N° d'ordre : 45/2012-M/ELN

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediène

Faculté d'Electronique et d'Informatique

MEMOIRE

Présenté pour l'obtention du Diplôme de Magister
en : ELECTRONIQUE

Spécialité : Radiofréquences et Micro-ondes

par

FARSI Narimane

THÈME

AMELIORATION DE LA DETECTION DE L'ACTIVITE VOCALE (VAD) DANS LES SYSTEMES DE COMMUNICATION UTILISANT LE CODEC G729

Soutenu publiquement, le 11 juin 2012, devant le Jury composé de :

M.	M. DEBYECHE	Maitre de Conférences/A à l'USTHB	Président
M.	A. AMROUCHE	Maitre de Conférences/A à l'USTHB	Directeur de mémoire
M.	B. BOUDRAA	Professeur à l'USTHB	Examinateur
M.	S. MEKAOUI	Maitre de Conférences/A à l'USTHB	Examinateur

Dédicaces

Je dédie ce mémoire

A mes très chers parents qui m'ont toujours soutenu et encouragé durant toutes mes années d'études.

A ma très chère tante qui nous a quittés trop tôt, mais sera à jamais présente dans mon cœur.

A mes amis qui ont toujours su trouver les mots pour me rassurer et m'encourager durant ce mémoire.

Remerciements

Ce travail de recherche a été effectué au sein de l'équipe reconnaissance vocale du laboratoire de communication parlée et de traitement du signal (LCPTS) dont je remercie vivement le directeur Monsieur *M. Debyeche,* Maitre de conférences à l'USTHB, pour son accueil et les moyens mis en œuvre, mais aussi l'honneur qu'il me fait en assumant la fonction de président de jury.

J'adresse mes plus vifs remercîments à mon directeur de mémoire Monsieur *A .Amrouche,* Maitre de conférences à l'USTHB pour m'avoir proposé ce sujet de recherche. Ses conseils avisés, sa disponibilité remarquable, ses encouragements incessants m'ont permis de mener à bien ce travail.

Je tiens également à remercier Monsieur *B. BOUDRAA* Professeur à l'USTHB ainsi que Monsieur *S. Mekaoui* maitre de conférences à la faculté d'Electronique et Informatique, pour avoir bien voulu examiner ce mémoire.

J'adresse un grand remerciement à tous ceux qui ont contribué de près ou de loin à la mise au point de ce travail.

Sommaire

Sommaire

Chapitre 2. Le codec G729

Chapitre 3. La Détection de l'Activité Vocale (VAD) de la norme G729

Chapitre 4. Evaluation et amélioration du module VAD de la norme G729 dans un milieu bruité

Liste des figures

Liste des figures

Chapitre 1. Caractéristiques du signal de parole et techniques de codage

Chapitre 2. Le Codeur G729

Chapitre 3. La détection d'Activité Vocale (VAD) de la norme G729B

Chapitre 4. La détection d'Activité Vocale (VAD) de la norme G729B

Annexe A. Schémas de principe du codec CS-ACELP G.729

Liste des tableaux

Liste des tableaux

Chapitre 1. Caractéristiques du signal de parole et techniques de codage

Chapitre 2. Le codeur G729

Chapitre 3. La Détection de l'Activité Vocale de la norme G729

Chapitre 4. Evaluation et amélioration du VAD de la norme G729 dans un milieu bruité

Liste des acronymes

Liste des acronymes

ACELP	Algebraic Code Excited Linear Predictive
ADPCM	Adaptive Differential Pulse Code Modulation
AR	Auto Regressive
ARMA	Auto Regressive à Moyenne Ajustée
BER	Bit Error Ratio
CELP	Code Excited Linear Prediction
CNG	Comfort Noise Generation
CS−ACELP	Conjugate Structure - Algebraic Code Excited Linear Predictive
DTX	Discontinuous Transmission
ETSI	European Telecommunications Standards Institute
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
IP	Internet Protocol
LP	Linear Prediction
LPC	Linear Predictive Coding
LSF	Line Spectrum Frequency
LSP	Line Spectrum Pair
MA	Moving Average
MIC	Modulation par Impulsion et Codage
MOS	Mean Opinion Score
PABX	Private Automatic Branch eXchange
PCM	Pulse Code Modulation
PESQ	Perceptual Evaluation of Speech Quality
QOS	Quality Of Service
QV	Vector Quantization
SID	Silence Insertion Descriptor
SNR	Signal to Noise Ratio
UIT	Union Internationale des Télécommunications
VAD	Voice Activity Detection

VoIP	Voice over IP
WAN	Wide Area Network

Introduction Générale

Introduction Générale

Avec le déploiement accéléré et à large échelle des systèmes de communications, se pose la problématique de l'utilisation appropriée de la capacité (exprimée en bps), pour une rentabilité accrue des infrastructures et équipements. Dans les réseaux de télécommunications mobiles (GSM, GPRS, Réseaux 3G, etc.), mais plus encore dans les applications multimédia en VoIP (Voice over IP), une économie dans la consommation de la bande passante (l'utilisation du terme capacité serait plus appropriée) devient une nécessité absolue, sous peine de provoquer des congestions susceptibles d'affecter sérieusement la qualité de service (QoS).

Pour les systèmes conversationnels, le flot de parole est souvent entrecoupé de nombreuses pauses, ainsi le taux moyen de périodes de silence est de l'ordre de 60 °/° du temps d'utilisation du canal. On peut donc judicieusement réduire la charge du réseau en limitant le débit de transmission sur le canal dès lors que la parole est absente dans la conversation. S'appuyant sur les caractéristiques du signal de parole, la détection d'activité vocale (VAD ; Voice Activity Detection) permet de faire la distinction entre les segments de parole et les moments de silence dans une conversation verbale. La VAD, présente dans les codecs vocaux (en particulier G711, G729, etc.), joue un rôle crucial dans les systèmes de communication vocale. En effet, la localisation des segments de non activité vocale (zone de silence) permet une réduction significative du débit pendant ces moments, donc une optimisation de la consommation de la bande passante. Ce qui donne comme avantage la possibilité de partage des canaux avec d'autres informations, en plus d'une faible consommation d'énergie dans les équipements portables permettant une autonomie accrue, etc.

Cependant, la fiabilité d'un module de détection d'activité vocale peut être altérée par les conditions adverses, comme par exemple un environnement acoustique bruité lors de l'émission. En effet, si certaines trames de parole sont détectées comme étant du bruit, l'intelligibilité sera sérieusement endommagée (coupure de la parole), tandis que si le bruit est détecté comme étant de la parole, alors les avantages de la détection de silence seront perdus.

Répondre à cette problématique est un véritable challenge, d'où notre motivation pour mener ce travail, afin d'améliorer la VAD dans les systèmes de communications utilisant le codec G729, dédié aux réseaux IP. Afin d'évaluer l'impact du bruit de fond sur la détection d'activité vocale, une évaluation des performances du dernier standard VAD (annexe B de la recommandation G729) de l'UIT-T, développé pour les communications multimédia, en particulier la VoIP, a été effectuée dans ce mémoire. Les performances ont été évaluées avec une variété de bruits tirés de la base de données bruités NOISEX-92 -NATO RSG 10, à différents rapport signal sur bruit (SNR). A l'issue de cette évaluation, des améliorations ont été apportées aux modules VAD au moyen d'algorithmes complémentaires que nous avons développés et implémentés sur la version de base du codec G729 annexe B.

Le chapitre 1 fait un rappel sur le codage de la parole. La modélisation du système phonatoire humain et les caractéristiques du signal de parole y sont d'abord présentées. Les principes des différentes techniques de compression du signal de parole y sont également rappelés.

Le chapitre 2 est focalisé sur les méthodes de codage normalisées par la recommandation G.729 de l'ITU. Des procédures relatives à la détermination des coefficients CS-ACELP y sont décrites. L'accent est ensuite porté sur la modélisation du signal résiduel d'excitation résultant du filtrage prédictif.

Dans le chapitre 3, nous abordons le problème de réduction du débit par compression des silences contenus dans le flux de parole. La généralisation de la VoIP restant confrontée à des problèmes de débit trop élevé, les discontinuités d'une conversation vocale sont considérées par les procédures de détection d'activité de parole (VAD) et de génération de bruit de confort (CNG : Comfort Noise Generation), ainsi que la transmission discontinue (DTX : Discontinuous Transmission), introduites par l'annexe B de la norme G.729.

Dans le chapitre 4, nous nous sommes intéressés à l'évaluation de la robustesse de l'algorithme VAD dans des conditions acoustiques adverses. Nous avons utilisé à cet effet des vecteurs de tests joints à la recommandation G729 annexe B et des enregistrements de la base de donnée ARADIGIT bruitée avec 5 types de bruit de la base NOISEX-92. En conséquence, des améliorations au module VAD du codec G729, seront proposées.

Enfin, une dernière partie conclut ce travail de mémoire et expose quelques perspectives à explorer susceptibles d'accroitre l'efficacité de la détection d'activité vocale, pour une utilisation optimale de la capacité des réseaux de communications.

Chapitre 1 :

Caractéristiques du signal de parole et techniques de codage

Chapitre 1

Caractéristiques du signal de parole et techniques de codage

Introduction

La détection de l'activité vocale (VAD) s'appuie sur une connaissance approfondie des caractéristiques, notamment acoustiques du signal de parole. Aussi, avant d'aborder l'aspect codage de parole, il est nécessaire de faire une étude sur le mode de production et les caractéristiques articulatoires et acoustiques des sons du langage.

Dans ce chapitre, l'appareil phonatoire humain est présenté et les notions de phonème, de formant, de son voisé ou non voisé et de pitch sont définies. La présentation de quelques exemples permettra de mieux appréhender les différentes techniques de codage présentées par la suite.

1.1. Production de la parole

En raison des caractéristiques du conduit vocal humain, le signal de parole est fortement redondant, ce qui explique sa robustesse. Ces redondances permettent aux algorithmes de codage de compresser le signal en enlevant l'information non pertinente contenue dans le signal. La connaissance du système vocal et des propriétés du signal de parole est essentielle pour concevoir des codeurs efficaces [1]. Les propriétés du système auditif humain peuvent également être exploitées pour améliorer la qualité perceptuelle du signal codé.

L'appareil phonatoire, dont une partie peut être vue à travers la coupe transversale de la tête en Figure 1.1 est l'organe responsable de la production de la parole. En effet, l'étude de ce système va nous permettre d'identifier les grandes classes de sons. Les principaux organes de ce système sont :

- Les poumons, qui constituent le système de production de l'air vibratoire et jouent ainsi le rôle de soufflerie, alimentent continuellement le système phonatoire pendant le processus de production de la parole. Ce mouvement de production de l'air phonatoire est favorisé par le diaphragme qui, par ses mouvements assure l'envoi de l'air vers la trachée artère située sur la partie sub-glottique de l'appareil phonatoire, en vue de son acheminement vers la partie glottique.
- Le larynx est une structure composée de cartilages, de ligaments et de muscles, placé sous l'os hyoïde, qui se trouve entre la trachée et le pharynx. Les cordes vocales constituent l'organe anatomique de la phonation. Il s'agit de replis des membranes muqueuses du larynx, dont la mise en vibration produit les sons vocaux différentiés, à l'origine des cris, du langage, des chants, etc.
- Le conduit vocal, forme la partie supra-glottique, qui est constitué de la cavité nasale, de la cavité buccale, de la cavité labiale, du pharynx, de la langue et de la mandibule.
    - La cavité nasale : le volume de cette cavité est fixe. La position du voile du palais (vélum) détermine si l'air expiré durant la phonation s'échappe uniquement par la bouche (vélum relevé), par la bouche et le nez (vélum partiellement baissé) ou uniquement par le nez avec le vélum totalement abaissé.
    - La cavité buccale : Cet espace est délimité par le palais. Sa forme et son volume peuvent être fortement modifiés par le mouvement de la langue, et des mâchoires.
    - La cavité labiale : L'existence de la cavité labiale dépend de la position des lèvres et de la position de la langue. En effet, lorsque les lèvres sont contre les dents (comme c'est le cas lors de la production de la voyelle */i/*), le volume de cette cavité est négligeable. A l'inverse, lors de la production de la voyelle /y/ les lèvres sont avancées et la cavité labiale intervient pleinement dans la production.
    - Le pharynx : il est la racine de la langue et il constitue la paroi antérieure de l'oropharynx. La paroi postérieure est également recouverte de muscles et de muqueuses. Il est le lieu d'articulation de certaines consonnes particulières ; les pharyngales, que l'on retrouve notamment en langue Arabe.
    - La langue : c'est un articulateur fondamental pour la parole. Il est constitué d'un système musculaire complexe composé de dix sept muscles qui lui assurent une grande mobilité, mais surtout d'adopter des positions et des configurations particulières pour produire les différents sons.

- Le mandibule : c'est un os, qui s'impose comme un articulateur à part entière. Il est entouré d'un système musculaire complexe, lui permettant d'avoir des interactions avec la langue et la lèvre inférieure.

Les modifications de formes et de dimensions que subit la bouche, le pharynx pendant l'émission de la voix, donnent à la voix un timbre, un cachet qui est particulier à chacun d'entre nous. La voix subit d'importantes altérations quand l'un des éléments du conduit vocal est malade, quand le nez est obstrué, quand les amygdales sont tuméfiées, quand la langue est malade. Même pour les personnes âgées, l'intensité et la hauteur de la voix diminuent à cause de la faiblesse des muscles laryngiens et du souffle respiratoire.

La parole est un signal qui véhicule des informations ne concernant pas uniquement la signification objective du message, il contient des données sur l'accent, le rythme et l'intonation du locuteur, qui renseignent, non seulement sur son origine, mais également sur ses caractéristiques intrinsèques et ses états émotionnels.

L'étude du système phonatoire, notamment ses configurations volumiques variées, va nous permettre d'identifier les grandes classes de sons. Nous verrons que le mécanisme articulatoire influe directement sur la production différenciée des sons. D'ailleurs, les phonéticiens se basent essentiellement sur le lieu et le mode d'articulation pour caractériser les sons du langage.

Le signal de parole étant un signal réel, continu, d'énergie finie et non stationnaire, les caractéristiques du signal de parole et du conduit vocal évoluent dans le temps. Les positions des composantes du système phonatoire (cordes vocales, langue, lèvres, dents, mâchoire, etc.) agissent comme dans une opération de filtrage et sont déplacées pour former différents sons, en augmentant certaines fréquences tout en atténuant d'autres.

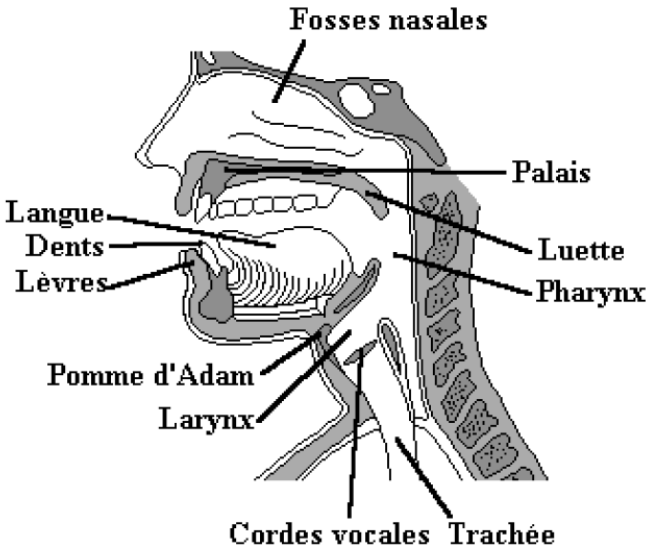

Figure 1.1 : Coupe longitudinale de l'appareil phonatoire humain.

1.2. Perception de la parole

Chez l'auditeur, la phase de perception se résume à la réception de l'onde acoustique au niveau de l'oreille, puis sa transformation en impulsions nerveuses et son transfert vers le cerveau à travers le nerf auditif. L'onde acoustique arrivant dans l'oreille de l'auditeur est transformée en vibrations mécaniques par le tympan et les chaînes des osselets [2].

Une première analyse, de nature essentiellement fréquentielle, est effectuée par la propagation le long de la cochlée. Le signal est ensuite transformé en un ensemble de stimuli neuronaux par les cellules ciliées de l'organe de Corti avec une sélectivité fréquentielle accrue. Cette activité neuronale est transmise au cortex à travers une organisation basée sur la sélection tonale du système auditif.

Des processus auditifs centraux, à la fois innés et acquis, interviennent en aval du traitement périphérique effectué par l'oreille pour assurer un décodage. Le processus de perception de la parole qui permet de décoder un message vocal à partir de la représentation neuronale des sons est beaucoup moins élucidé, mais nous pouvons penser qu'un phénomène de conceptualisation se met en place pour que le message transmis soit compris par l'auditeur qui se l'approprie. Par ailleurs, la psycho acoustique, discipline dont l'apport au traitement de la parole est inestimable, a réussi à établir des relations entre les stimuli acoustiques et les sensations perceptives. Des études récentes ont montré que le système auditif est organisé de

façon hiérarchique avec une complexité croissante de l'organisation des relais auditifs dans le cortex.

Les fréquences sonores restent associées de façon régulière à des positions localisées des fibres nerveuses au sein des cartes neuronales. Sachant que les pics de spectre sont plus importants pour la perception, les méthodes d'évaluation devront modéliser les formants du signal. Par ailleurs une différence de 20 dB entre deux signaux entrainera le masquage du signal le plus faible.

1.3. Classification des sons de la parole

Les sons produits par le système phonatoire humain peuvent être rattachés à différentes classes. Ces classes permettent de regrouper les sons selon leurs principales caractéristiques qui sont identifiables. À l'intérieur de ces classes sont regroupés des sons dont les dissimilarités peuvent être faibles.

1.3.1. Le phonème

Le phonème est la plus petite unité présente dans la parole et susceptible par sa présence de changer la signification d'un mot. Sur le plan linguistique, les similarités et les différences entre phonèmes peuvent être quantifiées en décrivant chaque phonème par un ensemble de traits distinctifs [3].

S'agissant d'une unité abstraite non observable, le phonème peut être réalisé acoustiquement sous forme allophonique multiple ou "phones". Les phonèmes successifs sont liés entre eux et s'influencent mutuellement, de sorte que les réalisations d'un même phonème peuvent largement différer sur le plan acoustique en fonction du contexte, mais aussi de la vitesse d'élocution, du style d'élocution, de l'origine géographique du locuteur, etc.

Le nombre de phonèmes est toujours limité (en général inférieur à 50) ; ainsi, la langue arabe comprend 28 phonèmes consonantiques (29 selon certains). Le système phonétique est un ensemble d'images acoustiques emmagasinées dans le cerveau du locuteur dans la mesure où celui-ci maîtrise la langue.

1.3.2. Sons voisés ou non-voisés

Une décomposition basée sur la périodicité et l'énergie d'un signal de parole fait ressortir deux types de sons :

Les sons voisés : Ils sont produits suite au passage de l'air des poumons à travers la trachée qui met en vibration les cordes vocales [4]. Ce mode, qui représente 80% du temps de phonation, est caractérisé en général par une quasi-périodicité et une énergie élevée.

Les sons non-voisés : Ils sont obtenues par resserrement du conduit vocal, les cordes vocales sont écartées et n'entrent pas en vibration. Ces sons sont apériodiques et ont habituellement une énergie inférieure aux sons voisés.

Les Figures 1.2 et 1.3 montrent des exemples de représentations spectrales de sons voisé et non voisé. Les formants sont plus marqués dans la représentation du son voisé, alors que le spectre du son non-voisé est beaucoup plus plat.

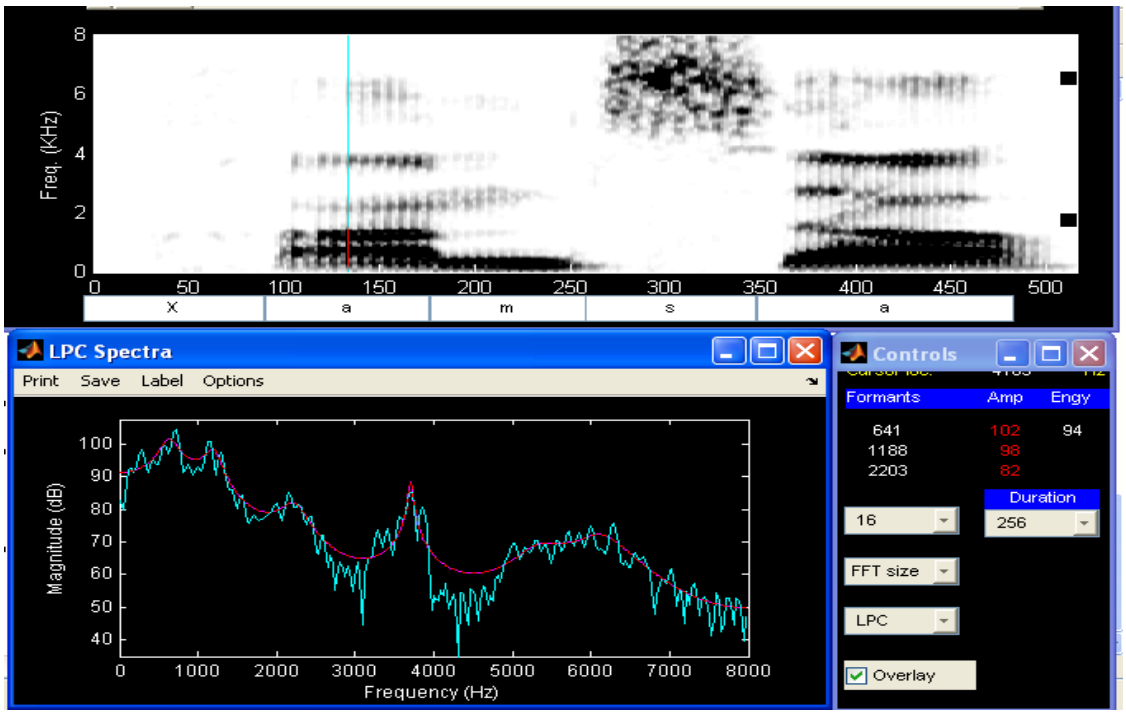

Figure 1.2 : Spectrogramme du mot arabe /Xamsa/ (chiffre arabe 5). Le spectre par LPC (en trait rouge) et le spectre par FFT (trait vert) calculés à partir d'une trame de 256 échantillons pris dans la partie stable de la première voyelle /a/ sont superposés. Le spectre LPC, qui a une forme lissée, permet de mettre en évidence les formants (pic de fréquences).

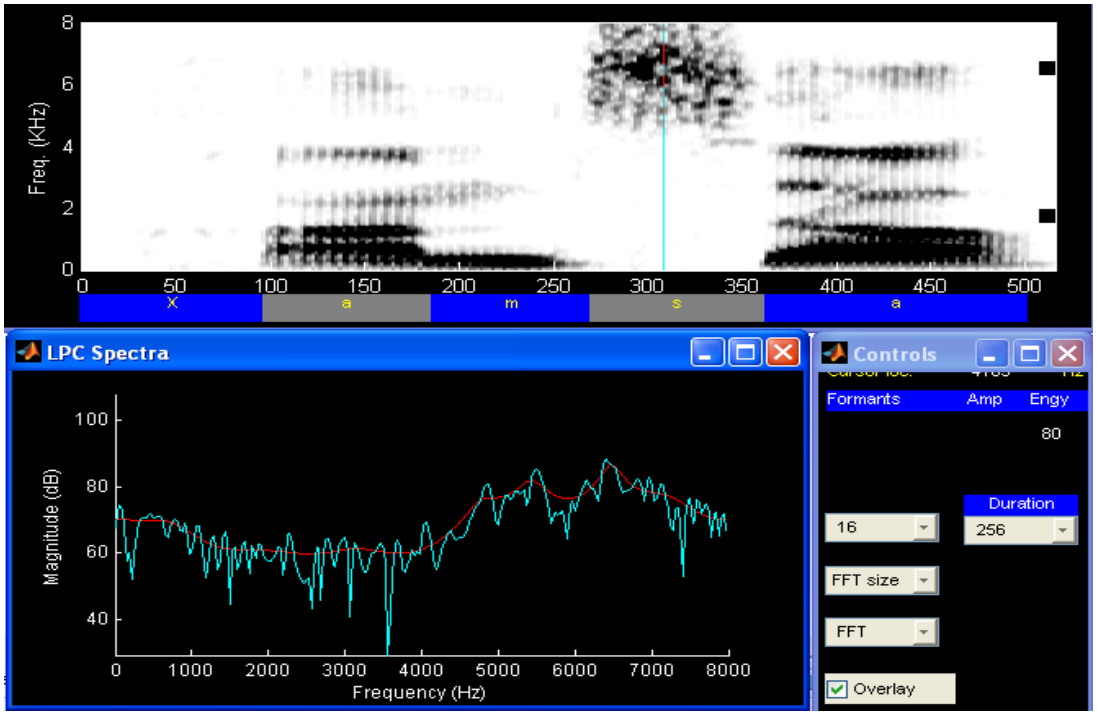

Figure 1.3 : Spectrogramme du mot arabe /Xamsa/ (chiffre arabe 5). Le spectre par LPC (en trait rouge) et le spectre par FFT (trait vert) calculés à partir d'une trame de 256 échantillons pris dans la partie stable de la fricative /s/ sont superposés. Le spectre LPC, qui a une forme lissée permet de mettre en évidence la bande de fréquence caractéristique des fricatives non voisées.

1.3.3. Classes phonétiques

Les différentes classes phonétiques existantes, correspondent à des regroupements qui suivent les modes de production. Traditionnellement, les sons sont divisés en voyelles et consonnes. Mais l'étude des sons de la parole a obligé à nuancer cette répartition et à créer d'autres classes subdivisant l'ensemble des consonnes [3]. Les différentes classes phonétiques présentes en français et en anglais sont :

Les voyelles

Elles sont caractérisées par le passage libre de l'air dans le conduit vocal, qui présente dans ce cas une configuration quasi-stable, la source d'excitation du conduit vocal est la vibration laryngienne. La fréquence fondamentale de cette vibration est appelée fréquence fondamentale F0 ou pitch. Les voyelles se différentient par leur lieu d'articulation, leur aperture ou degré d'ouverture.

Du point de vue acoustique, elles sont classées dans le plan des deux premiers formants F1-F2, dans une forme géométrique proche du triangle appelé triangle vocalique.

Elles se caractérisent principalement par le voisement qui crée des formants. Ces formants, qui sont des zones fréquentielles de forte énergie, correspondent à des résonances produites par une configuration particulière du conduit vocal. Ce sont principalement les formants en basses fréquences (F1, F2 et F3) qui caractérisent les voyelles.

Dans certaines langues, telle que le français, la nasalisation est également une caractéristique distinctive importante, car quatre voyelles de cette langue sont nasalisées.

Les consonnes

Ce sont des sons résultant d'une fermeture partielle (constriction) ou totale (occlusion) du conduit vocal lors du passage de l'air phonatoire. Elles peuvent être voisées ou non voisées, nasales ou orales. Les consonnes sont classées selon les principaux types suivants :

- **Les occlusives**

  Elles se caractérisent oralement par la fermeture du conduit vocal, fermeture précédant un brusque relâchement. Les occlusives peuvent être voisées comme */b/ et / d /*, ou sourdes, c'est à dire non voisées comme */ p / et / k /*.

- **Les fricatives**

  Cette classe regroupe les sons produits par la friction de l'air dans le conduit vocal lorsque celui-ci est rétréci au niveau des lèvres, des dents ou de la langue. Cette friction produit un bruit de hautes fréquences et peut être voisée ou sourde, par exemple entre 4 et 8 kHz pour la consonne non voisée */ s /,* ou la consonne voisée */ z /* qui comporte en plus une composante de voisement basse fréquence.

- **Les nasales**

  Elles résultent de l'obstruction du conduit vocal et de l'ouverture de vélum qui permet l'échappement de l'air par les cavités nasales. On distingue deux consonnes nasales, toutes les deux voisées :
  / m / : dont le lieu d'articulation est labial.
  / n / : dont le lieu d'articulation est dental.
  Il est à noter que certaines voyelles possèdent également un caractère de nasalité comme c'est le cas dans l'espagnol ou le français.

- **Vibrantes**

  Il s'avère qu'il en existe une seule : le / *r* / qui est produite par une vibration de la langue et qui est caractérisée par une structure de formants interrompus par des intervalles de silences très court, résultat du battement de la langue.

- **Liquides**

  Il en existe une seule / *l* /, produite par une obstruction partielle du conduit buccal et un écoulement latéral. Au plan spectral, elle est caractérisée par une structure de formant similaire à celle des voyelles.

- **Les semi-voyelles**

  Elles ont la structure acoustique des voyelles mais ne peuvent en jouer le rôle car elles ne sont que des transitions vers d'autres voyelles qui sont les véritables noyaux syllabiques. Les semi-voyelles sont évidement sonores.

  Nous distinguons les semi voyelles / *j* /, / *w* /

1.3.4. Fréquence fondamentale (F_0) ou pitch

Sous l'effet du passage de l'air phonatoire à travers la glotte, les cordes vocales peuvent entrer en vibration. La fréquence de vibration est appelée fréquence fondamentale ou pitch. Cette périodicité des vibrations peut être constatée sur la forme temporelle du signal vocal tel que donné par la Figure 1.4.

La fréquence fondamentale varie en fonction de chaque individu. En général, la variation s'étend de 80 à 200 Hz pour une voix masculine (voix grave), de 150 à 450 Hz pour une voix féminine et de 200 à 600 Hz pour une voix d'enfant (voix aiguës).

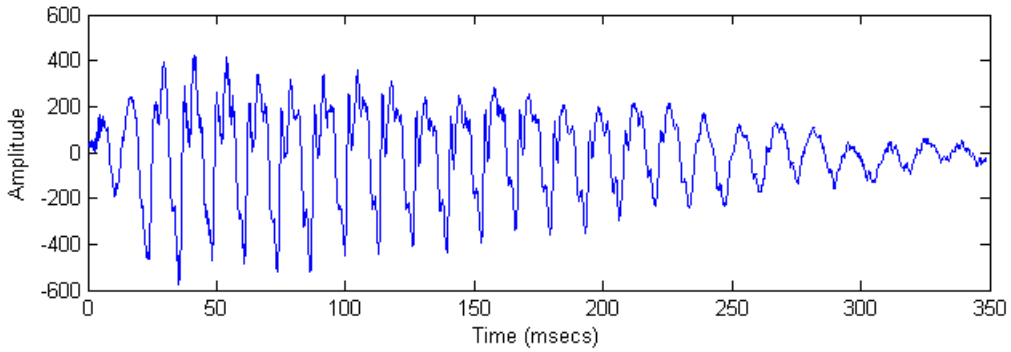

Figure 1.4 : Forme d'onde de la deuxième voyelle /a/ du mot arabe /Xamsa/ (chiffre arabe 5)

1.3.5. Les formants

De fait de sa configuration, le conduit vocal influe sur le signal vocal produit. Ainsi pour un son voisé, le conduit vocal se comporte comme un filtre tout pole, faisant ressortir les maximas d'amplitude ou formants. La Figure 1.2 - page 9, montre clairement la présence des formants (trait rouge) dans la voyelle voisée |*a*| du mot |*Xamsa*|.

Concernant les sons non voisés, le spectre du signal résultant d'une configuration particulière du conduit vocal donne une forme analogue à celle de la Figure 1.3 - page 10. Ainsi le spectre LPC (trait rouge) ne donne pas lieu à des pics de fréquences, mais plutôt à des plateaux.

1.4. Prétraitement Acoustique

Avant d'effectuer une analyse acoustique du signal de parole, il est nécessaire de procéder à un prétraitement acoustique sur ce signal [4].

