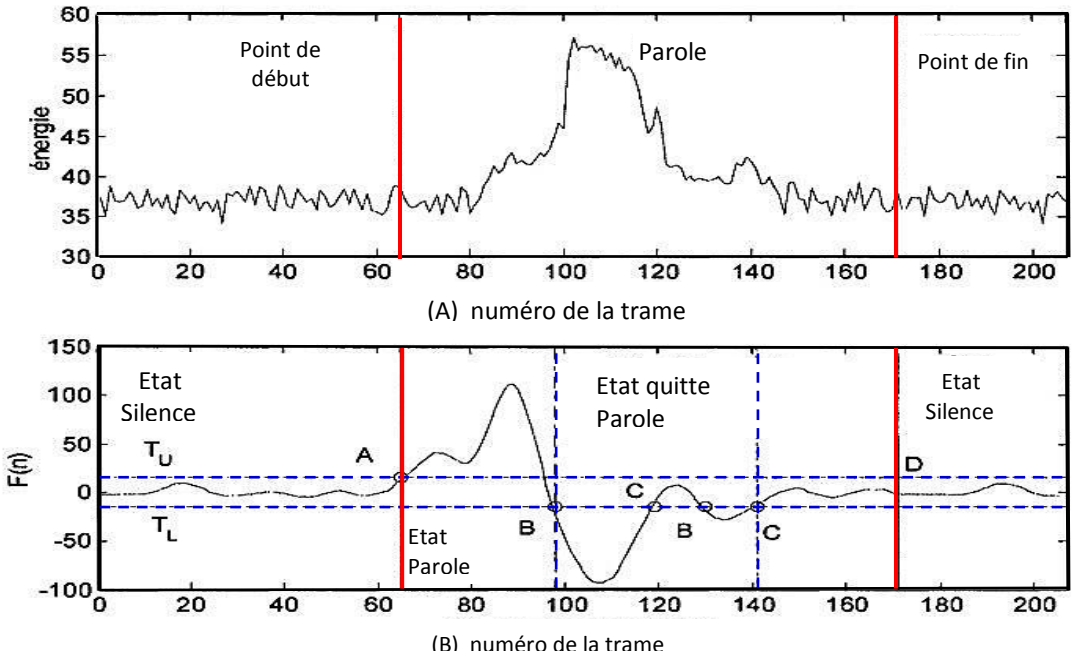


Figure 2. 4. Exemple (A) énergie du mot six et (B) la sortie du filtre F(n)

2.4 VAD basé sur l'analyse de l'énergie court-terme en sous bandes de fréquence [54]

2.4.1 Problématique

A partir de l'observation que les deux paramètres déjà utilisés pour le VAD, l'énergie court-terme et le taux de passage par zéro, ne sont pas suffisants pour une détection robuste même si l'on utilise des algorithmes élaborés pour la décision [55], un nouveau paramètre qui présente de meilleures caractéristiques a été proposé. Ce paramètre est la somme du logarithme de l'énergie court-terme lissé et normalisé et l'énergie du signal dans la bande de fréquence de 250 à 3500 Hz à son tour lissée et normalisée. Basé sur ce paramètre et certains seuils d'énergie, l'algorithme proposé [56] trouve tout d'abord les zones de signal vocal de haute énergie correspondant aux sons voisés. La décision finale est le résultat d'une procédure de raffinement qui utilise le taux de passage par zéro et des seuils de durée. Si les taux de passage par zéro moyenné pour une durée de 100 ms avant et 150 ms après les frontières déjà trouvés dépasse les taux de passage par zéro moyenné des 100 premières ms, ces régions sont classifiées comme étant des régions de parole. Sinon la décision initiale reste inchangée.

2.4.2 Définition des paramètres

L'idée d'un paramètre composé est aussi à la base de l'algorithme proposé en [54] qui essaie de résoudre le problème de VAD dans le cas d'un RSB variable. Pour cela deux paramètres sont proposés. Dans le cas d'un RSB variable l'énergie du bruit change en temps et les seuils basés sur l'analyse en début du signal doivent être adaptés pour tenir le pas avec l'évolution du bruit. Pour modéliser le fait que l'oreille humaine perçoit les sons non-linéaires en rapport avec leur fréquence réelle, a été introduite l'échelle de mel qui est une mesure subjective pour la hauteur d'un son. La relation entre le mel et la fréquence est donnée par la relation [49] :

$$mel = 2595 \log (1+f / 700) \qquad (2.6)$$

Où f est la fréquence. Une façon d'obtenir ce spectre subjectif est d'utiliser une banque de filtres passe-bande de gain constant et qui ont une largeur de bande en rapport avec l'échelle de mel. Dans la figure 2.5 on pressente une banque de 20 filtres triangulaires conçue pour le domaine de fréquence de 0 à 4000 Hz. Chaque filtre est multiplié par le module de la TFR de la trame courante pour générer le spectre subjectif qui est donc calculé en 20 points pour chaque trame.

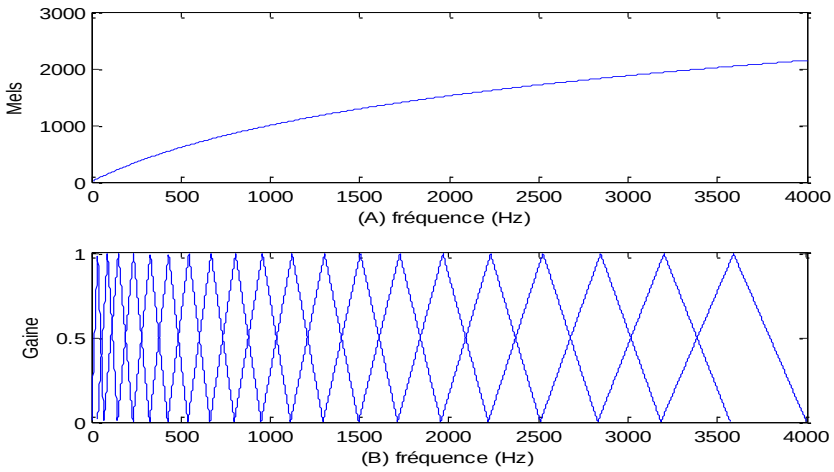


Figure 2. 5.(A) la relation entre le mel et la fréquence (B) le banque de 20 filtres triangulaires utilisés pour obtenir le spectre subjectif de l'oreille

On réalise un lissage temporel pour chacune de 20 bandes du spectre subjectif résultant utilisant un filtre médian en trois points. Le spectre ainsi obtenu est normalisé. Les valeurs moyennes pour chacune de 20 bandes de fréquence des 5

premières trames, considérées juste du bruit, sont soustraites de la bande correspondante pour le reste du signal.

$$X\,(n,i)\;=\;X_{lise}(n,i)-Noise\;=X_{lise}(n,i)-\frac{\sum_{j=0}^{4}X_{lise}(j,i)}{5} \tag{2.7}$$

où *X*(*n* ,*i*) est l'énergie lissée et normalisée de la *i*-ième bande de la *n*-ième trame. Etant donné que le spectre du bruit a été soustrait, on peut maintenant définir l'énergie du signal propre *E*(*i*) pour chaque bande *i*, comme suit :

$$E(i)=\sum_{n=0}^{N-1}X\,(n,i) \tag{2.8}$$

Où N est le nombre total des trames du signal.

Un exemple du spectre subjectif pour le mot six est représenté dans la figure 2.6. On a utilisé 128 points pour la TFR et une fenêtre temporelle de 15 ms.

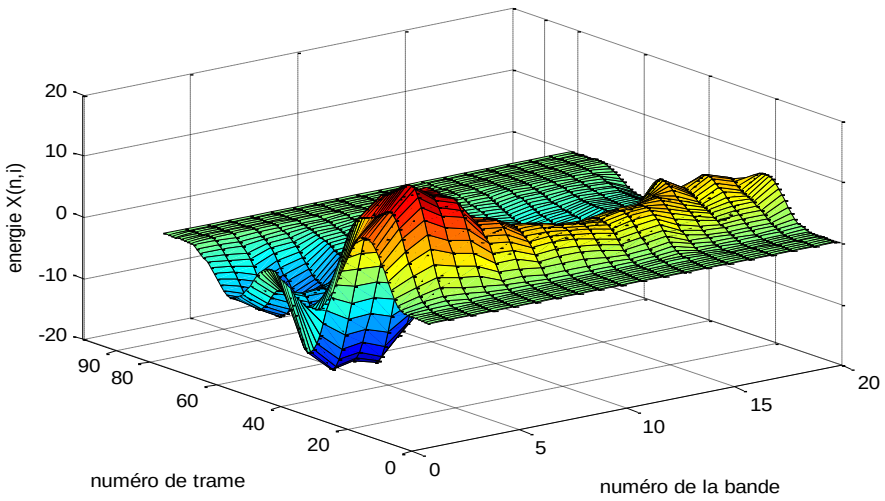


Figure 2. 6. Exemple de spectre subjectif pour le mot six

Ensuite on effectue un tri de ces vingt valeurs de *E*(*i*). On obtient un nouvel ensemble d'indices *I*(*i*) pour lesquels *E*(*I*(1)) est la bande qui contient le plus d'énergie et *E*(*I*(20)) est la bande qui contient le moins d'énergie. Lorsque E_{min}= *E*(*I*(20)) contiennent le moins d'énergie du signal vocal, son évolution temporelle est un bon indicateur pour l'évolution temporelle du bruit. Le paramètre appelé *VAR* est une mesure de la variation du bruit pour tout le signal :

$$VAR=\frac{\sum_{n=0}^{N-1}|X(n,I(20))|}{N} \tag{2.9}$$

Les essais ont montré que la plus grande partie de l'énergie du signal vocal est concentrée dans les six premières bandes du spectre subjectif trié. L'estimation de

l'énergie du signal se fait donc à l'aide d'un nouveau paramètre, l'énergie temps-fréquence, *ETF* qui est la somme lissée de l'énergie du domaine temps *T* et des six premières bandes du spectre subjectif :

$$ETF(n) = lisse\big(T(n) + cF(n)\big) \qquad avec \qquad \sum_{i=1}^{6} X\big(n, I(i)\big) \tag{2.10}$$

Où c ≅1.1 est une constante de proportionnalité

L'énergie du domaine temps *T*(*n*) est le logarithme de l'énergie court-terme pour chaque trame, lissé et normalisé, exactement comme dans le cas de chaque bande du spectre subjectif.

2.4.3 Algorithme de décision

Avec ces paramètres on peut passer à l'étape de décision résumée dans la figure 2.7. Quand dépasse un certain seuil *th1*, le paramètre *VAR* indique une variation importante dans le niveau de bruit et les seuils de décision *th2* et *th3* sont modifiés en conséquence. Dans la partie A du diagramme de décision, le paramètre *ETF* et le seuil conservateur *th1* sont utilisés pour trouver une première estimation des frontières pour les régions de parole qui possèdent plus d'énergie et dépassent un certain seuil de durée *th4* = 90 ms. Dans la partie B, le début de la région de parole est poussé vers le début du signal tant que *ETF* est plus grande que le deuxième seuil moins conservateur *th3* ou le taux de passage par zéro dépasse le seuil *th5* et le seuil de durée n'est pas dépassé. Le seuil de passage par zéro est fixé à partir du taux de passage par zéro moyen des 5 premières trames de signal. La partie C déplace la fin de la région de parole vers la fin du signal dans les mêmes conditions.

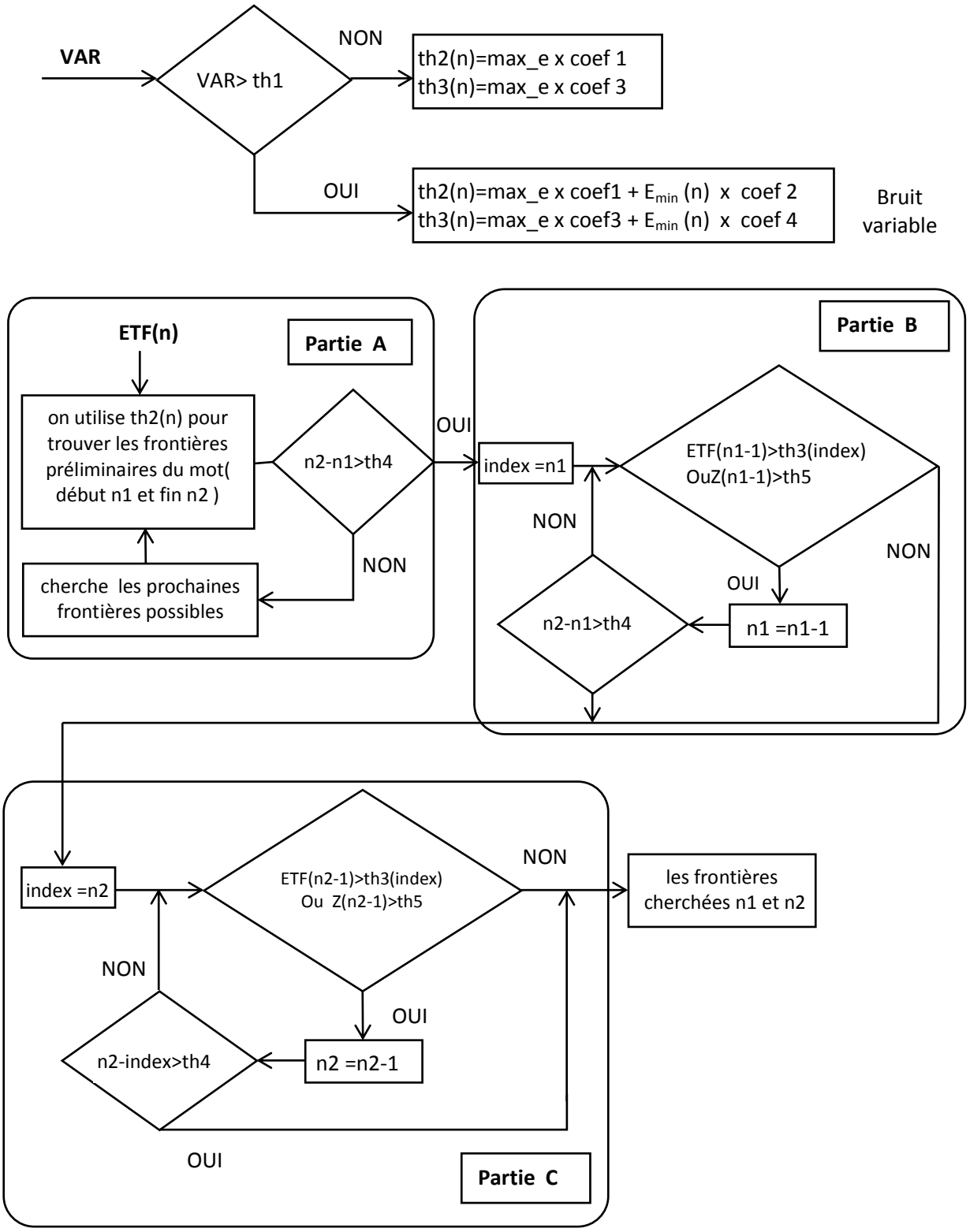


Figure 2. 7. Diagramme de décision

Le paramètre max_e est le maximum de $T(n)$ pour tout le signal. Les valeurs des paramètres *thl, coefi*, i=1 ,...,4 sont choisies pour optimiser un set d'exemples donnée pour lesquels les frontières ont été marquées manuellement.

2.5 L'algorithme de VAD de l'annexe G.729 B de l'ITU

Cet algorithme de VAD[49], utilise une trame de 10 ms donc 80 échantillons à une fréquence d'échantillonnage de 8000 Hz. Il effectue une décision pour chaque trame sans introduire aucun retard entre la décision pour la trame courante et l'arrivée d'une nouvelle trame. Si la décision du module de VAD indique la présence de la parole, le codeur de parole code la trame qui sera après transmise ; si l'absence de la parole est indiquée, les algorithmes de transmission discontinue et de génération de bruit de confort génèrent les trames d'inactivité vocale. L'algorithme adopte une approche de classification des formes. On utilise un ensemble de quatre paramètres qui décrit chaque trame de signal et un ensemble de frontières de décision dans l'espace 4D des paramètres. La décision initiale est lissée tenant compte d'un certain nombre de trames passées. Dans une dernière étape on procède à l'actualisation des moyennes courantes du bruit du fond si les seuils énergétiques imposés au bruit de fond sont dépassés. Le fonctionnement de l'algorithme est résumé dans le diagramme de la figure 2.8:

2.5.1 Extraction des paramètres

Dans une première étape on calcule les douze premiers coefficients d'autocorrélation court-terme R_n , n = 1,...12. A partir de ces coefficients on dérive les coefficients de prédiction linéaire et le deuxième coefficient de réflexion. A partir des coefficients de prédiction linéaire on obtient une série de 10 paramètres qui représentent le spectre du signal en terme paires de raies spectrales, notés LSP_i ,i = 1,...10, dont une méthode efficace de calcul est présenté en [57]. Ces paramètres représentent des angles ayant les valeurs dans l'intervalle de 0 à π.

On calcule l'énergie pleine bande E_f qui correspond au logarithme du *R1* normalisé par la taille de la fenêtre *L* :

$$\mathrm{E_f} = 10 \log_{10}\left(\frac{\mathrm{R_1}}{\mathrm{L}}\right) \qquad (2.11)$$

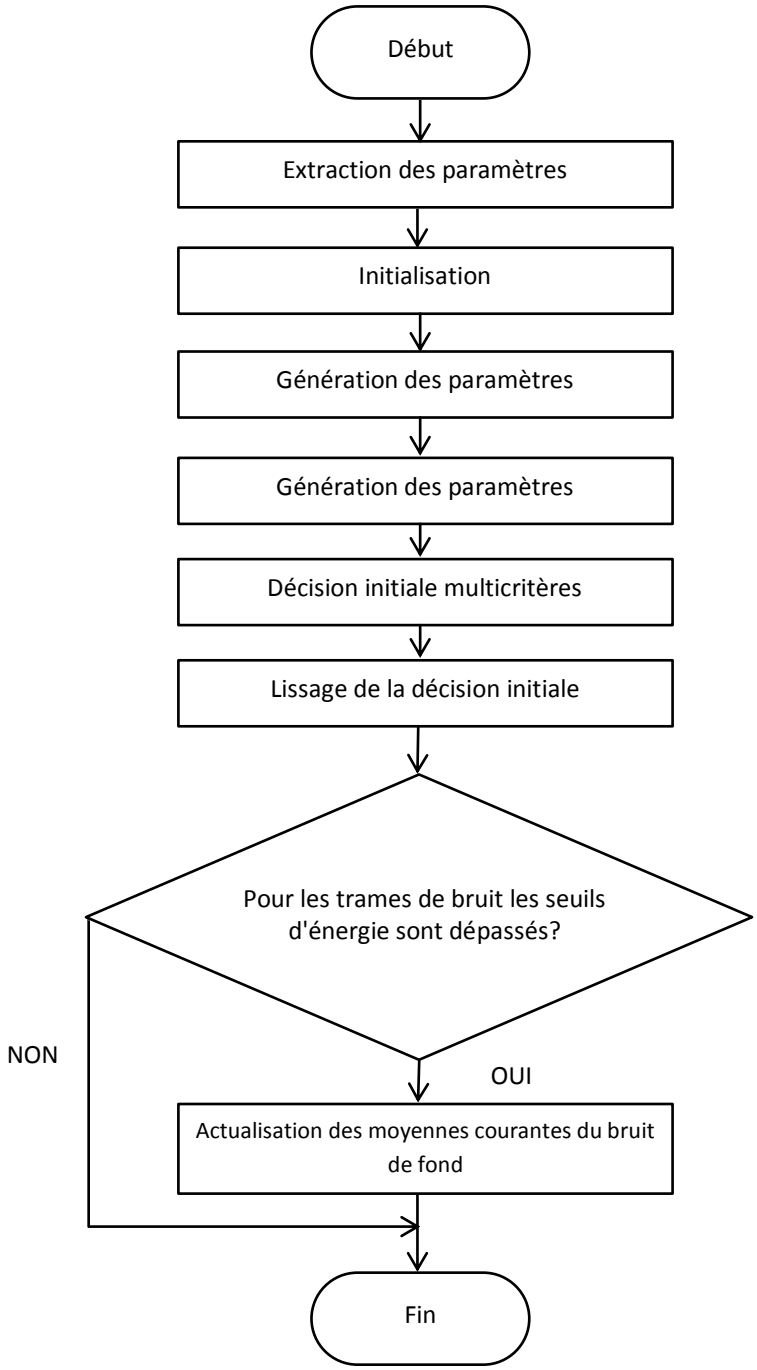


Figure 2. 8. Diagramme du fonctionnement de l'algorithme de VAD de l'annexeG.729 B de l'ITU

L'énergie basse fréquence E_b est obtenue par une multiplication de la matrice de Toeplitz, engendrée par les coefficients d'autocorrélation, avec les 13 premiers coefficients de la réponse impulsionnelle d'un filtre RIF ayant la fréquence de coupure 1000 Hz. Le dernier paramètre de la trame courante est le taux de passage par zéro *Z*.
On définit le minimum de l'énergie long-terme E_{min} comme étant le minimum de l'énergie pleine bande sur un intervalle d'une seconde.