1.4.1. Echantillonnage

Le signal de parole capté par le microphone étant analogique, il s'avère nécessaire de le numériser avant tout traitement. Une opération de filtrage analogique sur ce signal est d'abord effectuée au moyen d'un filtre passe bas ou passe bande. La fréquence de coupure basse doit être choisie inférieure à 80 Hz, afin d'éviter de filtrer le pitch. La fréquence de coupure haute dans le cas d'un filtrage par filtre passe bande est de 8 kHz pour la parole de bonne qualité et peut aller jusqu'à 11 kHz pour la parole de qualité élevée. Notons toutefois que le signal de la parole de qualité téléphonique requiert un filtrage dans la bande 300-3400 Hz.

La numérisation du signal consiste à échantillonner ce signal puis à quantifier chaque échantillon. D'après le théorème de Shannon, la perte d'information entre le signal analogique et le signal discret correspondant est nulle si et seulement si on a :

$$f_e \geq 2.f_{\max} \tag{1.1}$$

Où f_e est la fréquence d'échantillonnage, et $f_{\max}$ la fréquence maximale du signal à traiter.

La valeur de la fréquence d'échantillonnage doit être choisie en fonction du type d'analyse recherchée, mais surtout de l'application : voie téléphonique, système de haute

qualité vocale, système avec limitation en taille mémoire, etc. Dans les applications téléphoniques un échantillonnage de 8 kHz, avec une quantification de 8 bits est utilisé.

1.4.2. Préaccentuation et Fenêtrage

Préaccentuation

Avant d'aborder l'analyse acoustique, il est recommandé de faire subir au signal vocal un prétraitement, et ce, pour un bon conditionnement des algorithmes de résolution qui vont être utilisés dans les phases ultérieures. Le prétraitement acoustique se présente en deux opérations qui sont : la préaccentuation et le fenêtrage ou la fragmentation du signal en trames.

Si on prend en compte, l'atténuation de la source d'excitation et l'amplification par le rayonnement labial, il est nécessaire de traiter en conséquence le signal de parole résultant de cette association source-filtre. L'excitation subit une atténuation de -12dB/octave et le rayonnement aux lèvres amplifie de +6dB/octave. Ainsi, le spectre du signal en sortie est atténué de -6dB/octave. Pour retrouver la fonction de transfert initiale du conduit vocal et remédier à ce problème d'atténuation qui affecte le signal de parole, on procède à une préaccentuation numérique qui consiste à passer le signal dans un filtre du premier ordre dont la transmittance est *H(z)* telle que :

$$H(z) = 1 - a.z^{-1} \tag{1.2}$$

a : étant un coefficient de pondération compris entre 0.95 et 0.98**.**

Fenêtrage

Le fenêtrage est l'opération qui permet de délimiter la durée d'un signal pour s'assurer que l'on travaille sur des portions du signal stationnaire. Le signal de parole peut être considéré comme stationnaire sur des segments de courte durée, de l'ordre de 10 à 30 ms**.** Ainsi, pour limiter la durée d'un signal *x(n)*, on le multiplie par un autre signal *h(k)* possédant *N* échantillons unités. Un tel signal est souvent appelé fonction fenêtre ou plus simplement fenêtre temporelle glissante. Le signal *s(n)* de la trame résultante à *N* échantillons est calculé comme suit :

$$s(n) = x(n).h(n) \tag{1.3}$$

où *h(n)* a pour équation :

$$h(n) = \alpha + (1-\alpha)\cos(\frac{2\pi n}{N}) \quad (1.4)$$

α : étant un coefficient compris entre]0 ; 1] ;

Si $\alpha = 1$, on a une fenêtre rectangulaire.

Si $\alpha = 0.5$ on a une fenêtre de Hanning.

Si $\alpha = 0.54$ on a une fenêtre de Hamming (la plus utilisée en analyse du signal de la parole).

Les fenêtres doivent avoir une longueur suffisante si l'on veut que l'analyse ait un sens. En pratique, en prend 128, 256 ou 512 points, et les fenêtres successives sont glissantes et se recouvrent de moitié, parfois d'un tiers.

Rappelons que la recommandation ETSI ES 20108 V1.1.3 préconise l'utilisation de trames de 20 ms, soit 160 échantillons pour le signal de nature téléphonique.

1.5. Techniques de codage du signal de parole

Les algorithmes de codage de la parole peuvent être divisés en trois classes bien distinctes : les codeurs de forme d'onde, les codeurs paramétriques et les codeurs hybrides [5]. Les codeurs de forme d'onde ne sont pas fortement influencés par les modèles de production de la parole ; par conséquent, ils sont plus simples à mettre en œuvre. L'objectif de cette classe de codeurs est de produire un signal reconstruit qui correspond au signal original aussi fidèlement que possible. Le signal reconstruit converge vers le signal original avec un débit binaire élevé.

Cependant, les codeurs paramétriques reposent sur les modèles de production de la parole. Ils sont basés sur l'extraction des paramètres de modèle du signal de parole et ont pour objet leur codage. La qualité de ces codeurs de parole est limitée à cause du signal synthétique reconstruit. Toutefois, comme on le voit dans la Figure 1.5, ils fournissent des performances supérieures aux codeurs de forme d'onde pour des débits inférieurs. Beaucoup de codeurs à approximation de forme d'onde utilisent des modèles de production de la parole afin d'améliorer l'efficacité du codage. Ces codeurs se chevauchent dans les deux catégories et sont donc appelés codeurs hybrides.

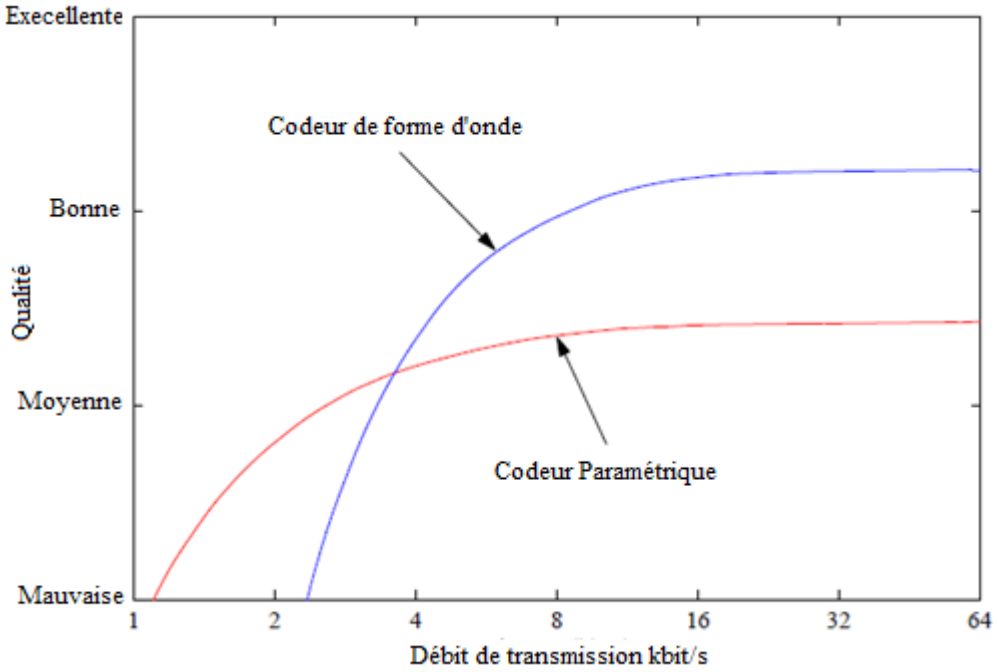

Figure 1.5 : Performances subjectives des codeurs de forme d'onde et paramétriques [5].

1.5.1. Les codeurs de forme d'onde

Les codeurs temporels codent directement la forme d'onde du signal, sans connaissance a priori de ses caractéristiques. Ce genre de codage est utilisé dans plusieurs standards destinés principalement aux réseaux téléphoniques filaires. Il conduit à un débit de codage élevé, typiquement compris entre 16 et 64 Kb/s, et permet de préserver une bonne qualité perceptuelle pour le signal vocal. La complexité de ces codeurs est minimale ainsi que le délai de codage.

Le codage porte directement sur les échantillons de la forme d'onde du signal de la parole qui sont quantifiés et transmis. Le signal variant peu d'un échantillon à un autre, le codage différentiel est introduit afin de réduire le débit. Dans ce cas, c'est la différence entre la valeur d'un échantillon et sa valeur prédite qui est quantifiée puis transmise. La prédiction est réalisée par une combinaison linéaire à partir des valeurs précédemment quantifiées.

La modulation par impulsion codées

La technique PCM (Pulse Code Modulation), adoptée par le standard G.711, effectue une quantification pseudo-logarithmique qui permet, à débit identique, d'obtenir une meilleure qualité perceptuelle qu'une quantification linéaire et ceci compte-tenu des propriétés de la perception auditive. Deux lois de compression pseudo-logarithmiques sont utilisées : la loi A et la loi μ. Chaque échantillon est codé sur 8 bits et la fréquence

d'échantillonnage est de 8 KHz, ce qui donne un débit de codage de 64 Kb/s. Le rapport signal sur bruit de ce codeur est de l'ordre de 35 dB sur une large plage d'amplitudes.

La modulation par impulsion et codage différentiel

L'ADPCM (Adaptive Differential Pulse Code Modulation) est une technique de codage différentiel qui utilise un pas de quantification adaptatif. Le pas est proportionnel à l'énergie à court terme normalisée du signal. La technique de codage permet d'obtenir des débits réduits allant de 16 à 40 Kb/s avec une faible dégradation dans la qualité du signal codé. Le codeur G.726 est basé sur la technique ADPCM.

1.5.2. Les codeurs paramétriques

Les codeurs paramétriques modélisent le signal vocal à l'aide d'un modèle paramétrique propre aux signaux de parole. Le codeur extrait les paramètres du modèle et les code sur un nombre limité de bits. Le décodeur utilise ces paramètres pour alimenter le modèle de production et synthétiser le signal de parole. Ces codeurs permettent d'obtenir de faibles débits de codage, typiquement inférieurs à 4 Kb/s, tout en conservant une bonne qualité de parole (MOS proche de 3,5) même si elle est généralement inférieure à celle des codeurs hybrides décrits dans le paragraphe 1.6.3. Les algorithmes de codage paramétriques sont destinés à des applications de sécurité et ils sont souvent développés et standardisés par des organismes militaires, comme le DoD ou l'OTAN.

Analyse par codage prédictif linéaire

Le codage par prédiction linéaire ou LPC (Linear Predictive Coding) est une technique de modélisation de la fonction de transfert du conduit vocal. La prédiction linéaire est une technique très utilisée dans les systèmes de codage et de compression du signal de parole, notamment dans les télécommunications. Cette méthode est considérée comme une technique prédominante pour l'estimation des paramètres de la parole. Son succès est dû au fait qu'elle représente une solution linéaire au problème de l'estimation des paramètres du modèle de la production de la parole.

Le principe fondamental de la prédiction linéaire est qu'un échantillon donné du signal peut être prédit à partir d'une combinaison linéaire des échantillons qui le précèdent. Un seul jeu de coefficients du prédicateur est déterminé en minimisant les différences entre les échantillons actuels et ceux prédits.

Un échantillon du signal de parole *s(n)* peut être modélisé comme la sortie d'un système auto régressif à moyenne ajustée (ARMA) avec une entrée *u(n)*. Son expression est alors :

$$s(n) = G.\sum_{i=0}^{q} b_i.u(n-i) - \sum_{k=1}^{p} a_k.s(n-k) \qquad b_0 = 1. \tag{1.5}$$

Où *G* est le gain, les coefficients $\{a_k\}$ et $\{b_i\}$ sont les paramètres du système, *p* et *q* sont les ordres des polynômes. L'équation (1.5) prédit la sortie courante en utilisant une combinaison linéaire des sorties précédentes et des entrées courantes et précédentes.

Dans le domaine fréquentiel, la fonction de transfert du modèle de prédiction linéaire de la parole est de la forme :

$$H(z) = \frac{S(z)}{U(z)} = \frac{G.[1 + \sum_{i=1}^{q} b_i.z^{-i}]}{1 + \sum_{k=1}^{p} a_k.z^{-k}} \tag{1.6}$$

S(z) et U(z) étant respectivement la transformée en *z* du signal de la parole et celui de l'excitation glottique, *H(z)* est la fonction de transfert du modèle pôle – zéro dans lequel les racines du dénominateur et du numérateur sont respectivement, les pôles et le zéros du système.

Si $a_k = 0$ pour $1 \le k \le p$, $H(z)$ devient un modèle tous – zéros ou modèle à moyenne ajustée (MA), par contre si $b_i = 0$ pour $1 \le i \le q$, $H(z)$ devient un modèle tous – pôles ou modèle auto régressive (AR), exprimé par l'Equation (1.7).

$$H(z) = G.\frac{1}{1 + \sum_{i=1}^{i=p} a_k z^{-k}} \tag{1.7}$$

La Figure 1.6 illustre le schéma simplifié de la production de la parole basée sur le modèle Auto Régressif :

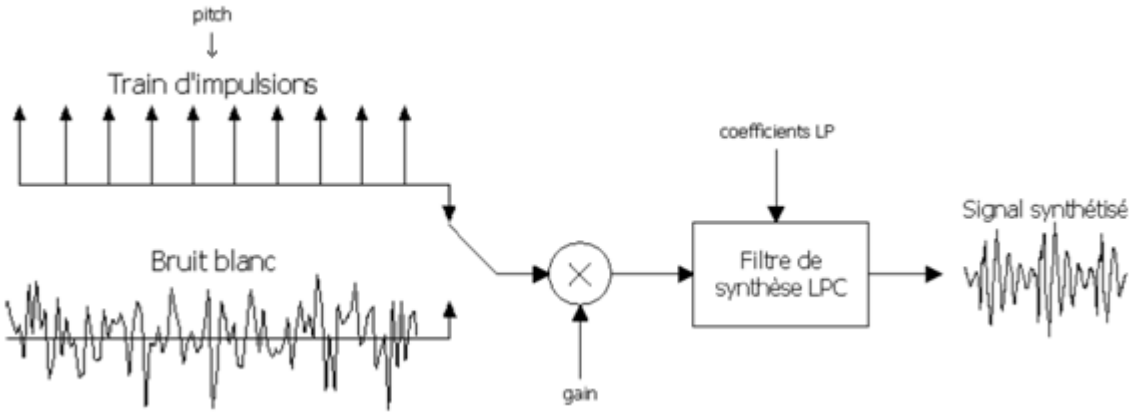

Figure 1.6 : Modèle simplifié de la production de la parole.

L'analyse spectrale montre que les pôles correspondent aux résonances du conduit vocal, c'est-à-dire aux *pics* que sont les *formants* ; tandis que les zéro correspondent aux antirésonances, c'est-à-dire aux *vallées*. Dans l'analyse de la parole, les classes de phonèmes comme les fricatives et les nasales, contiennent des *vallées* spectrales qui correspondent aux zéros dans *(Hz).* Par contre, les voyelles contiennent des résonances qui peuvent être modélisée par le modèle tous – pôle. Pour des raisons de facilité d'implémentation et de temps de calcul, le modèle AR est préféré pour l'analyse par prédiction linéaire de la parole, malgré qu'il ne reproduise pas correctement certains phonèmes tels que les nasales. Ainsi, le signal prédit est égal à :

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{1.8}$$

La différence entre l'échantillon original $s(n)$ et l'échantillon prédit $\hat{s}(n)$ est appelée erreur de prédiction (ou résidu), et elle est définie par :

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{1.9}$$

Le problème de l'analyse par prédiction linéaire se réduit donc à trouver un ensemble de coefficients a_k de façon à minimiser l'erreur de prédiction *e(n)* dans un certain intervalle. Les méthodes d'estimation des coefficients a_k sont nombreuses. Deux grandes approches sont utilisées pour l'analyse par prédiction linéaire LPC court terme : la méthode d'autocorrélation (algorithme de Levinson) et la méthode de covariance (algorithme de Cholesky).

1.5.3. Les codeurs hybrides

La qualité de la parole des codeurs de forme d'onde diminue rapidement pour des débits de moins de 16 kbps. Les codeurs hybrides sont donc utilisés pour combler cette lacune et offrir une parole de bonne qualité à des débits moyens. Toutefois, ces codeurs ont tendance à être plus exigeants en calcul. Pratiquement tous les codeurs hybrides reposent sur une analyse LPC pour obtenir les paramètres du modèle de synthèse et les techniques de codage de forme d'onde sont ensuite utilisées pour coder le signal d'excitation.

Le codage CELP

Les codeurs hybrides sont pour la plupart dérivés de l'algorithme de codage CELP (Code Excited Linear Prediction) introduit par Schroeder et Atal [9]. Cet algorithme et ses dérivés sont à la base de la majorité des standards de codeurs de parole utilisés dans les réseaux de téléphonie mobiles et les communications vocales sur Internet en voix sur IP. Les codeurs qui utilisent un algorithme CELP délivrent un débit moyen compris typiquement entre 4 et 16 Kb/s (4.8 Kb/s, 8 Kb/s, 16 Kb/s), avec une bonne qualité de codage allant de 3 à 4 en terme de note MOS. La Figure 1.7 représente le principe du codage CELP [6].

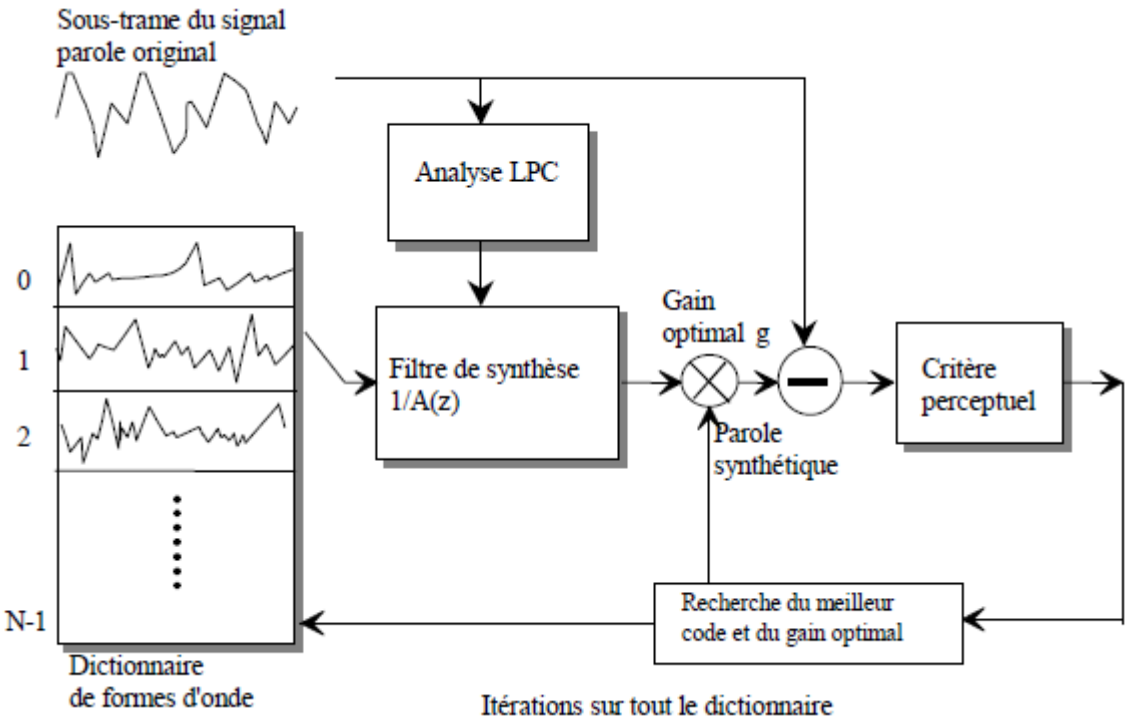

Figure 1.7 : Principe du codage CELP [6].

Dans chaque trame, une analyse spectrale par prédiction linéaire détermine le filtre de synthèse *1/A(z)*. On découpe chaque trame en sous-trames (durée typique 5 ms) sur

lesquelles on effectue une quantification vectorielle du signal par une technique d'analyse par synthèse. On compare à l'aide d'un critère dit perceptuel de type moindres carrés pondérés, le signal de parole original avec tous les signaux synthétiques possibles obtenus après quantification vectorielle. Ces signaux synthétiques sont générés en filtrant par le filtre de synthèse, un signal d'excitation choisi dans un dictionnaire de séquences d'excitation (on ajoute parfois la sortie de plusieurs dictionnaires) et en ajustant le signal résultant par le gain optimal. Le codeur transmet le ou les index des segments qui minimisent le critère ainsi que le ou les gains associés, les paramètres spectraux et le pitch fractionnaire. Le critère perceptuel prend en compte la propriété de masquage du bruit de quantification par les formants en pondérant plus fortement l'erreur de quantification dans les zones de faible amplitude du spectre et plus faiblement dans les zones de formants. Cette pondération s'effectue en filtrant le signal d'erreur par un filtre de type $A(z)/A(z/\gamma)$ où γ est compris entre 0 et 1 (typiquement γ = 0.85). Les dictionnaires utilisés sont appelés stochastiques ou adaptatifs selon qu'ils contiennent des séquences fixes de bruit ou bien les séquences d'excitation de trames précédentes. Le dictionnaire adaptatif permet de prendre en compte la redondance introduite par la quasi-périodicité des sons voisés.

1.6. Les attributs d'un codeur de parole

Les codeurs de parole ont plusieurs attributs, dont certains sont divergents. Les codeurs de parole sont généralement optimisés avec certains de ces attributs selon les besoins de l'application [5]. Les principaux attributs d'un codeur de parole sont :

1.6.1. Débit binaire

L'objectif d'un algorithme de codage est de réduire le débit binaire tout en maintenant la bonne qualité du signal de parole. Les normes existantes, spécifient des codeurs de parole à débit fixe ou variable. Les débits nécessaires sont souvent tributaires du canal de communication et de l'application prévue. Par exemple, pour la téléphonie par satellite et cellulaire, les débits binaires varient de 3,3 à 13 kbit / s, et pour le réseau téléphonique commuté, les débits en cours d'utilisation sont de 64 kbit / s.

1.6.2. Qualité de la parole

La qualité de la parole reconstruite est un attribut essentiel des codeurs de parole. Pour un débit fixé, le critère de qualité pourra alors être employé pour évaluer un système de codage. Deux types de mesures, peuvent permettre l'évaluation de la qualité de parole :

Les mesures subjectives

La qualité de restitution peut être déterminée par des tests d'écoute du signal de parole, original et synthétisé dans les conditions désirées, où des auditeurs jugeront, subjectivement, de la qualité globale et de l'intelligibilité de la parole. Pour ce genre d'étude, un grand nombre de personnes est nécessaire pour effectuer une analyse statistique de leur opinion moyenne (MOS : Mean Opinion Score) [7]. Une telle expérimentation est très contraignante à réaliser mais reste incontournable. Un échantillon de résultats de test simple, comme par exemple comparer le signal synthétisé au signal original, peut apporter de premières informations significatives sur les forces et les faiblesses de la procédure de codage. La perception des caractéristiques de la parole tend à changer considérablement entre les différents auditeurs, mais ces essais restent toutefois utiles pour pointer sur différents aspects déterminés unanimement. Les tests nécessaires à l'établissement d'un MOS sont spécifiés dans la recommandation « P.800 : Méthodes d'évaluation subjective de la qualité de transmission » de l'ITU-T. Le score MOS est une valeur numérique de 1 à 5, telle que donnée par le Tableau 1.2.

Tableau 1.2 : Mean Opinion Score – MOS

Score MOS	**Qualité perçue**	**Dégradation**
5	Excellente	Imperceptible
4	Bonne	Perceptible mais pas gênante
3	Moyenne	Légèrement gênante
2	Mauvaise	Gênante
1	Très mauvaise	Très gênante

Les mesures objectives

Ils utilisent des fonctions ou des critères mathématiques pour comparer les formes d'onde synthétisées et originales telles que des mesures de distorsion ou de gain. Certaines mesures donnent des informations utiles selon le type de codage testé. Par exemple, le rapport signal sur bruit (SNR : Signal to Noise Ratio) est représentatif pour les codeurs temporels et certains codeurs hybrides, tels que les codeurs de type CELP, qui incorporent des mécanismes de modélisation de forme d'onde. D'autres évaluent certains éléments des algorithmes de codage, tels que les distorsions cepstrales ou spectrales, employés pour transcrire la déformation introduite par la quantification des paramètres LPC.

En outre, des mesures qui conduisent à de bonnes corrélations avec la perception humaine ont été élaborés. La mesure la plus communément utilisée est le PESQ (Perceptual Evaluation of Speech Quality), normalisé par l'UIT-T sous le nom P.862 «Evaluation de la qualité vocale perçue: méthode objective d'évaluation de la qualité vocale de bout en bout des codecs vocaux et des réseaux téléphoniques à bande étroite». Cette méthode permet d'évaluer la qualité d'écoute dans de nombreuses conditions de dégradation (perte de paquets, distorsion due au codage, bruit ambiant du côté émission, variation du délai), aboutissant à une corrélation de 0,935 avec les notes subjectives. L'évaluation PESQ compare un signal original à un signal reconstitué produit par le passage du signal original à travers un système de communication. Elle permet de prévoir les notes de qualité de perception qu'attribueraient au signal restitué les sujets participant à un essai d'écoute subjectif [8]. Idéalement, les mesures objectives rejoignent les résultats obtenus par notre perception subjective de la parole. Toutefois, les tests subjectifs et objectifs peuvent produire des résultats légèrement différents.

1.6.3. Le délai

Le délai est important, surtout dans les communications full-duplex temps réel. Le seuil de retard est tributaire de la nature de l'application. Par exemple, pour des conversations hautement interactives, un retard au-dessus de 150 ms peut être perçu comme une défaillance. D'autre part, pour une conversation normale, un retard de 400 à 500 ms peut être toléré sans une réduction significative de la performance globale. Toutefois, il convient de noter que dans un système sans annulation d'écho, le seuil de retard peut être aussi bas que 100 ms. Le délai d'un coder est souvent composé de quatre segments. Le premier est le délai algorithmique, qui est un retard dû à l'accumulation d'une trame de parole avec un délai d'exploration. Le second est le temps de transmission dans la liaison de communication. Le troisième est le délai de multiplexage lors de la combinaison de données audio avec d'autres données. Le dernier est le temps de traitement nécessaire pour les opérations de codage et de décodage

1.6.4. Sensibilité aux erreurs de canal (Robustesse)

Les erreurs de canal sont divisées en deux principaux types. Le premier est celui des erreurs aléatoires, qui sont généralement dues au bruit de canal. Ceci est normalement spécifié comme taux d'erreur binaire (BER : Bit Error Ratio) et est limité à environ 1%. Pour contrer les erreurs aléatoires, un rapport signal sur bruit suffisant doit être réalisé. En outre, un codage de canal, qui est différent du codage de la parole (codage de source), est effectué. Le codage de canal est souvent mis en œuvre par l'ajout de redondances à l'information transmise afin de

la rendre plus robuste contre les erreurs de canal. L'inconvénient, cependant, est la surcharge supplémentaire encourue en raison de l'ajout des redondances. Le second type d'erreur est le train d'erreur, qui est plus commun dans les canaux mobiles et survient en raison de mécanismes tels que le fading. Pour se prémunir contre une telle erreur, des schémas de détection d'erreurs sont mises en œuvre.

Conclusion

Dans ce chapitre nous avons décrit le mécanisme de production de la parole, les caractéristiques du signal de la parole et les techniques de son codage. Ces éléments sont essentiels pour la détection de l'activité vocale (VAD) et aideront certainement à la bonne compréhension du principe de fonctionnement du codec G729, qui sera détaillé dans le prochain chapitre.

Chapitre 2 :

Le codec G729

Chapitre 2 :

Le Codec G729

Introduction :

Le but du codage est de diminuer le débit nécessaire à la transmission des informations de synthétisation de la parole à la réception. Basée sur l'étude des caractéristiques acoustiques et perceptives des sons de la parole, la compression (ou codage) des signaux à l'émission occupe donc une place importante dans les systèmes de communications. Etant réalisée par les codeur-décodeurs, ou codecs (codage à l'émission et décodage pour synthétiser la parole à la réception), cette opération peut contribuer à une utilisation optimale des réseaux de communications par réduction de la consommation de la bande passante.

Pour bien situer le contexte de ce mémoire, le présent chapitre est dédié au codec G729 qui a l'avantage de pouvoir être associé avec la détection de l'activité vocale, afin de diminuer le débit de transmission pendant les pauses de parole.

Après une brève présentation de la recommandation G729 et de ses différentes annexes, dont l'annexe B, nous décrirons le codec G729, de type CS-ACELP (Conjugate Structure - Algebraic CELP : prédiction linéaire à excitation par séquences codées à structure algébrique conjuguée), en soulignant ses principales caractéristiques et fonctionnalités. Dans sa version standard à 8 Kbps, le codeur opère sur des trames vocales de 10 ms correspondant à 80 échantillons. Le signal numérique injecté dans ce codeur est obtenu en effectuant d'abord un filtrage du signal analogique d'entrée dans la bande téléphonique, puis en l'échantillonnant à 8 kHz.

2.1. La recommandation G729 ITU-T

Cette recommandation contient la description d'un algorithme pour le codage des signaux de parole au moyen de la prédiction linéaire à excitation par séquences codées à structure algébrique conjuguée (CS-ACELP) [10].

Dans son mode de base, le codeur G.729 consiste en un codeur de parole à débit constant de 8 kbit/s, utilisant des opérations arithmétiques à virgule fixe. Les annexes A, B et D à J étendent ses fonctionnalités. L'annexe A présente une version simplifiée du codec vocal CS-ACELP à 8 kbit/s. La qualité du son est légèrement réduite mais reste proche de celle fournie par la version originelle. Cette version simplifiée du codec a été mise au point pour des applications de transmission simultanée de signaux vocaux et de données. L'annexe B définit un schéma de compression des silences pour la recommandation G729 incluant un VAD/DTX/CNG (détecteur d'activité vocale, dispositif de transmission discontinue et générateur de bruit de confort). Les annexes D, E et H définissent un fonctionnement à débit multiple et précisent les mécanismes de commutation de débit : l'annexe D spécifie une extension à faible débit (6,4 kbit/s) pour l'algorithme G.729, censée conférer à cet algorithme une plus grande souplesse, notamment dans des conditions de surcharge, L'annexe E décrit une extension à débit élevé (11,8 kbit/s) de l'algorithme G.729 conçue pour offrir une meilleure qualité de la voix en présence de bruits de fond, tandis que l'annexe H définit les mécanismes nécessaires aux opérations de commutation entre les systèmes conformes à l'annexe D (6,4 kbit/s) et à l'annexe E (11,8 kbit/s) de la recommandation G.729. Par conséquent, les annexes D, E et H ne mettent pas en œuvre le mode de transmission discontinue de l'annexe B. Pour cette fonctionnalité, de nouvelles annexes ont été développés.

Les annexes F et G définissent la fonctionnalité DTX pour les annexes D et E de la recommandation G.729 au moyen de l'algorithme de base présenté dans l'annexe B de cette même recommandation. L'annexe I fournit la fonctionnalité DTX à l'annexe H et décrit l'intégration du corps principal de G.729 avec les annexes B, D et E. L'annexe J fait référence à une extension à large bande modulable de 8 à 32 kbit/s interopérable avec le codeur G.729 et ses annexes A et B. Comme le corps principal G.729, les annexes A, B et D à J utilisent l'arithmétique à virgule fixe. Des implémentations alternatives basées sur des opérations arithmétiques à virgule flottantes qui sont mieux adaptée aux processeurs standards des ordinateurs type PC Pentium figurent dans l'annexe C pour le corps principal G729 et son

annexe A, et dans l'annexe C + pour l'annexe I. Ces informations sont résumées dans le Tableau 2.1 ci-dessous.

Tableau 2.1 : Les fonctionnalités du codec G729 suivant les différentes annexes ITU-T.

	Annexes											
Fonctionnalité	**-**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**C+**	**J**
Faible niveau de complexité		X	X									
virgule fixe	X	X	X		X	X	X	X	X	X		X
virgule flottante				X							X	
8 kbit/s	X	X	X	X	X	X	X	X	X	X	X	X
6.4 kbit/s					X		X		X	X	X	
11.8 kbit/s						X		X	X	X	X	
DTX			X				X	X		X	X	
Train binaire embarqué variable, large bande												X

2.2. Description du codeur vocal CS-ACELP

Dans le codeur CS-ACELP, le signal décodé localement est comparé au signal original et les paramètres de codage sont sélectionnés de telle sorte que l'erreur quadratique pondérée entre le signal original et le signal reconstruit soit minimisée. Le codeur CS-ACELP est conçu pour fonctionner avec un signal approprié à bande limitée, échantillonné à 8000 Hz. Les échantillons d'entrée et de sortie sont représentés en mots MIC linéaires de 16 bits. Le codeur opère sur des trames de 10 ms, en utilisant un délai d'exploration de 5 ms pour l'analyse prédictive linéaire (LP). Il en résulte un délai algorithmique global de 15 ms. Le principe de codage CS-ACELP est représenté dans la Figure 2.1.

Après le traitement des échantillons d'entrée de mots MIC de 16 bits à travers un filtre passe-haut de fréquence de coupure de 140 Hz, une analyse LP du dixième ordre est effectuée, et les paramètres LP sont quantifiés dans le domaine des paires de raies spectrales (LSP : Line Spectral Pair) sur 18 bits. La trame d'entrée est divisée en deux sous-trames de 5 ms chacune. L'utilisation de sous-trames permet de mieux localiser les paramètres de délai tonal et de gain et réduit la complexité d'exploration des répertoires codés. Les coefficients de filtre LP quantifiés et non quantifiés sont utilisés dans la deuxième sous-trame tandis que les coefficients de filtre LP interpolés (aussi bien quantifiés que non quantifiés) sont utilisés dans la première sous-trame. Pour chaque sous-trame l'excitation est représentée par la contribution

du répertoire codé adaptatif et du répertoire codé fixe. Les paramètres des répertoires codées adaptatifs et fixes sont transmis chaque sous-trame.

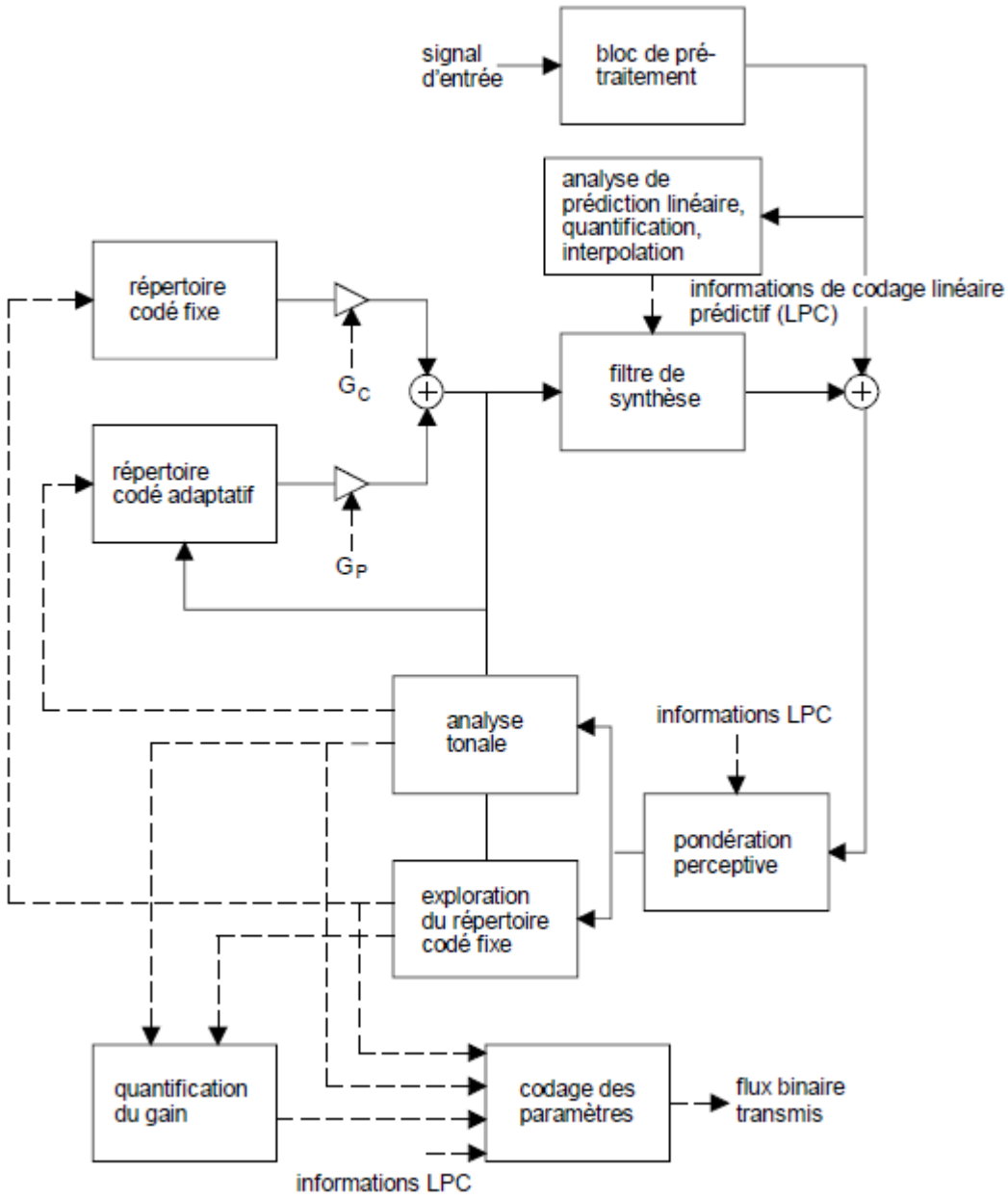

Figure 2.1 : Principe de codage dans l'algorithme CS-ACELP [10].

Le répertoire codé adaptatif représente la périodicité dans le signal d'excitation en utilisant un délai tonal fractionnaire de résolution 1/3. Le répertoire codé adaptatif est exploré au moyen d'une procédure en deux étapes. Un délai tonal en boucle ouverte est estimé une fois par trame en fonction du signal de parole pondérée. L'index et le gain du répertoire codé adaptatif sont trouvés par une recherche en boucle fermée autour du délai tonal en boucle

ouverte. Le signal à identifier, dénommé le signal cible, est calculé par filtrage du signal résiduel LP à travers le filtre de synthèse pondérée.

L'index du répertoire codé adaptatif est codé sur 8 bits dans la première sous-trame et sur 5 bits dans la seconde sous-trame. Le signal cible est mis à jour en enlevant la contribution du répertoire codé adaptatif. Ce nouveau signal cible est utilisé dans l'exploration du répertoire codé fixe. Le répertoire codé fixe est un répertoire algébrique de mots de 17 bits. Les gains des répertoires codés adaptatifs et fixes sont quantifiés vectoriellement sur 7 bits en utilisant un répertoire à structure conjuguée (avec une analyse à moyenne mobile (MA : Moving Average) appliquée au gain par répertoire codé fixe). L'allocation des bits pour une trame de 10 ms est montrée dans le Tableau 2.2.

Tableau 2.2 : Affectation des bits dans l'algorithme de codage CS-ACELP à 8 kbit/s.

Paramètre	Mot de code	Bits
Paires de raies spectrales	$L0$, $L1$, $L2$, $L3$	18
Délai du répertoire codé adaptatif	$P1$, $P2$	13
Parité du délai tonal	$P0$	1
Index de répertoire codé fixe	$C1$, $C2$	26
Signe de répertoire codé fixe	$S1$, $S2$	8
Gains de répertoire (étape 1)	$GA1$, $GA2$	6
Gains de répertoire (étape 2)	$GB1$, $GB2$	8

2.3. Description fonctionnelle du codeur CS-ACELP

Dans cette partie, nous allons détailler les différentes fonctions du codeur qui sont représentées par les blocs de la Figure 2.1.

2.3.1. Prétraitement

Avant le processus de codage, deux fonctions de prétraitement sont appliquées au signal de parole d'entrée, échantillonné à 8 kHz et quantifié sur 16 bits :

1) Une normalisation du signal qui consiste à diviser par un facteur de 2 l'énergie d'entrée afin de diminuer la probabilité de dépassements de capacité dans une réalisation en virgule fixe.

2) Un filtrage passe-haut qui sert de précaution à l'encontre de composantes parasites à basse fréquence. On fait appel à un filtre du deuxième ordre, dont la fréquence de coupure est de 140 Hz.

On combine les deux opérations, de normalisation moitié et de filtrage passe-haut, en divisant par 2 les coefficients figurant au numérateur de ce filtre, dont l'équation résultante est donnée par la formule suivante:

$$H(z) = \frac{0{,}46363718 - 0{,}92724705z^{-1} + 0{,}46363718z^{-2}}{1 - 1{,}9059465z^{-1} + 0{,}9114024\,z^{-2}} \tag{2.1}$$

Le signal d'entrée filtré par $H(z)$ est dénommé *s(n)*. Il sera utilisé dans toutes les opérations ultérieures du codeur.

2.3.2. Analyse par prédiction linéaire et quantification

L'analyse LP est effectuée à chaque trame vocale au moyen de la méthode d'autocorrélation avec une fenêtre asymétrique de 30 ms [14]. Tous les 80 échantillons (10 ms), les coefficients d'autocorrélation des données vocales fenêtrées sont calculés et convertis en coefficients LP par l'algorithme de Levinson-Durbin. Ensuite, les coefficients LP sont transformés en coefficients de paires de raies spectrales (LSP) à des fins de quantification et d'interpolation. Les coefficients LSP interpolés, quantifiés et non quantifiés, sont reconvertis en coefficients LP pour construire les filtres de synthèse et de pondération pour chaque sous-trame. L'analyse à court terme et les filtres de synthèse sont fondés sur des filtres LP du dixième ordre. Le filtre de synthèse LP est défini comme suit :

$$\frac{1}{\hat{A}(z)} = \frac{1}{1 + \sum_{i=1}^{10} \widehat{a_i}\, z^{-i}} \tag{2.2}$$

où $\hat{a}_i$, i = 1,..., 10, sont les coefficients LP (quantifiés).

a. Fenêtrage et calcul des coefficients d'autocorrélations

La fenêtre d'analyse LP se compose de deux parties: la première est une demi-fenêtre de Hamming et la seconde un quart de période d'une fonction cosinus. Cette fenêtre est donnée par l'équation suivante:

$$wlp(n) = \begin{cases} 0.54 + 0.46\cos\left(\frac{2\pi n}{399}\right) & n = 0,\ldots,199 \\ \cos\left(\frac{2\pi(n-200)}{159}\right) & n = 200,\ldots,239 \end{cases} \tag{2.3}$$

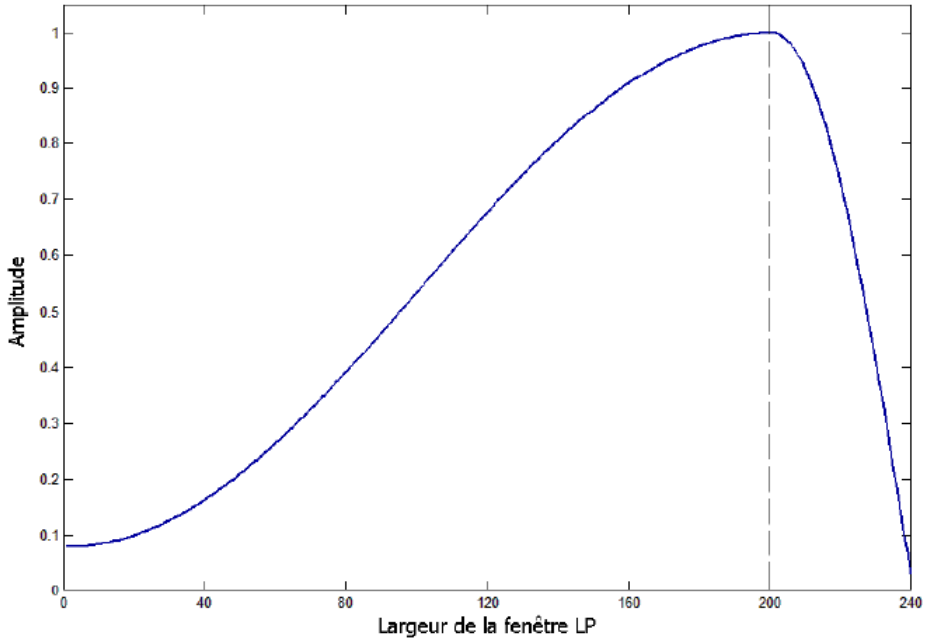

Figure 2.2: Fenêtre d'analyse LP [12]

L'analyse LP comporte une exploration de 5 ms, c'est-à-dire qu'il faut 40 échantillons issus de la prochaine trame vocale. Cela se traduit par un délai algorithmique supplémentaire de 5 ms au niveau du codeur. L'utilisation d'une fenêtre asymétrique permet une réduction du délai d'exploration sans compromettre la qualité du signal vocal reconstruit. La fenêtre d'analyse LP s'applique à 120 échantillons issus de trames vocales passées, à 80 échantillons issus de la trame vocale présente, et à 40 échantillons de la trame vocale future (Figure 2.2). L'utilisation d'une fenêtre de 30 ms permet une évolution lisse du filtre LP, fournissant ainsi une meilleure qualité de parole.

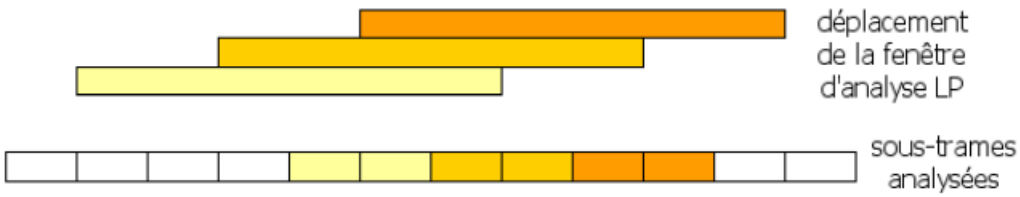

Figure 2.3 : Procédure de fenêtrage en analyse LP.

Les coefficients d'autocorrélation sont calculés à partir de la parole fenêtrée ainsi :

$$r(k) = \sum_{n=k}^{239} w_{lP}(n)s(n)\, w_{lP}(n-k)\, s(n-k), \quad k = 0, \ldots, 10 \qquad (2.4)$$

Les coefficients d'autocorrélation sont utilisés pour obtenir les coefficients de filtre LP, a_i, i = 1,.., 10, en utilisant l'algorithme de Levinson-Durbin.

b. Quantification des coefficients LSP

Les coefficients du filtre LP a_i, i = 1,..., 10 sont convertis en paires de raies spectrales (LSP) au moyen des polynômes de Tchebycheff. Dans cette procédure, les racines se trouvent dans le domaine cosinusoïdal. Puisque le quantificateur est basé sur une quantification vectorielle (VQ : Vector Quantization), il est plus commode de représenter les coefficients LSP dans le domaine fréquentiel normalisé [0, π]. La relation entre ces deux représentations est donnée par :

$$w_i = arc\cos(q_i)\; i = 1, \dots, 10 \qquad (2.5)$$

où q_i sont les coefficients LSP dans le domaine cosinusoidal, et w_i sont les coefficients LSP dans le domaine fréquentiel.

Une prédiction à moyenne mobile du quatrième ordre est utilisée pour prédire les coefficients LSP de la trame courante. La différence entre les coefficients calculés et prédits est quantifiée en utilisant un quantificateur vectoriel à deux étapes. La première étape est une quantification vectorielle à 10 dimensions utilisant un répertoire L_1 avec 128 entrées (7 bits). La deuxième étape est une quantification vectorielle sur 10 bits partagée entre deux répertoires à 5 dimensions, L_2 et L_3 contenant 32 entrées (5 bits) chacun. La raison d'utiliser une première étape non partagée est qu'elle permet l'exploitation des corrélations entre les 5 premiers et les 5 derniers coefficients LSP. Lors de la deuxième étape, ces corrélations sont moins fortes, et le partage réduit le temps de recherche et les besoins de stockage.

c. Interpolation des coefficients LSP

Les coefficients LP (quantifiés et non quantifiés) sont utilisés pour la deuxième sous-trame. Pour la première sous-trame, les coefficients LP (quantifiés et non quantifiés) sont obtenus par interpolation linéaire des paramètres correspondants dans les sous-trames adjacentes. Cette interpolation est appliquée aux coefficients LSP dans le domaine cosinusoïdal plutôt que dans le domaine fréquentiel. L'interpolation dans les deux domaines ne produit pas de différences audibles notables, le domaine cosinusoïdal a été choisi en raison de la facilité de mise en œuvre. Une fois quantifiés et interpolés, les coefficients LSP sont reconvertis en coefficients LP, $\hat{a}_i$.

2.3.3. Pondération perceptive

Le signal vocal pondéré $S_w(n)$ dans une sous trame est obtenu par filtrage de la parole à travers un filtre de pondération perceptive $W(z)$. Ce filtre de pondération perceptive est basé sur les coefficients non quantifiés a_i de filtre LP, et est donné par :

$$W(z) = \frac{A(\frac{z}{\gamma_1})}{A(\frac{z}{\gamma_2})} = \frac{\sum_{i=1}^{10} \gamma_1^i a_i z^{-i}}{\sum_{i=1}^{10} \gamma_2^i a_i z^{-i}} \quad (2.6)$$

L'utilisation des coefficients non quantifiées donne un filtre de pondération qui correspond mieux au spectre d'origine. Les coefficients de pondération γ_1 et γ_2 modifient la réponse en fréquence du filtre w(z). Il est difficile de trouver des valeurs fixes de γ_1 et γ_2 qui fournissent de bonnes performances pour des caractéristiques différentes du signal d'entrée. Par conséquent, les valeurs de γ_1 et γ_2 sont obtenus en fonction de la forme spectrale du signal d'entrée. Cette adaptation est effectuée à chaque trame de 10 ms mais on fait appel à une procédure d'interpolation pour chaque première sous-trame afin de lisser cette adaptation. La formes spectrale est obtenue au moyen d'un filtre de prédiction linéaire du 2^{e} ordre, construit en tant que sous produit de l'algorithme de récurrence de Levinson-Durbin.

2.3.4. L'analyse tonale

Un délai tonal en boucle ouverte T_o est estimé à chaque trame en utilisant le signal de parole pondéré $S_w(n)$. L'approche du répertoire codé adaptatif est utilisée pour représenter le composant périodique dans le signal d'excitation [13]. Le vecteur sélectionné du répertoire codé adaptatif est représenté par un index, qui correspond à une valeur déterminée du délai fractionnaire.

Pour chaque sous-trame le signal cible $x(n)$, et la réponse impulsionnelle $h(n)$, du filtre de synthèse pondéré sont calculés. Une recherche en boucle fermée du répertoire codé adaptatif est effectuée dans la première sous-trame autour de l'index correspondant à l'estimation (± 3) du délai tonal en boucle ouverte. Une résolution fractionnaire de 1/3 est utilisée dans la gamme $\left[19 + \frac{1}{3}, 85 - \frac{1}{3}\right]$ et seul des entiers sont utilisés dans la gamme 85 à 143. Il a été constaté que ce choix de résolution fournit un bon compromis entre la performance et le débit binaire. L'index du répertoire codé adaptatif dans la première sous-trame est codé sur 8 bits. Dans la deuxième sous-trame, une résolution fractionnaire de $\frac{1}{3}$ est utilisée dans la plage $[T_1 - (5 + \frac{2}{3}), T_1 + (4 + \frac{2}{3})]$ où T_1 est la partie entière du délai tonal

du répertoire codé adaptatif dans la première sous-trame. Cette gamme est adaptée pour les cas où T_1 chevauche les limites de la plage du délai tonal. Le délai tonal dans la deuxième sous-trame est codé sur 5 bits.

a. Estimation du délai tonal en boucle ouverte:

L'estimation du délai tonal en boucle ouverte utilise le signal vocal pondéré, et elle est réalisée comme suit: dans la première étape, nous recherchons 3 valeurs maximales de la corrélation donnée par l'équation (2.2) dans chacune des plages suivantes:

1) k = 80,…,143
2) k = 40,…,79
3) k = 20, …,39

$$R(k) = \sum_{n=0}^{79} s_w(n)s_w(n-k) \tag{2.7}$$

Notons que pour $n - k < 0$ les valeurs du signal de la trame précédente sont utilisées. Les maxima retenues $R(t_i)$, où t_i sont les valeurs de délai correspondant à des maximas dans les trois régions de délai $i = 1,\ldots,3$, sont normalisés par selon l'équation suivante :

$$R'(t_i) = \frac{R(t_i)}{\sqrt{\sum_n s_w^2(n-t_i)}}, \qquad i = 1,\ldots,3 \tag{2.8}$$

Le choix est réalisé parmi les trois corrélations normalisées en favorisant les délais avec des valeurs dans la gamme inférieure. Ceci est fait, en pondérant les corrélations normalisées qui correspondent aux plus longues valeurs de délai. Cette procédure qui consiste à diviser la gamme de délai en trois sections et favoriser les plus petites valeurs, est utilisée pour éviter de choisir des multiples de la hauteur tonale.

b. Calcul du signal cible

Le signal résiduel de prédiction linéaire est donné par :

$$r(n) = s(n) + \sum_{i=1}^{10} \hat{a}_i s(n-i), \quad n = 0,\ldots,39 \tag{2.9}$$

Le signal cible $x(n)$ pour l'exploration du répertoire codé adaptatif est calculé par filtrage du signal résiduel de prédiction linéaire $r(n)$ à travers la combinaison du filtre de synthèse $1/\hat{A}(z)$ et du filtre de pondération $A(z/\gamma_1) / A(z/\gamma_2)$. Après détermination de l'excitation pour la sous-trame, les états initiaux de ces filtres sont mis à jour par filtrage de la différence entre le signal résiduel et l'excitation.

c. Exploration du répertoire codé adaptatif

Les paramètres du répertoire codé adaptatif (ou paramètres tonaux) sont les index correspondant à un certain délai tonal et gain. Dans l'approche du répertoire codé adaptatif pour la mise en œuvre du filtre tonal, l'excitation est répétée pour les intervalles inférieurs à la longueur de la sous-trame. L'utilisation de délais fractionnaires rend ce processus de calcul coûteux lors de la phase de recherche. Ainsi, lors de la recherche, l'excitation au delà de la durée du délai tonal est prolongée par le signal résiduel LP. Cette procédure est plus simple, et il a été constaté qu'elle produit des résultats similaires à l'utilisation du répertoire codé adaptatif pour la sous-trame complète. Notons qu'une fois le délai déterminé, l'approche classique du répertoire codé adaptatif est utilisée pour générer le vecteur du répertoire codé adaptatif.

Pour chaque sous-trame de 5 ms, le délai est déterminé au moyen d'une analyse en boucle fermée qui minimise l'erreur quadratique pondérée. Dans la première sous-trame le délai T_1 est trouvé en explorant une petite étendue (six échantillons) des valeurs des délais autour du délai en boucle ouverte T_{op}. Pour la deuxième sous-trame, la recherche du répertoire adaptatif en boucle fermée est effectuée au tour du délai sélectionné dans la première sous-trame pour trouver le délai optimal T_2.

La recherche en boucle fermée minimise l'erreur quadratique pondérée entre la parole originale et reconstruite. Ceci est réalisé en maximisant le terme suivant :

$$R(k) = \frac{\sum_{n=0}^{39} x(n) y_k(n)}{\sum_{n=0}^{39} y_k(n) y_k(n)} \tag{2.10}$$

où $x(n)$ est le signal cible et $y_k(n)$ la dernière excitation filtrée au délai k (excitation passée convoluée avec $h(n)$, où $h(n)$ est la réponse impulsionnelle du filtre de synthèse pondéré $w(z)/\hat{A}(z)$).

Une fois que le délai tonal a été déterminé, on calcule le vecteur de répertoire codé adaptatif $v(n)$ en interpolant le précédent signal d'excitation $u(n)$ pour la valeur entière k du délai tonal indiqué et pour la fraction t. Quand au gain du répertoire codé adaptatif g_p on le calcule comme suit :

$$g_p = \frac{\sum_{n=0}^{39} x(n) y(n)}{\sum_{n=0}^{39} y(n) y(n)} \quad \text{borné par } 0 \leq g_p \leq 1.2 \tag{2.11}$$

où *x(n)* est le signal cible et *y(n)* est le vecteur *v(n)* filtré du répertoire adaptatif, obtenu par la convolution de *v(n)* avec *h(n)*. Le vecteur mis à l'échelle et filtré du répertoire codé adaptatif est soustrait de *x(n)* pour produire un nouveau signal cible *x'(n)*.

2.3.5. Répertoire codé (algébrique) fixe

Un répertoire algébrique de mots de 17 bits est utilisé pour le répertoire codé fixe [10]. Les répertoires algébriques sont des répertoires déterministes dans lesquels les vecteurs du répertoire codé sont déterminés à partir de l'indice transmis en utilisant une algèbre simple plutôt que des tables de correspondance. Cette structure présente des avantages en termes de stockage, de complexité de recherche, et de robustesse. Chaque vecteur du répertoire codé fixe contient quatre impulsions non nulles. Ces impulsions peuvent prendre les amplitudes et les positions indiquées dans le Tableau 2.3, et sont codés séparément en utilisant l'allocation de bits donnée dans ce tableau.

Tableau 2.3 : Structure du codebook algébrique

Impulsions	Amplitude	Positions	Bits
i_0	$S_0 : \pm 1$	m_0 :{0, 5, 10, 15, 20, 25, 30, 35}	4
i_1	$S_1 : \pm 1$	m_1 :{1, 6, 11, 16, 21, 26, 31, 36}	4
i_2	$S_2 : \pm 1$	m_2 :{2, 7, 12, 17, 22, 27, 32, 37}	4
i_3	$S_3 : \pm 1$	m_3 :{3, 8, 13, 18, 23, 28, 33, 38} {4, 9, 14, 19, 24, 29, 34, 39}	5

Soit c_k le vecteur du répertoire codé à index *k*. Le mot de code optimal est celui qui maximise le terme :

$$T_k = \frac{(d^t c_k)^2}{c_k^t \Phi \, c_k} \tag{2.12}$$

où *d* est le vecteur de corrélation entre le signal cible *x'(n)* et la réponse impulsionnelle *h(n)* du filtre de synthèse pondéré, et *Φ* est la matrice de corrélation de *h(n)*.

La structure du répertoire codé permet une procédure de recherche rapide puisque le vecteur du répertoire codé fixe ne contient que quatre impulsions non nulles dont les amplitudes sont $\pm$ 1.

Le répertoire codé fixe possède la caractéristique particulière que le code vectoriel choisi dans le répertoire passe par un préfiltre adaptatif *P*(*z*) qui renforce les composantes harmoniques pour améliorer la qualité du signal reconstitué. On utilise à cet effet le filtre suivant:

$$P(z) = 1/(1 - \beta z^{-T}) \tag{2.13}$$

où *T* est la partie entière du délai tonal pour la sous-trame courante, et β est le gain du répertoire codé adaptatif.

Cette modification des vecteurs du répertoire codé fixe est intégrée dans la recherche par filtrage de la réponse impulsionnelle *h(n)* avec *P(z)* avant la recherche du répertoire codé (pour les valeurs de délais inférieurs à la taille de la sous trame). Comme à ce moment le gain quantifié du répertoire adaptatif n'est pas connu, il a été constaté que le dernier gain tonal quantifié délimité par [0,2 - 0,8] fournit une bonne alternative.

2.3.6. Quantification des gains

Le gain par répertoire codé adaptatif et le gain par répertoire codé fixe sont quantifiés en vecteurs sur 7 bits. Cette quantification conjointe permet un gain d'environ 2 bits par rapport à la quantification scalaire.

L'exploration du répertoire codé de gain se fait en minimisant l'erreur quadratique pondérée entre la parole originale et reconstruite qui est donnée par

$$E_w = \sum_{n=0}^{39} \left(x(n) - g_p y(n) - g_c z(n) \right)^2 \tag{2.14}$$

où *x(n)* est le vecteur cible, et *y(n)* et *z(n)* sont respectivement, les vecteurs filtrés du répertoire codé adaptatif et du répertoire codé fixe.

a. Prédiction du gain par répertoire codé fixe :

Les gains par répertoire codé fixe dans des trames adjacentes sont corrélés. Un moyen efficace d'exploiter cette redondance est d'utiliser un prédicteur de gain d'énergie logarithmique. Ce prédicteur de gain permet, non seulement la réduction de la gamme dynamique du gain par répertoire codé fixe, mais il rend également cette gamme moins dépendante des variations de niveau d'entrée. L'utilisation d'un filtre à moyenne mobile réduit la propagation des erreurs de canal.

Le gain par répertoire codé fixe g_c peut être exprimé comme

$$g_c = \gamma\, \tilde{g}_c \tag{2.15}$$

où $\tilde{g}_c$ est un gain prédit sur la base d'énergies précédentes du répertoire codé fixe, et γ est un facteur de correction, qui est codé pour la transmission.

b. Exploration du répertoire pour quantifier les gains:

Le gain par répertoire codé adaptatif g_c, et le facteur γ sont quantifiés vectoriellement au moyen d'un répertoire codé à structure conjuguée en deux étapes. Le terme conjugué désigne le fait que chaque vecteur d'entrée est quantifié par une combinaison linéaire des deux dictionnaires codés. Une telle structure permet de réduire à la fois les besoins de calcul et de mémoire. La première étape consiste en un répertoire codé à deux dimensions sur 3 bits, *F*, et la deuxième étape consiste en un répertoire codé à deux dimensions sur 4 bits, *G*. Le premier élément de chaque répertoire codé représente le gain par répertoire codé adaptatif quantifié $\hat{g}_p$, et le second élément représente le facteur de correction $\hat{\gamma}$ du gain par répertoire codé fixe quantifié.

Si on a les index i_f et i_g, pour respectivement, les répertoire codé *F* et *G*, le gain par répertoire codé adaptatif quantifié est donné par :

$$\hat{g}_p = F_1(i_f) + G_1(i_g) \tag{2.16}$$

et le gain par répertoire codé fixe quantifié par :

$$\hat{g}_c = \tilde{g}_c\hat{\gamma} = \tilde{g}_c(F_2(i_f) + G_2(i_g)) \tag{2.17}$$

2.4. Description générale du décodeur :

La fonction du décodeur consiste en un décodage des paramètres transmis (paramètres LP, vecteur du répertoire codé adaptatif, vecteur du répertoire codé fixe, et les gains) et la synthèse pour obtenir la parole reconstruite, suivie d'une étape de post-traitement, qui comprend un post-filtre adaptatif et un filtre passe-haut [13].