2.5.2 Initiation

Les moyennes de paramètres déjà vus pour un nombre de N_i premières trames constituent l'initiation des paramètres du bruit. Pendant ce temps la décision de l'algorithme est basé sur un seuil d'énergie seulement.

2.5.3 Génération des paramètres

La différence entre les paramètres qui caractérisent la trame courante et les paramètres qui caractérisent le bruit de fond au même instant génère les quatre paramètres utilisés pour la décision initiale. Ces paramètres sont :

- **la distorsion spectrale**

$$\Delta S = \sum_{i=1}^{10} (LSP_i - \overline{LSP})^2 \qquad (2.12)$$

- **la différence d'énergie pleine bande**

$$\Delta E_f = \overline{E_f} - E_f \qquad (2.13)$$

- **la différence d'énergie basse fréquence**

$$\Delta E_b = \overline{E_b} - E_b \qquad (2.14)$$

- **la différence de taux de passage par zéro**

$$\Delta \mathrm{Z} = \bar{Z} - \mathrm{Z} \qquad (2.15)$$

2.5.4 Décision initiale multicritères

Les quatre paramètres de décision se trouvent dans une région de l'espace Euclidienne quadri- dimensionnelle. Les paramètres indiquant l'activité vocale se groupent dans un certain hyper-volume de l'espace lorsque les paramètres indiquant le manque

d'activité vocale se groupent dans un autre hyper-volume. Ces hyper-volumes ont été identifiés et séparés à l'aide de quatorze hyper plans définis dans des espaces tridimensionnels qui constituent donc les frontières pour la décision initiale. Pour chaque trame la décision initiale correspond à la région où le point, dont les coordonnés sont les quatre paramètres de décision, se retrouve.

2.5.5 Lissage de la décision initiale

Normalement les régions de parole ou de silence ont une longueur de quelques dizaines de trames, la décision initiale est lissée pour refléter cette caractéristique de stationnarité. Le lisage s'effectue en quatre étapes dérivées d'une observation approfondie d'une large base de données.

Dans la première étape la décision d'activité vocale est prolongée à la trame courante si l'énergie de la trame courante dépasse un certain seuil.

Dans un deuxième pas la décision d'activité vocale est prolongée à la trame courante si les deux trames précédentes sont des trames de parole et la différence entre l'énergie de la trame courante et l'énergie des deux trames précédentes est au-dessous d'un certain seuil.

Dans un troisième pas la décision d'inactivité vocale est prolongée à la trame courante si les dix trames précédentes sont des trames de silence et la différence entre l'énergie de la trame courante et l'énergie des dix trames précédentes est au-dessous d'un certain seuil. Dans une dernière étape la décision d'activité vocale et corrigée si l'énergie de la trame courante est inférieure à l'énergie du bruit avec un certain écart, le deuxième coefficient de réflexion est plus petit que 0.6 et aucun des deux premiers pas n'ont pas été exécutés.

Les tests ont montré qu'on obtient des meilleurs résultats lorsqu'une logique floue remplace la partie de décision de l'algorithme [57] et que même la langue utilisée pour tester l'algorithme a des influences sur le résultat de classification [58].

2.5.6 Actualisation des paramètres du bruit de fond

La dernière étape de l'algorithme est l'actualisation des moyennes courantes du bruit de fond. Pour cela on utilise une version simplifiée de l'algorithme de VAD qui utilise seulement l'énergie pleine bande, le coefficient de distorsion spectrale et le deuxième coefficient de réflexion. Si la décision est silence les paramètres du bruit de fond sont actualisés à l'aide d'un filtre autorégressif d'ordre un. Différents coefficients de régression sont utilisés pour chaque paramètre et une adaptation plus rapide est exécutée en début du signal et après chaque réinitialisation. Cet algorithme de VAD simplifié est plus conservateur en ce qui concerne la détection de bruit pour éviter l'adaptation des paramètres du bruit avec des valeurs qui en réalité proviennent de trames de parole. Il se comporte bien pour des bruits qui changent lentement dans le temps. Dans le cas des bruits moins stationnaires il perd plus de trames de bruit qui pourraient être utilisées pour l'adaptation des paramètres du bruit. Dans ces conditions le bruit estimé est loin du bruit réel ce qui augmente la probabilité de fausse alarme. S'il y a une augmentation brusque dans le niveau d'énergie du bruit, l'algorithme peut se bloquer dans l'état d'activité vocale, pour éviter cette possibilité on utilise un mécanisme de réinitialisation qui initialise le niveau du bruit de fond avec la valeur du E_{min}.

La figure 2.9 explique le fonctionnement du VAD G.729.

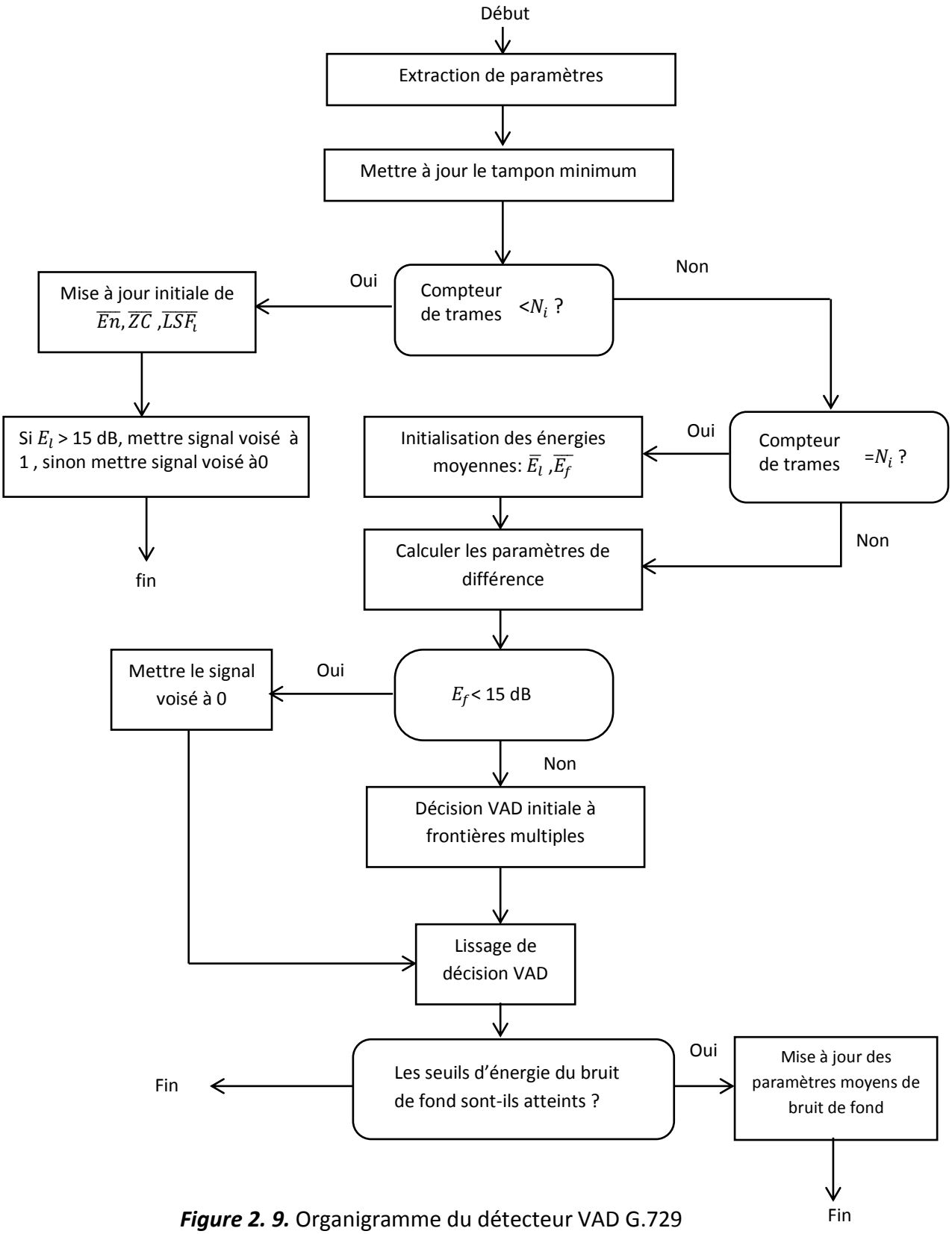


Figure 2. 9. Organigramme du détecteur VAD G.729

2.6 Conclusion

Dans ce chapitre nous avons développé quatre algorithmes de VADs. L'algorithme de l'annexe G.729 B de l'ITU, un algorithme basé sur un filtrage optimal de l'énergie court-terme, un troisième basé sur l'analyse de l'énergie court-terme en sous bandes de fréquence et enfin un algorithme VAD basé sur l'énergie courte-terme et le taux de passage par zéro. Ces 4 VAD ont été implémentés et comparés dans le chapitre 4.

Chapitre 3 Filtrage de Wiener et rehaussement de la parole multi-microphones

3.1 Introduction

Dans un précédent travail [3], nous nous sommes intéressées au développement d'un système de réduction de bruit basé sur le « Beamforming Adaptatif ». La solution proposée était à base d'annulateurs de lobe latéraux «Generalized Sidelobe Canceller (GSC)» [4].

Dans ce projet, nous avons développé un système de rehaussement de la parole utilisant deux techniques « GSC » et « SDW-MWF » intégrées dans un seul système. Ce dernier a été proposé dans [14] et nommé « Prétraitement Spatial – Distorsion Pondéré de la Parole – Filtre Multicanal de Wiener » ou en anglais « the Spatially Preprocessed - Speech Distortion Weighted- Multi-channel Wiener Filter » (SP-SDW-MWF). Il est constitué de deux parties, l'une est fixe (Prétraitement Spatial) et l'autre adaptative (SDW-MWF).

Pour réduire la distorsion de la parole, l' « ANC » est adapté pendant les périodes de bruit seulement [5], [6], [7], [8].Pour mettre au point une telle solution, on a implémenté 4 algorithmes de détection d'activité vocale (VAD) afin de pouvoir détecter les périodes de silence, ainsi que les zones d'activité vocales.

La dernière étape est l'intégration d'une matrice de blocage variable pour une meilleure estimation des interférences dans le système proposé.

Pour cela des notions sur le « beamforming » fixe et adaptatif sont exposées. Les notations adoptées pour le développement de notre application, ainsi que méthodes de filtrage par blocs « overlap-save », consistant à calculer la convolution via la transformée de Fourier discrète, sont explicitées dans la suite.

3.2 Le Beamforming

Le « Beamforming » consiste à combiner les signaux à la sortie des microphones, qui seront convolués à des filtres pondérés optimaux (gain, retard) et additionnés pour obtenir un « faisceau » dans une direction d'intérêt spécifique. Ce faisceau rend le réseau de microphone fortement directif figure.3.1.

La direction d'intérêt s'appellera direction de visée. Elle peut être la direction d'une source acoustique dans un environnement bruyant et/ou réverbérant par exemple. Le réseau peut être utilisé afin de localiser une source acoustique mobile ou un point d'une source d'énergie élevé. Le « beamforming » peut être fixe ou adaptatif. Ces deux notions sont développées à la suite. Les approches de « beamforming » conventionnelles ont pour but de trouver un filtre w complexe linéaire et invariant dans le temps TI (Time Invariant), tel que sa sortie est de la forme:

$$\boldsymbol{y}(t) = \boldsymbol{w}^H \boldsymbol{x}(t) \tag{3.1}$$

Et optimise un critère au second ordre sous d'éventuelles contraintes, où x(t) désigne le vecteur des signaux observés en sortie des microphones. Cette sortie correspond ainsi à une estimée au second ordre du signal utile venant d'une direction particulière et potentiellement corrompu par des interférences et d'un bruit de fond.

3.2.1 Le Beamforming fixe

Dans le « beamforming » fixe les poids optimaux sont prédéterminés et stockés pour traiter les signaux aux sorties des microphones. Ils sont des donnés indépendantes et basées sur un modèle (ou des normes) du bruit ambiant et de la source dans l'environnement du réseau : bureau, pièce de téléconférence…

Le « beamforming » fixe peut être employé dans des algorithmes adaptatifs, tel que l'annulateur de lobe latéral généralisé (GSC) pour la suppression du bruit.

3.2.2 Le Beamforming Adaptatif

Dans ce cas la pondération optimale pour diriger et former un faisceau dans la direction de visé est exécutée en temps réel pendant que le signal est capturé et stocké dans une mémoire tampon (Buffer). Les algorithmes peuvent tenir compte de

l'environnement sonore réel (bruit) et s'adapté en poursuivant des sources acoustiques mobile et éliminer le bruit venant soudainement dans une direction spécifique.

Soit le réseau « beamformer » de N microphone suivant figure 3.1:

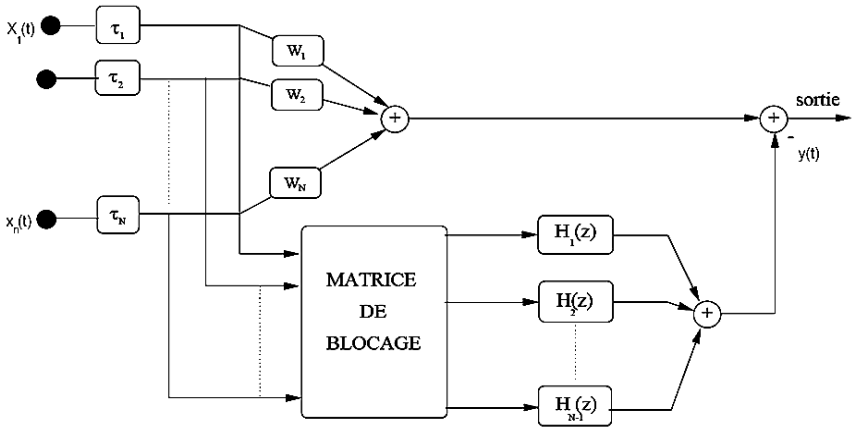


Figure 3. 1. Exemple d'un « beamformer » avec un réseau de N microphones

La sortie du réseau s'écrit :

$$\boldsymbol{y}(t) = \left[w_{1,}w_{1,} \dots \dots \dots \dots w_{N,}\right]\begin{bmatrix} x_1(t) \\ x_2(t) \\ . \\ . \\ x_N(t) \end{bmatrix} = \boldsymbol{w}^H(t)\boldsymbol{x}(t) \tag{3.2}$$

Où $\mathbf{w^H}$désigne le conjugué de la transposée complexe de pondération du vecteur de poids.

Si nous supposons que le premier élément du réseau est la référence de phase, le déphasage relatif du signal reçu au nième élément est :

$$\emptyset_n = \left[2\pi \frac{d(n-1)}{\lambda}\right] sin\theta_s \tag{3.3}$$

Où $\lambda = c/f$, θ_s la direction du signal incident et d la distance inter-microphones.

Un réseau adaptatif modifie continuellement le faisceau, selon la manière désirée au moyen d'un algorithme d'optimisation adaptatif. Il est optimisé de sorte qu'un gain maximum soit offert dans une direction spécifique correspondant au signal désiré, alors que l'atténuation est maximale ailleurs (correspond au signal des interférences indésirables ou aux brouilleurs).

La méthode de formation de voies adaptative « Adaptative Beamforming » est une technique puissante qui consiste à améliorer un signal d'intérêt en supprimant le bruit d'interférence à la sortie d'un réseau de capteurs.

3.2.3 La formation de voie par délais-somme (FVDS)

La formation de voie est une technique très connue et largement utilisable, elle se base sur l'ITD (Interaural Time Difference) classique, mais la différence est que la formation de voie ne donne pas des calculs exactes, elle estime la position de la source en maximisant sa sortie énergétique **E**.

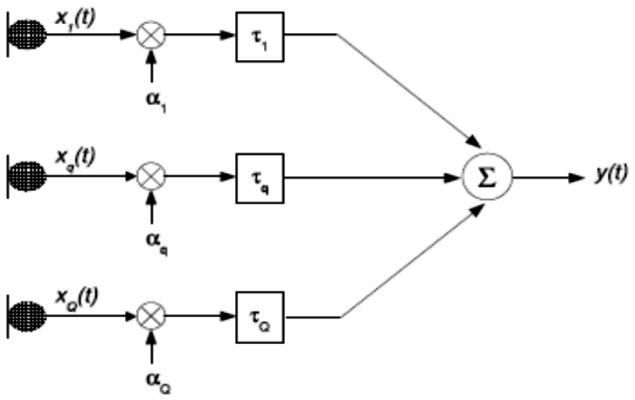


Figure 3. 2. La structure de formateur de voie par somme-délais du M microphone

Dans la FVDS tous les signaux reçus sont alignés (correction de la phase) et sommés figure 3.2. Le bruit, supposé blanc, se neutralise [59]. La sortie du formateur de voie par délai-somme comme le montre la figure 3.2 avec M microphones est définie par :

$$y(n) = \sum_{m=0}^{M-1} x_m(n - \tau_m) \tag{3.4}$$

Où $x_m(n)$ représente le signal reçu pour chaque m^{ieme} microphone, τ_m correspond aux retards d'arrivés.

La recherche du maximum peut se faire par le calcul d'énergie sur une fenêtre de longueur L.

$$E = \sum_{m=0}^{L-1} y(n)^2 \tag{3.5}$$

$$E = \sum_{m=0}^{L-1} [x_0(n - \tau_0) + \cdots\cdots + x_{M-1}(n - \tau_{M-1})]^2 \quad (3.6)$$

Après le développement de l'équation précédente, nous trouvons :

$$E = \sum_{m=0}^{M-1} \sum_{n=0}^{L-1} {x_m}^2(n - \tau_m) + 2 \sum_{m_1=0}^{M-1} \sum_{m_2=0}^{m_1-1} \sum_{n=0}^{L-1} x_{m_1}(n - \tau_{m_1})\, x_{m_2}(n - \tau_{m_2}) \quad (3.7)$$

Et avec l'intercorrélation :

$$E = K + 2 \sum_{m_1=0}^{M-1} \sum_{m_2=0}^{m_1-1} R_{x_{m_1}, x_{m_2}}(\tau_{m_1} - \tau_{m_2}) \quad (3.8)$$

K est supposé constant et négligeable quand la sortie énergétique E est maximisée. Dans le domaine fréquentiel, l'intercorrélation est définie par :

$$R_{ij}(\tau) \approx \sum_{k=0}^{L-1} X_i(K) X_j(K)^* e^{\frac{j2\pi k\tau}{L}} \quad (3.9)$$

Où $X_i(k)$ est la transformée de Fourier discrète de $x_i(n_t)$, $X_i(k)X_j(k)^*$ est la densité spectrale de $x_i(n_t)$et$x_j(n_t)$.