Le principe du décodeur est représenté sur la Figure 2.3:

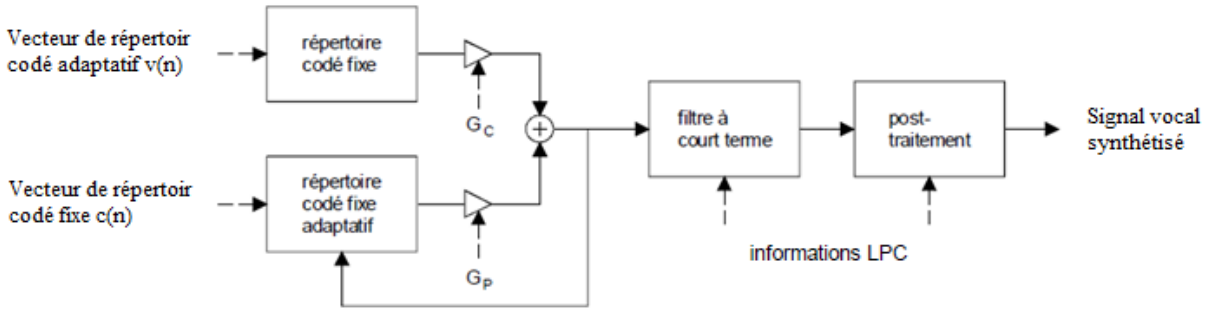

Figure 2.4 : Principe du décodeur CS-ACELP (Recommandation G729).

Les index paramétriques sont d'abord extraits du flux binaire reçu. Ces index sont ensuite décodés pour obtenir les paramètres de codage correspondant à une trame vocale de 10 ms. Ces paramètres sont les coefficients convertis en paires de raies spectrales (LSP), les 2 délais tonaux fractionnaires, le 2 vecteurs de répertoire codé fixe et les deux séries de gains par répertoire codé adaptatif et par répertoire codé fixe.

Les coefficients en paires LSP sont interpolés et reconvertis en coefficients de filtre de prédiction linéaire pour chaque sous-trame de 5 ms, qui passe par les étapes suivantes:

- l'excitation est construite par combinaison des codes vectoriels adaptatifs et fixes, normalisés par leur gain respectif;
- le signal vocal est reconstitué par filtrage de l'énergie d'excitation dans le filtre de synthèse du codage prédictif linéaire;
- le signal vocal reconstitué est amélioré par une opération de post-traitement qui met en œuvre un post-filtre adaptatif utilisant la sortie des filtres de synthèse à court et à long terme, suivi d'un filtre passe-haut et d'un échantillonneur-normalisateur.

2.5. Masquage des trames effacées

Une procédure de masquage des erreurs a été incorporée dans le décodeur afin de réduire la dégradation dans le signal vocal reconstitué en raison d'effacements de trame dans le flux binaire [10]. Ce processus de masquage des erreurs est fonctionnel lorsque la trame des paramètres du codeur (correspondant à une trame de 10 ms) a été identifiée comme étant effacée. La stratégie de masquage consiste à reconstruire la trame actuelle sur la base d'informations déjà reçues. Cette méthode remplace le signal d'excitation manquant par un signal de caractéristiques similaires, tout en diminuant progressivement son énergie. Une trame effacée hérite sa classe (périodique ou apériodique) de la trame vocale (reconstituée)

précédente. On notera que la classification des éléments voisés est mise à jour en permanence sur la base de ce signal reconstitué. Les étapes précises à suivre pour masquer une trame effacée sont les suivantes:

1) répétition des paramètres du filtre de synthèse;
2) affaiblissement de gains de répertoire codé adaptatif et de répertoire codé fixe;
3) affaiblissement de l'énergie mémorisée par le prédicteur de gain;
4) production de l'excitation de remplacement.

2.6. Utilisation du codec G729

Le codec G.729 est utilisé pour obtenir une téléphonie de qualité. Il est :

- Supporté par la majorité des IPBX (ou PABX IP),
- Utilisé pour le codage de la partie audio d'une visioconférence,
- Rencontré aussi pour transporter de la voix sur IP sur les WAN,
- Sera utilisé préférentiellement par les opérateurs de téléphonie IP en raison de son faible débit.

Conclusion

Une description du corps principal de la recommandation G729 pour le codage de signaux vocaux à 8 kbit/s au moyen de l'algorithme CS-ACELP a été présentée dans ce chapitre. Les différentes fonctionnalités de ce codeur ont été détaillées. Elles aideront à mieux comprendre le fonctionnement de ce codeur et la nécessité de l'intégration de la détection de l'activité vocale (VAD) dans les réseaux de communication. Dans le prochain chapitre nous nous focaliserons sur l'annexe B de la recommandation G729 relative à la détection de l'activité vocale. Elle concerne une combinaison d'algorithmes conçus et optimisés pour être employés en association avec le codec G729 afin de diminuer le débit de transmission pendant les pauses de parole.

Chapitre 3 :

La détection d’activité vocale (VAD) de la norme G.729B

Chapitre 3 :

La détection d'activité vocale (VAD) de la norme G.729B

Introduction

Pour des applications multimédia ou nécessitant un faible débit, il est impératif de réduire au maximum le nombre de bits transmis en utilisant des techniques de compression de silence. Lorsque la parole n'est pas présente dans le signal à coder, le débit peut être optimisé pour libérer le canal à d'autres applications, effectuer une transmission simultanée de données ou, dans le cas de transmission sur réseau IP, limiter le nombre de paquets à envoyer [16]. En effet, un pourcentage considérable du temps d'une conversation est rempli de zones de silence ou de pauses. L'utilisation de la détection de silence et de l'insertion de bruit de confort permet d'augmenter l'efficacité du codage.

Le concept de détection de silence et d'insertion de bruit de confort conduit à des techniques bimodales de codage de la parole. Les différents modes du signal d'entrée, désignés par la voix active pour la parole et la voix inactive pour le silence ou le bruit de fond, sont déterminés par un classificateur de signal. Ce classificateur peut fonctionner à l'extérieur du codeur de parole ou en interne. Le codeur de parole plein débit est opérationnel pendant le signal vocal actif, mais un système de codage différent est utilisé pour le signal vocal inactif, utilisant moins de bits et menant à un rapport de compression globale en moyenne plus élevé. Le classificateur est communément appelé un détecteur d'activité vocale (VAD : Voice Activity Detection), et sa sortie est appelée une décision d'activité vocale. La décision d'activité vocale est 1 ou 0, indiquant respectivement la présence ou l'absence d'activité vocale.

L'algorithme VAD et le codeur vocal inactif, ainsi que le codeur de parole G.729 fonctionnent sur des trames de parole numérisées. Pour des raisons de compatibilité, la taille

des trames est la même pour tous les régimes, et aucun retard supplémentaire n'est introduit par l'algorithme VAD ou le codeur vocal inactif. Une représentation schématique d'un système de communication vocale utilisant un VAD et un codeur vocal inactif est présentée dans la Figure 3.1.

Pour chaque trame du signal de parole d'entrée le VAD fournit une décision d'activité vocale, qui est utilisée comme un commutateur entre le codeur vocal actif et inactif. Lorsque le codeur vocal actif est opérationnel, un train de bits de voix active est envoyé au décodeur actif pour chaque trame. Toutefois, pendant les périodes d'inactivité vocale, le codeur vocal inactif peut choisir d'envoyer une information de mise à jour appelée un descripteur d'insertion de silence (SID : Silence Insertion Descriptor) au décodeur inactif, ou de ne rien envoyer. Cette technique est appelée transmission discontinue (DTX). Pour chaque trame, la sortie de chaque décodeur est traitée en tant que signal reconstruit.

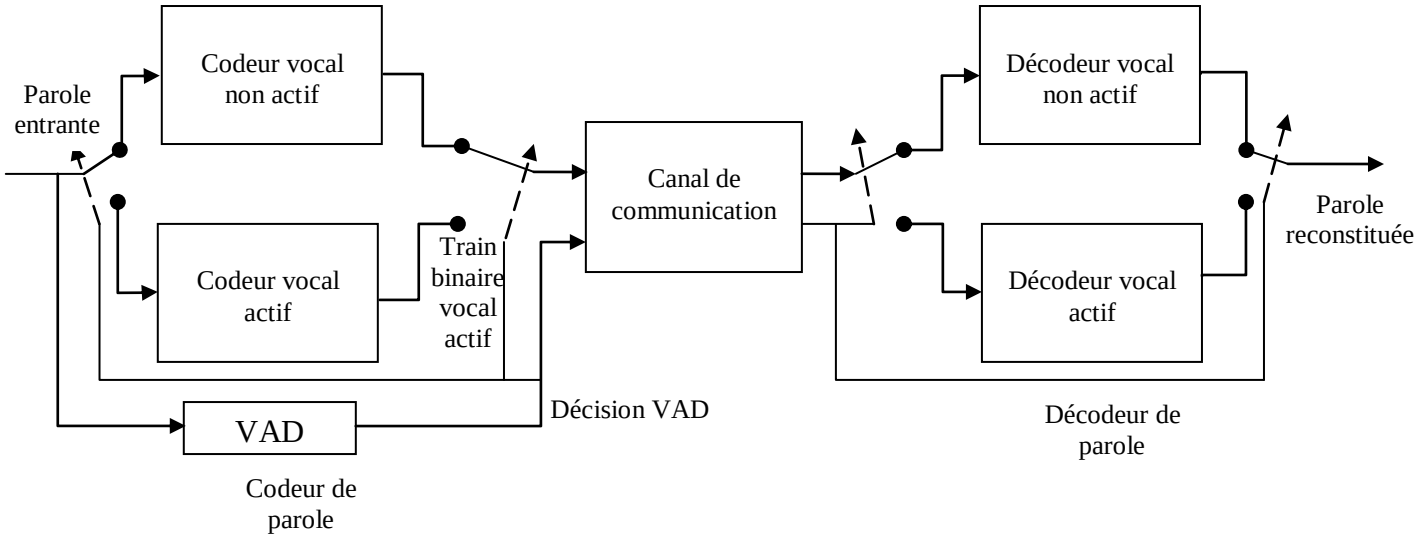

Figure 3.1 : Système de communication de parole avec VAD (Recommandation G.729B ITU)

Dans ce chapitre, nous allons décrire l'algorithme de détection d'activité vocale de l'annexe B de la recommandation G.729 de l'Union Internationale des Télécommunications [17]. Cet algorithme, associé aux procédures de Transmission Discontinue et de génération de bruit de confort (CNG : Comfort Noise Generation), est utilisé pour réduire le débit de transmission pendant les pauses d'une conversation. Ces trois procédures ont été conçues et adaptées pour interagir et être intégrées au processus de codage de la norme G.729, développé dans le chapitre précédent.

3.1. Description générale de l'algorithme VAD

L'organigramme fonctionnel de l'algorithme VAD est représenté par la Figure 3.2. Le rôle de la VAD est de faire la distinction entre la voix active et inactive. Cette classification est un problème bien connu. Cependant, dans la pratique, la diversité et la nature variable de la voix active et du bruit ambiant (voix inactive) augmente la complexité de cette classification [18]. Par exemple, un simple détecteur de niveau d'énergie peut fonctionner de manière satisfaisante dans des conditions de rapport signal sur bruit (SNR) élevé, mais échoue lorsque le SNR chute. Les paramètres utilisés pour la classification doivent être extraits et une fonction discriminante doit être conçue. Un ensemble de paramètres décrivant le contenu énergétique et spectral du signal est sélectionné, comme c'est le cas pour la plupart des applications VAD. Le choix a été dicté par la contribution de chaque paramètre au résultat de la classification, de sa robustesse et de sa complexité de calcul. Le choix final a abouti à l'ensemble des paramètres suivants:

- Les fréquences de raies spectrales (LSF : Line Spectral Frequencies) ;
- L'énergie dans la pleine bande de fréquences ;
- L'énergie dans la bande de fréquences basses ;
- Le taux de passage par zéro.

Un ensemble de paramètres instantanés est calculé pour chaque trame. Une grande partie de la charge de calcul est partagée avec le codeur de parole, évitant ainsi une augmentation supplémentaire de la complexité.

Le bruit de fond peut changer considérablement entre différentes conversations ainsi que durant une même conversation. Ainsi, une estimation des caractéristiques variables du bruit de fond est nécessaire. Un ensemble de paramètres similaires à l'ensemble instantané est utilisé pour décrire les statistiques du bruit. Une mise à jour autorégressive de cet ensemble est basée sur un algorithme VAD simplifié qui utilise un sous ensemble de paramètres d'énergie et de spectre. Ce VAD simplifié n'a pas à répondre aux exigences élevées de la VAD finale, car il est uniquement utilisé pour mettre à jour les paramètres du bruit.

La décision VAD est obtenue en deux étapes. Tout d'abord, une décision initiale est prise sur la base des paramètres de la trame courante. Puis, la décision initiale est lissée, en tenant compte des dernières trames voisines.

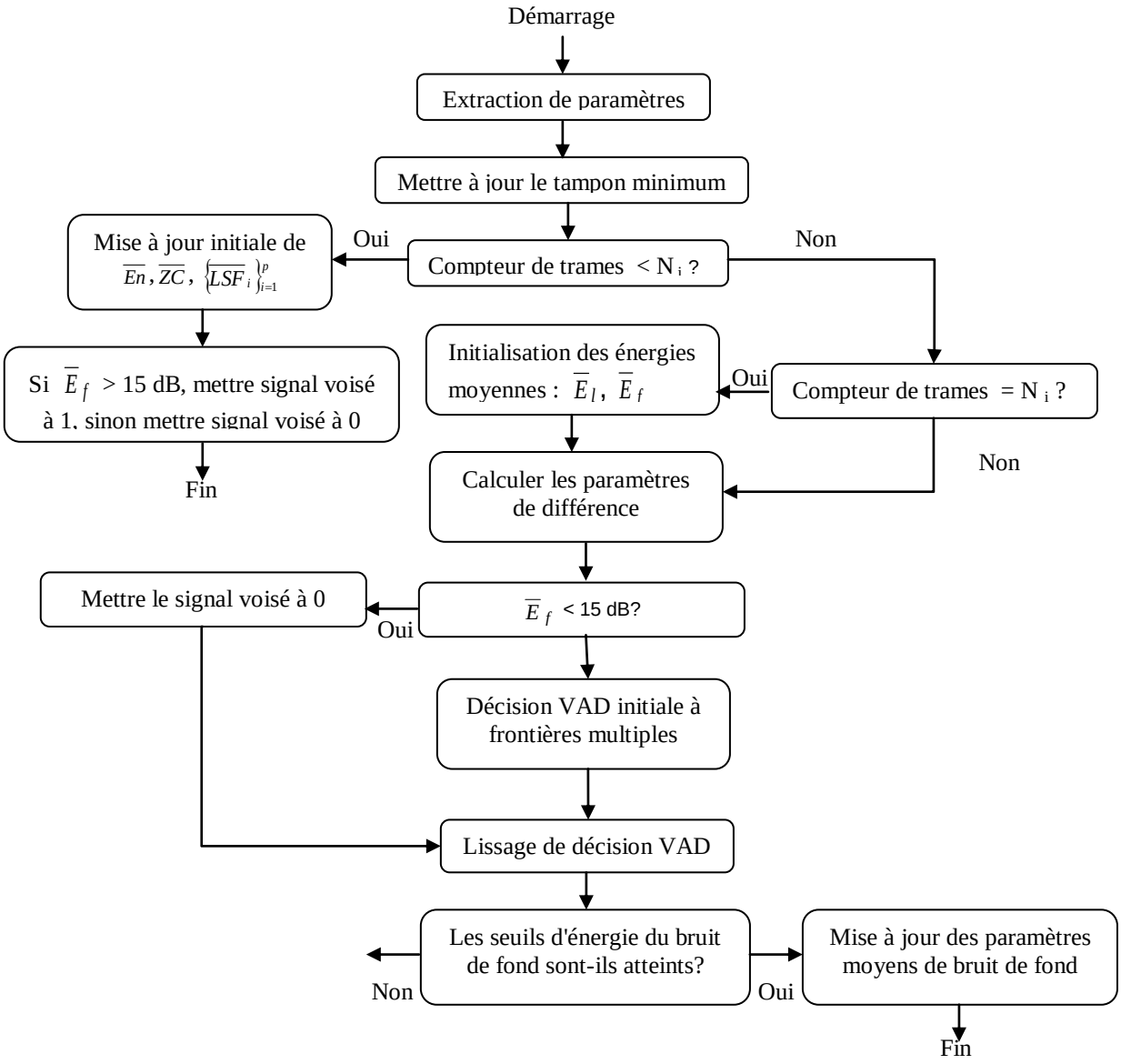

Figure 3.2 : Organigramme fonctionnel du VAD de la recommandation G.729B de l'ITU.

L'ensemble des paramètres utilisés pour prendre une décision VAD initiale est constitué de quatre mesures de différence entre les séries de paramètres instantanés et estimés du bruit. D'autres algorithmes VAD utilisent des approches similaires, où la décision VAD est prise en comparant une fonction de mesures de différence à un seuil global. Toutefois, l'approche de la reconnaissance des formes a été adoptée pour la classification. Les quatre paramètres occupent une région spécifique d'un espace Euclidien à quatre dimensions. Les paramètres de la voix active sont regroupés dans un certain hyper-volume de cet espace, tandis que les paramètres de la voix inactive sont regroupés dans un autre hyper-volume. Ces régions ont été identifiées et une frontière linéaire de décision à trois dimensions est utilisée pour séparer un hyper-volume de l'autre, générant la décision VAD initiale.

La décision initiale est locale, c.-à-d. qu'elle ne prend pas en compte la nature stationnaire à court terme à la fois de la parole et du bruit. Un lissage en quatre étapes est employé en utilisant les dernières trames voisines. Un mécanisme de réinitialisation a été conçu pour empêcher le blocage de l'algorithme si le niveau du bruit change brusquement, en utilisant une estimation de l'énergie minimale sur une longue période de temps.

3.2. Description fonctionnelle de l'algorithme VAD

Dans cette partie, nous allons détailler les différentes fonctions de l'algorithme VAD qui sont représentées par les blocs de la Figure 3.2.

3.2.1. Extraction de paramètres

Un ensemble de paramètres est extrait du signal d'entrée à chaque trame. Le module d'extraction de paramètres est partagé entre le VAD, le codeur vocal actif, et le codeur vocal inactif pour améliorer l'efficacité des calculs. L'ensemble de base des paramètres est l'ensemble des coefficients d'autocorrélation, qui est noté par $\{R(i)\}_{i=0}^{q}$, avec $q = 12$. Perceptiblement, seul 11 coefficients d'autocorrélation sont nécessaires pour le codeur plein débit G.729, deux coefficients d'autocorrélation supplémentaires doivent être calculés pour l'algorithme VAD afin d'obtenir une meilleure estimation de l'énergie dans la bande de fréquences basses [17].

a. Fréquences de raies spectrales (LSF)

Un ensemble de coefficients de prédiction linéaire est dérivé des 11 premiers termes d'autocorrélation en utilisant les procédures du G.729. Ces coefficients sont convertis en un ensemble de $\{LSF_i\}_{i=1}^{p}$, où $p = 10$. (Le second coefficient de réflexion est également calculé à cette étape).

b. Énergie dans la pleine bande de fréquences

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

$$E_f = 10 \cdot \log_{10}\left[\frac{1}{N} R(0)\right] \tag{3.1}$$

où $N = 240$ est la taille de la fenêtre d'analyse LP dans des échantillons de parole.

c. Énergie dans la bande de fréquences basses

L'énergie dans la bande de fréquences basses E_l, mesurée sur la bande 0-1 kHz, est calculée comme suit:

$$E_l = 10 \cdot \log_{10}\left[\frac{1}{N}\mathbf{h}^{\mathbf{T}}\mathbf{R}\mathbf{h}\right] \tag{3.2}$$

où h est la réponse impulsionnelle d'un filtre FIR dont la fréquence de coupure est de 1 kHz, R est la matrice d'autocorrélation de Toeplitz de taille 13 x 13 avec les coefficients d'autocorrélation sur chaque diagonale.

d. Taux de passage par zéro

Le taux de passage par zéro normalisé pour chaque trame est calculé par :

$$ZC = \frac{1}{2M}\sum_{i=1}^{M}\left[\left|\mathrm{sgn}\left[x(i)\right] - \mathrm{sgn}\left[x(i-1)\right]\right|\right] \tag{3.3}$$

où *{x(i)}* est le signal d'entrée prétraité, et $M = 80$ est la taille d'une trame.

3.2.2. Initialisation des moyennes courantes des caractéristiques de bruit de fond

En ce qui concerne les premières trames N_i, les moyennes des paramètres spectraux du bruit de fond, indiqués par $\left\{\overline{LSF}_i\right\}_{i=1}^{p}$ sont initialisés à la moyenne $\left\{LSF_i\right\}_{i=1}^{p}$ des trames. La moyenne des passages par zéro du bruit de fond, indiquée par $\overline{ZC}$, est initialisée à la moyenne ZC du nombre de passage par zéro des trames.

Les moyennes courantes de l'énergie de bruit de fond, désignée par $\overline{E}_f$, et l'énergie dans la bande de fréquences basses du bruit de fond, désignée par $\overline{E}_l$, sont initialisées de la façon suivante :

Tout d'abord, la procédure d'initialisation utilise $\overline{En}$, définie comme moyenne de l'énergie de trame E_f sur les premières trames N_i. Ces trois opérations de moyenne ($\overline{En}$, $\overline{ZC}$, et $\left\{\overline{LSF}_i\right\}_{i=1}^{p}$) n'incluent que les trames qui ont une énergie *E* supérieure à 15 dB. En second lieu, la procédure d'initialisation se poursuit comme suit:

si $\overline{En} \leq T_1$ alors

$$\overline{E}_f = \overline{En} + K_0$$
$$\overline{E}_l = \overline{En} + K_1$$

sinon, si $T_1 < \overline{En} < T_2$ alors

$$\overline{E}_f = \overline{En} + K_2$$
$$\overline{E}_l = \overline{En} + K_3$$

sinon

$$\overline{E}_f = \overline{En} + K_4$$
$$\overline{E}_l = \overline{En} + K_5 \qquad (3.4)$$

Voir le Tableau 3.1 pour les valeurs des constantes.

Tableau 3.1 : Tableau des constantes [17]

Nom	Constante	Nom	Constante
N_i	32	N_1	4
N_0	128	N_2	10
K_0	0	T_1	671088640
K_1	–53687091	T_2	738197504
K_2	–67108864	T_3	26843546
K_3	–93952410	T_4	40265318
K_4	–134217728	T_5	40265318
K_5	–161061274	T_6	40265318
a_1	23488	b_1	28521
a_2	–30504	b_2	19446
a_3	–32768	b_3	–32768
a_4	26214	b_4	–19661
a_5	0	b_5	–30802
a_6	28160	b_6	–19661
a_7	0	b_7	30199
a_8	16384	b_8	–22938
a_9	–19065	b_9	–31576
a_{10}	0	b_{10}	–17367
a_{11}	22400	b_{11}	–27034
a_{12}	30427	b_{12}	29959
a_{13}	–24576	b_{13}	–29491
a_{14}	23406	b_{14}	–28087

La décision de détection d'activité vocale est forcée à 1 si l'énergie de trame obtenue à partir de l'analyse LPC est supérieure à 15 dB. Sinon, la décision de détection d'activité vocale est forcée à 0.

Quand le numéro de trame est égal à *Ni* une étape d'initialisation pour les énergies caractéristiques du bruit de fond intervient.

3.2.3. Calcul de l'énergie minimale à long terme

Le paramètre d'énergie minimale à long terme $E_{\min}$ est calculé comme le minimum de E_f sur N_0 trames antérieures. Étant donné que la valeur N_0 est relativement élevée, $E_{\min}$ est calculé en utilisant des valeurs enregistrées du minimum E_f sur des segments antérieurs courts.

3.2.4. Calcul des paramètres de différence

Les paramètres instantanés doivent être comparés avec les paramètres du bruit de fond. La comparaison est effectuée sur quatre mesures de différence, qui sont calculés à partir des paramètres de la trame courante et des moyennes courantes des caractéristiques du bruit de fond.

a. La distorsion spectrale ΔS

La mesure de la distorsion spectrale se calcule par la somme des carrés de la différence entre le vecteur $\{LSF_i\}_{i=1}^{p}$ de la trame courante et les moyennes courantes du bruit de fond $\{\overline{LSF}_i\}_{i=1}^{p}$:

$$\Delta S = \sum_{i=1}^{p} \left(LSFi - \overline{LSF}_i\right)^2 \tag{3.5}$$

b. Différence d'énergie dans la pleine bande de fréquences ΔE_f

La mesure de la différence d'énergie dans la pleine bande de fréquences se calcule par la différence entre l'énergie de la trame courante E_f, et la moyenne courante de l'énergie du bruit de fond, $\overline{E}_f$:

$$\Delta E_f = \overline{E}_f - E_f \tag{3.6}$$

c. Différence d'énergie dans la bande de fréquences basses ΔE_l

La mesure de la différence d'énergie dans la bande de fréquences basse se calcule par la différence entre l'énergie dans la bande de fréquences basses de la trame courante E_l, et la moyenne courante de l'énergie dans la bande de fréquences basses du bruit de fond, $\overline{E}_l$:

$$\Delta E_l = \overline{E}_l - E_l \tag{3.7}$$

d. Différence de passages par zéro ΔZC

La mesure de la différence de passage par zéro se calcule par la différence entre le taux de passage par zéro de la trame courante *Zc*, et la moyenne courante du taux de passe par zéro du bruit de fond $\overline{ZC}$:

$$\Delta ZC = \overline{ZC} - ZC \tag{3.8}$$

3.2.5. Décision initiale de détection d'activité vocale à frontières multiples

Les quatre paramètres de différence résident dans une région de l'espace Euclidien quadridimensionnel. Chaque vecteur possible des paramètres de différence définit un point dans cet espace. Les points générés par les trames vocales actives sont regroupés dans une certaine région (hyper-volume) de l'espace, tandis que les points générés par les trames vocales inactives sont regroupés dans une autre région. Ces hyper-volumes ont été identifiés et séparés à l'aide de quatorze hyper plan définis dans des espaces tridimensionnels qui constituent donc les frontières pour la décision initiale. Pour chaque trame, la décision VAD initiale correspond à la région où le vecteur des quatre paramètres de différence réside.

La prise de décision initiale sur l'activité vocale est indiquée par I_{VD}. Les quatorze frontières de décision dans l'espace quadridimensionnel sont définies comme suit:

1) $si\ \Delta S > a_1 \cdot \Delta ZC + b_1\ alors\ I_{VD} = 1$
2) $si\ \Delta S > a_2 \cdot \Delta ZC + b_2\ alors\ I_{VD} = 1$
3) $si\ \Delta E_f < a_3 \cdot \Delta ZC + b_3\ alors\ I_{VD} = 1$
4) $si\ \Delta E_f < a_4 \cdot \Delta ZC + b_4\ alors\ I_{VD} = 1$
5) $si\ \Delta E_f < b_5\ alors\ I_{VD} = 1$
6) $si\ \Delta E_f < a_6 \cdot \Delta S + b_6\ alors\ I_{VD} = 1$
7) $si\ \Delta S > b_7\ alors\ I_{VD} = 1$

8) $si\ \Delta E_l < a_8 \cdot \Delta ZC + b_8\ alors\ I_{VD} = 1$

9) $si\ \Delta E_l < a_9 \cdot \Delta ZC + b_9\ alors\ I_{VD} = 1$

10) $si\ \Delta E_l < b_{10}\ alors\ I_{VD} = 1$

11) $si\ \Delta E_l < a_{11} \cdot \Delta S + b_{11}\ alors\ I_{VD} = 1$

12) $si\ \Delta E_l > a_{12} \cdot \Delta E_f + b_{12}\ alors\ I_{VD} = 1$

13) $si\ \Delta E_l < a_{13} \cdot \Delta E_f + b_{13}\ alors\ I_{VD} = 1$

14) $si\ \Delta E_l < a_{14} \cdot \Delta E_f + b_{14}\ alors\ I_{VD} = 1$ (3.9)

Si aucun des quatorze états n'est "VRAI", $I_{VD} = 0$.

Voir le Tableau 3.1 pour les valeurs constantes.

3.2.6. Lissage de la décision de détection d'activité vocale

La décision initiale de détection d'activité vocale est lissée (temps de maintien) pour refléter la nature stationnaire du signal de parole et du bruit de fond [18]. La prise en compte de l'énergie, de même que les décisions antérieures sur les trames voisines, sont utilisées pour le lissage de décision.

Le lissage s'effectue en quatre étapes dérivées d'une observation approfondie d'une large base de données. Un fanion indiquant que le temps de maintien est intervenu est défini par v_flag . Il est mis chaque fois à zéro avant que le lissage de la décision de détection d'activité vocale ne soit effectué. On indiquera la décision d'activité vocale lissée de la trame courante, de la trame précédente et de la trame antérieure par respectivement S_{VD}^0, S_{VD}^{-1} et S_{VD}^{-2}. Les paramètres S_{VD}^{-1} et S_{VD}^{-2} sont initialisés à 1. Pour commencer, $S_{VD}^0 = I_{VD}$.

Dans la première étape de lissage, une décision d'activité vocale est étendue à la trame courante si l'énergie de la trame courante est au-dessus d'un certain seuil.

$$\text{si } \left(I_{VD} = 0\right) \text{ et } \left(S_{VD}^{-1} = 1\right) \text{ et } \left(E > \overline{E}_f + T3\right) \text{ alors } S_{VD}^0 = 1 \text{ et } v\_flag = 1 \quad (3.10)$$

Dans la seconde étape de lissage, une décision d'activité vocale est étendue à la trame actuelle, si les deux trames précédentes étaient des trames vocales actives, et la valeur absolue de la différence d'énergie entre la trame actuelle et précédente est au-dessous d'un certain seuil.

Ainsi pour cette seconde étape de lissage, on définit un paramètre booléen F_{VD}^{-1} et un compteur de lissage C_e.

Le paramètre F_{VD}^{-1} est initialisé à 1 et C_e est initialisé à 0. On indiquera l'énergie de la trame précédente par E_{-1}.

$$
\begin{aligned}
&\text{si}\left(F_{VD}^{-1}=1\right)\text{et}\left(I_{VD}=0\right)\text{et}\left(S_{VD}^{-1}=1\right)\text{et}\left(S_{VD}^{-2}=1\right)\text{et}\left(\left|E_f-E_{-1}\right|\le T_4\right)\{\\
&\qquad S_{VD}^{0}=1\\
&\qquad v\_flag=1\\
&\qquad C_e=C_e+1\\
&\qquad \text{si}\left(C_e\le N_1\right)\{\\
&\qquad\qquad F_{VD}^{-1}=1\\
&\qquad \}\\
&\qquad \text{sinon}\{\\
&\qquad\qquad F_{VD}^{-1}=0\\
&\qquad\qquad C_e=0\\
&\qquad \}\\
&\}\\
&\text{sinon}\\
&F_{VD}^{-1}=1
\end{aligned}
\tag{3.11}
$$

Dans la troisième étape de lissage, une décision d'inactivité vocale est prolongée à la trame courante si les dix trames précédentes sont des trames vocales inactives, et la différence entre l'énergie de la trame actuelle et les dix trames antérieures est au dessous d'un certain seuil.

Ainsi pour cette troisième étape de lissage, on définit un compteur de continuité de bruit C_s qui est initialisé à 0.

$$\text{si}\left(S_{VD}^{0}=0\right)C_s = C_s + 1$$

$$\begin{aligned}&\text{si}\left(S_{VD}^{0}=1\right)\text{et}\left(C_s > N_2\right)\text{et}\left(E_f - E_{-1} \le T_5\right)\{\\ &S_{VD}^{0}=0\\ &C_s = 0\\ &\}\end{aligned}$$

$$\text{si}\left(S_{VD}^{0}=1\right)C_s = 0 \qquad (3.12)$$

Dans la dernière étape, la décision d'activité vocale de la trame est corrigée à une décision d'inactivité vocale si l'énergie de la trame actuelle est au-dessous de l'énergie du bruit avec un certain écart, le second coefficient de réflexion est plus petit qu'un certain seuil, et aucune des deux premières étapes de lissage n'a été exécutée.

$$\text{si}\left(\left(E_f < \overline{E}_f + T_6\right)\text{et}\left(frm\_count > N_0\right)\text{et}\left(v\_flag = 0\right)et(\text{rc} < |b_6|)\right)\text{alors}\, S_{VD}^{0} = 0. \qquad (3.13)$$

3.2.7. Mise à jour des moyennes courantes des caractéristiques du bruit de fond

Les moyennes courantes des caractéristiques du bruit de fond sont mises à jour pendant la dernière étape du module VAD. Les moyennes courantes doivent être mises à jour uniquement en présence de bruit de fond et non en présence de parole. Pour cela on utilise une version simplifié de l'algorithme VAD qui utilise uniquement l'énergie, la distorsion spectrale, et le second coefficient de réflexion. 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 tend à perdre 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 classification erronée.

Si la décision de l'algorithme VAD simplifié indique le silence ou le bruit de fond, l'ensemble des paramètres caractéristiques du bruit, est mis à jour, en utilisant un algorithme autorégressif (AR) du premier ordre.

Si ($(E_f < \overline{E}_f + 614)$ et $(rc < |b_6|)$ et $(\Delta S < 83)$) (3.14)

alors mettre à jour les caractéristiques du bruit de fond.

Différents coefficients AR ont été fixés pour chaque paramètre, et une adaptation plus rapide est effectuée au début de la conversation ou de l'enregistrement et après une réinitialisation.

Soit β_{E_f} le coefficient AR pour la mise à jour de $\overline{E}_f$, soit β_{E_l} le coefficient AR pour la mise à jour de $\overline{E}_l$, soit β_{ZC} le coefficient AR pour la mise à jour de $\overline{ZC}$ et soit β_{LSF} le coefficient AR pour la mise à jour de $\left\{\overline{LSF}_i\right\}_{i=1}^{p}$. Le nombre total de trames pour lesquelles la condition de mise à jour a été satisfaite est compté par C_n. Un ensemble différent de coefficients β_{E_f}, β_{E_l}, β_{ZC}, et β_{LSF} est utilisé selon la valeur de C_n.

La mise à jour AR est effectuée selon:

$$\begin{aligned}
\overline{E}_f &= \beta_{E_f} \cdot \overline{E}_f + \left(1-\beta_{E_f}\right) \cdot E_f \\
\overline{E}_l &= \beta_{E_l} \cdot \overline{E}_l + \left(1-\beta_{E_l}\right) \cdot E_l \\
\overline{ZC} &= \beta_{ZC} \cdot \overline{ZC} + \left(1-\beta_{ZC}\right) \cdot ZC \\
\overline{LSF}_i &= \beta_{LSF} \cdot \overline{LSF}_i + \left(1-\beta_{LSF}\right) \cdot LSF_i \quad i = 1,\ldots,p
\end{aligned} \tag{3.15}$$

Dans le cas d'une augmentation brutale du niveau de bruit, l'algorithme peut se bloquer en mode ‘’activité vocale‘’ sans aucune autre mise à jour des caractéristiques du bruit. Un mécanisme de remise à zéro, qui réinitialise le niveau d'énergie du bruit ambiant avec la valeur de E_{min}, empêche un tel blocage.

$\overline{E}_f$ et C_n sont en outre mis à jour selon la relation suivante :

$$\begin{aligned}
&\text{si } \left(\text{comptage de trame} > N_0\right) \text{ et } \left(\overline{E}_f < E_{\min}\right)\{ \\
&\quad \overline{E}_f = E_{\min} \\
&\quad C_n = 0 \\
&\}
\end{aligned} \tag{3.16}$$

3.3. Description des algorithmes DTX/CNG

Les algorithmes DTX/CNG décrits dans l'annexe B de la recommandation G.729 sont utilisés pour coder et décoder les trames vocales non actives [17]. Les codeurs et décodeurs

classiques de parole utilisent le bruit de confort pour simuler le bruit de fond dans les trames vocales non actives. Si le bruit de fond n'est pas stationnaire, une simple insertion du bruit de confort n'apporte pas le naturel du bruit de fond initial. Il est donc souhaitable d'envoyer de façon intermittente des informations relatives au bruit de fond afin d'obtenir une meilleure qualité quand des trames vocales non actives sont détectées. L'efficacité de codage des trames vocales non actives peut être obtenue par le codage de l'énergie de la trame et de son spectre avec un nombre de bits aussi réduit que quinze bits. Ces bits ne sont pas transmis automatiquement toutes les fois qu'une détection vocale non active intervient. Ils sont plutôt transmis uniquement quand un changement notable a été détecté par rapport à la dernière trame vocale non active qui a été transmise. Du côté du décodeur, le module CNG est appelé pour reproduire les trames vocales non actives. Le schéma fonctionnel d'un système de communication avec codage des périodes de silence est représenté dans la Figure 3.3.

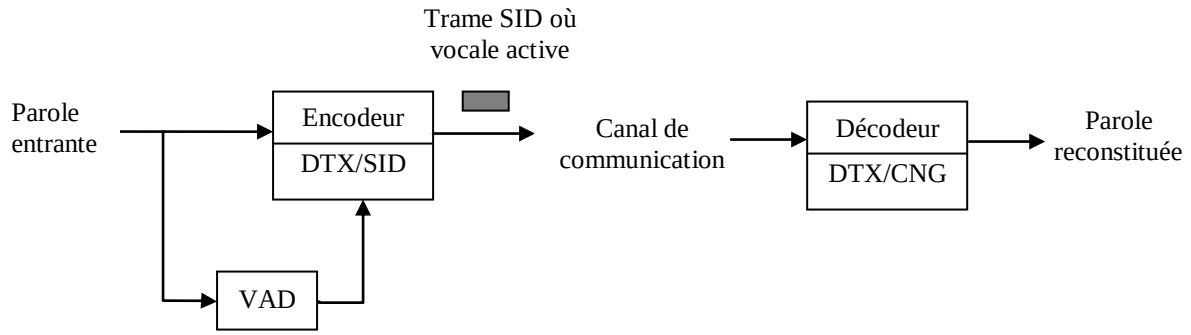

Figure 3.3 : Système de communication de parole avec codage des périodes de silence [19].

3.3.1. Transmission discontinue (DTX)

L'algorithme DTX détermine, pour chaque trame vocale inactive, le besoin d'envoyer un ensemble de paramètres de mise à jour de signaux vocaux non actifs vers le décodeur de parole [18]. Lors d'une transition de trame vocale active à une trame vocale inactive, une trame SID est toujours transmise, initialisant les paramètres CNG. Pour les trames suivantes, le module DTX mesure les changements spectraux et énergiques dans les caractéristiques du bruit de fond depuis la dernière trame SID transmise pour prendre la décision de transmission. Des seuils absolus et adaptatifs de l'énergie de trame et de la mesure de distorsion spectrale sont utilisés pour obtenir la décision de mise à jour. Si une mise à jour est nécessaire, le codeur vocal non actif envoie les informations nécessaires pour générer un signal dont la perception est similaire au signal vocal non actif initial. Ces informations comprennent un niveau d'énergie et une description de l'enveloppe spectrale. Si aucune mise à jour n'est

nécessaire, le signal vocal non actif est généré par le décodeur non actif selon les informations reçues relatives à l'énergie et à la forme spectrale de la dernière trame vocale non active.

Un intervalle minimal de N_{min} = 2 trames est cependant nécessaire entre deux trames SID consécutives, c'est-à-dire que si un changement de niveau ou de forme de spectre se produit avec $n < N_{min}$ trames après une trame SID, l'émission SID est retardée.

Le module DTX reçoit les informations vocales actives et non actives du module VAD, et en provenance des modules de codage, la fonction d'autocorrélation du signal de parole calculée pour chaque trame, de même que l'échantillon d'excitation antérieur. Pour chaque trame, la décision DTX $Ftyp_t$ (type de trame pour la trame numérotée t), est une valeur de sortie prise parmi l'une des trois valeurs 0, 1 ou 2 correspondant respectivement à la trame non transmise, à la trame de parole active, ou à la trame SID.

3.3.2. Structure d'une trame SID

Lorsque la décision de transmettre des informations SID est validée, le codeur de parole inactive évalue les paramètres représentant le spectre et le niveau d'énergie du bruit de fond, les quantifie sur 15 bits et les transmet. Le Tableau 3.2 détaille le train binaire relatif à la transmission d'une trame SID. Les coefficients de lignes de raies spectrales (LSP) sont codés par une quantification vectorielle à deux étapes sur 10 bits et le terme d'énergie suit une quantification scalaire sur 5 bits.

Tableau 3.2 : Index paramétriques transmis pour une trame SID

Description des paramètres	Bits
Indice de prédiction pour la quantification des LSP	1
Vecteur pour la première étape de quantification des LSP	5
Vecteur pour la seconde étape de quantification des LSP	4
Gains des énergies	5

3.3.3. Génération de bruit de confort (CNG)

Le bruit de confort est généré en introduisant un signal d'excitation pseudo-blanc de niveau contrôlé dans des filtres LP interpolés, de la même manière que le décodeur produit de la parole active en filtrant l'excitation décodée. Les coefficients LP sont obtenus à partir des coefficients LSP de la dernière trame SID décodée. L'interpolation avec les coefficients LSP de la trame précédente est effectuée comme pour les trames actives.

L'excitation pseudo-blanche *ex(n)* est un mélange entre une excitation du même type que celle de la parole active *ex*1(*n*) et une excitation gaussienne blanche *ex*2 (*n*). L'excitation *ex*1(*n*) selon le G.729 est composée d'une excitation adaptative avec un petit gain et d'une excitation fixe ACELP, qui améliore la transition entre les trames vocales actives et non actives. L'ajout d'une excitation gaussienne *ex*2 (*n*) permet de générer un signal plus blanc.

3.3.4. Dissimulation de l'effacement de trame

Dans de mauvaises conditions de transmission, un système utilisant l'algorithme de transmission discontinue décrit précédemment tend à être moins sensible aux erreurs de transmission qu'un système où toutes les trames sont actives, à condition que le décodeur soit informé quand une trame a été effacée.

Dans ce cas, l'effacement de trames non transmises n'affecte pas le décodeur. En outre, l'effacement de trames SID n'entraine pas une dégradation sévère quand ceci se produit à l'intérieur d'une période vocale inactive dès lors que les paramètres des trames SID précédentes remplacent les paramètres manquants [17]. L'effacement de la première trame SID d'une période vocale inactive pourrait causer une certaine dégradation, puisque les trames vocales inactives ultérieures ont besoin de paramètres pour générer le bruit de confort. Pour remédier à cela, un ensemble de paramètres LP et un niveau d'énergie d'excitation sont stockés durant chaque trame vocale active par le décodeur vocal actif. Ces paramètres de sécurité sont utilisés si l'effacement de la première trame SID d'une période inactive est détecté.

Conclusion

Ce troisième chapitre, inspiré de l'annexe B de la recommandation G.729 de l'ITU décrit un schéma de compression des silences qui comprend un module de détection d'activité vocale (VAD), un module de transmission discontinue (DTX), et un module de génération de bruit de confort (CNG). Ces trois procédures sont utilisées pour réduire le débit de transmission pendant les pauses d'une conversation.

Après avoir présenté l'architecture et le fonctionnement de la détection d'activité vocale dans les réseaux de communications utilisant le codec G.729, nous nous attèlerons dans le prochain chapitre à évaluer la robustesse de cet algorithme VAD dans des conditions acoustiques adverses et à proposer quelques améliorations dans sa prise de décision.

Chapitre 4 :

Evaluation et amélioration du module VAD de la norme G.729B en milieu bruité

Chapitre 4 :

Evaluation et amélioration du module VAD de la norme G.729B en milieu bruité

Introduction

La détection d'activité vocale permet de faire la distinction entre les trames vocales actives et les trames vocales inactives. Il est plutôt simple d'obtenir une performance élevée de détection parole / silence dans un milieu isolé. Toutefois, dans des environnements réels, le signal d'entrée est généralement mélangé avec les caractéristiques du bruit ambiant qui peuvent être inconnus et variables dans le temps. Dans le cas où le bruit de fond est élevé, la parole peut être noyée dans du bruit. Particulièrement les sons non voisés, qui sont importants pour l'intelligibilité de la parole et qui peuvent être mal détectés dans de tels environnements bruyants [21].

Une classification erronée des régions vocales actives peut produire des sons saccadés conduisant à une dégradation considérable de la qualité vocale. D'autre part, l'augmentation d'une mauvaise classification des périodes d'inactivité vocale fait perdre les avantages de la compression des silences. Ainsi, les performances du module VAD sont tributaires d'une maximisation du taux de détection de la parole active, accompagnée d'une minimisant du taux de classement erroné des régions vocales inactives.

Sachant que les bruits environnants perturbent la parole, et peuvent mettre à défaut la discrimination silence / parole, nous nous sommes intéressés à l'évaluation de la robustesse de l'algorithme VAD décrit dans l'annexe B de la norme G.729 dans des conditions acoustiques adverses. Un protocole expérimental a été suivi à cet effet.

4.1. Le Protocole Expérimental

Dans cette partie, nous détaillerons la méthodologie suivie pour l'élaboration des ressources nécessaires à la réalisation de notre travail. Ensuite, nous décrirons les étapes nécessaires pour la mise en œuvre du codec G.729B de l'ITU, puis nous présenterons les différentes études comparatives effectuées.

Une série d'expériences a été menée pour évaluer l'efficacité du module VAD de la norme G.729B de l'ITU. La procédure expérimentale que nous avons mise en place consiste en une comparaison des performances du module VAD dans différents milieux bruités.

4.1.1. Elaboration d'une base de données parlée bruitée

Le premier problème auquel nous étions confrontés au début de ce travail, était l'indisponibilité de bases de données sonores bruitées pour réaliser l'évaluation du codec G.729B.

L'élaboration, à partir de la base ARADIGIT, et des vecteurs de test joints à ce codec [17], d'une base de données parlée bruitée avec la base NOISEX-92 a été entreprise. Cinq situations environnementales adverses ont été retenues : bruit de chahut dans une cantine, bruit dans une usine de production de véhicule, bruit dans le cockpit d'un avion de combat en vol, bruit dans la cabine d'un camion militaire en marche, et enfin le bruit HF dans une chaine radio après démodulation.

4.1.2. Vecteurs de test de l'ITU

L'annexe B de la recommandation G.729 de l'ITU fournit un ensemble de vecteurs de test. Ces vecteurs sont un outil qui peut donner une indication du succès de la mise en œuvre du codec. L'usage de l'ensemble des vecteurs de tests est présenté sur le Tableau 4.1.

Tableau 4.1 : Les vecteurs de test de l'ITU-T G.729B :

Entrée codeur	Sortie codeur Entrée décodeur	Sortie décodeur
tstseq1.bin	tstseq1.bit	tstseq1.out
tstseq2.bin	tstseq2.bit	tstseq2.out
tstseq3.bin	tstseq3.bit	tstseq3.out
tstseq4.bin	tstseq4.bit	tstseq4.out
	tstseq5.bit	tstseq5.out
	tstseq6.bit	tstseq6.out

Les vecteurs d'entrée et de sortie du codec correspondent à des fichiers de données échantillonnées contenant des signaux MIC à mots de 16 bits, alors que les vecteurs intermédiaires correspondent à un flux de données codées.

4.1.3. Mise en œuvre du codec G.729B

Le code source qui est une partie intégrante de l'annexe B de la recommandation G.729 de l'ITU, reflète la description binaire exacte en arithmétique à virgule fixe et mots de 16 bits de l'algorithme de codage des périodes de silence.

La compilation de ce code qui simule le codec a été réalisée sur une plate forme cible de mots de 32 bits. Une vérification du succès de la mise en œuvre s'en est suivie en utilisant les vecteurs de test prévus à cet effet.

L'exécution du codeur peut se faire sous deux modes : un mode avec le VAD activé, et un autre avec le VAD désactivé (ce qui revient à se limiter à la recommandation G.729 de l'ITU**)**.

Après le traitement du signal d'entrée par le codeur, un flux binaire codé est récupéré dans un fichier tampon. Un point de sortie a été inséré dans le codeur pour récupérer les prises de décisions du module VAD dans un fichier de données. Le flux binaire codé est passé à travers le décodeur. Le signal d'entrée synthétisé par le décodeur est récupéré dans un fichier de sortie.

Pour pouvoir traiter les enregistrements de la base de données ARADIGIT à travers ce codec une normalisation est requise :

- Un sous-échantillonnage à la fréquence de 8 kHz.
- Une conversion des fichiers wav en bin.

4.2. Résultats Expérimentaux

Dans cette section, nous allons présenter les résultats expérimentaux obtenus avec le codec G.729B mis en œuvre. Les résultats ont été obtenus avec les expériences suivantes :

- En milieu calme
- En environnement bruité (5 conditions de bruits)

Pour chaque type d'expérience un ensemble de courbes qui mettent en avant les paramètres sur lesquels s'est appuyé le module VAD pour sa prise de décision

(spectrogramme, courbe d'énergie, courbe de taux de passage par zéro), ainsi que la courbe VAD, la forme d'onde du signal d'entrée et celle du signal synthétisé, ont été extraites.

Les études comparatives seront suivies de discussions sur l'efficacité du codec et sa robustesse aux bruits.

4.2.1. Evaluation du module VAD du codec G.729B en milieu calme

Dans un premier temps, nous allons évaluer la prise de décision du module VAD dans un milieu non bruité. L'un des vecteurs de test joint à la recommandation et deux enregistrements de la base de données ARADIGIT seront utilisés à cet effet.

a. Cas du vecteur de test

Le vecteur de test non bruité tstseq1.bin fourni par l'ITU a été choisi pour notre étude comparative. Les Figures 4.1 - 4.4 fournissent les éléments d'analyse.

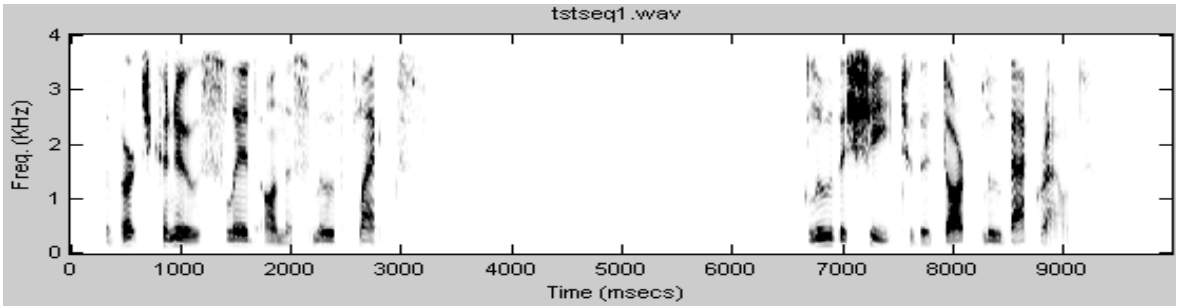

Figure 4.1 : Spectrogramme du vecteur de test "tstseq1.wav"

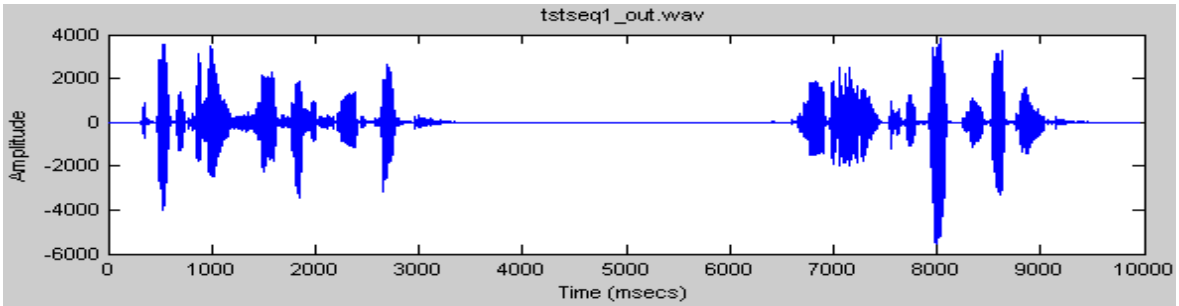

Figure 4.2 : Forme d'onde du vecteur de test "tstseq1.wav" synthétisé.

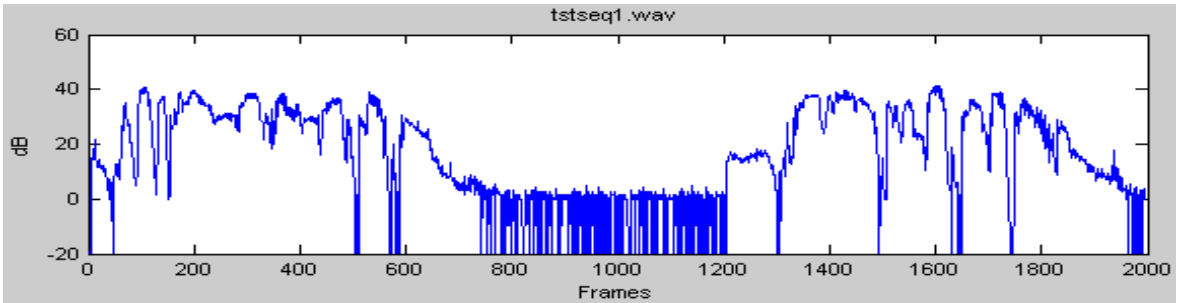

Figure 4.3 : Courbe d'énergie du vecteur de test "tstseq1.wav"

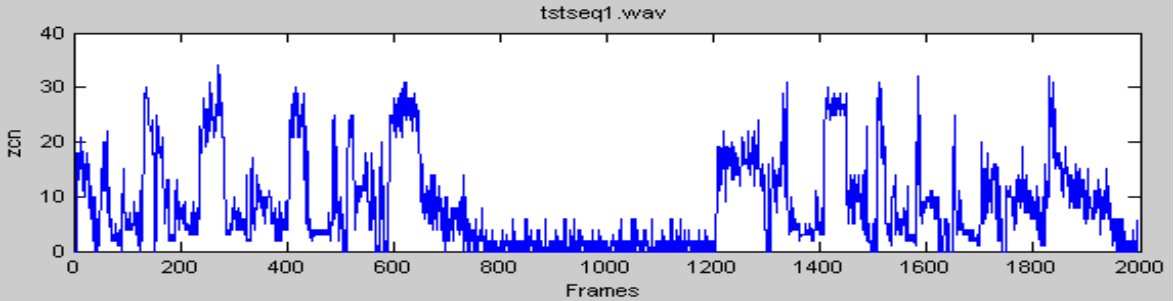

Figure 4.4 : Courbe de TPZ du vecteur de test "tstseq1.wav"

En observant la forme temporelle du signal vocal (Figure 4.5) et la courbe de détection d'activité vocale (Figure 4.6) ainsi que celle du signal vocal synthétisé (Figure 4.2) à la sortie du décodeur, on remarque que le module VAD du schéma de compression des silences de la norme G.729B donne de bonnes décisions en milieu calme (Figure 4.6), néanmoins la présence de segments d'amplitude nulle donne une allure hachée à la courbe VAD.

Amélioration de la VAD par temps de maintien

L'observation de la Figure 4.6 représentant la courbe VAD du vecteur de test tstseq1.bin, nous amène à proposer une méthode d'amélioration de la détection d'activité vocale. Ainsi, un lissage plus maintenu de cette décision peu être envisagé afin d'éviter que la décision oscille entre la parole et le bruit, et obtenir ainsi un signal de sortie plus cohérent. Une substitution de la méthode de lissage actuelle par un simple retard qui correspond à N décisions BRUIT détectés consécutivement a été réalisée. Ce retard est introduit avant le passage du mode PAROLE au mode BRUIT.

/ / Initialisation du compteur de continuité de bruit dans
/ / le cas d'une décision PAROLE.
Si $(I_{VD} = 1)$ alors $C_s = 0$
/ / Si on est en présence d'une décision BRUIT avec
/ / un compteur de continuité de bruit inférieur à un certain seuil ;
Si ($(I_{VD} = 0)$ et (C_s < N)) {
/ / Incrémentation du compteur de continuité de bruit,
Count ++ ;
/ / Maintien de la décision PAROLE.
$S_{VD}^{0} = 1$;
} (4.1)

Pour rappel, I_{VD} correspond à la décision initiale de détection d'activité vocale, S_{VD}^{0} à la décision lissée de la trame actuelle et C_s au compteur de continuité de bruit.

La Figure 4.7, qui est à comparer avec la Figure 4.6 montre clairement la bonne séparation silence / parole pour le vecteur de test testseq1.wav avec la solution que nous avons proposée.

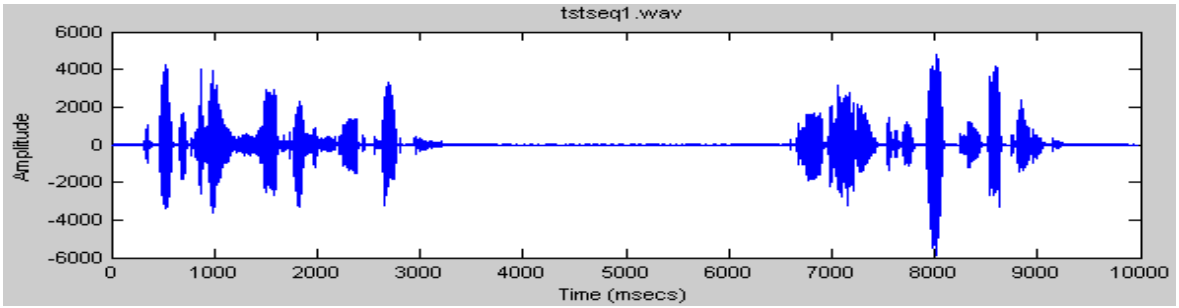

Figure 4.5 : Forme d'onde du vecteur de test "tstseq1.wav"

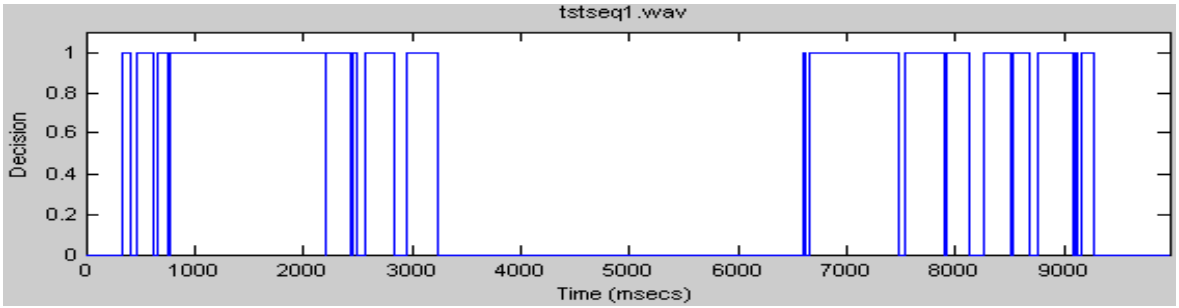

Figure 4.6 : Courbe VAD du vecteur de test "tstseq1.wav"

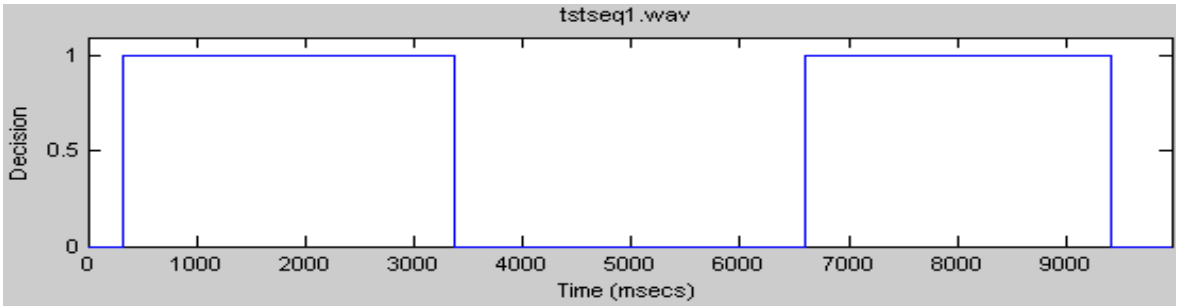

Figure 4.7 : Courbe VAD du vecteur de test "tstseq1.wav" avec temps de maintien

b. Cas d'enregistrements de la base de données ARADIGIT

L'enregistrement fa035m1.wav de la base de données ARADIGIT qui correspond à une suite de chiffres arabes prononcés par une locutrice dans un ordre aléatoire à été choisi (voir Figure 4.12). Les Figures 4.8 - 4.11 procurent les éléments d'analyse.

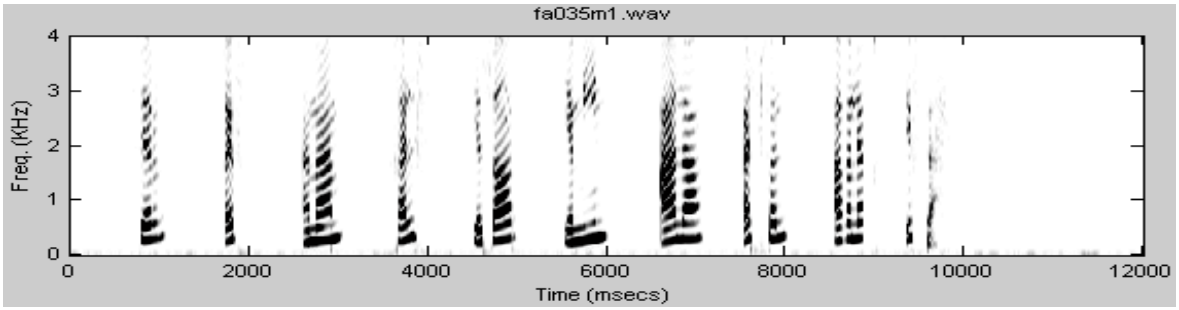

Figure 4.8 : Spectrogramme de l'enregistrement "fa035m1.wav"

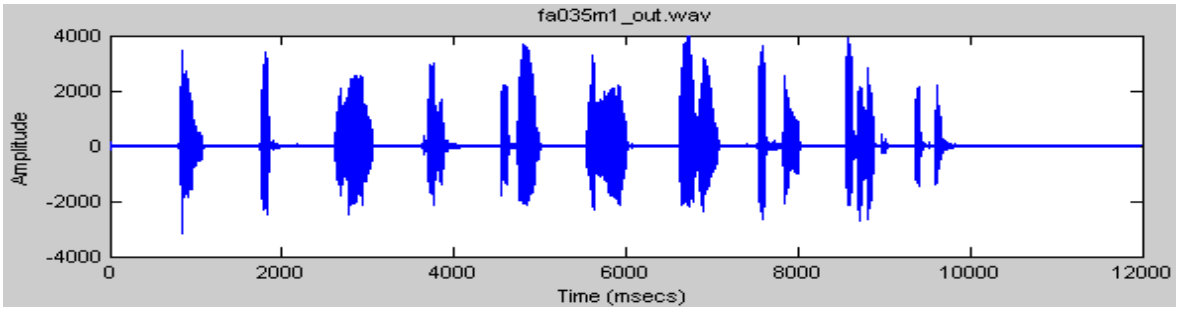

Figure 4.9 : Forme d'onde de l'enregistrement "fa035m1.wav" synthétisé

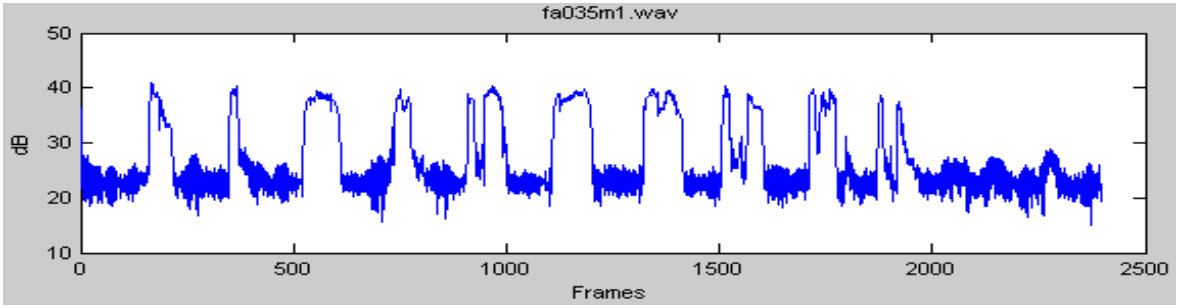

Figure 4.10 : Courbe d'énergie de l'enregistrement "fa035m1.wav"

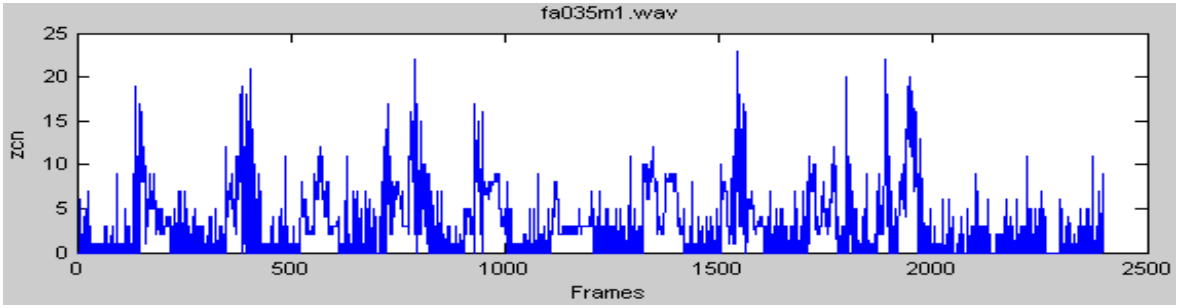

Figure 4.11 : Courbe du TPZ du vecteur de l'enregistrement "fa035m1.wav"

En plus de l'oscillation de la décision entre parole et bruit, on peut remarquer que certaines portions de bruit sont assimilées à de la voix (voir Figure 4.13). Cela tendra à s'accentuer dans un environnement fortement bruité. Une solution à ce problème est proposée dans la seconde partie expérimentale.