3.3 Notation

Dans le domaine temporel les variables sont désignées par des lettres minuscules, et dans le domaine fréquentiel les variables sont désignées par des lettres majuscules, les vecteurs et les matrices sont écrits en caractères gras et les scalaires sont écrits en italique. La première ligne ou colonne d'une matrice sera d'indice0. Les principales notations utilisées dans ce chapitre sont les suivantes :

$\mathbf{F}_N$	la matrice carrée d'ordre *N* associée à la TFD directe *N* points,
$\mathbf{F}_N^{-1}$	la matrice inverse de $\mathbf{F}_N$ associée à laTFD inverse *N* points,
$\mathbf{0}_{m\times n}$	la matrice nulle de taille *m*×*n*,
$\mathbf{I}_m$	la matrice identité d'ordre *m*.
S	la matrice sélection de taille *L*×*M*. C'est une matrice binaire qui permet de sélectionner *L* éléments centraux d'un vecteur formé de *M* éléments. En notant,$d = \frac{M-L}{2}$, on a :$\mathbf{S} = [\mathbf{0}_{L\times d}\mathbf{I}_m\ \mathbf{0}_{L\times d}]$
$\mathbf{S}^T$	la matrice transposée de **S**. C'est une matrice binaire qui permet de former un vecteur de taille *M* à partir d'un vecteur de taille *L*=*M*− 2*d* en ajoutant 2*d*

zéros, *d* zéros de chaque côté, au vecteur initial :

$$\mathbf{S}^T = \begin{bmatrix} \mathbf{0}_{d\times L} \\ \mathbf{I}_L \\ \mathbf{0}_{d\times L} \end{bmatrix}$$

$\lvert\cdot\rvert$	le module d'un nombre complexe
$\boldsymbol{A}^T$	la matrice transposée de **A**,
A^{H}	la matrice hermitienne de A,
*	le conjugué complexe
k	l'indice de temps discret.
$\varepsilon\{\cdot\}$	l'espérance mathématique

3.4 Filtrage par transformée de Fourier rapide

L'idée de base des deux méthodes de filtrage par blocs « overlap-save » et « overlap-add » consiste à calculer la convolution via la transformée de Fourier discrète. En effet, d'après les propriétés de la TFD, le produit de convolution circulaire de deux séquences, $\{x_n\}$ et $\{h_n\}$, de longueur *M* chacune, peut être obtenu indirectement via la transformée de Fourier discrète selon l'équation suivante :

$$\{y_n\} = \{x_n\} * \{h_n\} = TFDI\,(TFD\,(\{x_n\}) \bullet TFD(\{h_n\})\,) \qquad (3.10)$$

Où $(*)$ et $(\bullet)$ désignent respectivement les opérations du produit de convolution circulaire et de multiplication terme à terme.

Ce calcul de produit de convolution nécessite :

- le calcul de deux transformées directes : $\{\overline{x_n}\} = TFD(\{x_n\})$ et $\{\overline{h_n}\} = TFD(\{h_n\})$,
- la multiplication terme à terme des séquences ainsi obtenues : $\{\overline{y_n}\} = \{\overline{x_n}\} \bullet \{\overline{h_n}\}$,
- le calcul de la transformée inverse de la séquence finale : $\{y_n\} = TFDI\,(\{\overline{y_n}\})$.

L'intérêt de ce calcul indirect de convolution est la possibilité de calculer la TFD et la TFDI au moyen de la TFR, permettant ainsi une large réduction du coût de calcul de la convolution circulaire, comparé à l'algorithme de calcul direct de cette convolution.

Cependant, les avantages de ce calcul ne peuvent pas être exploités directement dans le filtrage numérique. En effet, dans la plupart des applications de filtrage, le problème qui se pose est d'implanter la convolution linéaire de deux séquences plutôt que la convolution circulaire. Le traitement au moyen de la TFD d'un signal décomposé en

blocs disjoints est entaché d'un défaut systématique lié aux discontinuités induites par le découpage. Les études faites pour pallier ce défaut ont abouti à deux types de mise en œuvre des méthodes rapides de filtrage, devenus classiques et qui sont [60] :

- la méthode d' « overlap-save » (recouvrement avec mémorisation des blocs d'entrée),
- la méthode d' « overlap-add » (recouvrement avec addition des blocs de sortie).

3.5 Overlap-save (recouvrement des blocs d'entrée)

3.5.1 Relation entre la Convolution linéaire & la convolution circulaire

Pour illustrer la différence entre une convolution linéaire et une convolution circulaire, nous prenons le cas d'un signal *x*(*n*) de taille N_x = 8, $x(n) = (1,2,\dots,8)$ et d'un filtre *h*(*n*) d'ordre *P*= 2, $h(n) = (9,10,11)$. Pour la convolution linéaire, le signal résultant y(n), de taille N_y=$N_x + \mathrm{P} = 10$, est donné par :

$$y(n) = \sum_{m=0}^{2} h(n)x(n-m)\,, \qquad n = 0,1,\dots,9$$

$$\begin{cases} y(0) = & 9*1 & = 9 \\ y(1) = & 9*2+10*1 & = 28 \\ y(2) = & 9*3+10*2+11*1 & = 58 \\ y(3) = & 9*4+10*3+11*2 & = 88 \\ y(4) = & 9*5+10*4+11*3 & = 118 \\ y(5) = & 9*6+10*5+11*4 & = 148 \\ y(6) = & 9*7+10*6+11*5 & = 178 \\ y(7) = & 9*8+10*7+11*6 & = 208 \\ y(8) = & 10*8+11*7 & = 157 \\ y(9) = & 11*8+ & = 88 \end{cases}$$

Une façon d'illustrer ce calcul est de considérer la fenêtre f = (11, 10, 9), formée à partir des coefficients de la séquence *h* prise en sens inverse, et de la faire glisser le long du signal *x* pour obtenir le signal de sortie *y* comme montré dans la figure 3.3.

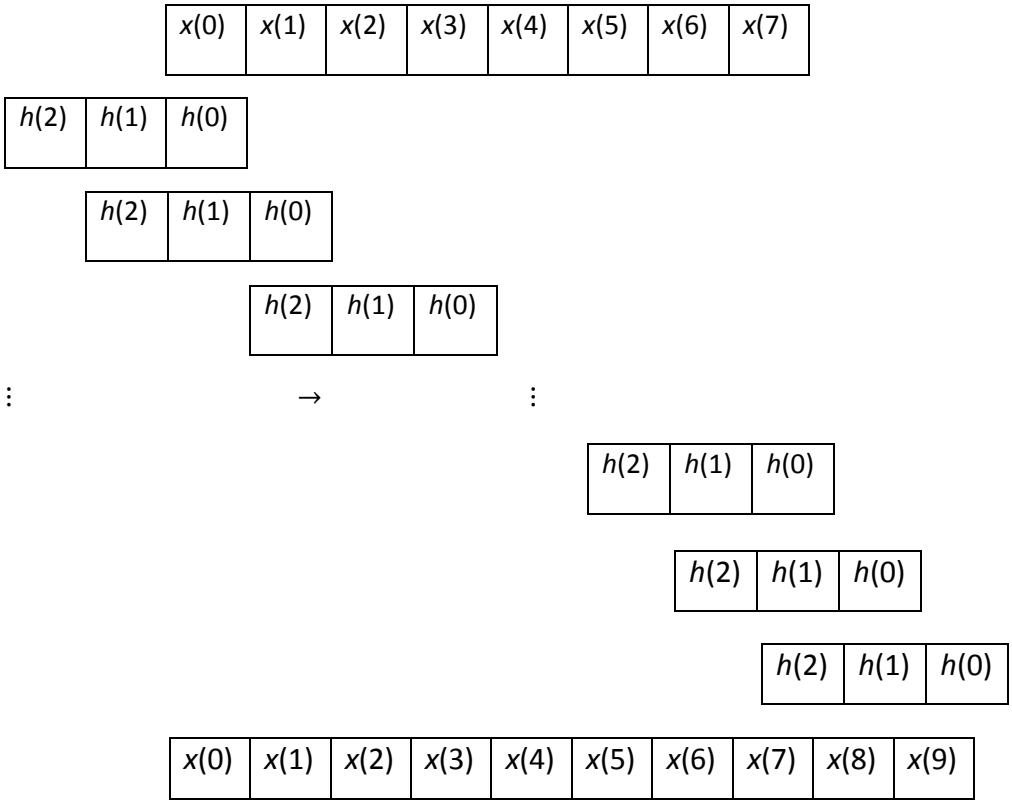


Figure 3. 3.Schéma de calcul de la convolution linéaire

En a joutant des zéros à la suite de *h*(*n*), pour obtenir la séquence *hc* de taille 8, le signal *yc*(*n*) résultant de la convolution circulaire est de taille M= 8 et il est donné par :

$$y_c(n) = \sum_{m=0}^{7} h_c(m)x(\langle n-m\rangle_8)$$

Où $\langle . \rangle_q$ désigne l'opérateur de calcul modulo q.

Nous avons donc :

$$\begin{cases} y_c(0) = & 9*1+10*8+11*7+0*6+0*5+0*4+0*3+0*2 & = 166 \\ y_c(1) = & 9*2+10*1+11*8+0 & = 116 \\ y_c(2) = & 9*3+10*2+11*1+0 & = 58 \\ y_c(3) = & 9*4+10*3+11*2+0 & = 88 \\ y_c(4) = & 9*5+10*4+11*3+0 & = 118 \\ y_c(5) = & 9*6+10*5+11*4+0 & = 148 \\ y_c(6) = & 9*7+10*6+11*5+0 & = 178 \\ y_c(7) = & 9*8+10*7+11*6+0 & = 208 \end{cases}$$

Nous pouvons remarquer que l'on a *y*(*i*) =y_c(*i*) pour *i* = 2,3,…,7. Ce résultat peut être généralisé pour des tailles quelconques N_x de la séquence *x* et un ordre quelconque *P* du filtre *h*. Cette généralisation peut s'écrire :

$$y(i) = y_c(i) \qquad \text{pour} \quad i = P, P+1, \ldots, N_x - 1 \tag{3.11}$$

3.5.2 La méthode Overlap-save

Pour obtenir une égalité (3.11) entre la convolution linéaire et la convolution circulaire, la méthode « overlap-save » consiste à subdiviser le signal d'entrée *x* en blocs $x_k$de taille *M* présentant un recouvrement partiel de *P* échantillons.

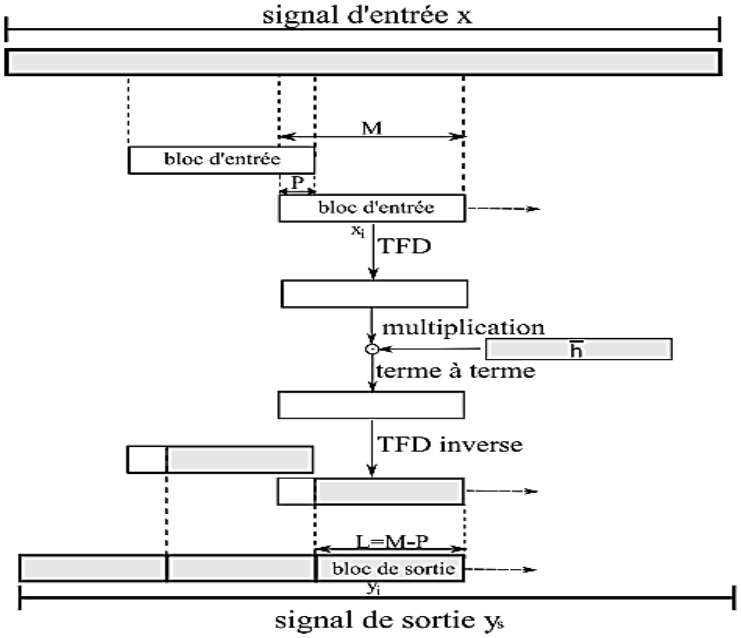


Figure 3. 4. Schéma d'overlap-save

Le recouvrement des blocs permet de remédier aux effets non désirés qui se produisent sur les premiers *P* éléments de chaque bloc. A la fin du traitement de chaque bloc par la convolution circulaire, nous ne gardons que les *L=M-P* échantillons à droite. Le principe de cette méthode est représenté par la figure 3.4.

Sachant que la convolution linéaire de deux séquences n'est pas modifiée si l'on complète les séquences par des zéros à gauche ou à droite. Nous introduisons *P* zéros au début du premier bloc d'entrée et nous complétons si nécessaire par des zéros le dernier bloc pour avoir un bloc de taille *M*. Ainsi le premier bloc x_1 est donné par :

$$x_1(n) = \begin{cases} 0 & 0 \le n \le P-1 \\ & \\ x(n-p)P \le n \le M-1 \end{cases}, \quad n = 0,1,\ldots.,M-1$$

Les autres blocs x_k sont tels que :

$$x_k(n) = x(n + (k-1)L - P)\,, \quad n = 0,1,\ldots,M-1, \quad k = 2,3,\ldots.$$

3.6 L'algorithme (SP-SDW-MWF)

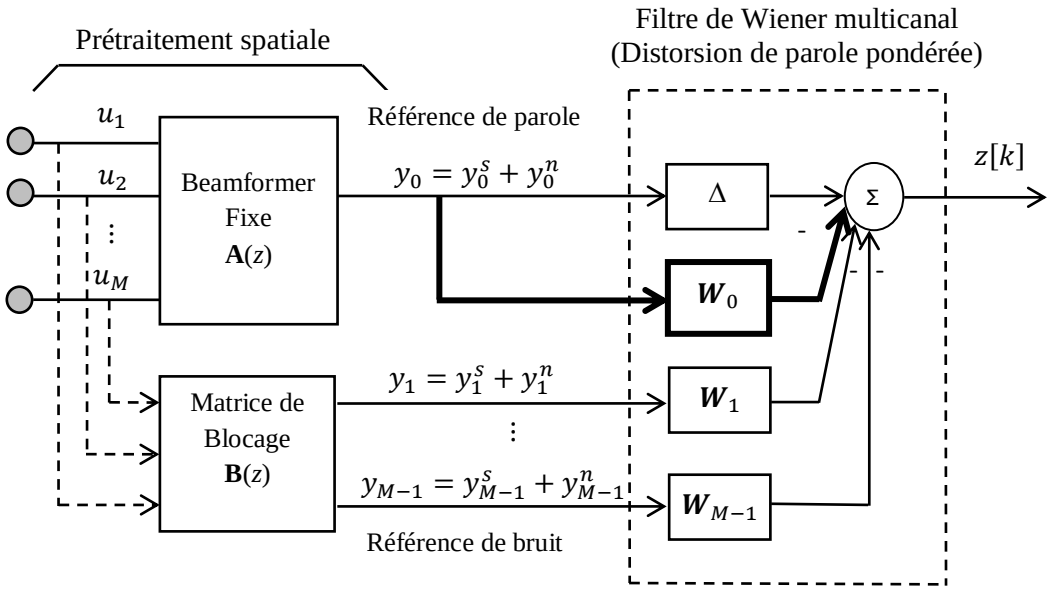


Figure 3. 5. Structure de l'algorithme « prétraitement spatiale distorsions pondéré de parole filtre de Wiener multicanal » (SP-SDW-MWF).

3.6.1 Constitution de l'algorithme SP-SDW-MWF

L'algorithme SP-SDW-MWF [36], présenté à la figure 3.5 est constitué :

- D'une partie fixe, « prétraitement spatiale » à savoir un formateur de faisceau fixe ***A***(*z*). La technique utilisée pour ce bloc est la formation par délais-somme (FVDS), voir le chapitre (3).

$$y_0(n) = \frac{1}{M}\sum_{i=1}^{M} u_i(n) \tag{3.12}$$

Et une matrice de blocage **B**(*z*),

$$\boldsymbol{B} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \tag{3.13}$$

Le produit de la matrice de blocage et de la matrice des données reçues est :

$$\boldsymbol{y}(n) = \boldsymbol{B}\boldsymbol{u}(n) \tag{3.14}$$

- Et une partie adaptative « SDW-MWF » [36, 37, 38].

Etant donnée M signaux de microphone.

$$\boldsymbol{u}_i[k] = \boldsymbol{u}_i^s[\mathrm{k}] + \mathbf{u}_\mathrm{i}^\mathrm{n}[\mathrm{k}] \text{ ,} i = 1, \ldots\ldots, M \tag{3.15}$$

Avec : *s* et *n* utilisé pour référer entre le signal parole et le bruit.

Le formateur de faisceaux fixe **A**(*z*) crée une référence de la parole :

$$\boldsymbol{y}_0[\mathrm{k}] = \mathbf{y}_0^\mathrm{S}[\mathrm{k}] + \mathbf{y}_0^\mathrm{n}[\mathrm{k}],$$

En dirigeant le faisceau vers l'avant et la matrice de blocage **B** (z) crée *M* - 1 références de bruit :

$$\boldsymbol{y}_i[k] = \boldsymbol{y}_i^s[k] + \boldsymbol{y}_i^n[k], i=1, \ldots.,M\text{-}1.$$

Au cours des périodes de parole, les références $\boldsymbol{y}_i\ [k]$ se composent de la parole et du bruit,

$$\boldsymbol{y}_i[k] = \boldsymbol{y}_i^s[k] + \boldsymbol{y}_i^n[k], i=0, \ldots.,M\text{-}1.$$

Pendant les périodes de bruit, il n'y a que la composante de bruit $\mathbf{y}_\mathrm{i}^\mathrm{n}$ [k] qui est observée. On suppose que les statistiques du deuxième ordre du bruit sont suffisamment stationnaires de sorte qu'ils puissent être estimés pendant les périodes de bruit seulement.

Le filtre « SDW-MWF » $\boldsymbol{W}_k \in \mathbb{R}^{\mathrm{M\,Lx1}}$ fournit une estimation $\mathbf{W}_k^T\mathbf{y}_k$ de la contribution du bruit $\mathbf{y}_0^\mathrm{n}$[k-Δ] dans la référence de la parole, en minimisant la fonction de coût **J** ($\mathrm{W_k}$) [36].

$$\mathbf{J}\ (\mathrm{w_k}) = \frac{1}{\mu}\underbrace{\varepsilon\ \{\ |\mathrm{w_k^T y_k^s}|^2\ \}}_{\varepsilon_\mathrm{d}^2} + \underbrace{\varepsilon\ \{\ |\mathrm{y_0^n[k-\Delta]} - \mathrm{w_k^T y_k^n}|^2\ \}}_{\varepsilon_\mathrm{n}^2}. \tag{3.16}$$

Avec

$$\boldsymbol{w}_k^T = [\ w_0^T[k] \quad w_1^T[k] \quad \ldots w_{M-1}^T[k]\ ], \tag{3.17}$$

$$\boldsymbol{w_i}[k] = [\ w_i[0] w_i[1] \ldots\ w_i[L-1]]^T, \tag{3.18}$$

$$\boldsymbol{y}_k^T = [\ y_0^T[\mathrm{k}] \quad y[k] \quad \ldots y_{M-1}^T[k]\ ], \tag{3.19}$$

$$\boldsymbol{y}_i[k] = [\ y_i[0] y_i[1] \quad \ldots \quad y_i[L-1]]^T, \tag{3.20}$$

Cette estimation est alors soustraite de la référence de la parole, comme indiqué dans la figure3.5, pour obtenir un meilleur signal de la parole ***z*** [*k*]. Le terme ε_d^2 représente l'énergie de distorsion de la parole et ε_n^2 l'énergie du bruit résiduel. Le paramètre μ ∈ [0, ∞) négocie entre la réduction du bruit et de distorsion de la parole. Selon le réglage de $\frac{1}{\mu}$ et la présence de filtre $\mathbf{w_0}$ sur la référence de la parole, le « GSC, (SDW) MWF » ou le « SDR-GSC » est obtenue [36].