Application de la VAD avec temps de maintien à l'enregistrement fa035m1.wav

La nouvelle allure de la courbe de détection d'activité vocale (Figure 4.14) après introduction du temps de maintien de la décision voix, présente une meilleure distinction parole / silence pour l'enregistrement fa035m1.wav.

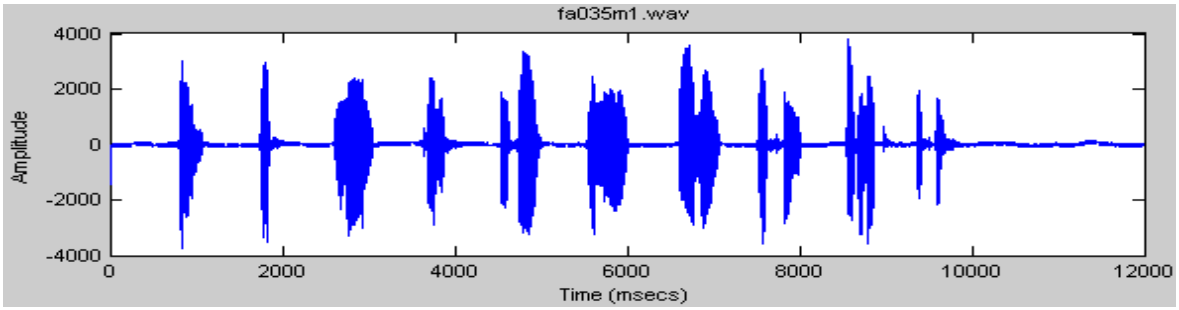

Figure 4.12 : Forme d'onde de l'enregistrement "fa035m1.wav"

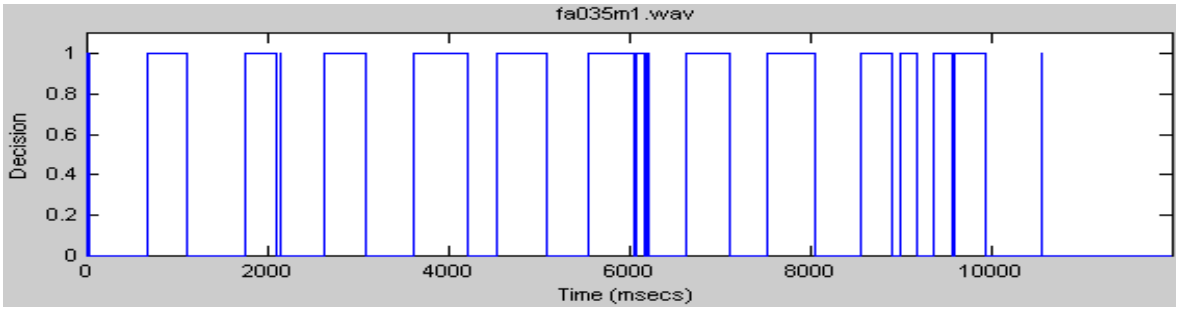

Figure 4.13 : Courbe VAD de l'enregistrement "fa035m1.wav"

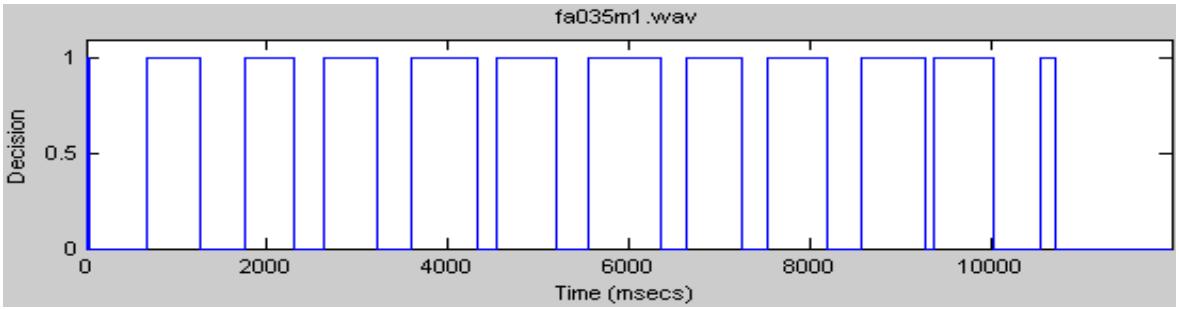

Figure 4.14 : Courbe VAD de l'enregistrement "fa035m1.wav" avec temps de maintien

Enfin, l'enregistrement ma042m1.wav de la base de données ARADIGIT qui correspond à une suite de mots arabes prononcés par un locuteur, clôt notre sélection d'enregistrements. Les Figures 4.15 - 4.18 constituent les éléments d'analyse.

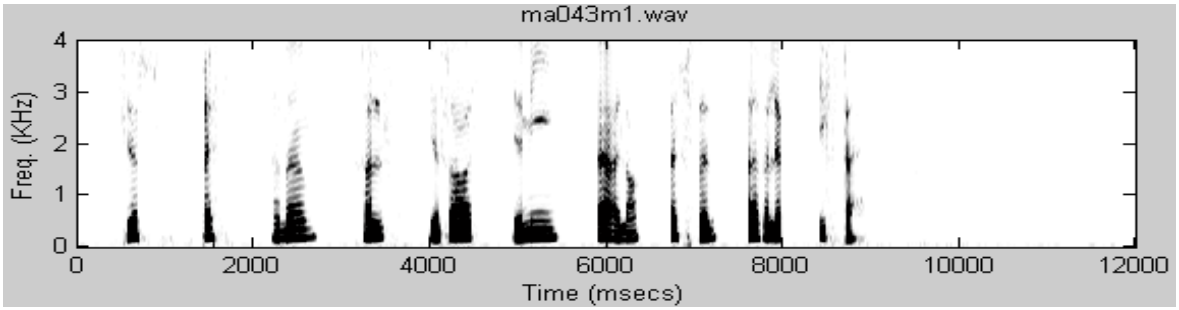

Figure 4.15 : Spectrogramme de l'enregistrement "ma042m1.wav"

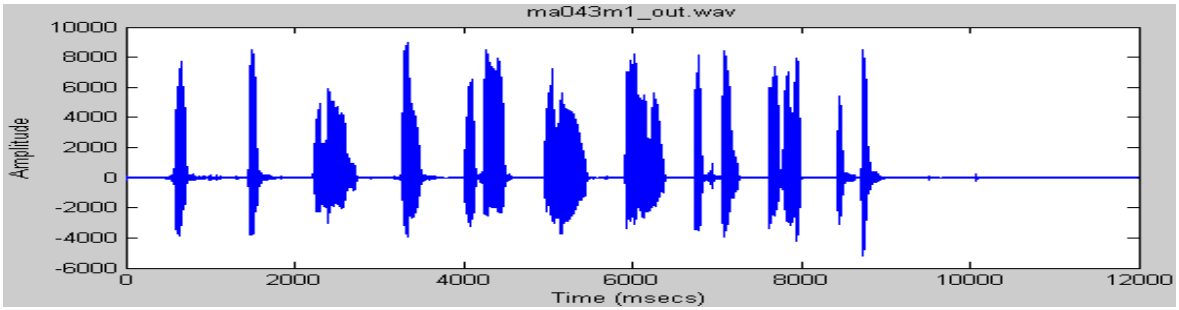

Figure 4.16 : Forme d'onde de l'enregistrement "ma042m1.wav" synthétisé

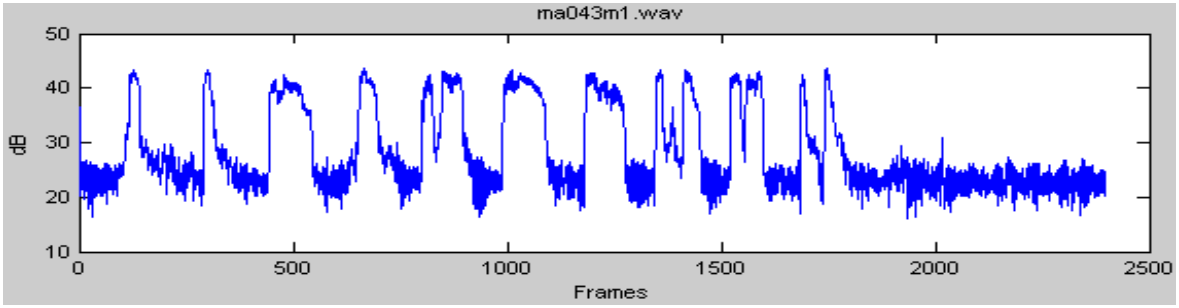

Figure 4.17 : Courbe d'énergie de l'enregistrement "ma042m1.wav"

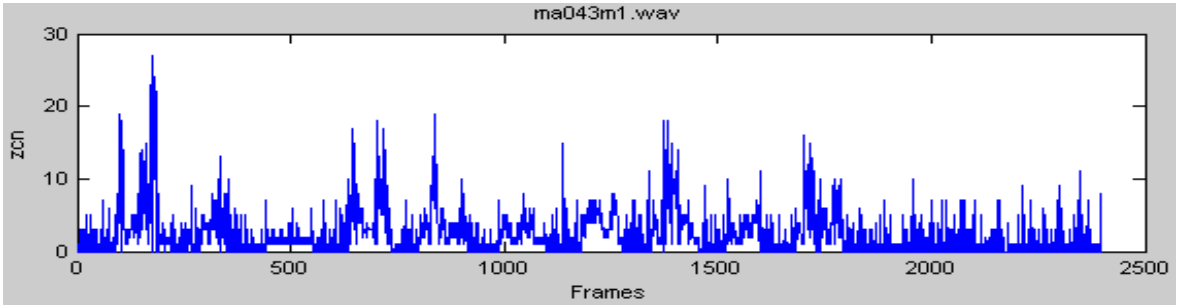

Figure 4.18 : Courbe TPZ de l'enregistrement "ma042m1.wav"

Application de la VAD avec temps de maintien à l'enregistrement ma042m1.wav

Le passage de cet enregistrement à travers le codeur G.729B modifié (VAD avec temps de maintien), donne lieu à une nouvelle courbe de détection d'activité vocale (Figure 4.21) dont l'allure révèle une meilleure classification silence / parole.

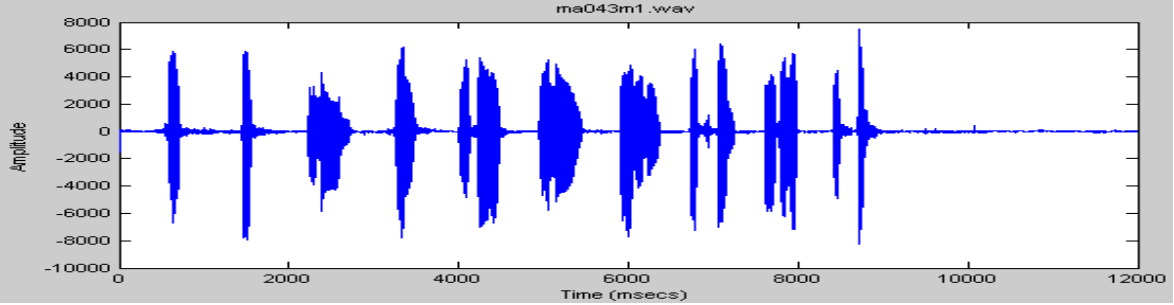

Figure 4.19 : Forme d'onde de l'enregistrement "ma042m1.wav"

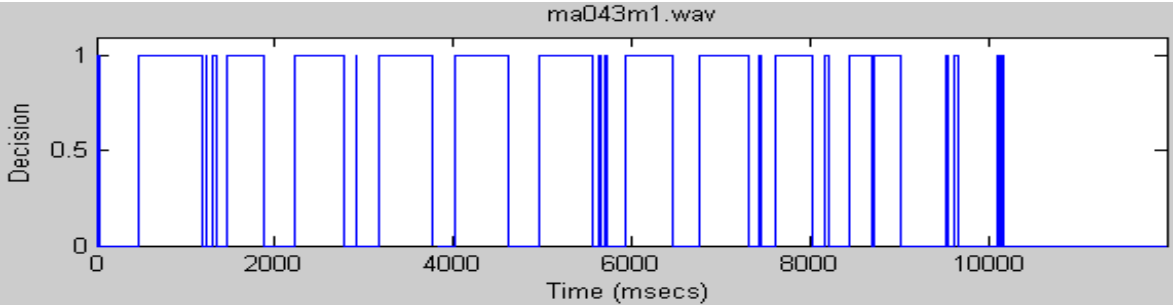

Figure 4.20 : Courbe VAD de l'enregistrement "ma042m1.wav"

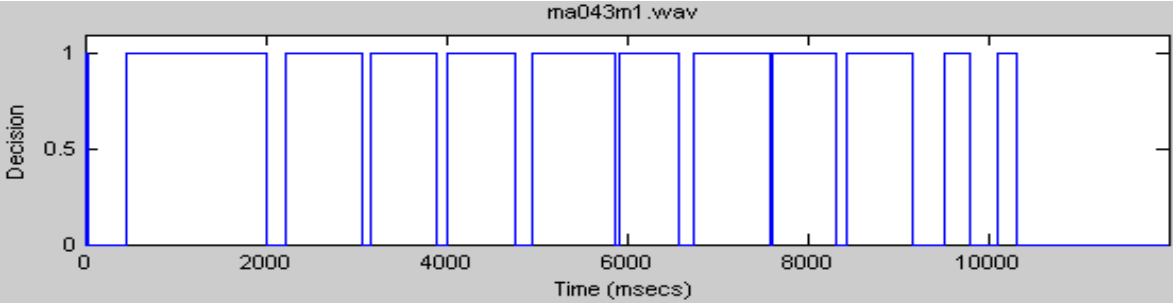

Figure 4.21 : Courbe VAD de l'enregistrement "ma042m1.wav" avec temps de maintien

Evaluation objective de la qualité du codec G.729B modifié (VAD avec temps de maintien)

Les performances de notre algorithme en termes de qualité perçue du signal vocal restitué, sont évaluées au travers d'une mesure objective de la qualité fournie par l'outil *PESQ* de l'ITU. Cet algorithme compare le signal vocal restitué et le signal vocal original sans compression, et attribue une note *PESQ* entre -0,5 et 4,5, traduisant la distorsion introduite par le codec G.729B.

L'outil PESQ de la recommandation P.862 a été mis en œuvre à cet effet. Une vérification du succès de son implémentation s'en est suivie en utilisant une base de données jointe à cette même recommandation [8]. Le tableau 4.2 regroupe les notes objectives attribuées par l'outil PESQ à nos signaux vocaux restitués à travers les codecs G.729B et G.729B modifié (VAD avec temps de maintien).

Tableau 4.2 : Evaluation PESQ pour les codecs G.729B et G.729B modifié (VAD avec temps de maintien):

Signaux vocaux	PESQ (G.729B)	PESQ (G.729B modifié)
Vecteur de test (tstseq1.bin - G.729B)	3.599	3.757
Enregistrement féminin (fa035m1.wav - ARADIGIT)	3.483	3.525
Enregistrement masculin (ma042m1.wav - ARADIGIT)	3.690	3.757

En analysant les notes objectives fournies par l'outil PESQ (Tableau 4.2), on constate une amélioration de la qualité perçue de nos signaux vocaux synthétisés à travers le codec G.729B modifié (VAD avec temps de maintien) comparée à celle obtenue avec le codec G.729B.

Evaluation de l'impact de l'amélioration du VAD sur la capacité (bande passante)

Pour évaluer l'impact de l'amélioration de l'efficacité du VAD sur la capacité, nous avons fait une analyse approfondie portant sur :

- Le nombre de trames vocales ;
- Le nombre de trames SID (Silence Insertion Descriptor) ;
- Le nombre de trames non transmises.

Les pourcentages respectifs de ces trames, notamment les trames non transmises, renseignent sur l'optimisation de la capacité. L'amélioration que nous proposons dans le codec G.729B occasionne une diminution des trames non transmises et des trames SID au profit des trames vocales. Ce qui augmente légèrement la capacité (consommation de la bande passante). Ceci est compensé par l'amélioration de l'intelligibilité et de la qualité de la parole transmise (voir Tableau 4.2). Aussi, une légère augmentation de la consommation de la bande passante peut être tolérée si la qualité de la parole s'en trouve améliorée. Le Tableau 4.3 donne les proportions des différents types de trames pour les 3 séries d'expériences.

Tableau 4.3 : Optimisation de la largeur de bande pour les codecs G.729B et G.729B modifié (VAD avec temps de maintien):

Signaux vocaux	VAD	Trames vocales	Trames SID	Trames NT
Vecteur de test (tstseq1.bin - G.729B)	G.729B	485(48,60%)	85(8,52%)	428(42,88%)
	G.729 modifié	587(58,82%)	46(4,61%)	365(36,57%)
Enregistrement féminin (fa035m1.wav - ARADIGIT)	G.729B	510(42,54%)	90(7,50%)	599(49,96%)
	G.729 modifié	668(55,71%)	66(5,50%)	465(38,78%)
Enregistrement masculin (ma042m1.wav - ARADIGIT)	G.729B	589(49,12%)	82(6,84%)	528(44,04%)
	G.729 modifié	817(68,14%)	43(3,59%)	339(28,27%)

4.3.2. Evaluation du module VAD du codeur G.729B dans un milieu bruité

La base de donnée NOISEX 92, qui fournit une variété de bruits enregistrés, est utilisée pour bruiter artificiellement le vecteur de test et les enregistrements de la base de données ARADIGIT, afin d'évaluer le codec, et ce en comparant les différentes courbes de détection d'activité vocale sur une gamme relativement large de rapport signal sur bruit (0, 5, 10, 15 dB). Avant le bruitage de notre base à différents SNR, un sous-échantillonnage à 8 kHz est opéré sur les différents bruits.

a. Bruit de chahut

a.1 Cas du vecteur de test

Dans nos expériences, nous avons procédé au bruitage du vecteur de test tstseq1.bin avec le bruit de chahut de la base de données NOISEX92 à différents SNR. Les Figures 4.22 - 4.25 illustrent les éléments d'analyse à un niveau SNR = 15 dB.

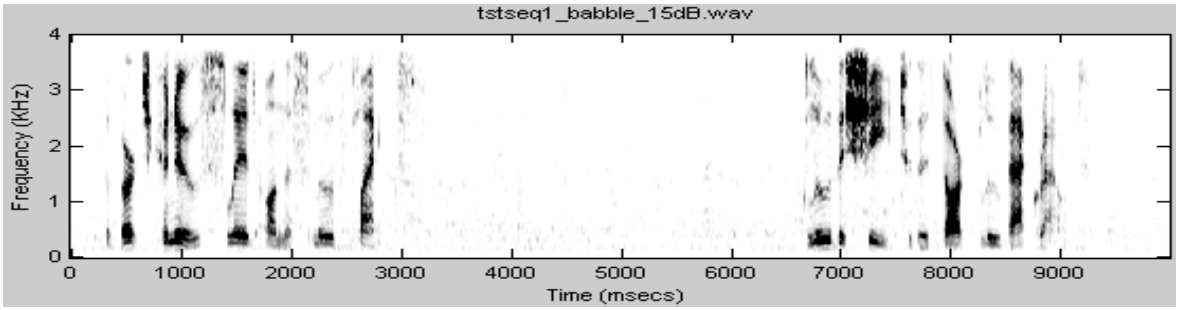

Figure 4.22 : Spectrogramme du vecteur de test "tstseq1.wav" bruité à 15dB avec du bruit de chahut

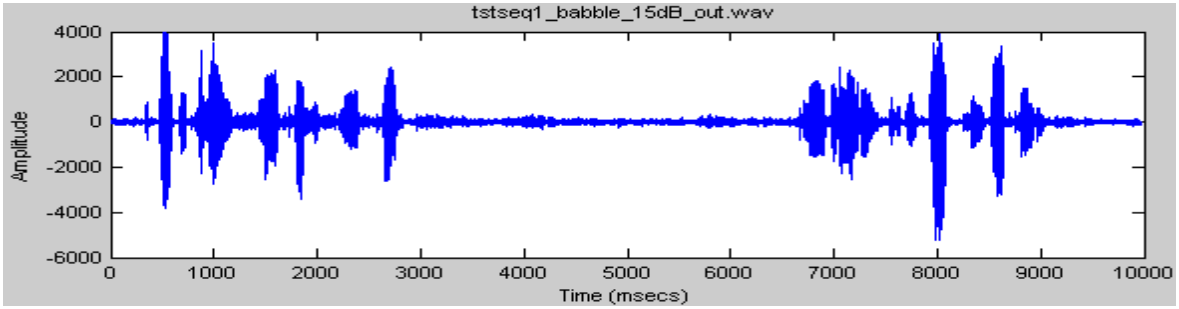

Figure 4.23 : Forme d'onde du vecteur de test synthétisé "tstseq1.wav" bruité à 15dB avec du bruit de chahut

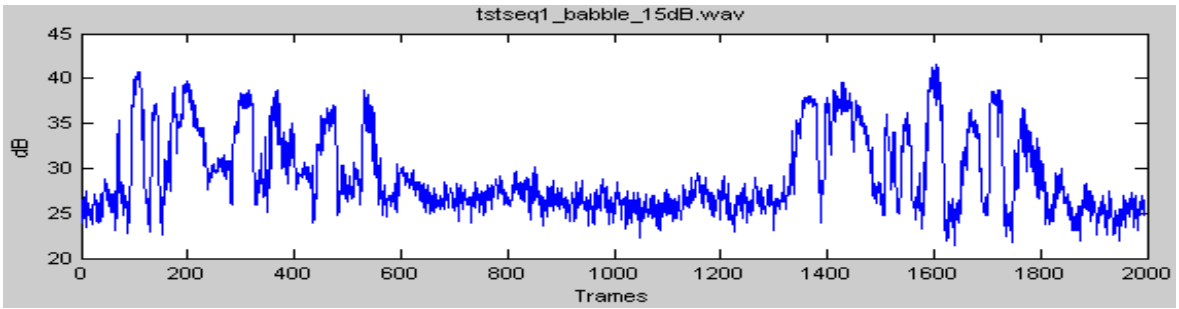

Figure 4.24 : Courbe d'énergie du vecteur de test "tstseq1.wav" bruité à 15dB avec du bruit de chahut

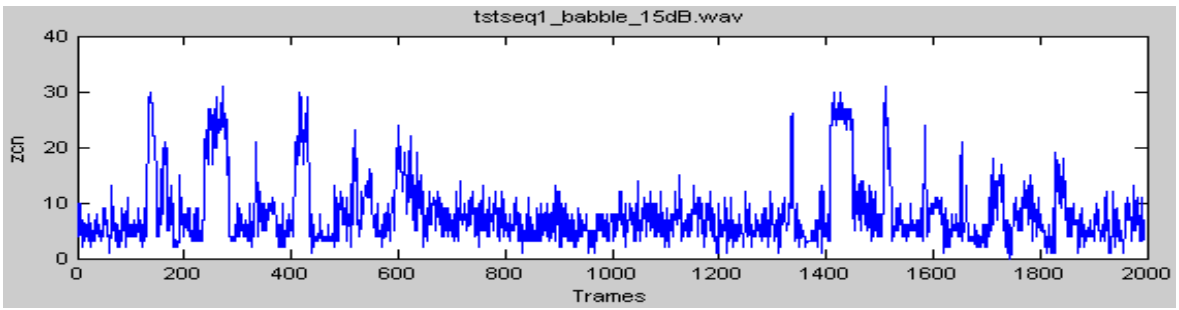

Figure 4.25 : Courbe du TPZ du vecteur de test "tstseq1.wav" bruité à 15dB avec du bruit de chahut

Même à ce niveau relativement élevé de SNR, la courbe VAD (voir Figure 4.31) fait apparaitre des zones d'oscillation parole / bruit dans une région normalement occupée par le

silence. Ceci s'accompagne d'une consommation excessive de la bande passante et d'une dégradation de la qualité de la parole transmise.

Les Figures 4.26 - 4.29 donnent les éléments d'analyse pour le vecteur de test tstseq1.bin bruité à 5 dB avec du bruit de chahut.

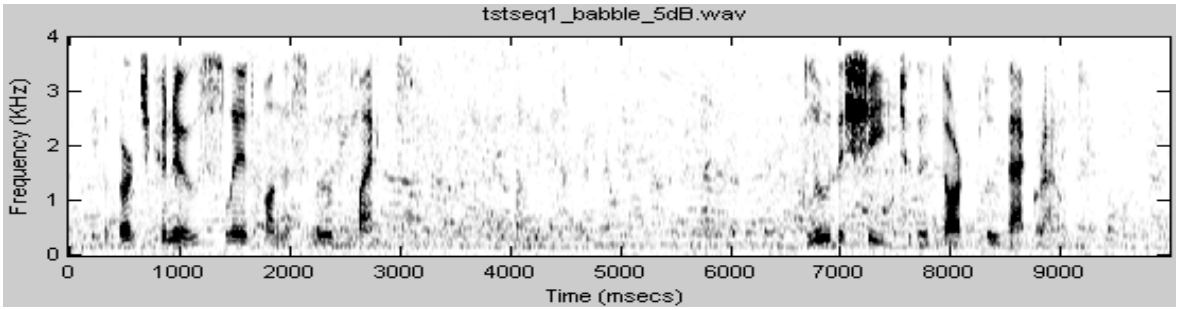

Figure 4.26 : Spectrogramme du vecteur de test "tstseq1.wav" bruité à 5dB avec du bruit de chahut

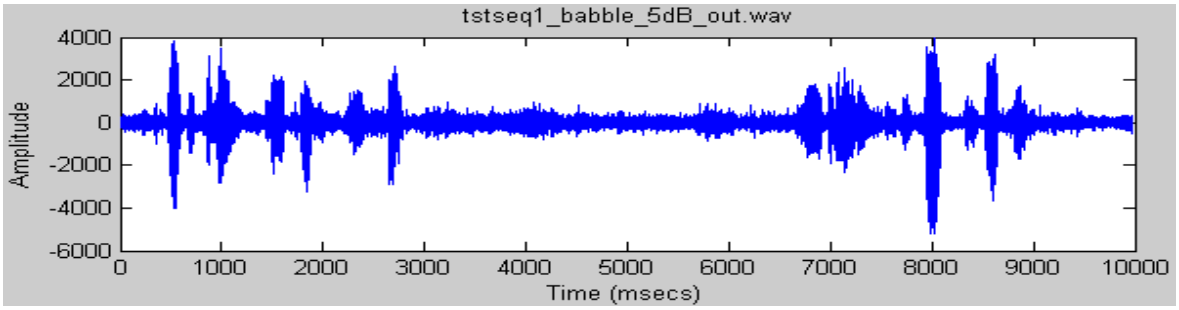

Figure 4.27 : Courbe d'onde du vecteur de test synthétisé "tstseq.wav" bruité à 5dB avec du bruit de chahut

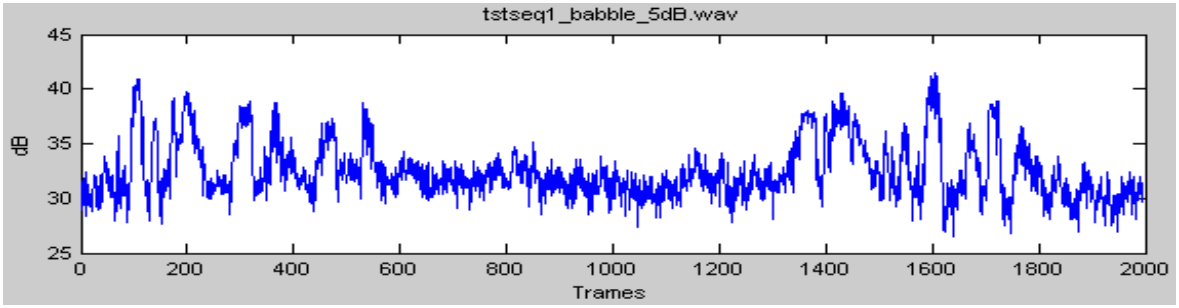

Figure 4.28 : Courbe d'énergie du vecteur de test "tstseq1.wav" bruité à 5dB avec du bruit de chahut

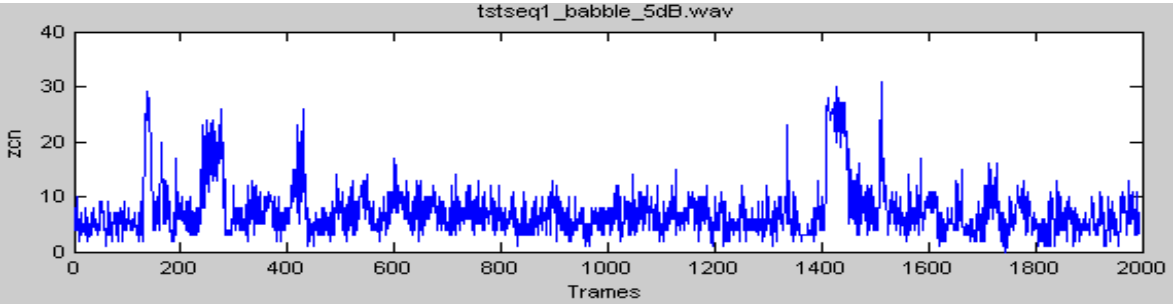

Figure 4.29 : Courbe du TPZ du vecteur de test "tstseq1.wav" bruité à 5dB avec du bruit de chahut

Lorsque le rapport signal sur bruit est inférieur à 15 dB, deux problèmes se posent au cours de la détection VAD courante : d'une part, la décision oscille entre parole / bruit, en réduisant les avantages en termes d'économie de largeur de bande, au moyen des mises à jour du descripteur SID et, d'autre part, la décision en faveur du bruit est souvent prise pendant le signal vocal.

Amélioration de la VAD par ajustement du seuil de détection

On remarque que l'allure de la courbe de détection d'activité vocale du signal est considérablement altérée en milieu bruité. Le seuil de détection, qui est défini aux alentours de 15 dB dans l'actuelle recommandation G.729 annexe B, a un poids important dans l'algorithme de détection. Durant les premières trames, une étape d'initialisation des énergies et des caractéristiques du bruit de fond ainsi qu'une prise de décision définitive (voix ou bruit) intervient suivant que l'énergie moyenne de la trame courante soit supérieure ou inferieure à ce seuil. De même, à partir d'un nombre N de trames, la décision de détection d'activité vocale initiale est forcée à 0, ou passe par des régions de décision à frontières multiples suivant que l'énergie moyenne de la trame actuelle soit supérieure ou inferieure à ce seuil.

Afin de parer à la perte d'efficacité de détection en milieu bruité, nous proposons un ajustement du seuil de détection.

L'ajustement de ce seuil dans un premier temps a été fait de manière statique, c.-à-d. que sa valeur à été fixée suivant le SNR du signal. Dans un second temps, une sélection plus dynamique de ce seuil a été élaborée en se basant sur la valeur moyenne de l'énergie de trame dans la pleine bande de fréquences calculée durant les 32 premières trames (temps d'acquissions des caractéristiques du bruit de fond).

Si ($\overline{E}_f$ < $\overline{En}$) {

$I_{VD} = 0$;

}

Sinon {

Décision VAD initiale à frontières multiples ;

} (4.2)

L'ajustement seul de ce seuil ne fournit pas un résultat concluant, il doit être accompagné d'un calibrage du temps de maintien de la décision voix (que nous avons proposé comme première amélioration) pour éviter l'oscillation de la décision VAD entre bruit et voix qui tend à s'accentuer avec la diminution du rapport signal sur bruit.

Les nouvelles allures des courbes VAD du vecteur de test bruité successivement à 15 dB (Figure 4.32) puis à 5 dB (Figure 4.35) avec du bruit de chahut, présentent une meilleure prise de décision après ajustement du seuil de détection.

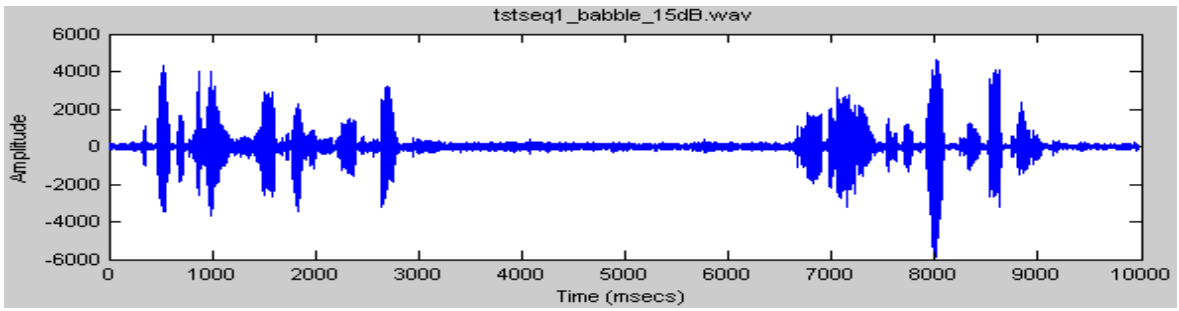

Figure 4.30 : Forme d'onde du vecteur de test "tstseq1.wav" bruité à 15dB avec du bruit de chahut

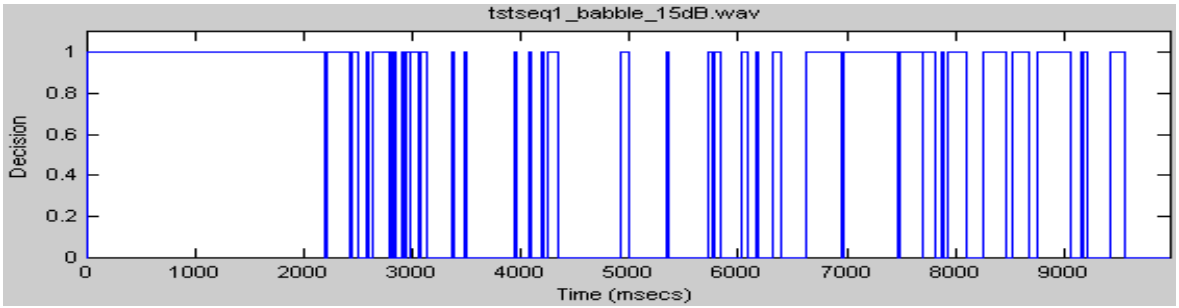

Figure 4.31 : Courbe VAD du vecteur de test "tstseq1.wav" bruité à 15dB avec du bruit de chahut

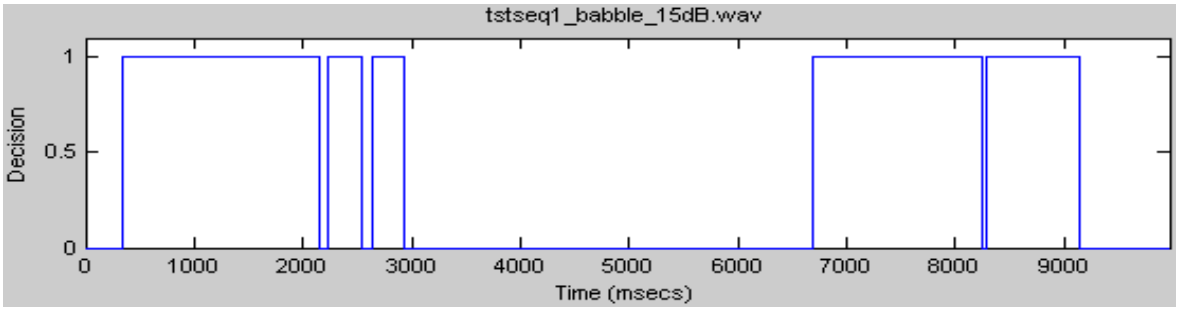

Figure 4.32 : Courbe VAD du vecteur de test “tstseq1.wav” bruité à 15dB avec du bruit de chahut avec temps de maintien et ajustement du seuil

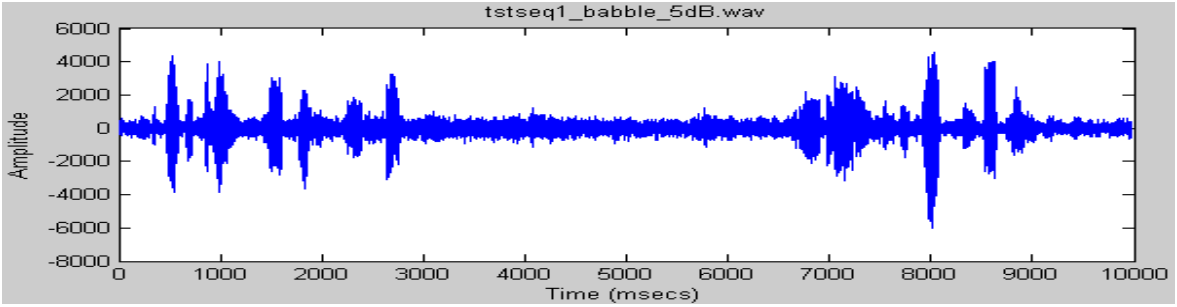

Figure 4.33 : Courbe d’onde du vecteur de test “tstseq1.wav” bruité à 5dB avec du bruit de chahut

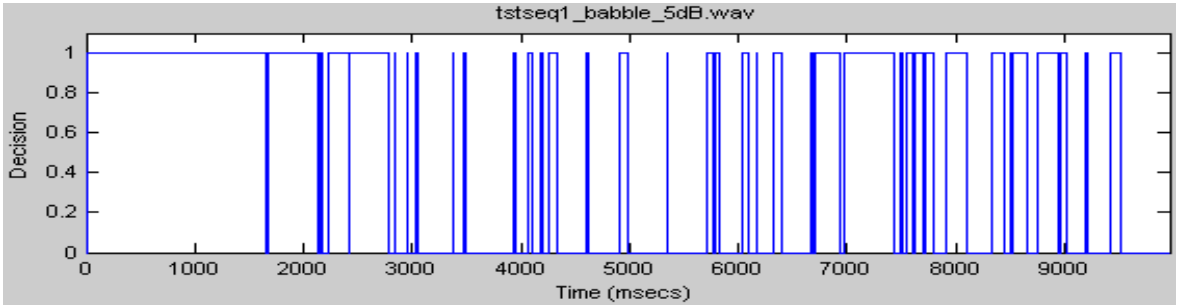

Figure 4.34 : Courbe VAD du vecteur de test “tstseq1.wav” bruité à 5dB avec du bruit de chahut

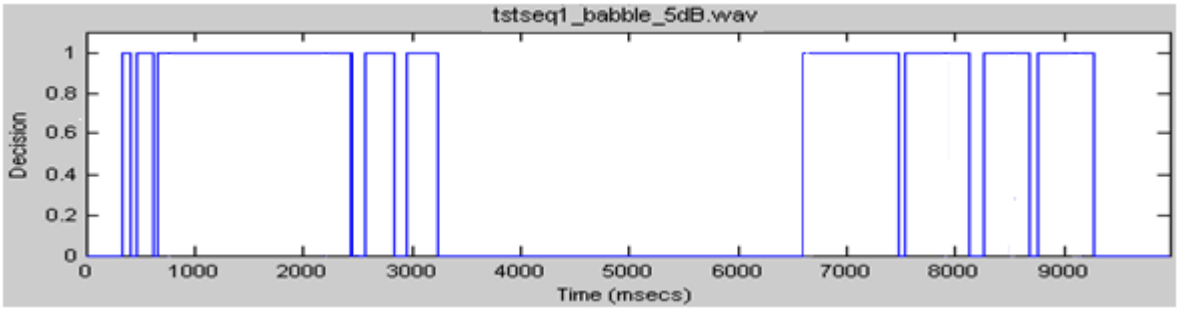

Figure 4.35 : Courbe VAD du vecteur de test “tstseq1.wav” bruité à 5dB avec du bruit de chahut avec temps de maintien et ajustement du seuil

a.2 Cas d'enregistrements de la base de données ARADIGIT

On procède également au bruitage de l'enregistrement fa035m1.wav avec le bruit de chahut à différents SNR. Les Figures 4.36 - 4.39 présentent les éléments d'analyse à un niveau SNR = 15 dB.

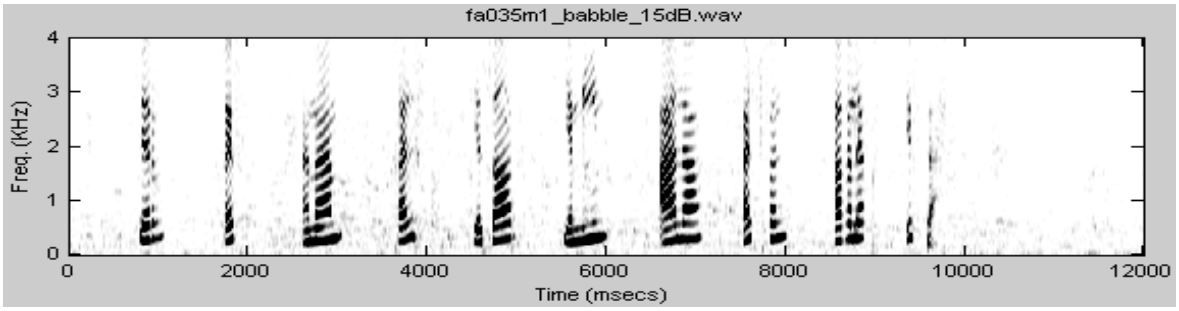

Figure 4.36 : Spectrogramme de l'enregistrement "fa035m1.wav" bruité à 15dB avec du bruit de chahut

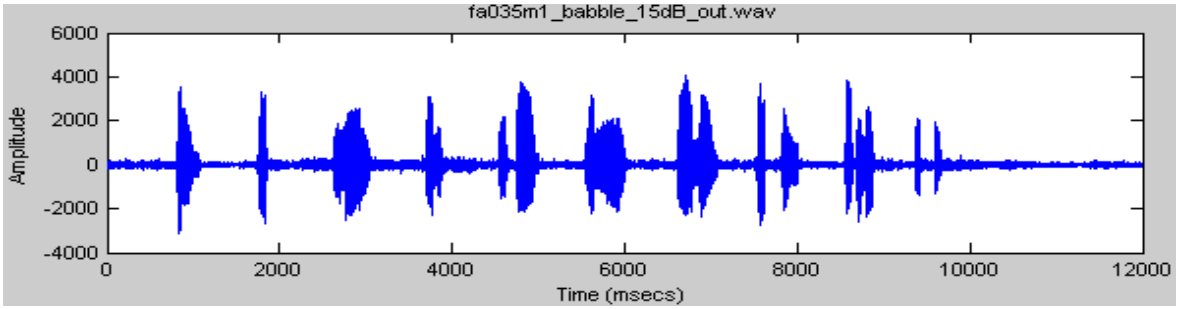

Figure 4.37 : Forme d'onde de l'enregistrement synthétisé "fa035m1.wav" bruité à 15dB avec du bruit de chahut

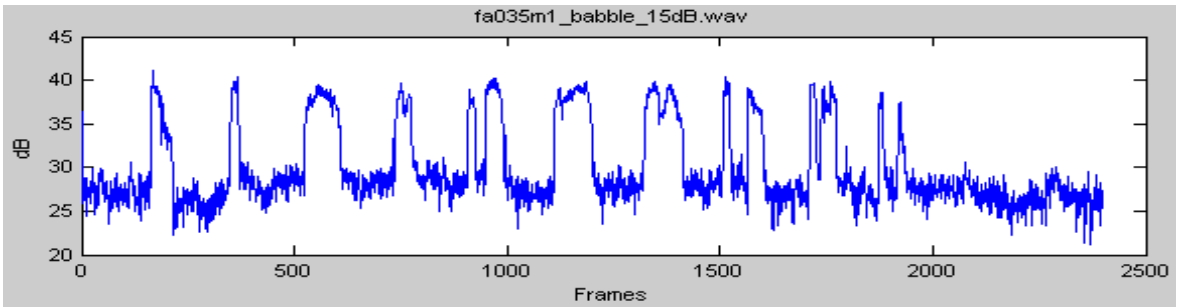

Figure 4.38 : Courbe d'énergie de l'enregistrement "fa035m1.wav" bruité à 15dB avec du bruit de chahut

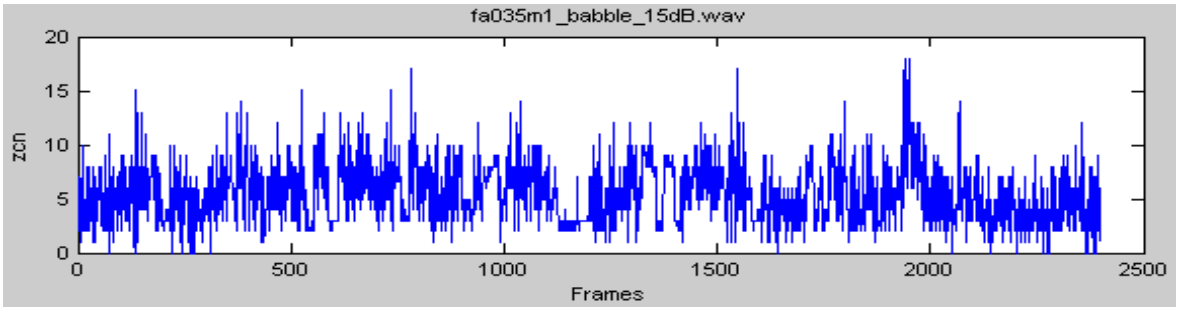

Figure 4.39 : Courbe TPZ de l'enregistrement "fa035m1.wav" bruité à 15dB avec du bruit de chahut

Les Figures 4.40 - 4.43 donnent les éléments d'analyse pour l'enregistrement fa035m1.wav bruité à 5 dB avec du bruit de chahut.

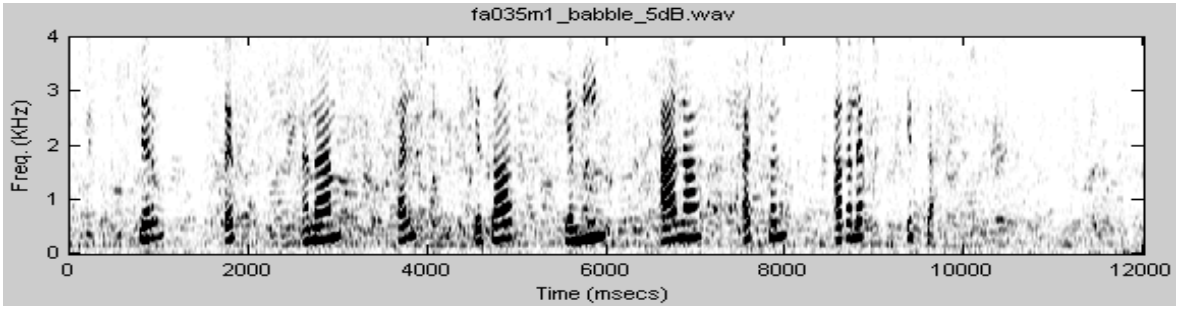

Figure 4.40 : Spectrogramme de l'enregistrement "fa035m1.wav" bruité à 5dB avec du bruit de chahut

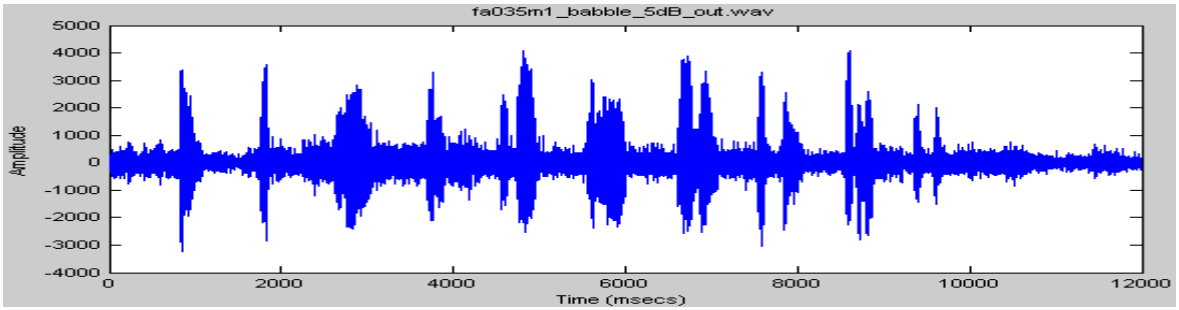

Figure 4.41 : Forme d'onde de l'enregistrement synthétisé "fa035m1.wav" bruité à 5dB avec du bruit de chahut

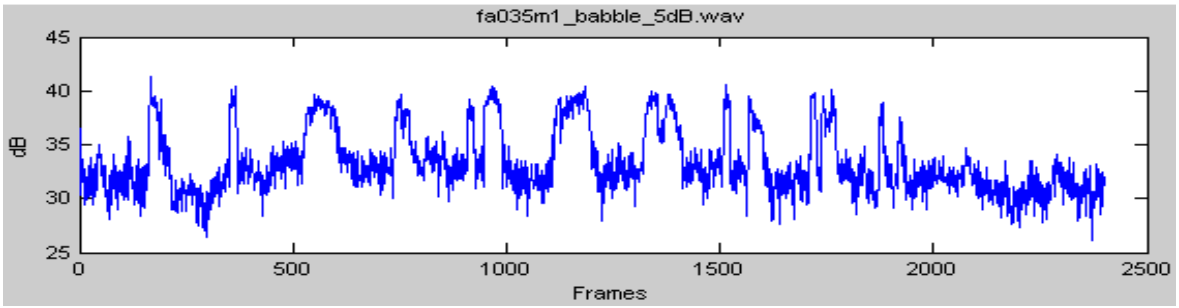

Figure 4.42 : Courbe d'énergie de l'enregistrement "fa035m1.wav" bruité à 5dB avec du bruit de chahut

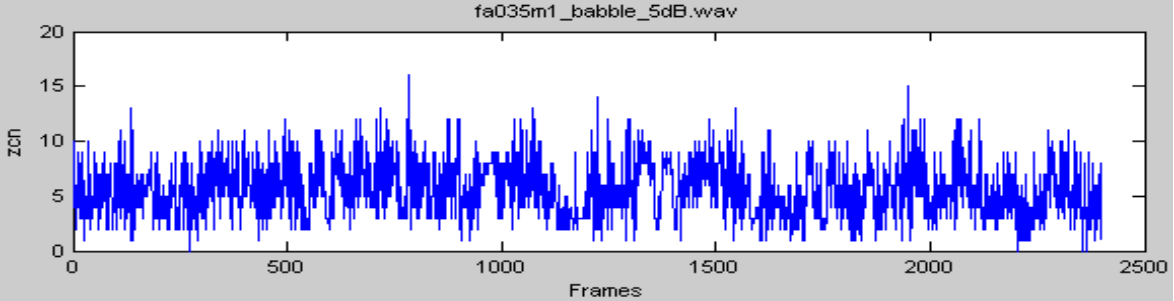

Figure 4.43 : Courbe TPZ de l'enregistrement "fa035m1.wav" bruité à 5dB avec du bruit de chahut

Application de la VAD avec temps de maintien et ajustement du seuil à l'enregistrement fa035m1.wav

Les nouvelles allures des courbes VAD de l'enregistrement fa035m1.wav bruité successivement à 15 dB (Figure 5.46) puis à 5 dB (Figure 5.49) avec du bruit de chahut, affichent une meilleure prise de décision après ajustement du seuil de détection.

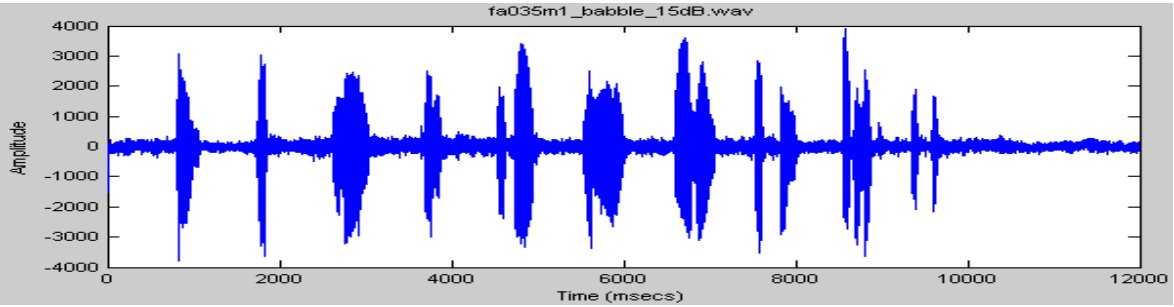

Figure 4.44 : Forme d'onde de l'enregistrement "fa035m1.wav" bruité à 15dB avec du bruit de chahut

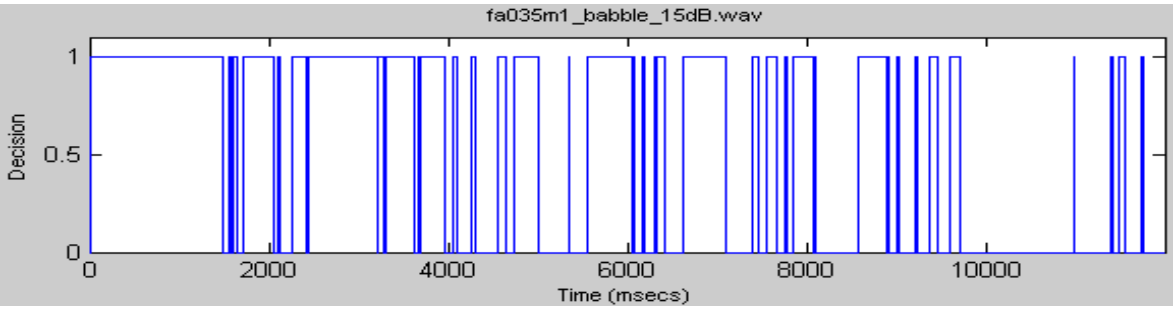

Figure 4.45 : Courbe VAD de l'enregistrement "fa035m1.wav" bruité à 15dB avec du bruit de chahut

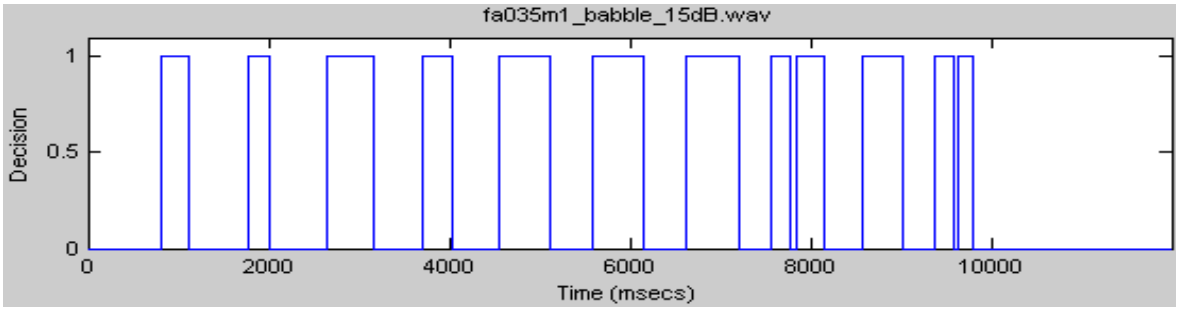

Figure 4.46 : Courbe VAD de l'enregistrement "fa035m1.wav" bruité à 15dB avec du bruit de chahut avec temps de maintien et ajustement du seuil

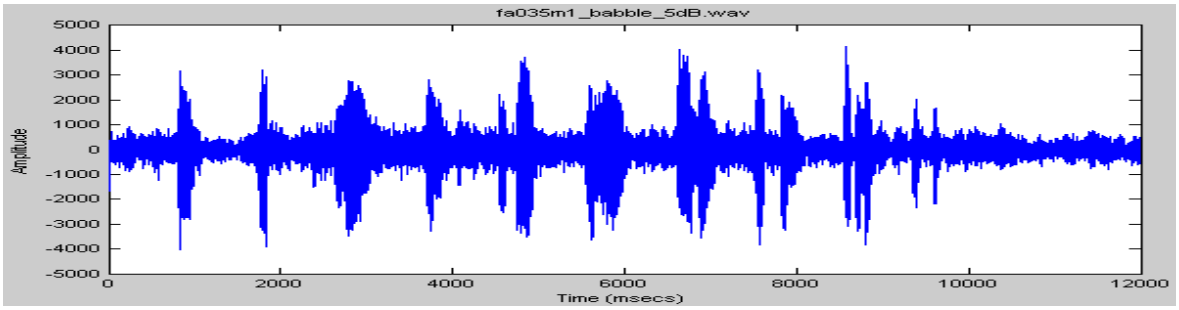

Figure 4.47 : Forme d'onde de l'enregistrement "fa035m1.wav" bruité à 5dB avec du bruit de chahut

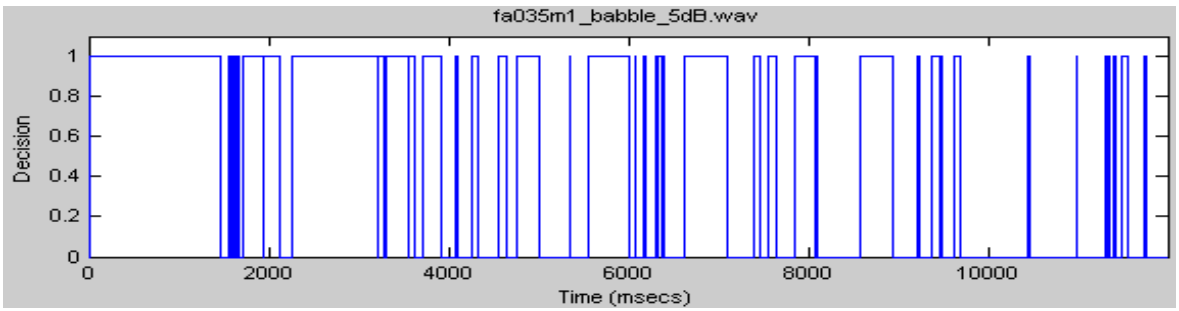

Figure 4.48 : Courbe VAD de l'enregistrement "fa035m1.wav" bruité à 5dB avec du bruit de chahut

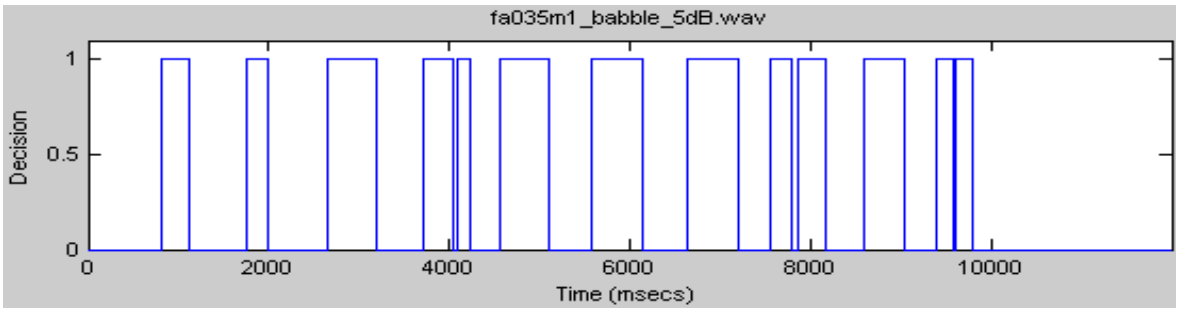

Figure 4.49 : Courbe VAD de l'enregistrement "fa035m1.wav" bruité à 5dB avec du bruit de chahut avec temps de maintien et ajustement du seuil

Enfin, on effectue le bruitage de l'enregistrement ma042m1.wav avec le bruit de chahut à différents SNR. Les figures 4.50 - 4.53 illustrent les éléments d'analyse à un niveau SNR = 15 dB.