- En l'absence du filtre w_0, l'algorithme « SP-SDW-MWF » correspond à un « SDR-GSC » : le critère de conception du bloc « ANC » est complété avec un terme de régularisation $\frac{1}{\mu}\varepsilon_d^2$ limitant la distorsion de la parole due au modèle du signal d'erreurs.

  Pour $\mu = \infty$, l'algorithme GSC est obtenu. Par rapport à l'algorithme « QIC-GSC », le « SDR-GSC » obtient une meilleure réduction du bruit pour les petites erreurs de modélisation du signal, tout en garantissant une robustesse contre les grandes erreurs.

- Comme le « SP-SDW-MWF » prend la distorsion de la parole explicitement en compte dans le critère de conception, un filtre w_0 sur la référence de la parole peut être ajouté. Pour $\mu = 1$, on obtient l'algorithme « MWF ». Comparé au « SDR-GSC », la performance est moins affectée par des modèles d'erreurs.

3.6.2 Algorithme de gradient stochastique (GS)

a Mise en œuvre dans le domaine temporel

Un algorithme de gradient stochastique approxime l'algorithme du gradient classique :

$$w_{n+1} = w_n + \rho \left(\frac{\partial J(w)}{\partial w} \right)_{w=wn} \quad , \tag{3.21}$$

En utilisant une estimation du gradient instantanée, remplacement l'indice d'itération *n* par un indice temps *k* et en laissant de côté les valeurs moyennes, on obtient l'équation de mise à jour suivante pour la fonction de coût (3.16) :

$$w_{k+1} = w_k + \rho \{Y_k^n (y_0^n[k-\Delta] - y_k^{n,T} w_k) - r_k\} \ , \tag{3.22}$$

$$r_k = \frac{1}{\mu} y_k^s y_k^{s,T} w_k \ , \tag{3.23}$$

Avec : $w_k, y_k \in \mathbb{R}^{NL \times 1}$, où N désigne le nombre de canaux d'entrée de filtre adaptatif (N = M si w_0 est présent, N = M - 1 si w_0 est absent). Pour $\frac{1}{\mu} = 0$ et le filtre w_0 absent, l'équation (3.22) se réduit à une mise à jour de type LMS, formule souvent utilisée dans GSC. Elle est sollicitée pendant les périodes de bruit seulement.

Le terme additionnel r_k dans (3.22) limite la distorsion de la parole due au modèle de signal d'erreurs y_i avec i = 0, ..., M-1

L'équation (3.22) nécessite la connaissance de la matrice de corrélation $y_k^S y_k^{S,T}$ ou $E\{ y_k^S y_k^{S,T}\}$ de la parole propre. Dans la pratique, cette information n'est pas disponible. Pour éviter la nécessité d'étalonnage, les (L ×1) dimensions de parole ainsi que les vecteurs de signaux de bruit $y_i[k]$, i = M - N, ..., M - 1 sont stockées dans un tampon circulaire contenant le signal de parole plus le bruit noté $B_1 \in \mathbb{R}^{L_{buf_1} \times N}$ au cours du traitement, comme dans [37]. Pendant les périodes de bruit uniquement (c'est à dire lorsque $y_i[k] = y_i^n[k]$, i = 0, ..., M-1), le filtre w_k est mis à jour en utilisant l'approximation suivante pour (3.23):

$$w_{k+1} = w_k + \rho \{y_k^n (y_0^n[k- \Delta] - y_k^{n,T} w_k) - r_k\}, \quad (3.24)$$

$$r_k = \tilde{\lambda} r_{k-1} + (1-\tilde{\lambda}) \frac{1}{\mu} (y_k^{buf_1} y_k^{buf_1,T} - y_k^n y_k^{n,T}) w_k \quad (3.25)$$

Où $y_k^{buf_1}$ un vecteur contenant le signal de parole + du bruit construit à partir de données du tampon B_1.

Dans la suite un pas normalisée ρ est utilisée:

$$\rho = \frac{\rho'}{\zeta_k + y_k^{n,T} y_k^n + \delta} \quad (3.26)$$

$$\zeta_k = \tilde{\lambda} \zeta_{k-1} + (1- \tilde{\lambda}) \frac{1}{\mu} | y_k^{buf_1,T} y_k^{buf_1}[k] - y_k^{n,T} y_k^n |. \quad (3.27)$$

Un stockage supplémentaire de vecteurs bruit seul y_i^n, i = 0, ……, M - 1 dans un deuxième tampon $B_2 \in \mathbb{R}^{L_{buf_2} \times M}$ permet d'adapter w_k aussi pendant les périodes de la parole + bruit, en utilisant

$$w_{k+1} = w_k + \rho \{y_k^{buf_2} (y_0^{buf_2}[k- \Delta] - y_k^{buf_2,T} w_k) - r_k\}, \quad (3.28)$$

$$r_k = \tilde{\lambda} r_{k-1} + (1-\tilde{\lambda}) \frac{1}{\mu} (y_k y_k^T - y_k^{buf_2} y_k^{buf_2,T}) w_k \quad (3.29)$$

Avec $y_k^{buf_2}$ un vecteur de bruit construit à partir de données dans le tampon B_2.

Remarque : Pour λ = 0 et μ > 1, un algorithme de gradient stochastique similaire à [37] peut être dérivé de (3.24) - (3.29) en invoquant des hypothèses d'indépendance.

Pour λ = 0, l'estimation (3.25), (3.29) de r_k est très mauvaise en raison de grandes différences entre les matrices de rang un $y_i^n y_i^{n,T}$ et $y_j^n y_j^{n,T}$ à des instants de temps différents i et j. Il en résulte de grandes erreurs en excès, en particulier pour un μ

faible et un pas ρ'' grand [61]. L'utilisation d'une estimation de la moyenne de la matrice de corrélation $E\{y_k^s\ y_k^{s,T}\}$ dans (3.23), permettrait d'améliorer sensiblement la performance, mais nécessite beaucoup d'opérations matricielles. Soit :

$$r_k = \frac{1}{\mu}\frac{1}{K}\left(\sum_{l=k-K+1}^{k} y_l^{buf_1} y_l^{buf_1,T} - \sum_{l=k-K+1}^{k} y_l^{n} y_l^{n,T}\right) w_k \qquad (3.30)$$

Par conséquent, en supposant que w_k varie lentement dans le temps, (3.25), (3.31) et en particulier pour de faible $\tilde{\lambda}$, on obtient une bonne approximation de (3.32) sans opérations matricielles. Pour un bruit stationnaire, un faible K ou $\tilde{\lambda}$ suffit (avec K = 1 / (1-$\tilde{\lambda}$) ~ ML) [61]. Dans la pratique, les signaux de parole et de bruit sont souvent spectralement fortement non stationnaire (présence de multi-locuteur), tandis que leurs caractéristiques spectrales et spatiales à long terme telles que les positions des sources varient généralement plus lentement dans le temps.

b Algorithme(GS) la mise en œuvre domaine fréquentiel

- **Initialisation et définitions des matrices :**

$\boldsymbol{W}_i$ [0] = [0 . . . 0]^T, *i*=*M*−*N*, …, *M*−1

$\boldsymbol{P}_m$ [0] = δ_m , *m* = 0, …., 2*L*−1 ;

F = 2 L × 2 L DFT matrice;

g= $\begin{bmatrix} I_L & 0_L \\ 0_L & 0_L \end{bmatrix}$; **k**= [$0_L I_L$] ;

$\mathbf{0}_L$ = L ×L matrice de zéros ; $\mathbf{I}_L$ = L × L matrice d'identité

- **Pour chaque nouveau bloc de ML d'échantillons d'entrée :**

Si le bruit détecté:

d[*k*] = $[y_0[kL-\Delta]$ …. $y_0[kL-\Delta+L-1]]^T$

$\mathbf{Y}_i^n[k] = diag\{F\ [y_i[kL-L]\ \ldots y_i[kL+L-1]\ ]^T\}$

Données d'entrée $\mathbf{Y}_i^n[k]$, ***d*** [*k*] en tampon de bruit B_2.

Créer $\boldsymbol{Y}_i[k]$ à partir de données de la parole + tampon de bruit B_1.

Si la parole détecté:

$\mathbf{Y}_i[k] = diag\{F\ [y_i[kL-L]\ \ldots y_i[kL+L-1]\ ]^T\}$

Données d'entrée $\mathbf{Y}_{\mathrm{i}}[k]$ dans la parole + tampon de bruit B_1.

Créer $\mathbf{Y}_{\mathrm{i}}^{\mathrm{n}}[k]$, $\mathbf{d}[k]$ à partir de données tampon de bruit B_2.

Le diagramme suivant correspond Schéma bloc de l'algorithme « SDW-MWF » dans le Domaine fréquentiel.

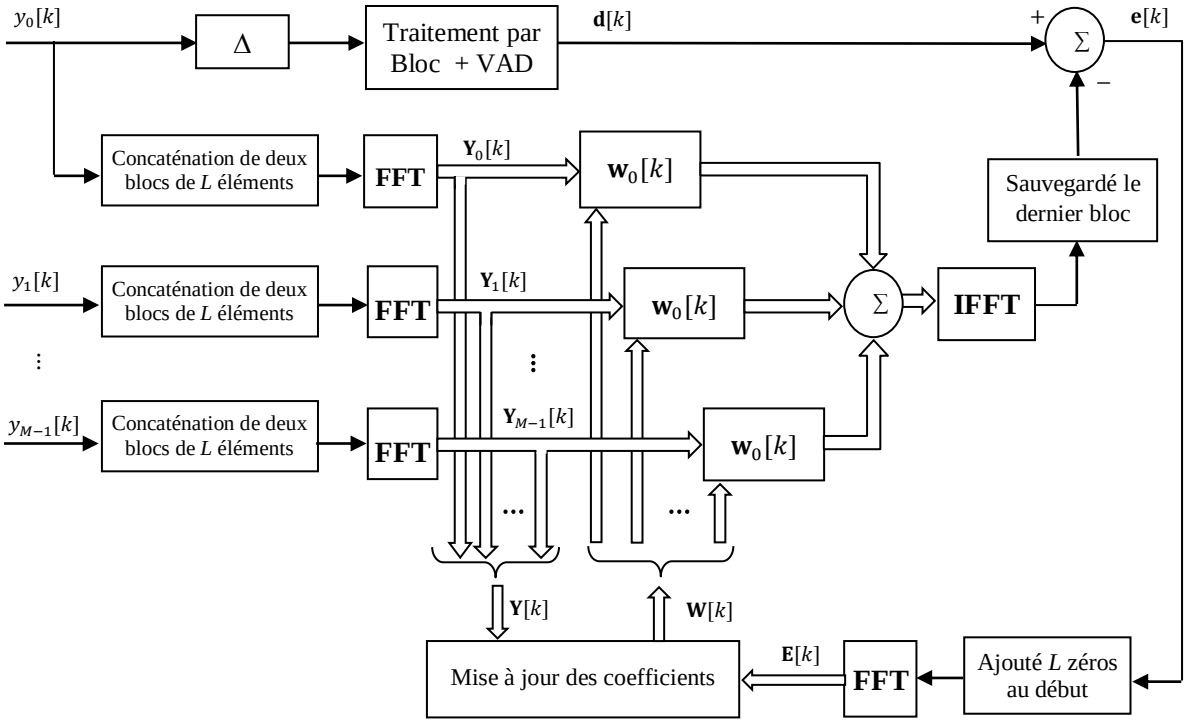


Figure 3. 6. Schéma bloc de l'algorithme SDW-MWF dans le domaine fréquentiel

- **Mise à jour de la formule:**

$\mathrm{W_i}[k+1] = \mathbf{W}_{\mathrm{i}}[k] + \mathbf{F}\mathbf{g}\mathbf{F}^{-1}\mathbf{\Lambda}[k]\left\{\mathbf{Y}_{\mathrm{i}}^{\mathrm{n,H}}[k]\mathbf{E}[\mathrm{k}] - \mathbf{R}_{\mathrm{i}}[\mathrm{k}]\right\}$,

$\mathbf{R}_{\mathrm{i}}[k] = \lambda\mathbf{R}_{\mathrm{i}}[k-1] + (1-\lambda)\frac{1}{\mu}(\mathbf{Y}_{\mathrm{i}}^{\mathrm{H}}[k]\ \mathbf{E}_2[k] - \mathbf{Y}_{\mathrm{i}}^{\mathrm{n,H}}[k]\ \mathbf{E}_1[k])$

Avec

$\mathbf{E}[k] = \mathbf{F}\mathbf{k}^{T}\left(\mathbf{d}[k] - \mathbf{k}\mathbf{F}^{-1}\sum_{\mathrm{j=M-N}}^{\mathrm{M-1}}\mathbf{Y}_{\mathrm{j}}^{\mathrm{n}}[k]\mathbf{W}_{\mathrm{j}}[k]\right)$

$\mathbf{E}1[k] = \mathbf{F}\mathbf{k}^{T}\mathbf{F}\mathbf{k}^{-1}\sum_{\mathrm{j=M-N}}^{\mathrm{M-1}}\mathbf{Y}_{\mathrm{j}}^{\mathrm{n}}[k]\mathbf{W}_{\mathrm{j}}[k] = \mathbf{F}\mathbf{k}^{T}e_1[k]$

$\mathbf{E}2[k] = \mathbf{F}\ \mathbf{k}^{T}\ \mathbf{F}\ \mathbf{k}^{-1}\sum_{\mathrm{j=M-N}}^{\mathrm{M-1}}\mathbf{Y}_{\mathrm{j}}[k]\mathbf{W}_{\mathrm{j}}[k] = \mathbf{F}\mathbf{k}^{T}e_2[k]$

- **Le pas $\mathbf{\Lambda}$ [k]:**

$\mathbf{\Lambda}\ [k] = \frac{2\rho'}{\mathrm{L}}diag\ \{\mathbf{P}_0^{-1}[k],\ \ldots.,\mathbf{P}_{2\mathrm{L}-1}^{-1}[k]\}$

$\mathbf{P}_{\mathrm{m}}[k] = \gamma\mathbf{P}_{\mathrm{m}}[k\text{-}1] + (1\text{-}\gamma)\ (\mathbf{P}_{1,\mathrm{m}}[k] + \mathbf{P}_{2,\mathrm{m}}[k])$

$\mathbf{P}_{1,m}[k] = \sum |Y_{j,m}^{n}|^2$

$\mathbf{P}_{2,m}[k] = \lambda \mathbf{P}_{2,m}[k-1] + (1-\lambda)\frac{1}{\mu}\left|\sum(|Y_{j,m}|^2 - |Y_{j,m}^{n}|^2)\right|$

Sortie z [k]:

$\mathbf{y}_0[k] = [y_0[kL - \Delta] \ldots . y_0[kL - \Delta + L - 1]]T$

- Si le bruit détecté: $\mathbf{z}[k] = \mathbf{y}_0[k] - \mathbf{e}_1[k]$
- Si la parole détecté: $\mathbf{z}[k] = \mathbf{y}_0[k] - \mathbf{e}_2[k]$

3.7 Algorithme SP-SDW-MWF et matrice de blocage adaptative (MBA)

La figure 3.7 montre un GSC robuste dans le domaine temporel avec MBA « matrice de blocage adaptative». La description commence par la formation de faisceau fixe BFF « Beamformer Fixe » qui est implémenté dans le domaine temporel. Et ensuite l'annulation adaptative des lobes secondaires dans le domaine fréquentiel. Le BFF, dans ce cas, est un formateur de faisceau par délai et somme, qui améliore le signal désiré et est utilisé comme référence pour l'adaptation du GSC.

Dans le domaine temporel discret, le retard est calculé en détectant la direction d'arrivé (DA), et est généralement réaliser par des filtres de retard fractionnaire court.

Par conséquent, les deux modules, le BFF et la direction du faisceau DA sont réalisées plus efficacement dans le domaine temporel que dans le domaine fréquentiel.

La sortie de BFF est obtenue à partir de la sommation simple des signaux reçus par des microphones constituants le réseau telle que :

$$y_0(n) = \frac{1}{M}\sum_{i=1}^{M} u_i(n) \qquad (3.33)$$

Avec : *M* le nombre de microphones.

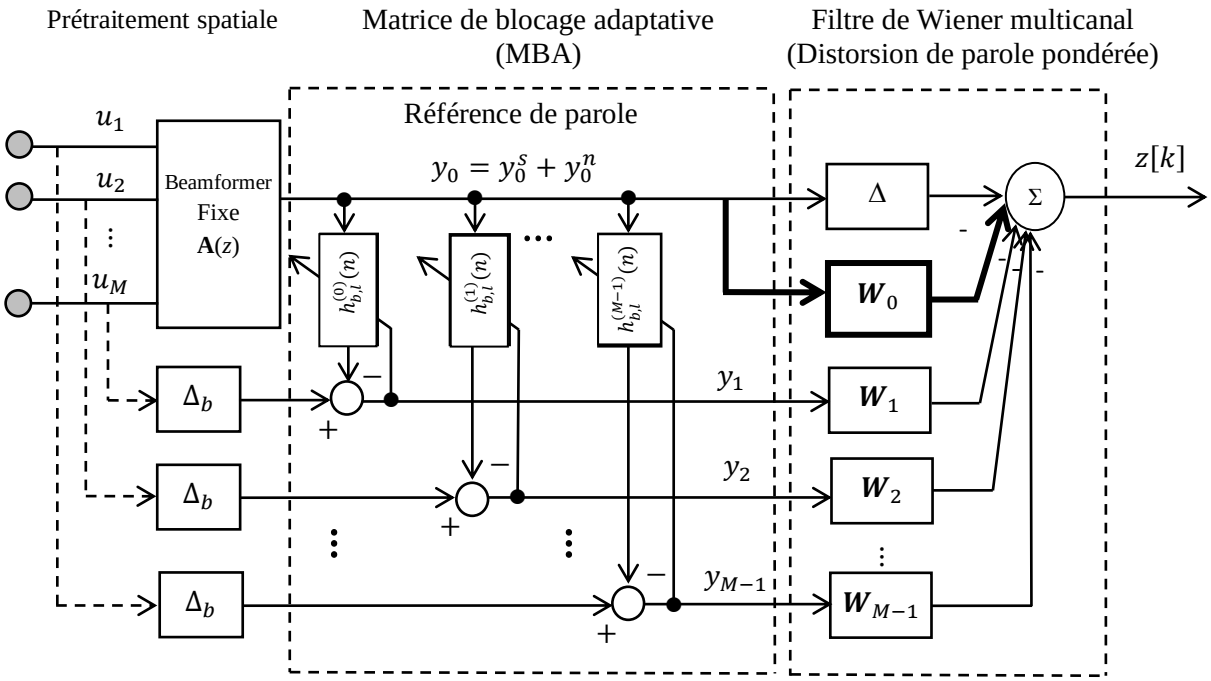


Figure 3. 7. Structure de l'algorithme « prétraitement spatiale – matrice de blocage adaptative -distorsions pondéré de parole - filtre de Wiener multicanal » (SP-ABM-SDW-MWF).

La matrice MBA constituée de filtres adaptatifs entre la sortie de BFF et les sorties des microphones figure 3.8. Le retard Δ_b assure la causalité des filtres adaptatifs. Des contraintes robustesse [62], ne sont pas mises en œuvre.

Dans la suite, on suppose que la taille de bloc est *L* échantillons, la taille de la transformé de Fourier discrète est $2L$ est utilisé pour éviter les effets de convolution circulaire.

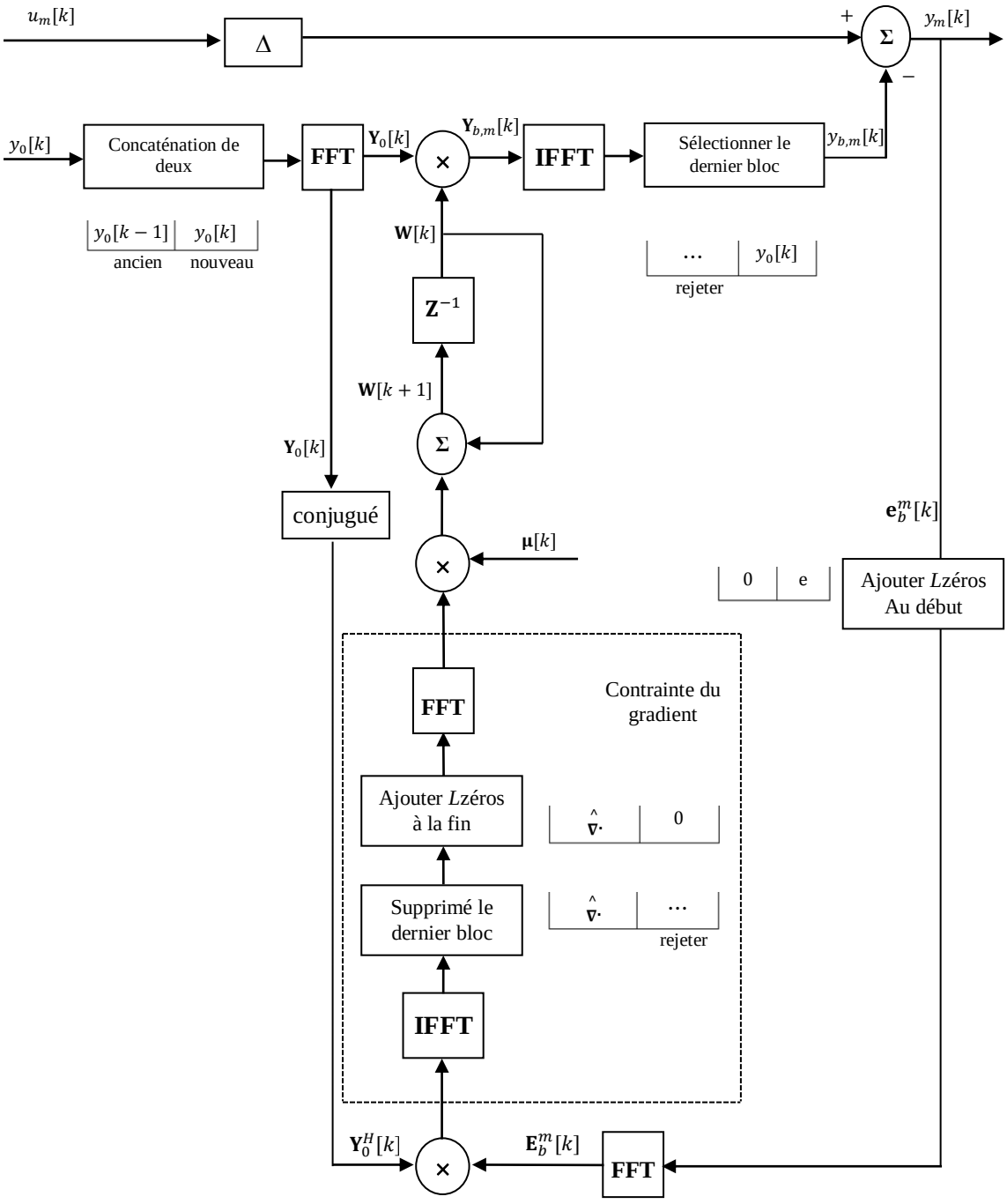


Figure 3. 8. La matrice de blocage adaptative en domaine fréquentiel

La DFT de la sortie de BFF, $\mathbf{Y}_0[k]$ peut s'écrit comme une matrice diagonale :

$$\mathbf{Y}_0[\mathrm{k}] = diag\{\mathbf{F}\ [\mathrm{y_i}[kL-L]\ \ldots \mathrm{y_i}[kL+L-1]\ ]^T\}\,, \tag{3.34}$$

Où k est l'indice de bloc courant, et F est la matrice de TFR 2L×2L

Dans le domaine fréquentiel, $\mathbf{H}_b^{(m)}[k]$sont les fonctions de transfert du filtre adaptatif, avec m=0, ⋯,M−1, sont calculé à partir des filtres $\boldsymbol{h}_{b,l}^{(m)}[k]$dans le domaine temporel avec l = 0, ⋯,L−1, c'est-à-dire :

$$\mathbf{H}_b^{(m)}[k] = \mathbf{F}\left(h_{b,0}^{(m)}[k], \cdots, h_{b,L-1}^{(m)}[k], \quad 0_{1\times L}\right)^T \tag{3.35}$$

La sortie de BFF, $\mathbf{Y}_0[k]$ est filtrée par les filtres adaptatifs $\mathbf{H}_b^{(m)}[k]$qui donnent :

$$\mathbf{Y}_{\mathrm{b,m}}[\mathrm{k}] = \mathbf{Y}_0[\mathrm{k}]\ \mathbf{H}_{\mathrm{b}}^{(\mathrm{m})}[\mathrm{k}]$$

L'adaptation de l'algorithme MBA nécessite des signaux d'erreur qui sont sans effets de convolution circulaire, par conséquent, dans le domaine temporel les signaux $\mathbf{e}_b^{(m)}[k]$ doivent être coupés de telle sorte que les L premier échantillons du bloc sont rejetés et les L dernier restent inchangés :

$$\mathbf{e}_b^{(m)}[k] = \mathbf{u}_m[k-\Delta_b] - \mathbf{w}\mathbf{F}^{-1}\mathbf{Y}_0[k]\mathbf{H}_{\mathrm{b}}^{(\mathrm{m})}[\mathrm{k}] \tag{3.36}$$

Avec : **w** =*diag*{ ($\mathbf{0}_{1\times L}$, $\mathbf{1}_{1\times L}$)} est une matrice diagonale constitué par des zéros sur la moitié supérieure de la diagonale principale et avec des uns sur la moitié inférieure de la diagonale principale .

Le vecteur $\mathbf{u}_m[k]$ est définit par :

$$\mathbf{u}_m[k] = [\mathbf{0}_{1\times L}\ , (\ u_m[kL]\ , \ldots, u_m[kL+L-1])]^T \tag{3.37}$$

La mise à jour des filtres MBA est donnée par la formule suivante :

$$\mathbf{H}_b^{(m)}[k+1] = \mathbf{H}_{\mathrm{b}}^{(\mathrm{m})}[\mathrm{k}] \ + \ \boldsymbol{\mu}[k]\ \mathbf{Y}_0^H[k]\mathbf{E}_{\mathrm{b}}^{(\mathrm{m})}[\mathrm{k}] \tag{3.38}$$

La matrice du pas normalisée $\boldsymbol{\mu}[k]$ est définie par :

$$\boldsymbol{\mu}[\mathrm{k}] = 2\mu\ \mathrm{diag}\ \{\mathbf{P}_0^{-1} \cdots \mathbf{P}_{2\mathrm{L}-1}^{-1}\} \tag{3.39}$$

Avec μ le pas d'adaptation fixe, et P_l est l'énergie du l-ième bin de fréquence :

$$\boldsymbol{P}_l[k] = \lambda \boldsymbol{P}_l[k-1] + (1-\lambda)\left|\boldsymbol{Y}_{0,l}\right|^2, \qquad l = 0, \ldots, 2L-1 \tag{3.40}$$

Où $\boldsymbol{Y}_{0,l}$[k] est la le l-ième bin de fréquence de $\mathbf{Y}_0[k]$.

3.8 Conclusion

Dans ce chapitre, nous avons développé un système de rehaussement de la parole utilisant les deux techniques « GSC » et « SDW-MWF » intégrées dans un seul système. Ce dernier a été proposé dans [14] et nommé « Prétraitement Spatial – Distorsion Pondéré de la Parole – Filtre Multicanal de Wiener » ou en anglais « the Spatially Preprocessed - Speech Distortion Weighted- Multi-channel Wiener Filter » « SP-SDW-MWF ». Il est constitué de deux parties, l'une est fixe (Prétraitement Spatial) et l'autre adaptative « SDW-MWF ». A la technique proposée dans [14] nous avons intégré une matrice de blocage adaptative pour une meilleure estimation des interférences dans le système proposé. Les résultats sont présentés dans le chapitre suivant.

Chapitre 4 Implémentation et résultats

4.1 Introduction

Les applications développées dans le cadre de ce projet se divisent en deux parties figure 4.1. La première partie est l'interfaçage des éléments matériels au micro-ordinateur, alors que la deuxième partie concerne le travail algorithmique la réduction du bruit en utilisant les algorithmes développés.

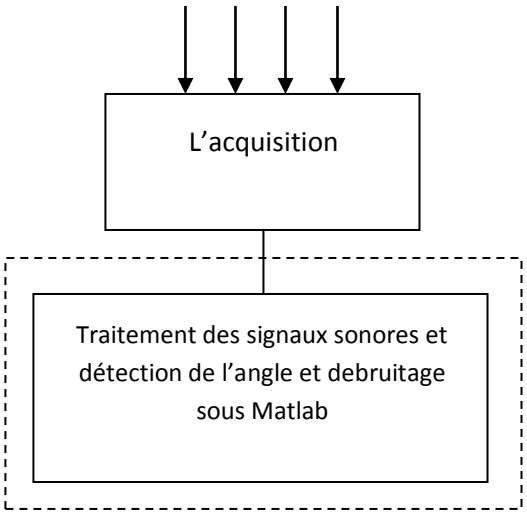


Figure 4. 1.Interfaçage et traitement

À l'entrée du système, on retrouve plusieurs sources sonores qui peuvent prendre diverses formes : clappement de mains, parole, bruit de sirène, etc. Ces sources sonores sont entremêlées entre elles et le bruit ambiant peut provenir de diverses sources : système de ventilation, bruit des moteurs du robot, etc. Le son environnant est capturé par une série de microphones. Quatre microphones omnidirectionnels peu dispendieux sont utilisés.

La première partie représente l'acquisition qui consiste en l'utilisation d'une série de quatre microphones omnidirectionnels identique (Annexe 1). Les fichiers issus de cette dernière sont utilisés pour l'estimation de la direction de l'angle d'incidence et le rehaussement des signaux au moyen d'algorithmes implémentés sous Matlab (ver 7.8.0).

4.2 Le module d'acquisition des sons

Le NI PCI-4472B est un module d'acquisition de données DAQ haute précision spécifiquement conçu pour les applications acoustiques et vibrations à grand nombre de voies. 8 entrées analogiques optimisées pour la mesure de vibrations et échantillonnées simultanément (jusqu'à 102,4 kéch./s) (voir annexe) . Quatre microphones de type WM-54B sont utilisés.

Figure 4. 2. La carte NI PCI-4472B et les microphones WM-54B

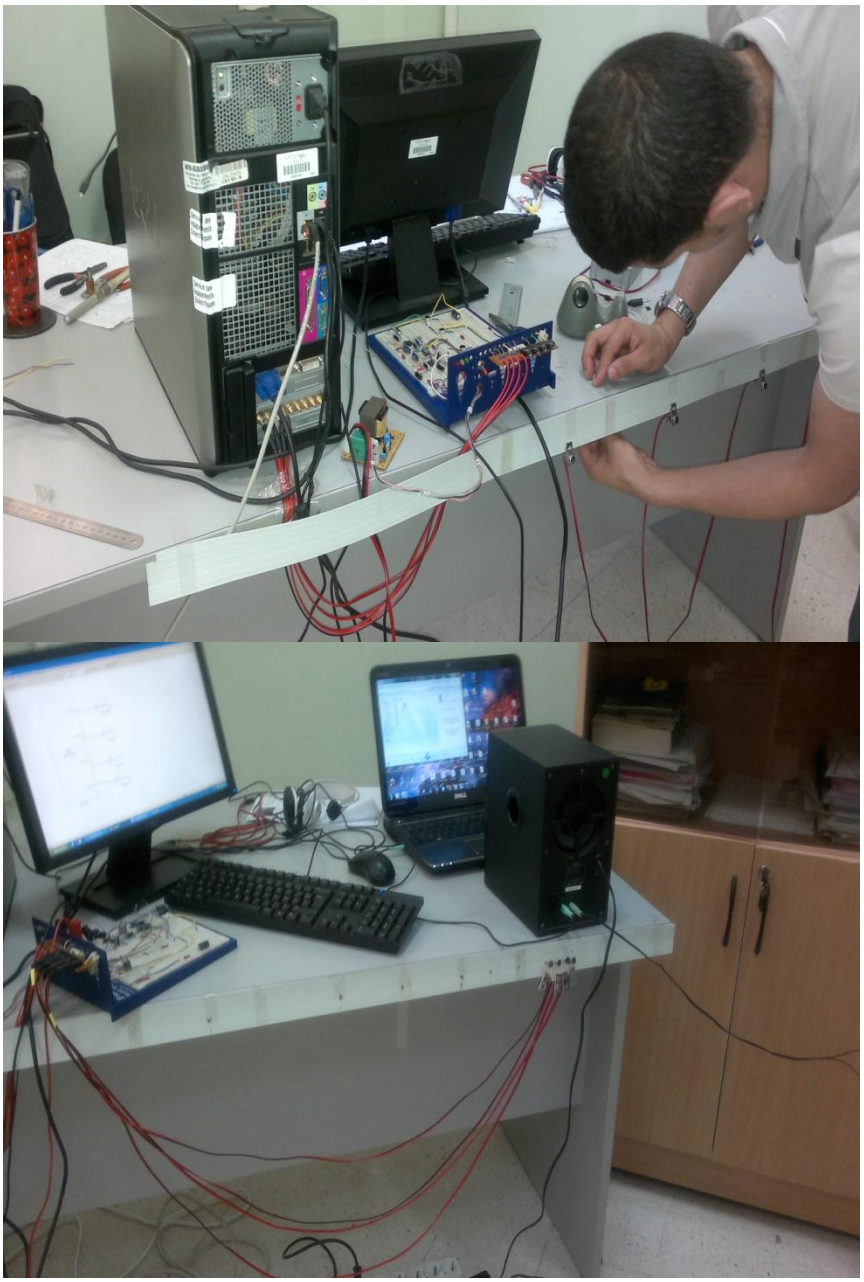


Figure 4. 3. Dispositif expérimental de l'acquisition avec 4 microphones

4.3 Les critères utilisés pour mesurer les performances

Les critères temporels et fréquentiels se basent essentiellement sur l'évaluation de la qualité en termes de comparaison de distorsion de formes entre signal de référence et signal débruité, sans tenir compte de l'aspect perceptif. Certes, c'est une condition nécessaire mais non suffisante dans la mesure où deux signaux pratiquement de même forme peuvent être perçus différemment [63], d'où l'intérêt d'introduire le facteur psychoacoustique pour tout système ayant pour objectif de conserver la qualité de la parole. Diverses mesures objectives perceptuelles sont élaborées conduisant à de bonnes corrélations avec la perception humaine. Elles sont essentiellement dédiées au codage de la parole, mais trouvent leur application en débruitage de la parole ([64], [65], [66]). A part le fait qu'elles donnent une meilleure corrélation avec la qualité vocale, leur application en débruitage n'a pas été justifiée jusqu'à présent. En guise d'illustration, citons la mesure de la qualité de la parole perçue (PSQM) (Perceptual Speech Quality Measure) [67] et sa version améliorée PESQ (Perceptual Evaluation of Speech Quality) [68], le BSD (Bark Spectral Distortion)[63] et sa version améliorée, MBSD (Modified Bark Spectral Distortion) [69-70].

4.3.1 SNR segmental (segSNR)

Le SNR (Signal to Noise Ratio) segmental segSNR est la mesure de qualité objective la plus utilisée dans le domaine temporel. Il définit la moyenne des SNRs issus de plusieurs segments de courte durée (15 à 20 ms) :

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{10}{\mathrm{M}} \sum_{\mathrm{m}=0}^{\mathrm{M}-1} \log_{10} \frac{\sum_{\mathrm{i}=\mathrm{mN}}^{\mathrm{mN}+\mathrm{N}-1} \mathrm{S}^2(\mathrm{i})}{\sum_{\mathrm{i}=\mathrm{mN}}^{\mathrm{mN}+\mathrm{N}-1} (\mathrm{s}(\mathrm{i}) - \hat{\mathrm{s}}(\mathrm{i}))^2} \tag{4.1}$$

Où s(i), $\hat{s}$(i), N et M sont respectivement le signal de référence, le signal débruité, la longueur d'un segment et le nombre total de segments.

Le SNR segmental souffre de deux limitations : d'abord si le signal de parole contient des segments de silence, ce qui est très probable, le s(i) sera nul et n'importe quelle quantité de bruit entraînera un SNR en dB négatif pour ce segment ; du coup le SNR total sera biaisé par cette quantité. Ce problème peut être résolu partiellement en choisissant un seuil d'énergie au-delà duquel le SNR segmental sera calculé. Ensuite, il

faut nécessairement que les deux signaux comparés soient alignés temporellement car ce critère est très sensible aux déphasages.

4.3.2 SNR par bande de fréquence

On a découpé le spectre en 20 bandes de fréquence, dont la largeur de bande est en rapport avec l'échelle de mel. Cette dernière est une échelle de fréquences basé sur la perception humaine. Pour chaque bande on a calculé le rapport signal/bruit. Cette information nous permettra d'obtenir l'amélioration du signal débruité pour chaque bande de fréquence.

4.4 Résultats expérimentaux et choix du VAD

Pour la détection des zones de silence et de parole, nous avons implémenté et comparé quatre VADs.

1- VAD-1 basé sur l'énergie courte-terme et le taux de passage par zéro
2- VAD-2 basé sur un filtrage optimal de l'énergie court-terme
3- VAD-3 basé sur l'analyse de l'énergie court-terme en sous bandes de fréquence
4- VAD-4L'algorithme de VAD de l'annexe G.729 B de l'ITU

Les résultats obtenus nous permettrons de choisir le meilleur VAD.
Les signaux utilisés pour tester les VAD sont de deux types.

1- signaux simulés pour un VAD idéal
2- signaux réels enregistrés localement (six + base de données)

Le VAD ainsi choisi, est utilisé dans nos implémentations pour le rehaussement des signaux de parole.
Les tableaux suivants, représentent trois métriques, à savoir le taux de bonne détection de signal de parole, le taux de fausse décisions et le manque à détecter. Les fichiers test sont issus de la base de données IEEE (720 phrases) [73]. Elle a été utilisée car elle contient des phrases phonétiquement équilibrés. Les signaux ont été à l'origine échantillonnés à 25 kHz et sous-échantillonné à 8 kHz. Les phrases ont été produites par trois hommes et trois femmes.

Dans de base de données le bruit est artificiellement ajouté au signal de parole comme suit. Le niveau vocal actif du signal de parole propre filtré est d'abord déterminée en utilisant la méthode B de l'UIT-T P.56 [74].

Un segment de bruit de la même longueur que le signal de parole est découpée de manière aléatoire des enregistrements sonores, et ajusté de manière appropriée pour atteindre le niveau SNR désiré et il est finalement ajouté au signal de parole propre filtré.

Les signaux de bruit ont été pris à partir de la base de données AURORA [75] et inclus dans les enregistrements pour simuler différents scénarios: « Babble (crowd of people), Car, Exhibition hall, Restaurant, Street, Airport, Train station, Train ».

Chaque contexte contient 30 enregistrement, nous avons pris le premier enregistrement de chacune des situations, les résultats sont transcrits dans les tableaux suivants : les situations choisies sont : « airport, street, babble, restaurant et bruit blanc ».

SNR	**VAD1**			**VAD2**			**VAD3**			**VAD4**		
	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)
0dB	70.00	16.43	13.57	55.36	26.43	18.21	55.00	26.07	18.93	61.07	37.14	1.79
5dB	70.00	29.29	0.71	55.00	27.50	17.50	61.43	23.39	15.18	64.64	35.00	0.36
10dB	85.00	11.43	3.57	56.07	27.14	16.79	59.64	18.39	21.96	60.71	37.86	1.43
15dB	87.86	10.71	1.43	56.07	26.79	17.14	57.50	18.39	24.11	59.29	40.71	0.00

Tableau 4. 1. Qualité de détection en fonction de SNR pour différents VADs pour « Airport SP 01 »

SNR	**VAD1**			**VAD2**			**VAD3**			**VAD4**		
	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)
0dB	72.14	26.43	1.43	56.43	26.07	17.50	51.43	35.54	13.04	66.79	32.50	0.71
5dB	77.14	20.71	2.14	55.71	26.79	17.50	47.86	33.04	19.11	61.79	38.21	0.00
10dB	87.86	7.86	4.29	54.64	27.14	18.21	56.96	18.93	24.11	61.43	38.57	0.00
15dB	83.57	15.00	1.43	56.43	27.14	16.43	58.04	17.32	24.64	58.92	40.36	0.71

Tableau 4. 2. Qualité de détection en fonction de SNR pour différents VADs pour « Restaurant SP 01 »

SNR	VAD1			VAD2			VAD3			VAD4		
	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)
0dB	67.86	29.29	2.85	67.14	23.57	9.29	48.04	40.71	11.25	60.71	37.50	1.79
5dB	67.86	32.14	0.00	54.29	28.57	17.14	63.39	17.86	18.75	58.93	40.00	1.07
10dB	80.00	15.71	4.29	56.43	26.79	16.79	58.40	18.75	22.85	61.43	38.57	0.00
15dB	74.29	25.71	0.00	55.36	26.79	17.85	56.61	19.11	24.28	61.79	38.21	0.00

Tableau 4. 3.Qualité de détection en fonction de SNR pour différents VADs pour « Babble SP 01 »

SNR	VAD1			VAD2			VAD3			VAD4		
	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)
0dB	65.71	32.14	2.14	59.29	17.86	22.85	55.71	27.86	16.43	65.36	34.64	0.00
5dB	65.00	30.71	4.29	55.71	27.14	17.15	47.86	18.93	33.21	59.29	40.71	0.00
10dB	92.14	5.71	2.14	55.00	27.50	17.50	57.50	19.46	23.04	58.57	40.00	1.43
15dB	78.57	21.43	0.00	56.07	26.07	17.86	57.68	18.04	24.28	59.29	40.71	0.00

Tableau 4. 4. Qualité de détection en fonction de SNR pour différents VADs pour « Street SP 01 »

SNR	VAD1			VAD2			VAD3			VAD4		
	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)	Bon (%)	Fausse (%)	Manque (%)
0dB	67.14	32.86	0.00	43.21	37.86	18.93	59.29	40.71	0.00	61.07	34.64	4.29
5dB	60.72	32.14	7.14	63.21	36.79	0.00	53.39	40.70	5.89	61.79	30.36	7.86
10dB	60.71	30.00	9.29	55.36	38.57	6.07	53.39	40.18	6.43	57.50	38.57	3.93
15dB	64.29	32.86	2.85	52.86	40.00	7.14	68.75	31.25	0.00	62.86	31.43	5.71

Tableau 4. 5. Qualité de détection en fonction de SNR pour différents VADs pour Bruit blanc

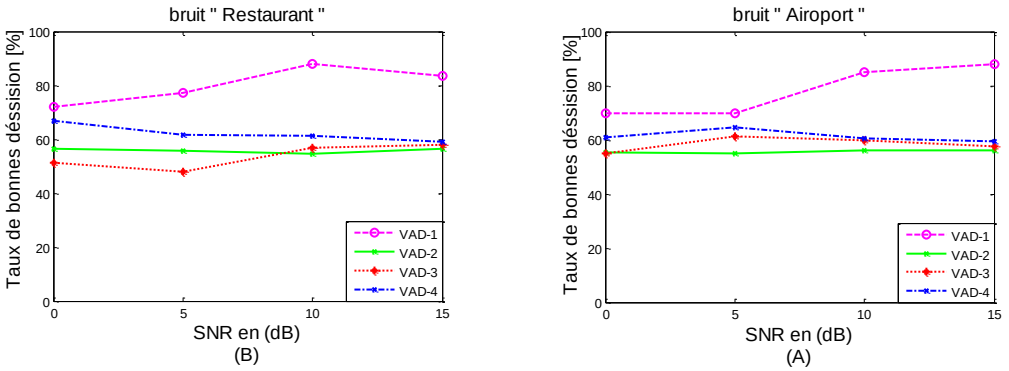


Figure 4. 4. (A) :Les 4 algorithmes dans un milieu « aéroport », (B) :dans un milieu « restaurants »

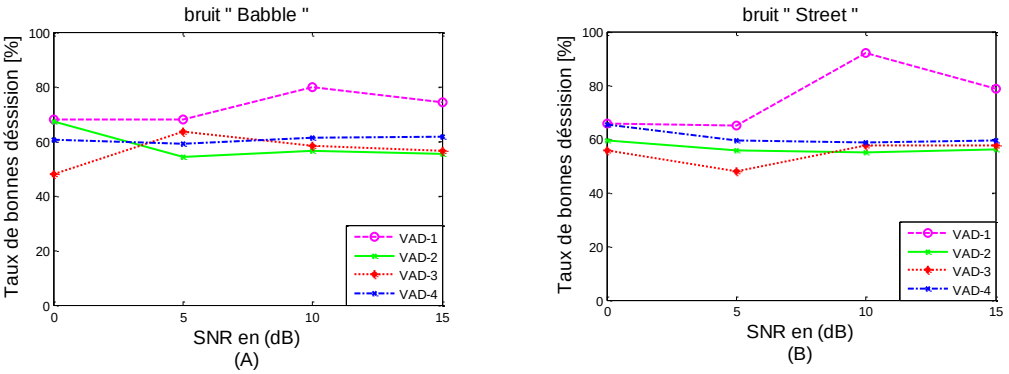


Figure 4. 5.(A) **:**Les 4 algorithmes dans un milieu « Babble », (B) : dans un milieu «street »

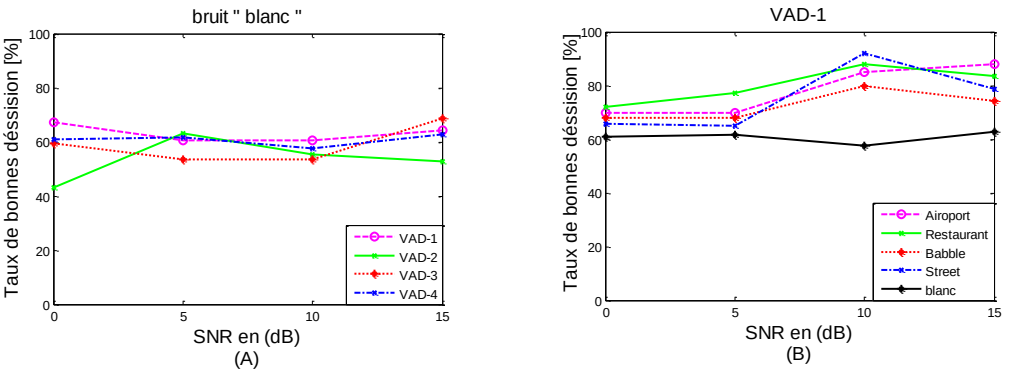


Figure 4. 6.(A) :Les 4 VAD bruité avec un bruit blanc, (B) :le VAD-1 dans les différents milieux

Dans la série de figures nous avons reporté les résultats des tests dans le but de pouvoir choisir le VAD donnant de meilleurs scores de bonnes décisions.

Pour une situation de type « airport » figure 4.4.(A) c'est-à-dire le contexte du fichier testé se passe dans un aéroport, le vad1 semble donner de meilleurs résultats. On remarque sur la courbe que pour un rapport signal/bruit de 0dB, on a un taux de bonne décisions 70%, et plus le rapport signal/bruit augmente plus le taux de bonnes décisions augmente jusqu'à atteindre 90% correspondant à 15dB.

Pour les autres VADs, le taux de bonnes décisions, varient entre 50 % et 60 % et on remarque une dégradation du taux de bonnes décisions concernant les VADs 2 et 3 au-dessus de 5 dB.

Pour une situation de type « restaurant » figure 4.4.(B) c'est-à-dire le contexte du fichier testé se passe dans un restaurant, les meilleurs résultats concernent aussi le vad1. On remarque sur la courbe que pour un rapport signal/bruit de 0dB, on a un taux de bonne décisions > 70%, et plus le rapport signal/bruit augmente plus le taux de bonnes décisions augmente jusqu'à atteindre 90% correspondant à 15dB.

Pour les autres VADs, le taux de bonnes décisions, varient entre 50% et 60% et on remarque une dégradation du taux de bonnes décisions concernant les VADs 2 et 3 au-dessus de 5 dB.

La courbe 4.6.(B) illustre les résultats obtenus pour le VAD-1 pour différentes situations. On remarque que le taux de bonne décision se situe globalement entre 60 à 90%, ce qui renforce notre choix du VAD-1

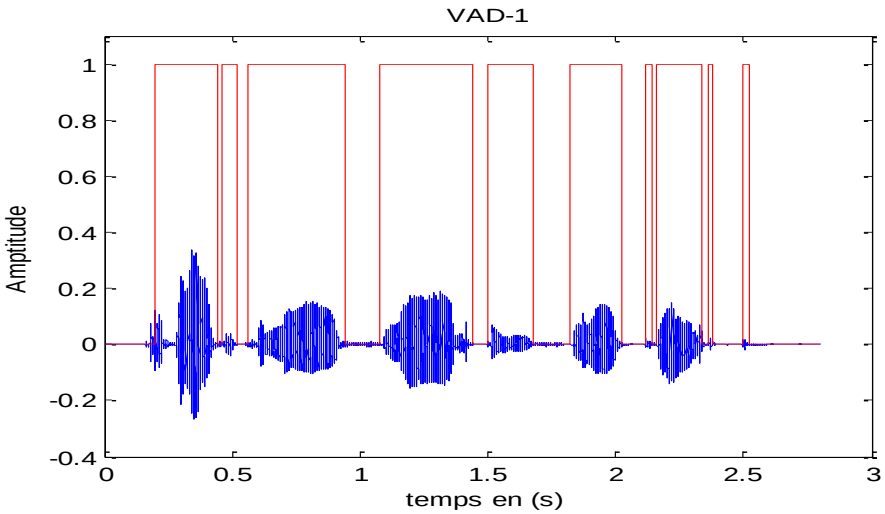


Figure 4. 7. Exemple de découpage silence/parole pour le VAD-1

La figure 4.7 donne un exemple de découpage silence/parole pour le VAD-1

4.5 Résultats expérimentaux : algorithmes de réhaussement du signal

Dans cette section nous présentons les résultats experimentaux pour les algorithmes devellopés.

Nous avons réaliser une base de données de signaux acquis par la carte NI PCI 4472B, le nombre de microphones est de 4 (voir annexe). Nous avons fait varier l'angle d'acquisition de la source , ainsi que les distances intermicrophones et la fréquence d'échantillonnage .

4.5.1 Temps d'execution

On a mesuré le temps d'exucution pour dix itérations et on a calculé la moyenne. Cette mesure est faite dans le domaine temporel et le domaine fréquentiel en utilisant la technique de filtrage « overlap-save ». On obtient les temps d'execution suivants :

Dans le domaine temporel la moyenne d'une seule itération qui contient 32 échantillons de signal filtré est de 39.3ms. Par contre, en utilisant la méthode « overlap-save » dans le domain fréquentiel, on obtient 8.11 ms par étiration. Donc ce dernier est à peu près 5 fois plus rapide. Avec ce résultat on a justifié le chois de l'approche d' « overlap-save » utilisée.

4.5.2 Les variations de diffirents SNR en fonction de l'angle incident

On a choisi pour ce test, des signaux de fréquence d'échantillonnage fe = 16 kHz et la distance inter microphone d= 20cm. L'algorithme utilise un beamformer fixe avec une matrice de blocage adaptative. Dans le bloc « ANC », on fixe le pas d'adaptation ρ = 0.03 , avec un facteur de compromis μ = 300.

Le tableau 4.6 contient les mesures des différents SNR (in, out, seg, bande). On notera que le SNRbande représente la moyenne des SNR de chaque bande fréquentielle.

Angle en degré	Angle détecté	SNRin en dB	SNRout en dB	SNRseg en dB	SNRbande en dB
0	1	1.5347	6.5719	5.9507	15.7521
20	1	0.0188	3.4718	2.1068	8.0553
30	27	0.4661	8.1308	4.7476	12.2191
45	39	0.4377	12.8124	10.4785	19.3577
60	56	0.0323	3.5757	2.6478	8.4154
80	70	0.0154	19.5213	15.1187	26.4793
90	76	0.0539	3.4287	3.0475	7.7821
120	107	0.0604	3.8305	2.1575	6.9282
140	127	0.2545	8.9672	3.5427	11.7169
160	144	0.1999	22.8988	22.6716	29.2962
180	156	0.2780	6.0603	4.3653	12.1529

Tableau 4. 6. les différente SNR en fonction de l'angle incident

La figure 4.8 représente les SNRin et SNRout en fonction de l'angle d'incidence de la source.

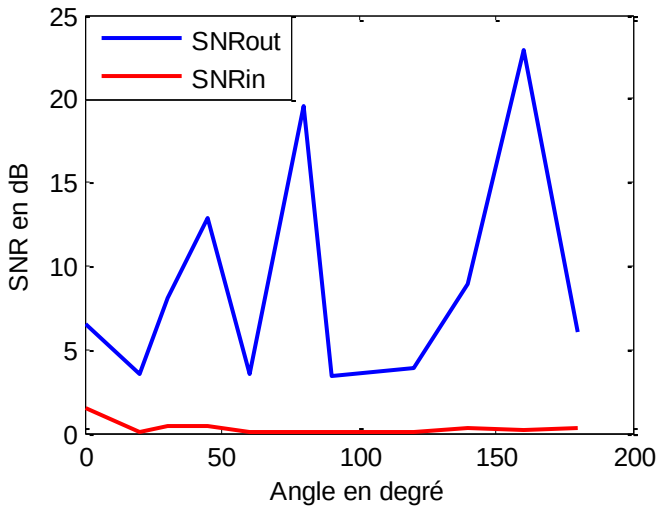
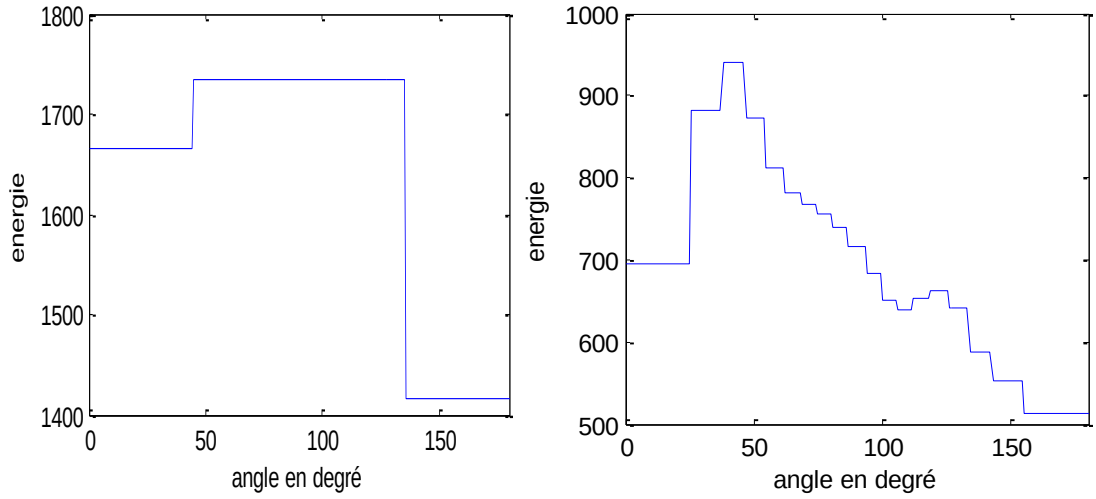


Figure 4. 8. Variation de SNRin et SNRout en fonction de l'angle de la source

On remarque que quelque soit l'angle en a un bon rehaussement de la parole.

4.5.3 Influence de distance inter-microphones sur la détection de l'angle d'incidence de la source

On a choisi pour ce test, un signal de fréquence d'échantillonnage fe = 16 kHz, et l'angle d'incidence de 45°. Pour une distance inter-microphone d=1cm correspondant à la figure 4.9 (A) et d=20cm correspondant à la figure 4.9 (B), on observe une bonne précision de détection de la source pour une distance de 20 cm.

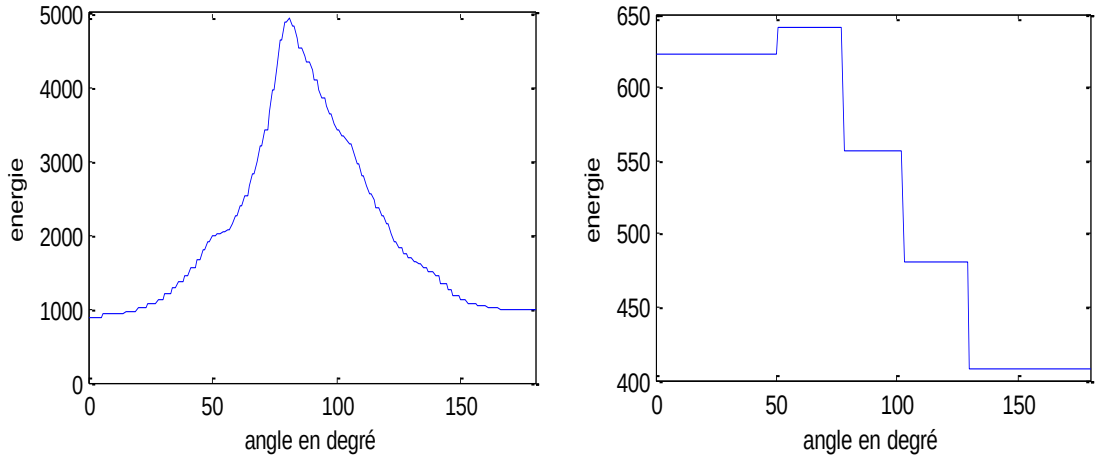


(A) (B)

Figure 4. 9. Influence de distance inter-microphones « (A) : d=1cm et (B) :d=20cm » sur la détection de l'angle d'incidence de la source

4.5.4 Influence de la fréquence d'échantillonnage sur la détection de l'angle d'incidence de la source

On a choisi pour ce test, une distance inter-microphone d=5cm, et l'angle d'incidence de 60°. Pour un signal de fréquence d'échantillonnage fe=64kHz correspondant à la figure 4.10 (A) et fe=16kHz correspondant à la figure 4.10 (B), on observe une bonne précision de détection de la source pour une fréquence de 64kHz.



(A) (B)

Figure 4. 10. Influence de la fréquence d'échantillonnage « (A) : fe=64kHz et (B) : fe=16kHz » sur la détection de l'angle d'incidence de la source

4.5.5 Evolution du SNR en fonction des bandes de fréquence et des angles incidents

On a choisi pour ce test, des signaux de fréquence d'échantillonnage fe = 16 kHz et la distance inter microphone d= 20cm. L'algorithme utilise un beamformer fixe avec une matrice de blocage adaptative. Dans le bloc « ANC », on fixe le pas d'adaptation ρ = 0.03 , avec un facteur de compromis μ = 300.

Le tableau 4.6 contient les mesures SNR pour chaque bande fréquentielle en fonction de l'angle d'incidence de la source (0-180°).

Bande de fréquence en Hz	Angle en degré										
	0°	20°	30°	45°	60°	80°	90°	120°	140°	160°	180°
0-69	4.3	2.1	9.8	7.9	1.5	11.1	2.6	2.1	8.1	20.4	4.4
34-146	7.5	4.0	11.5	10.2	3.3	13.9	4.0	3.9	10.4	22.8	5.6
69-231	8.5	1.5	11.9	11.7	1.6	15.8	1.2	2.1	9.8	22.4	4.0
146-324	8.3	2.4	11.5	12.3	1.3	17.9	2.4	3.0	10.2	22.0	3.1
231-426	6.0	2.6	12.0	11.8	1.6	18.1	2.6	3.1	8.9	19.8	3.8
324-539	3.3	3.1	13.6	11.0	1.9	19.8	2.5	4.1	11.4	19.1	5.7

426-663	3.7	5.4	17.4	11.7	5.1	22.8	3.7	5.0	11.8	23.2	5.4
539-799	5.3	7.0	13.5	17.5	8.5	25.3	8.3	7.5	13.5	30.1	5.9
663-949	16.3	6.2	6.7	25.3	9.0	25.5	10.0	8.3	9.1	31.8	11.6
799-1113	18.3	5.6	4.8	21.5	9.8	29.5	12.4	6.6	7.8	31.5	12.7
949-1295	19.6	6.8	3.6	19.4	9.8	30.4	12.1	9.1	5.9	37.7	14.4
1113-1494	21.7	6.1	3.3	22.6	8.9	31.1	11.3	9.9	7.1	37.1	16.2
1295-1713	23.0	5.1	7.1	18.5	9.4	30.9	11.0	4.6	8.5	33.1	18.2
1494-1954	23.1	11.0	11.3	26.8	10.9	34.2	14.2	7.7	13.7	33.5	15.2
1713-2219	25.8	13.0	16.2	22.5	15.5	34.8	16.6	12.0	14.5	28.4	18.6
1954-2511	22.4	14.3	13.0	28.3	16.0	32.7	8.8	9.8	16.2	35.2	18.4
2219-2832	20.8	13.6	18.7	25.9	12.5	32.0	6.6	9.8	15.0	36.5	14.0
2551-3185	21.8	15.5	16.4	29.0	10.1	36.8	6.7	7.8	15.4	38.8	19.9
2832-3573	28.0	19.6	18.5	28.4	16.3	31.9	13.1	13.2	22.4	26.9	23.8
3185-3999	28.5	15.7	24.0	24.9	15.3	34.8	5.4	8.7	14.4	35.6	22.9

Tableau 4. 7. SNR en fonction des bandes de fréquence et des angles incidents

La figure 4.11 représente la moyene des SNR de tous les angles par bande fréquentielle

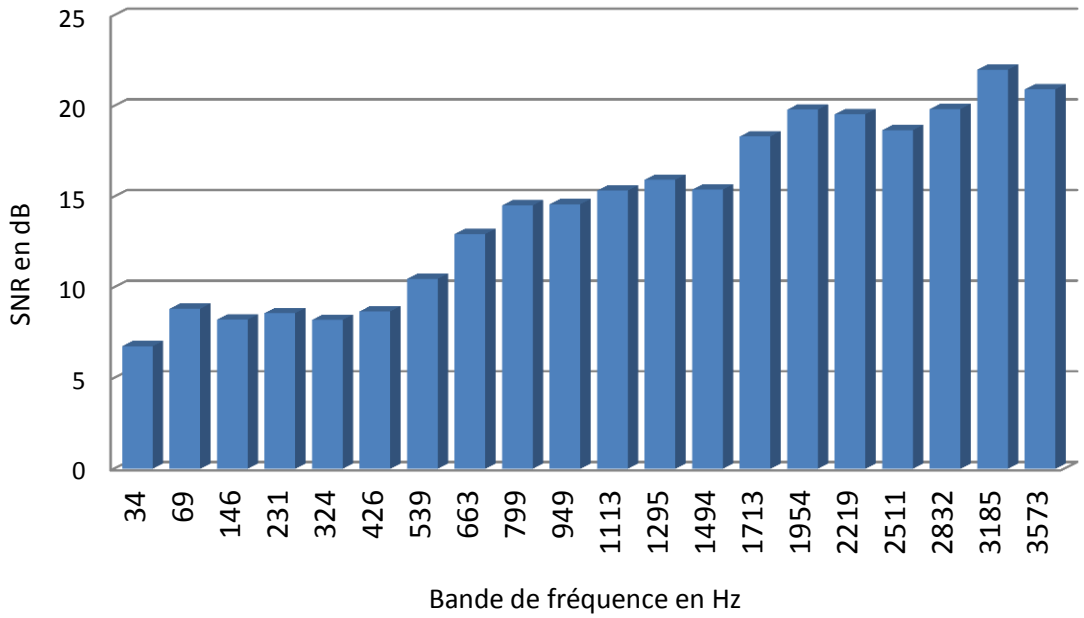


Figure 4. 11. Le SNR en fonction des bondes de fréquence

On remarque que pour toutes les bandes fréquentielle le SNR est superieur à 5 dB, il est compris entre 12 et 22dB dans la game 500-4000 Hz, qui correspond globalement à l'essentiel de l'energie de la parole 200-5000 Hz.

4.5.6 Influence de la matrice de blocage sur le réhaussement du signal de parole

On a fixé pour ce test l'angle d'incidence Θ=60°, le facteur de compromis μ =300 et on fait varier le pas d'adaptation ρ, la fréquence d'échantillonnage (fe=16 et 64kHz), ainsi que les distance inter-microphone (d=1,5 et 20cm).

Le tableau 4.8 représente les variations des SNR en fonction des distances, des fréquences d'échantillonnage, pour les différents algorithmes en faisant varier le pas d'adaptation.

Distance inter-microphone		Fréquence d'échantillonnage fe=64kHz						Fréquence d'échantillonnage fe=16kHz					
		MB adaptative			MB fix			MB adaptative			MB fix		
		ρ=0.3	ρ=0.03	ρ=0.003	ρ=0.3	ρ=0.03	ρ=0.003	ρ=0.3	ρ=0.03	ρ=0.003	ρ=0.3	ρ=0.03	ρ=0.003
d=1cm	SNR	**0.9**	**1.7**	**8.67**	**0.4**	**0.7**	**7.01**	**5.39**	**8.97**	**19.7**	**0.5**	**4.37**	**17.9**
	Θ	**59°**	**59°**	**59°**	**59°**	**59°**	**59°**	**46°**	**46°**	**46°**	**46°**	**46°**	**46°**
d=5cm	SNR	**0.3**	**0.2**	**1.2**	**0.03**	**0.2**	**1.12**	**0.68**	**3.12**	**9.29**	**0.62**	**3.02**	**9.08**
	Θ	**56°**	**56°**	**56°**	**56°**	**56°**	**56°**	**52°**	**52°**	**52°**	**52°**	**52°**	**52°**
d=20cm	SNR	**0.09**	**0.3**	**2.1**	**0.09**	**0.36**	**1.91**	**0.87**	**3.57**	**9.31**	**0.67**	**3.04**	**8.87**
	Θ	**51°**	**51°**	**51°**	**51°**	**51°**	**51°**	**56°**	**56°**	**56°**	**56°**	**56°**	**56°**

Tableau 4. 8. variation des SNR en fonction des distances, fréquence d'échantillonnage, pour les différents algorithmes en faisant varier le pas d'adaptation

On représente dans la figure 4.12 les résultats des tests utilisant les techniques de réhaussement à savoir l'algorithme utilisant la matrice de blocage adaptative et celui dont la matrice de blocage est fixe.

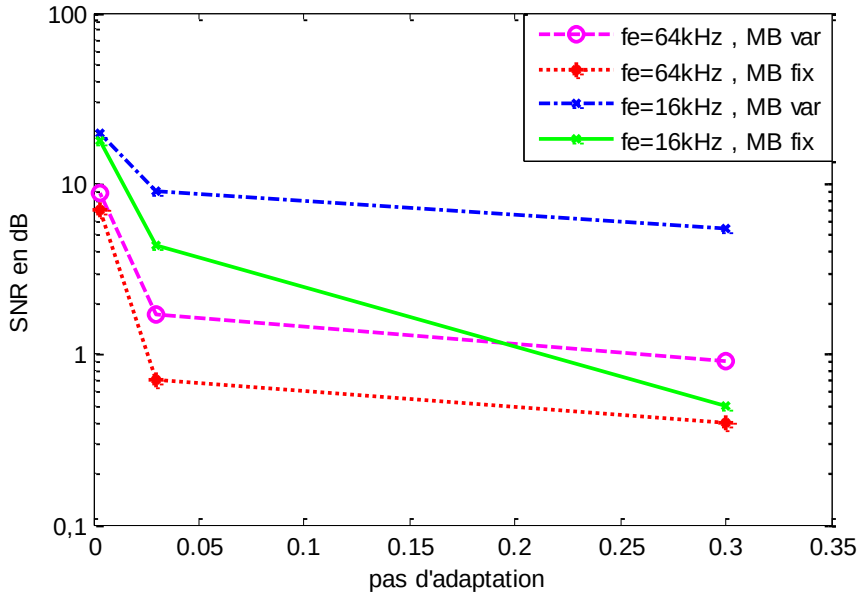


Figure 4. 12. SNR en fonction du pas d'adaptation pour les différentes situations

On obtient un bon SNR pour : une fréquence d'échantillonnage fe=16kHz , un pas d'adaptation $\rho < 0.05$, et avec une matrice de blocage adaptative.

4.5.7 Cas de deux sources dont l'une est assimilée à du bruit

La figure 4.15 représente le tracé énergitique de deux sources l'une à 45° considérée comme parole, l'autre à 140° assimilée à du bruit.

On remarque sur la courbe deux pics dont celui de plus grande énergie correspond à la source de parole puisque sa puissance était plus grande.

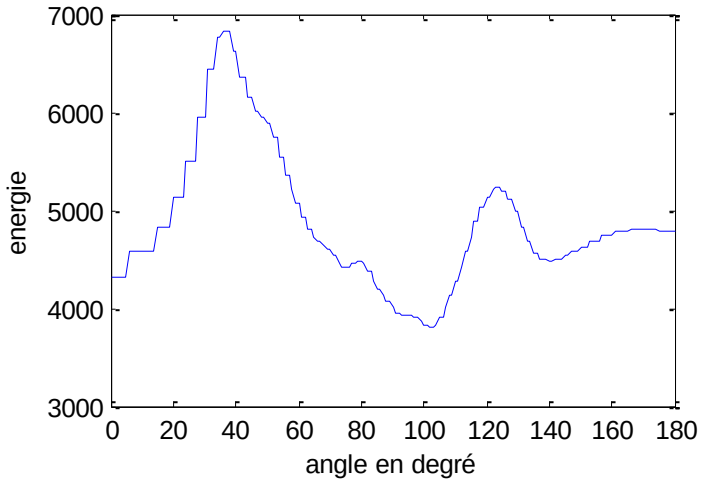


Figure 4. 13. Détection de deux sources

La figure 4.16 illustre un exemple : (A) : signal enregistré, (B) : signal de sortie de « beamformer », (C) : signal débruité.

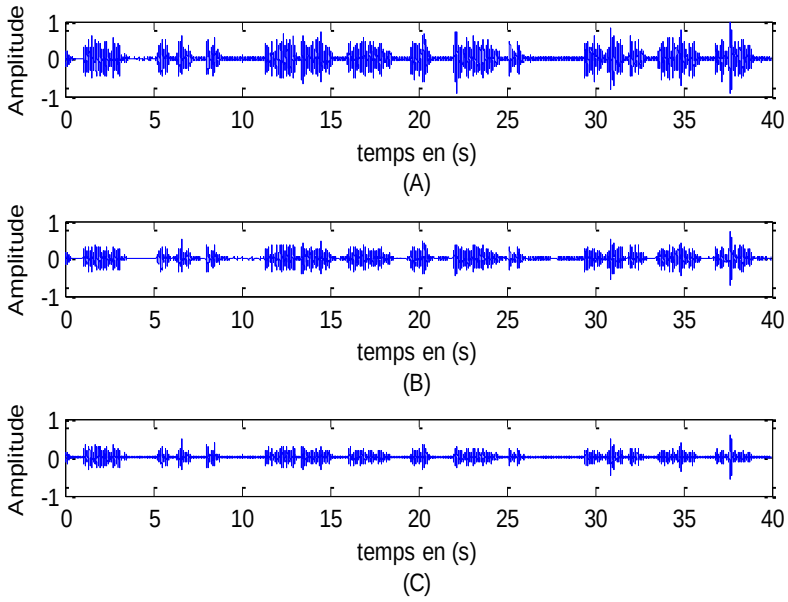


Figure 4. 14. (A) : Signal enregistré, (B) : Signal de sortie de « beamformer », (C) : Signal débruité

4.6 Conclusion

Dans ce chapitre nous avons établi, dans une première partie, quelques éclaircissements sur un problème très souvent rencontré dans le traitement du signal vocal qui est la détection d'activité vocale VAD. Autrement dit, la discrimination entre les régions où la parole est présente et les régions où la parole est absente dans le signal analysé. Ainsi, nous avons développé et implémenté quatre algorithmes de détection d'activité vocale, dont la comparaison a été illustrée dans les chapitres précédents. Nous en avons choisi un seul, dont l'adaptation pour notre application était naturelle. Nous avons développé, par la suite, un système de rehaussement de la parole utilisant les deux techniques GSC et SDW-MWF intégrées dans un seul système. Ce dernier a été proposé dans [14] et nommé « Prétraitement Spatial – Distorsion Pondéré de la Parole – Filtre Multicanal de Wiener » ou en anglais « the Spatially Preprocessed - Speech Distortion Weighted- Multi-channel Wiener Filter » (SP-SDW-MWF). Il est constitué de deux parties, l'une est fixe (Prétraitement Spatial) et l'autre adaptative (SDW-MWF).

L'adaptation de l'algorithme SDW-MWF nécessite des signaux d'erreur qui sont sans effets de convolution circulaire, par conséquent, dans le domaine temporel les signaux doivent être coupés de telle sorte que les *L* premier échantillons du bloc sont rejetés et les *L* dernier restent inchangés. Notre algorithme se base sur les remarques suivantes : Au cours des périodes de parole, les références se composent de la parole et du bruit. Alors que, pendant les périodes de bruit, il n'y a que la composante de bruit qui est observée (en supposant que les statistiques du deuxième ordre du bruit sont suffisamment stationnaires, de sorte qu'elles puissent être estimées pendant les périodes de bruit seulement).

Les résultats obtenus sont encourageant. En prenant un réseau de microphones de 4 capteurs, et en adoptant une configuration linéaire nous avons pu comparer les différentes architectures en calculant pour chacune d'entre elles les SNR et SNR segmental, et SNR par bande. Ainsi, pour des fréquences d'échantillonnage faibles, le débruitage est bon et la vitesse d'exécution est faible (bonne vitesse d'exécution). En contrepartie on obtient une mauvaise précision de détection de la source donc les distorsions augmentent.

La technique de filtrage par bloc dans le domaine fréquentiel «overlap-save », nécessite moins d'opérations par itérations, c'est à dire elle est plus rapide que la mise en œuvre dans le domaine temporel. Donc elle est appropriée pour des exécutions en temps réel.

Plus le pas d'adaptation ρ augmente, plus la vitesse de convergence augmente mais on a aussi une augmentation des distorsions.

Plus ρ diminue plus le SNR augmente.

La matrice de blocage adaptative introduit un bon compromis entre intelligibilité et débruitage mais rallonge le temps d'exécution.

Conclusion générale

Dans le cadre de cette étude, nous avons établi, dans une première partie, quelques éclaircissements sur un problème très souvent rencontré dans le traitement du signal vocal qui est la détection d'activité vocale VAD. Autrement dit, la discrimination entre les régions où la parole est présente et les régions où la parole est absente dans le signal analysé. Ainsi, nous avons développé et implémenté quatre algorithmes de détection d'activité vocale, dont la comparaison a été illustrée dans les chapitres précédents. Nous en avons choisi un seul, dont l'adaptation pour notre application était naturelle. Nous avons développé, par la suite, un système de rehaussement de la parole utilisant les deux techniques GSC et SDW-MWF intégrées dans un seul système. Ce dernier a été proposé dans [14] et nommé « Prétraitement Spatial – Distorsion Pondéré de la Parole – Filtre Multicanal de Wiener » ou en anglais « the Spatially Preprocessed - Speech Distortion Weighted- Multi-channel Wiener Filter » (SP-SDW-MWF). Il est constitué de deux parties, l'une est fixe (Prétraitement Spatial) et l'autre adaptative (SDW-MWF). C'est dans ce contexte que notre contribution a porté sur l'introduction d'une matrice de blocage adaptative pour une meilleure estimation des interférences dans le système proposé. Les applications développées sont proposées en approche temporelle et fréquentielles. Pour préparer une implantation de nos algorithmes sur un processeur DSP, dans les prochains travaux, nous avons opté pour une approche de programmation par bloc. Et partant de la remarque que, dans la plupart des applications de filtrage, le problème qui se pose est d'implanter la convolution linéaire de deux séquences plutôt que la convolution circulaire. Le traitement au moyen de la TFD d'un signal décomposé en blocs disjoints est entaché d'un défaut systématique lié aux discontinuités induites par le découpage. Les études faites pour pallier ce défaut ont abouti à deux types, devenus classiques, de mise en œuvre des méthodes rapides de filtrage qui sont : la méthode d' « overlap-save » et la méthode d' « overlap-add ». Or, l'intérêt de ce calcul indirect de convolution est la

possibilité de calculer la TFD et la TFDI au moyen de la TFR, permettant ainsi une large réduction du coût de calcul de la convolution circulaire, comparé à l'algorithme de calcul direct de cette convolution. Notre algorithme se base sur les remarques suivantes : Au cours des périodes de parole, les références se composent de la parole et du bruit. Alors que, pendant les périodes de bruit, il n'y a que la composante de bruit qui est observée (en supposant que les statistiques du deuxième ordre du bruit sont suffisamment stationnaires, de sorte qu'ils puissent être estimés pendant les périodes de bruit seulement). L'adaptation de l'algorithme MBA nécessite des signaux d'erreur qui sont sans effets de convolution circulaire, par conséquent, dans le domaine temporel les signaux doivent être coupés de telle sorte que les *L* premier échantillons du bloc sont rejetés et les *L* dernier restent inchangés.

Les résultats obtenus sont encourageant, en prenant un réseau de microphones de 4 capteurs, et en adoptant une configuration linéaire nous avons pu comparer les différentes architectures en calculant pour chacune d'entre elles les SNR et SNR segmental, et SNR par bande. Ainsi :

- Pour des fréquences d'échantillonnage faible, le débruitage est bon et la vitesse d'exécution est faible (bonne vitesse d'exécution). En contrepartie, on obtient une mauvaise précision de détection de la source donc les distorsions augmentent.
- La technique de filtrage par bloc dans le domaine fréquentiel «overlap-save », nécessite moins d'opérations par itérations, c'est à dire elle est plus rapide que la mise en œuvre dans le domaine temporel. Donc elle est appropriée pour des exécutions en temps réel.
- Plus le pas d'adaptation ρ augmente, plus la vitesse de convergence augmente, mais on a aussi une augmentation des distorsions.
- Plus ρ diminue plus le SNR augmente.
- La matrice de blocage adaptative introduit un bon compromis entre intelligibilité et débruitage mais rallonge le temps d'exécution.

En perspectives, pour les prochains travaux, nous pensons, qu'un développement des algorithmes proposés dans ce travail, devraient être testés sur DSP pour estimer leur comportement en temps réel. De plus des microphones de meilleure qualité devraient être utilisés.

Carte d'acquisition

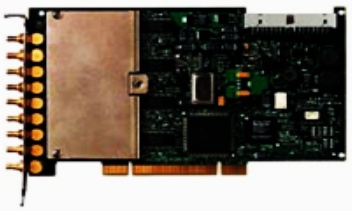
NATIONAL INSTRUMENTS

Représentant technique
Algérie
961 1 33 28 28
ni.arabia@ni.com

NI PCI-4472

Carte d'acquisition de signaux dynamiques 8 voies

- 8 voies d'entrées analogiques à échantillonnage simultané
- Conditionnement IEPE configurable par logiciel
- Vitesse d'échantillonnage jusqu'à 102,4 Kéch./s
- Bande passante 45 kHz sans repliement
- 24 bits de résolution, gamme dynamique de 110 dB
- Gamme de tension de ±10 V



Panorama

La carte PCI-4472, de National Instruments, est une carte d'acquisition de signaux dynamiques 8 voies pour les mesures de domaines de fréquence haute précision. Les voies d'entrées intègrent le conditionnement de signaux IEPE pour les accéléromètres et les microphones. Les 8 voies d'entrées de la carte numérisent simultanément les signaux sur une bande passante de tension DC jusqu'à 45 kHz. Vous pouvez synchroniser les cartes PCI-4472 pour les applications à grand nombre de voies ou avec d'autres cartes en utilisant le bus RTSI. Lorsque vous les utilisez avec le LabVIEW Sound and Vibration Toolkit ou d'autres logiciels d'analyse, la carte PCI-4472 peut acquérir une variété de mesures de domaines de temps ou de fréquence pour votre application.

Caractéristiques

Documents de spécifications

- Spécifications détaillées
- Fiche technique (angl.)

Résumé des spécifications

Général	
Format	PCI
Système d'exploitation / cible	Windows, Temps réel, Linux
Types de mesure	Accéléromètre IEPE, Tension
Support de LabVIEW RT	Oui
Conformité RoHS	Oui
Entrée analogique	
Nombre de voies	8 MC
Fréquence d'échantillonnage	102.4 kéch./s/voie
Résolution	24 bits
Échantillonnage simultané	Oui
Gamme de tension maximum	-10..10 V
Sensibilité de gamme	1.19 µV
Gamme de tension minimum	-10..10 V
Sensibilité de gamme	1.19 µV
Nombre de gammes	1
Mémoire embarquée	1023 échantillons
Conditionnement de signaux	Excitation de courant, Filtre anti-repliement
Sortie analogique	
Nombre de voies	0
E/S numériques	

Microphones

Panasonic Microphone Cartridges

Omnidirectional Electret Condenser Microphone Cartridge

Series: **WM-52B/54B** (pin type)

■ **Features**

- High-cost performance, electret condenser microphone cartridge
- WM-52B series are low-voltage, 1.5V operation microphones and suitable for all telephone applications, intercoms, and computers.
- WM-54B series are low-voltage 2.5V operation microphones and suitable for tape recorders, telephone devices, and toys

■ **Sensitivity**

WM-52B

Vs = 1.5V RL = 3.0kΩ	
T	−44±3dB
Y	−42±3dB
U	−40±3dB
M	−44±2dB
H	−42±2dB
P	−40±2dB

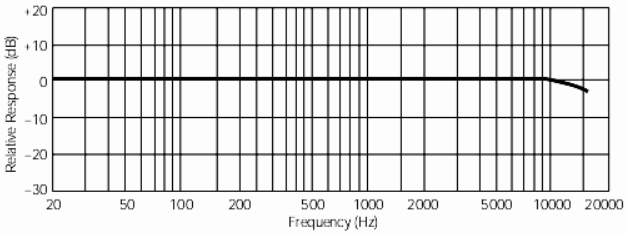
WM-54B

Vs = 2.5V RL = 2.2kΩ	
X	−46±3dB
U	−44±3dB
Y	−42±3dB
U	−40±3dB
M	−44±2dB
H	−42±2dB

■ **Specifications**

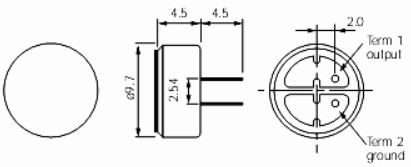
Sensitivity	See next page (0dB = 1V/pA, 1kHz)
Impedance	Less than 3.0k Ω [2.2k Ω]
Directivity	Omnidirectional
Frequency	20–16,000 Hz
Max. operation voltage	10V
Standard operation voltage	1.5V [2.5V]
Current consumption	Max. 0.3mA [0.6mA]
Sensitivity reduction	Within −3 dB at 1.1V [2V]
S/N ratio	More than 60dB

■ **Typical Frequency Response Curve**

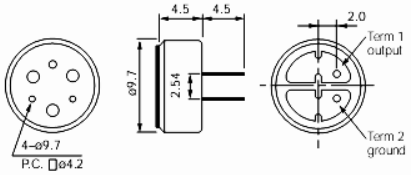


■ **Dimensions in mm** (not to scale)

WM-52B



WM-54B



Design and Specifications are subject to change without notice. Ask factory for technical specifications before purchase and/or use.
Whenever a doubt about safety issues arises from this product, please inform us immediately for technical consultation.

Bibliographie

[1] R. W. Stadler and W. M. Rabinowitz: 'On the potential of fixed arrays for hearing aids', J. Acoust. Soc. Amer., vol. 94, no. 3, pp. 1332–1342, September 1993.

[2] P. M. Peterson: ' Adaptive array processing for multiple microphone hearingaids', PhD. thesis, Dept. Elect. Eng. And Comp. Sci., M.I.T, Cambridge, MA, 1989, available as Res. Lab. Elect. Techn. Rept. 541.

[3] M. IRKI et I. ELGHRIBI: 'Exploitation des statistiques des signaux pour la localisation sonore utilisant la méthode de formation de voies', Thèse d'ingénieur d'état, Option communication, Université SAAD Dahleb ,Blida, 2011.

[4] Griffiths L.J. and Jim C.W.: 'An alternative approachto linearly constrained adaptive beamforming', IEEETrans. Antennas Propag., Vol. 30, No. 1, pp. 27–34,1982.

[5] J. E. Greenberg and P. M. Zurek:'Evaluation of an Adaptive Beamforming Method for Hearing Aids', J. Acoust. Soc. Amer., vol. 91, no. 3, pp. 1662–1676, March 1992.

[6] O. Hoshuyama, A. Sugiyama and A. Hirano:'A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters', IEEE Trans. Signal Processing, vol. 47, pp. 2677–2683, 1999.

[7] D. Van Compernolle:'Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings', in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Albuquerque, vol. 2, pp. 833–836, April. 1990.

[8] W. Herbordt and W. Kellermann:'Frequency-domain integration of acoustic echo cancellation and a generalized sidelobe canceller with improved robustness', European Transactions on Telecommunications, vol. 13, no. 2, pp. 123–132, Mar.-Apr. 2002.

[9] A. Spriet, M. Moonen, and J. Wouters:'Robustness analysis of GSVD based optimal Filtering and generalized Sidelobe Canceller for Hearing Aid Applications', inProc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 31–34, NY, USA, October 2001.

[10] A. Spriet, M. Moonen, and J.Wouters:'Robustness Analysis of Multi-channel Wiener Filtering and Generalized Sidelobe Cancellation for Multi-microphone Noise Reduction in Hearing Aid Applications', To appear in IEEE Trans. Speech, Audio Processing, May 2005. Available at ftp://ftp.esat.kuleuven.ac.be/pub/sista/spriet/reports/02-81.pdf.

[11]S. Doclo and M. Moonen, GSVD-Based Optimal Filtering for Multi-Microphone Speech Enhancement, chapter 6 in 'Microphone Arrays: Signal Processing Techniques and Applications' (Brandstein, M. S. and Ward, D. B., Eds.), pp. 111–132, Springer-Verlag, May 2001.

[12]A. Spriet, M. Moonen, and J. Wouters:'A multichannel subband GSVD based approach for speech enhancement in hearing aids', inProc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC), Darmstadt, pp. 187–191, Germany, September 2001.

[13] S. Doclo and M. Moonen:'GSVD-Based Optimal Filtering for Single and Multimicrophone Speech Enhancement', IEEE Trans. Signal Processing, vol. 50, no. 9, pp. 2230–2244, September 2002.

[14] A. Spriet, M. Moonen and J. Wouters: ' stochastic gradient implementation of spatially pre-processed multi-channel wiener filtering for noise reduction in hearing aids ', ICASSP, 2004.

[15] M. Berouti, R. Schwartz and J. Makhoul:' Enhancement of Speech Corrupted by Acoustic Noise', ICASSP, pp. 208- 211, 2-4,April 1979.

[16] S.F. Boll:'Suppression of Acoustic Noise in Speech Using Spectral Subtraction', IEEE Trans. on ASSP, vol. ASSP-27, n°2, pp. 113-120, April 1979.

[17] Y. Ephraim and D. Malah:'Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator', IEEE Trans. on ASSP, vol. ASSP-32, n°6, pp. 1109-1121, December 1984.

[18] K.Buckley:'Broadband beamforming and the generalizedsidelobe canceller', IEEE Trans. Acoust. SpeechSignal Process., Vol. 34, No. 5, pp. 1322–1323,1986.

[19] O. Frost:'An algorithm for linearly constrained adaptive array processing', Proc. IEEE, Vol. 60, No. 8, pp. 926–935,1972.

[20] B. Van Veen and K. Buckley: 'Beamforming: A versatile approach to spatial filtering ', IEEE ASSP Mag., Vol. 5,No. 4, pp. 4–24, 1988.

[21] O. Hoshuyama and A. Sugiyama:'A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters', in Proc. Int. Conf. Acoustics, Speech, Signal Process., pp. 925–928, Atlanta, GA, May, 1996.

[22] S. Nordholm, I. Claesson and B. Bengtsson:'Adaptive array noise suppression ofhands free speaker input in cars',IEEE Trans. Veh. Technol., vol. 42, pp. 514–518, November 1993.

[23] J. Meyer and C. Sydow:'Noise cancelling for microphone array',in Proc. Int. Conf. Acoust., Speech, Signal Process., pp. 211–214,Munich, Germany,April 1997.

[24] O. Hoshuyama, A. Sugiyama and A. Hirano:' A robustadaptivebeamformer with a blocking matrix using coefficientconstrained adaptive filters ', IEICE Trans. Fundament., Vol. E82-A, No. 4, pp. 640–647, 1999.

[25] S. Gannot, D. Burshteinand E. Weinstein:'Signal enchancement using beamforming and non stationarity with applications to speech» IEEE Trans. Signal Process., Vol. 49, No. 8, pp. 1614–1626, 2001.

[26] L. Li, B. Jeffs, A. Poulsen and K. Warnick: ' Analysis ofadaptive array algorithm performance for satellite interference cancellation in radio astronomy', Proc. XXVIIURSI General Assembly, Maastricht, The Netherlands, 2002.

[27] A. Leshem, A. van der Veen and A. Boonstra:' Multichannel interference mitigation techniques in radio astronomy', Astrophys. J. Suppl., Vol. 131, pp. 355–374.2000.

[28] A. Boukalov and S. Haggman: ' System aspects of smartantennatechnology in cellular wireless communications An overview ', IEEE Trans. Micr. Theory Techn., Vol. 48, No. 6, pp. 919–929,2000.

[29] L. Godara:'Applications of antenna arrays to mobilecommunications, Part II: Beamforming and direction-ofarrival considerations', Proc. IEEE, Vol. 85, No. 8, pp. 1195–1245,1997.

[30] M. W. Hoffman and K. M. Buckley:'Robust time-domain processing of broadband acoustic data', IEEE Trans. Speech, Audio Processing, vol. 3, pp. 193–203, May 1995.

[31] H. Cox, R. M. Zeskind and M. M. Owen:'Robust Adaptive Beamforming', IEEE Trans. Acoust., Speech, and Signal Processing, vol. 35, no. 10, pp. 1365–1376, 1987.

[32] N. K. Jablon:'Adaptive beamforming with the Generalized Sidelobe Canceller in the presence of array imperfections', IEEE Trans. Antennas Propag., vol. 34, pp. 996–1012, Aug. 1986.

[33] Z. Tian, K.L. Bell and H.L. Van Trees:'A Recursive Least Squares Implementation for LCMP Beamforming Under Quadratic Constraint', IEEE Trans. Signal Processing, vol. 49, no. 6, pp. 1138–1145, June 2001.

[34] A. Spriet, M. Moonenand J. Wouters:'A multi-channel subband generalized singular value decomposition approachto peech enhancement',European Transactions on Telecommunications, vol. 13, no. 2, pp. 149–158, Mar.-Apr. 2002.

[35] D. A. Florêncio and H. S. Malvar:'Multichannel filtering for optimum noise reduction in microphone arrays', inProc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Salt Lake City, Utah, May 2001.

[36] A. Spriet, M. Moonen, and J. Wouters:'Stochastic gradient based implementation of spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction in hearing aids', Tech. Rep. ESAT-SISTA/TR 03-47a, ESAT/SISTA, K.U. Leuven

(Belgium), 2003, available at ftp://ftp.esat.kuleuven.ac.be/pub/sista/spriet/reports/03-47a.pdf.

[37] S. Doclo and M. Moonen:'Multi-microphone noise reduction using recursive GSVD-based optimal filter-ing with ANC postprocessing stage',To appear in IEEE Trans. Speech, Audio Processing. Available at: ftp://ftp.esat.kuleuven.ac.be/pub/sista/doclo/reports/02-04.ps.gz.

[38] G. Rombouts and M. Moonen: 'QRD-based unconstrained optimal filtering for acoustic noise reduction', Signal Processing, vol. 83, no. 9, pp. 1889–1904, September 2003.

[39] A. Spriet, M. Moonen and J. Wouters:'A multi-channel subband generalized singular value decomposition approachto peech enhancement', European Transactions on Telecommunications, vol. 13, no. 2, pp. 149–158, Mar.-Apr. 2002.

[40] A. Spriet, M. Moonen and J. Wouters:'Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction', in Signal Processing, vol. 84, no. 12, pp. 2367–2387, December 2004.

[41] L. R. Rabiner, R. W. Schafer: 'Digital Processing of Speech Signals', (1e éd.), New Jersey : Prentice-Hall,1978.

[42] J. R. Deller, J. H. L. Hansen andJ.G. Proakis:'Discrete Time Processing of Speech Signals', (3e éd.), New Jersey, IEEE Press, 1999.

[43] D.G. Childers:' Speech processing and synthesis toolboxes', (1e éd.),John Wiley & Sons, Inc, New York,2000.

[44] J. G. Proakis andD. G. Manolakis:'Digital Signal Processing: Principles, Algorithms, and Applications', (3e éd.), Upper Saddle River,Prentice-Hall,New Jersey, 1996.

[45] F. Alexa: 'Introducere in tehnicasunetului', Timisoara, Editura de vest, 1999.

[46] Calliope : ' La parole et son traitement automatique',Masson, Paris,1989.

[47] C. S. Gargour : 'Traitement numérique de signaux', (1e éd), École de technologie supérieure, 2001.

[48] R. Boite and M. Kunt: 'Traitement de la Parole', (1e éd.), Lausanne, Presses Polytechniques Romandes, 1987.

[49] A. Benyassine, E. Sholomot, H. Y. Su, D. Massaloux and al.:'ITU-T Recommendation G.729 Annex B: A silence compression schema for use with G.729 optimized for V.70 digital simultaneous voice and data application', IEEE Communication Magazine, 35(9), 64-73, 1997.

[50] L. R. Rabiner andM. R. Sambur: 'An algorithm for determining the endpoints of isolated utterances'. Bell Syst. Tech. J, 54(2), 297-315, 1975.

[51] Q. Li, J. Zheng, A. Tsai and Q. Zhou: 'Robust endpoint detection end energy normalization for real time speech and speaker recognition', IEEE Tran. on Speech and Audio Processing, 10(3), pp. 146-157, 2002.

[52] J. Canny: 'A computational approach to edge detection', IEEE Trans. Pattern Anal. Machine Intell., PAMI-8, pp. 679-698, November 1986.

[53] M. Petrou, J. Kittler:'Optimal edge detectors for ramp edges', IEEE Trans. Pattern Anal. Machine Intell., 13(5),pp. 483-491, 1991.

[54] C. T. Lin,J. Y. Lin and G. D. Wu: ' A robust word boundary detection algorithm for variable noise-level environment in cars', IEEE Tran. on Intelligent Transportation Systems, 3, pp. 89 - 101, 2002.

[55] L. F. Lamel, L.R. Rabiner, A.E. Rossenberg andJ. G. Wilpon: 'An improved end point detector for isolated Word Recognition',IEEE Transaction on Acoustics, Speech and Signal Processing, ASSP-29(4), pp. 777-785, 1981.

[56] J.C.Junqua, B. Mak andB. Reaves: 'A robust algorithm for word boundary detection in presence of noise', IEEE Tran. on Speech and Audio Processing, 2(3), pp. 406-412, 1994.

[57] A. Cavallaro, L. Beritelli andS. Casale: 'A robust robust voice activity detector for wireless communication using soft computing', IEEE J Select. Areas. Commun., 16(12), pp. 1818-1829, 1998.

[58] L. Beritelli, S. Casale, G. Ruggeri and S. Serrano :'Performance Evaluation and Comparison of G.729/AMR/Fuzzy Voice Activity Detectors', IEEE Signal Processing Letters, 9(3), pp. 85-88, 2002.

[59] C. Lenz: ' Localization of sound sources ', Autonomous systems lab, ETH Zùrich, Spring,2009.

[60] A. V. Oppenheim and R. W. Schafer:' Discrete- Time Signal Processing', Engle-wood Cliffs, NJ: Prentice-Hall, 1989.

[61] A. Spriet, M. Moonen, and J. Wouters:'Stochastic gradient based implementation of spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction in hearing aids', Tech. Rep. ESAT-SISTA/TR 03-47, K.U. Leuven (Belgium), 2003, availableat ftp://ftp.esat.kuleuven.ac.be/pub/sista/spriet/reports/03-47.pdf

[62] O. Hoshuyama and A. Sugiyama:'A robust adaptive beamformer for micro-phone arrays with a blocking matrix using constrained adaptive filters', IEEE Trans. on Signal Processing, 47(10), October 1999.

[63] S. Wang, A. Sekey andA.Gersho:' Modified bark spectral distortion measure which uses noise masking threshold',IEEE Speech Coding Workshop, vol. SAC-10, 1997.

[64] Y. Hu and P. Loizou:'Evaluation of objective quality Measures for speech enhancement, Evaluation of objective Measures for speech enhancement', vol. 16, pp. 229–238, January 2008.

[65] N. Ma, M. Bouchard and R. A. Goubran: 'Perceptual Kalman filtering for speech enhancement in colored noise', In Proc. ICASSP'04, volume 4, pp. 1045–1048, Montreal, Canada 2004.

[66] Y. Hu and P. Loizou:'Evaluation of objective quality Measures for speech enhancement, Evaluation of objective Measures for speech enhancement', vol. 16, pp. 229–238, January 2008.

[67] UIT-T P.861. Objective quality measurement of telephone band (300-3400 Hz) speech codecs. 1998.

[68] UIT-T P862. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. 2000.

[69] S. Wang, A. Sekey and A.Gersho:'Modified bark spectral distortion measure which uses noise masking threshold', IEEE Speech Coding Workshop, vol. SAC-10, 1997.

[70] W. Yang, M. Dixon, and R. Yantorno:'A modified bark spectral distortion measure which uses noise masking threshold', IEEE Speech Coding Workshop, pp. 55-56, Pocono Manor, 1997.

[71] P. Loizou:'Speech enhancement: Theory and practice',CRC, 1 edition, 2007.

[72] W. Yang, M. Dixon and R.Yantorno:'Enhanced modified Bark spectral distorsion (EMBSD) :An objective speech quality measure based on audible distorsion and cognition model', PhD thesis, Temple University Graduate Board, May 1999.