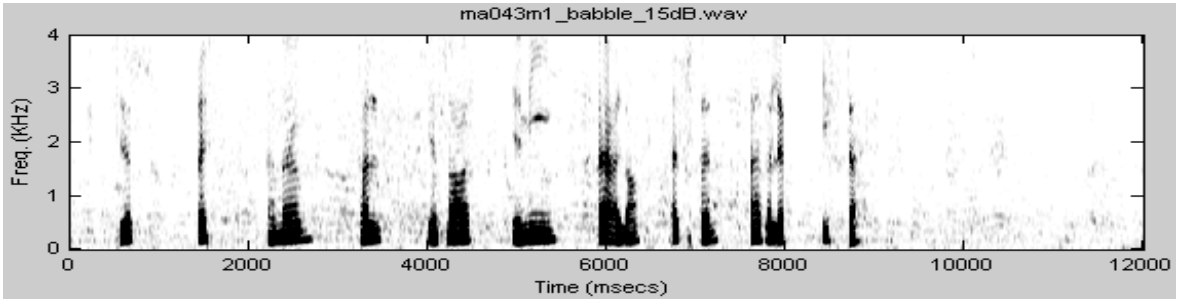

Figure 4.50 : Spectrogramme de l'enregistrement "ma042m1.wav" bruité à 15dB avec du bruit de chahut

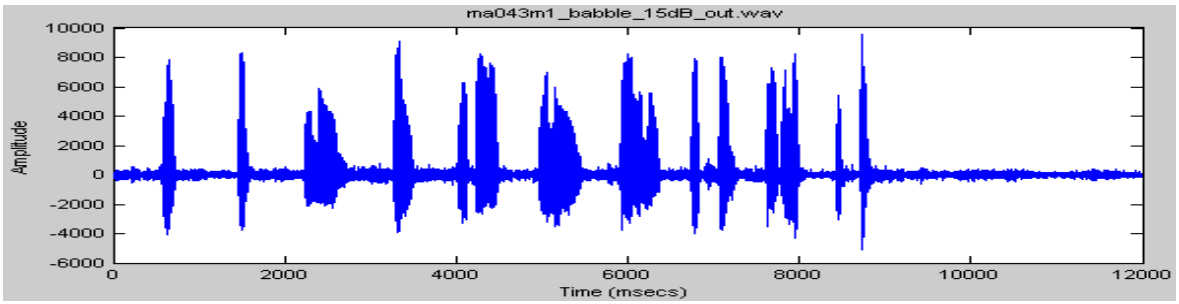

Figure 4.51 : Forme d'onde de l'enregistrement synthétisé "ma042m1.wav" bruité à 15dB avec du bruit de chahut

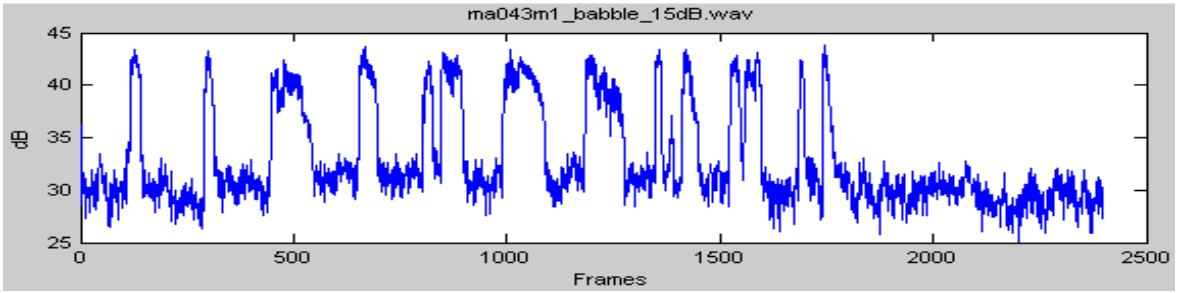

Figure 4.52 : Courbe d'énergie de l'enregistrement "ma042m1.wav" bruité à 15dB avec du bruit de chahut

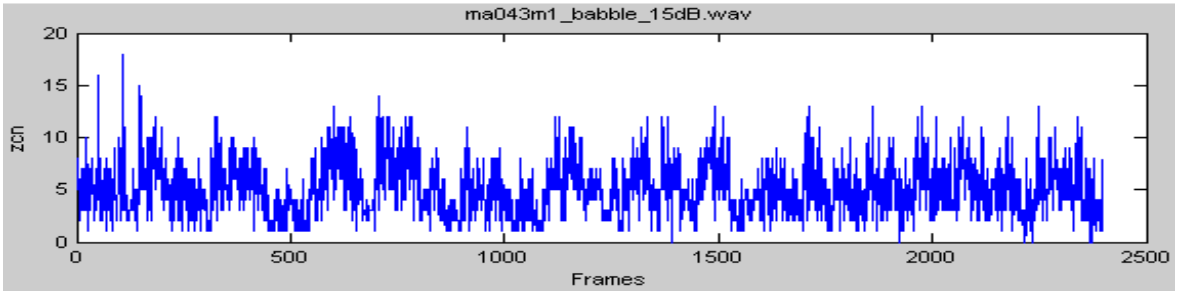

Figure 4.53 : Courbe TPZ de l'enregistrement "ma042m1.wav" bruité à 15dB avec du bruit de chahut

Les Figures 4.54 - 4.57 fournissent les éléments d'analyse pour l'enregistrement ma042m1.wav bruité à 5 dB avec du bruit de chahut.

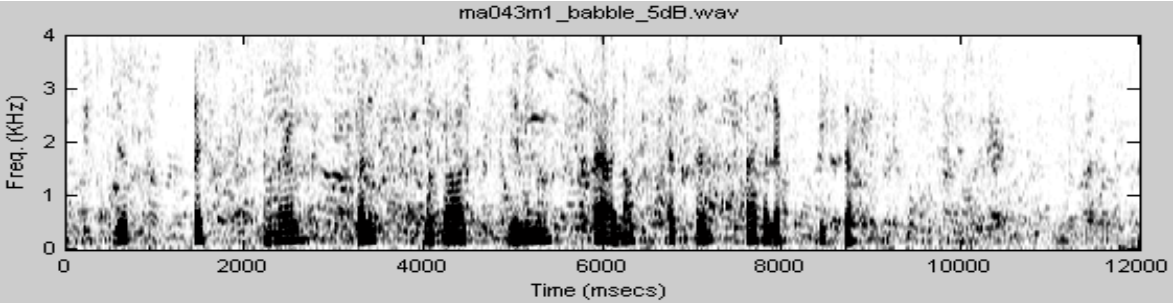

Figure 4.54 : Spectrogramme de l'enregistrement "ma042m1.wav" bruité à 5dB avec du bruit de chahut

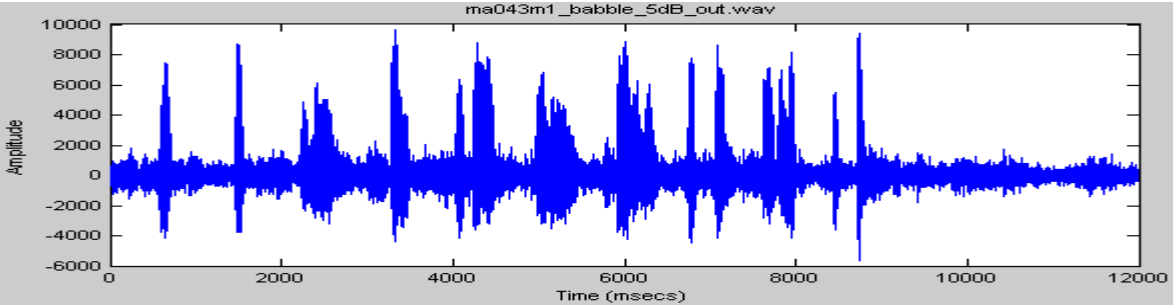

Figure 4.55 : Forme d'onde de l'enregistrement synthétisé "ma042m1.wav" bruité à 5dB avec du bruit de chahut

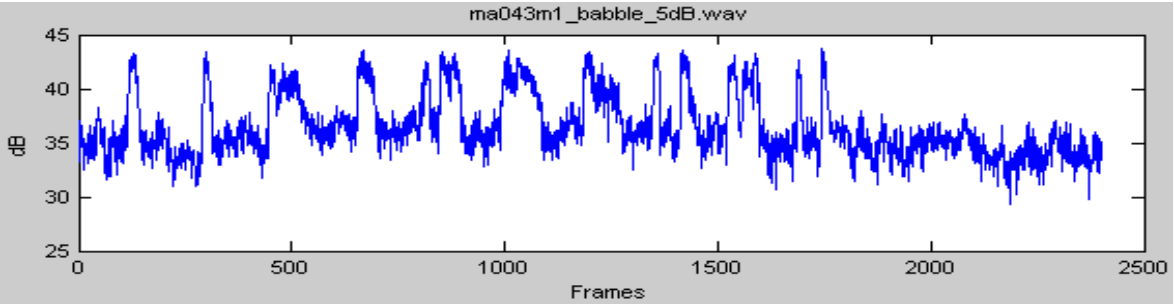

Figure 4.56 : Courbe d'énergie de l'enregistrement "ma042m1.wav" bruité à 5dB avec du bruit de chahut

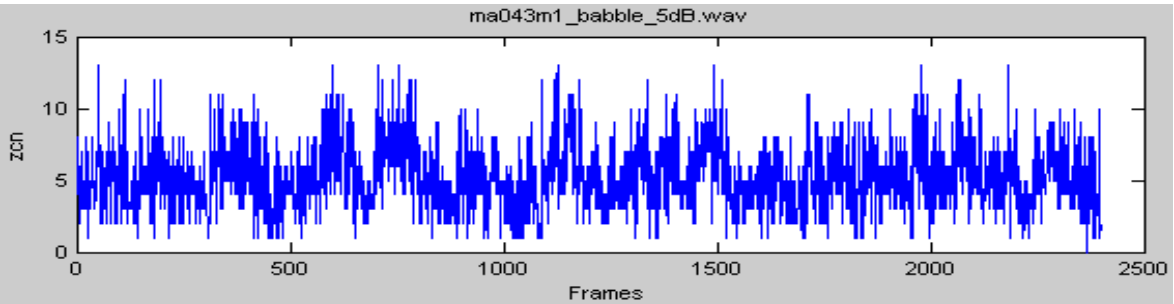

Figure 4.57 : Courbe TPZ de l'enregistrement "ma042m1.wav" bruité à 5dB avec du bruit de chahut

Application de la VAD avec temps de maintien et ajustement du seuil à l'enregistrement ma042m1.wav

Les nouvelles allures des courbes VAD de l'enregistrement ma042m1.wav bruité successivement à 15 dB (Figure 5.60) puis à 5 dB (Figure 5.63) avec du bruit de chahut, montrent une meilleure prise de décision après ajustement du seuil de détection.

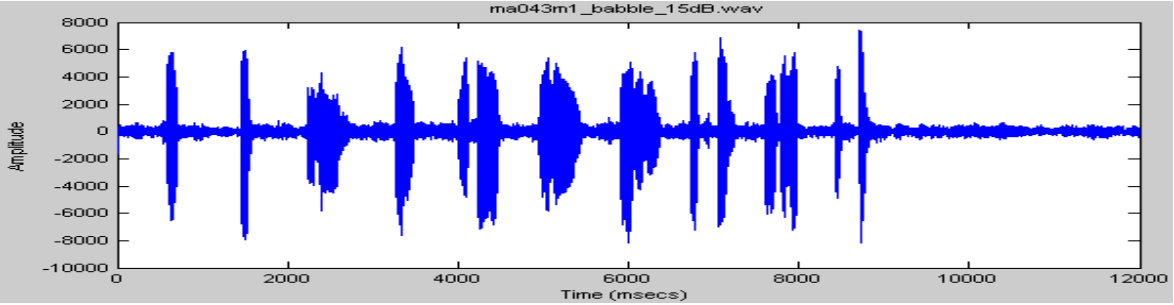

Figure 4.58 : Forme d'onde de l'enregistrement "ma042m1.wav" bruité à 15dB avec du bruit de chahut

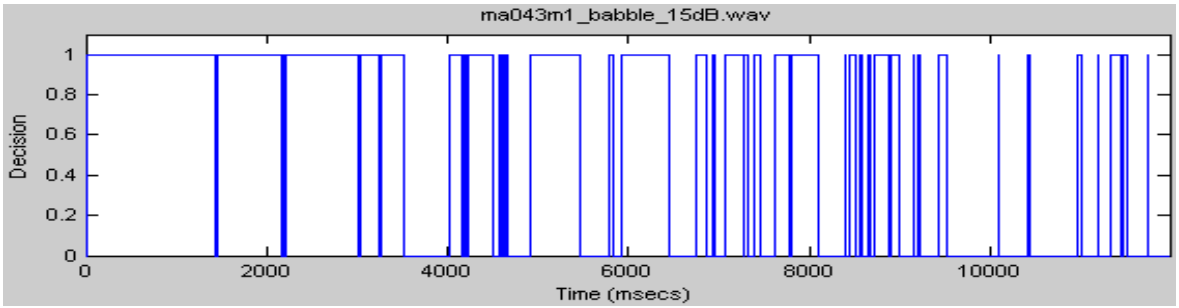

Figure 4.59 : Courbe VAD de l'enregistrement "ma042m1.wav" bruité à 15dB avec du bruit de chahut

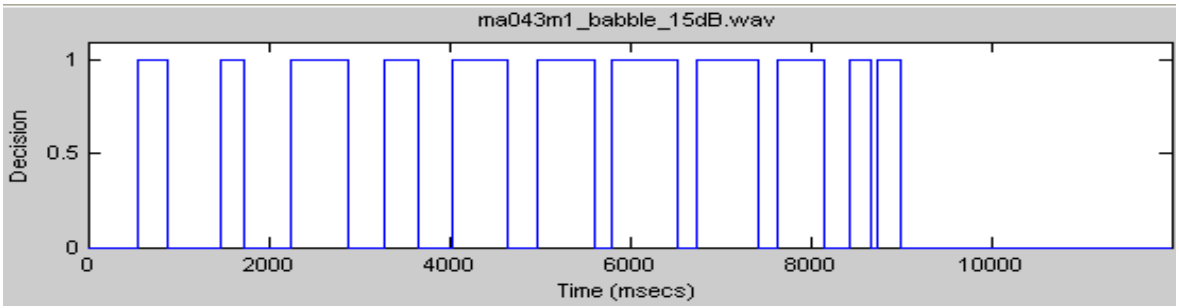

Figure 4.60 : Courbe VAD de l'enregistrement "ma042m1.wav" bruité à 15dB avec du bruit de chahut avec temps de maintien et ajustement du seuil

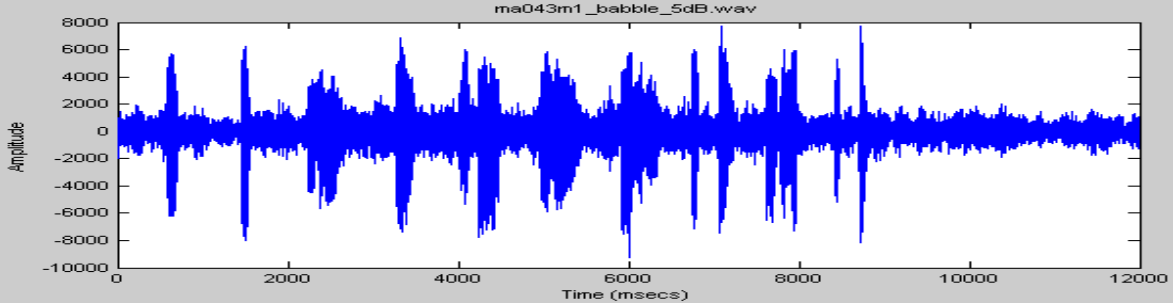

Figure 4.61 : Forme d'onde de l'enregistrement "ma042m1.wav" bruité à 5dB avec du bruit de chahut

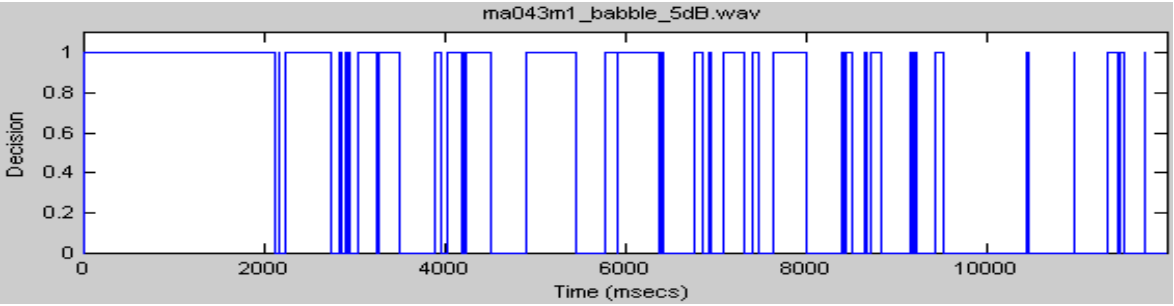

Figure 4.62 : Courbe VAD de l'enregistrement "ma042m1.wav" bruité à 5dB avec du bruit de chahut

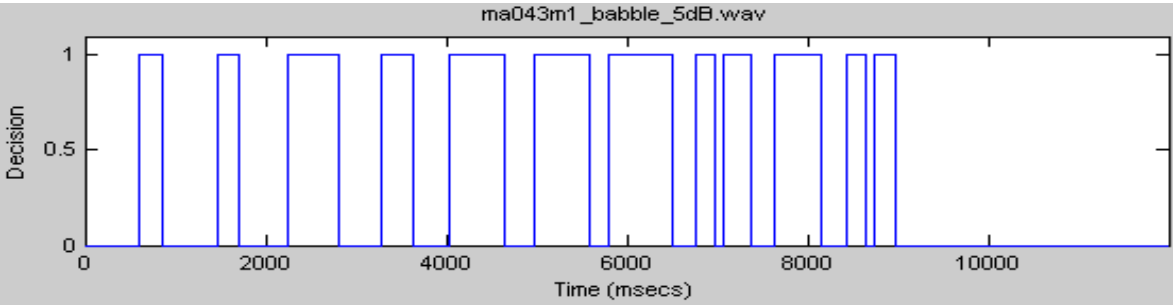

Figure 4.63 : Courbe VAD de l'enregistrement "ma042m1.wav" bruité à 5dB avec du bruit de chahut avec temps de maintien et ajustement du seuil

Evaluation objective de la qualité du codec G.729B modifié (VAD avec temps de maintien et ajustement du seuil)

Le Tableau 4.4 regroupe les notes objectives attribuées par l'outil PESQ à nos signaux vocaux restitués à travers les codecs G.729B et G.729B modifié (VAD avec temps de maintien et ajustement du seuil). Nous constatons une amélioration de la qualité objective de la parole transmise.

Tableau 4.4 : Evaluation PESQ pour les codecs G.729B et G.729B modifié (VAD avec temps de maintien et ajustement du seuil):

Signaux vocaux bruités (bruit de chahut - 15dB)	**PESQ (G.729B)**	**PESQ (G.729B modifié)**
Vecteur de test (tstseq1.bin - G.729B)	3.092	3.133
Enregistrement féminin (fa035m1.wav - ARADIGIT)	2.884	2.890
Enregistrement masculin (ma042m1.wav - ARADIGIT)	2.880	2.912

Evaluation de l'impact de l'amélioration du VAD sur la capacité (bande passante)

Le Tableau 4.5 donne les pourcentages respectifs des trames, notamment les trames non transmises qui renseignent sur l'optimisation de la capacité. Les méthodes d'ajustement du seuil et de l'introduction d'un temps de maintien que nous proposons dans le codec G.729B occasionne une légère diminution des trames SID au profit des trames vocales. Par contre les trames non transmises restent approximativement dans la même proportion. Ce qui améliore l'intelligibilité et la qualité de la parole transmise (voir Tableau 4.4) tout en gardant les mêmes performances concernant la bande passante.

Tableau 4.5 : Optimisation de la largeur de bande pour les codecs G.729B et G.729B modifié (VAD avec et temps de maintien et ajustement du seuil):

Signaux vocaux bruités (bruit de chahut - 15dB)	**VAD**	**Trames vocales**	**Trames SID**	**Trames NT**
Vecteur de test (tstseq1.bin - G.729B)	B729B	548(54,91%)	140(14,03%)	310(31,06%)
	G.729 modifié	553(55,41%)	131 (12,84%)	331 (31,76%)
Enregistrement féminin (fa035m1.wav - ARADIGIT)	B729B	620(51,71%)	174(14,51%)	405(33,78%)
	G.729 modifié	682(56,88%)	138(11,51%)	379(31,61%)
Enregistrement masculin (ma042m1.wav - ARADIGIT)	B729B	671(55,96%)	158(13,18%)	370(30,86%)
	G.729 modifié	679(56,63%)	147(12,26%)	373(31,11%)

b. Les autres bruits sélectionnés

b.1 Cas du vecteur de test

Le vecteur de test tstseq1.bin a été bruité avec le reste des bruits sélectionnés de la base de données NOISEX92, à différents SNR. Les Figure 4.64 – 4.71 soulignent l'apport de nos modifications (introduction d'un temps de maintien et ajustement du seuil) sur la prise de décision VAD pour le vecteur de test tstseq1.bin bruité à un SNR = 5 dB.

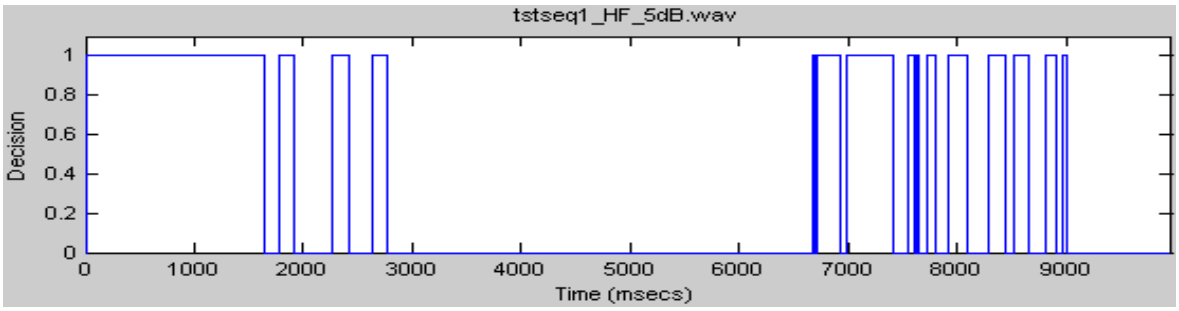

Figure 4.64 : Courbe VAD du vecteur de test "tstseq1.wav" bruité à 5dB avec du bruit de canal radio HF

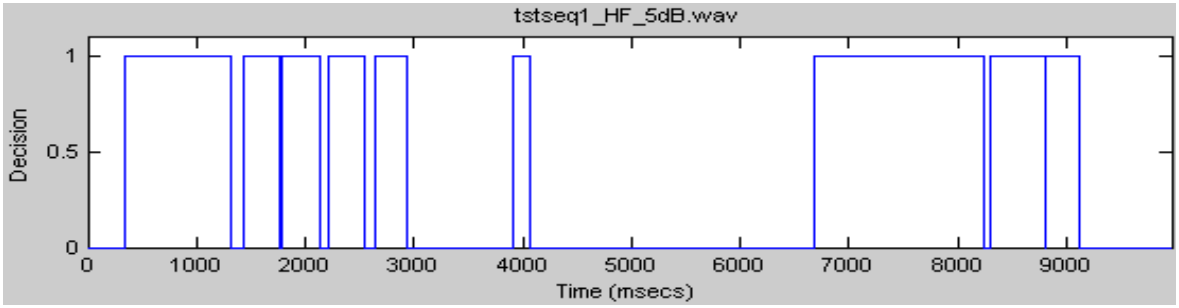

Figure 4.65 : Courbe VAD du vecteur de test "tstseq1.wav" bruité à 5dB avec du bruit de canal radio HF avec temps de maintien et ajustement du seuil

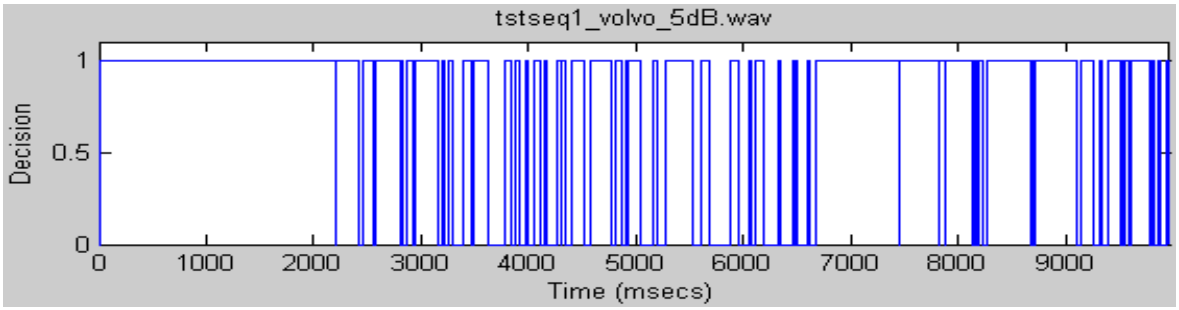

Figure 4.66 : Courbe VAD du vecteur de test "tstseq1.wav" bruité à 5dB avec du bruit de voiture

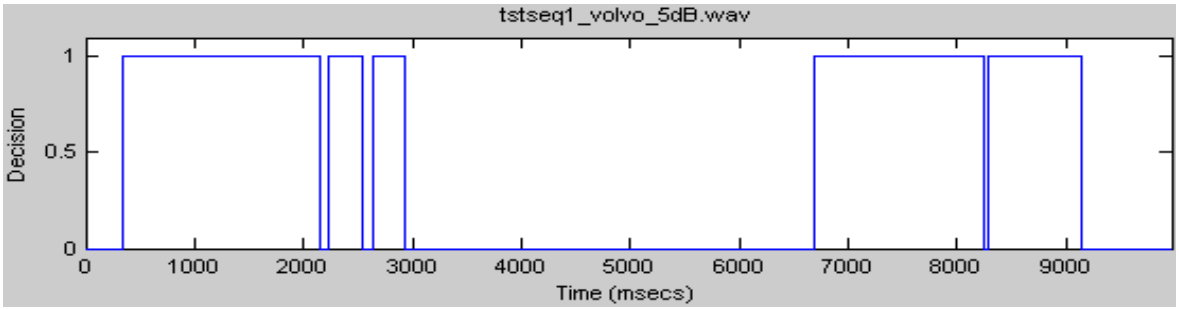

Figure 4.67 : Courbe VAD du vecteur de test "tstseq1.wav" bruité à 5dB avec du bruit de voiture avec temps de maintien et ajustement du seuil

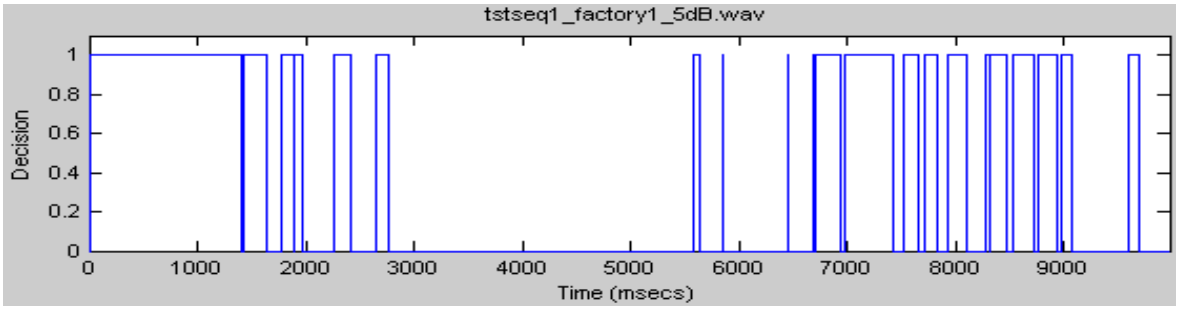

Figure 4.68 : Courbe VAD du vecteur de test "tstseq1.wav" bruité à 5dB avec du bruit d'usine

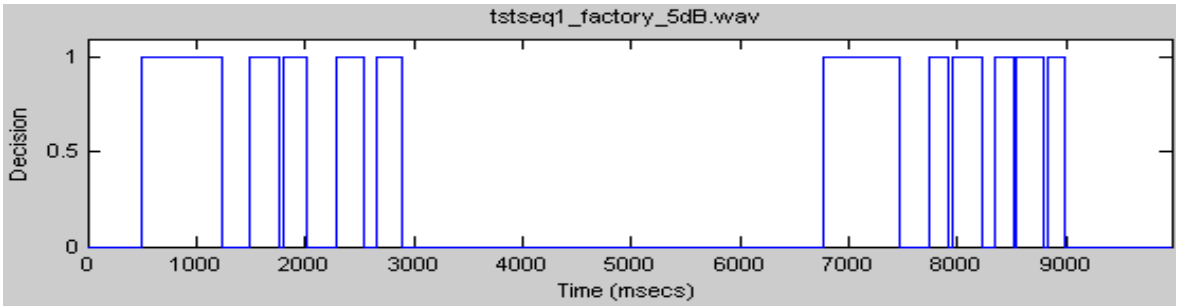

Figure 4.69 : Courbe VAD du vecteur de test "tstseq1.wav" bruité à 5dB avec du bruit d'usine avec temps de maintien et ajustement du seuil

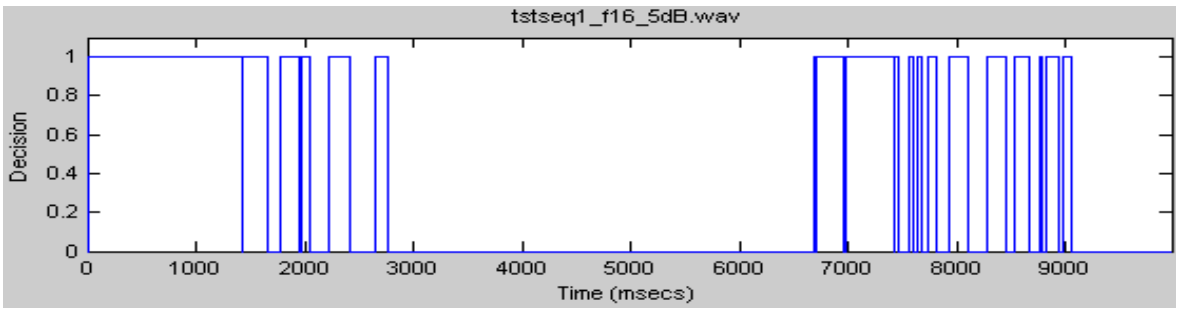

Figure 4.70 : Courbe VAD du vecteur de test "tstseq1.wav" bruité à 5dB avec du bruit d'avion

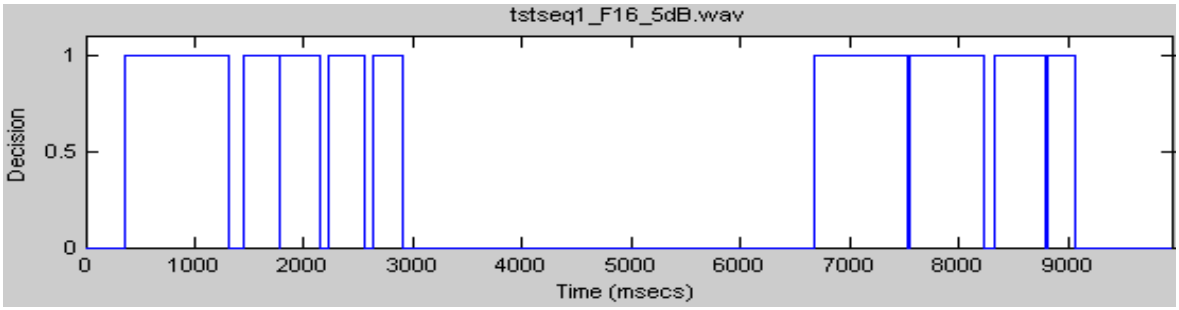

Figure 4.71 : Courbe VAD du vecteur de test "tstseq1.wav" bruité à 5dB avec du bruit d'avion avec temps de maintien et ajustement du seuil

b.2 Cas d'enregistrements de la base de données ARADIGIT

On a procédé au bruitage de l'enregistrement fa035m1.wav avec le reste des bruits sélectionnés, à différents SNR. Les Figure 4.72 – 4.79 illustrent l'apport de nos modifications (introduction d'un temps de maintien et ajustement du seuil) sur la prise de décision VAD pour l'enregistrement fa035m1.wav bruité à un SNR = 5 dB.

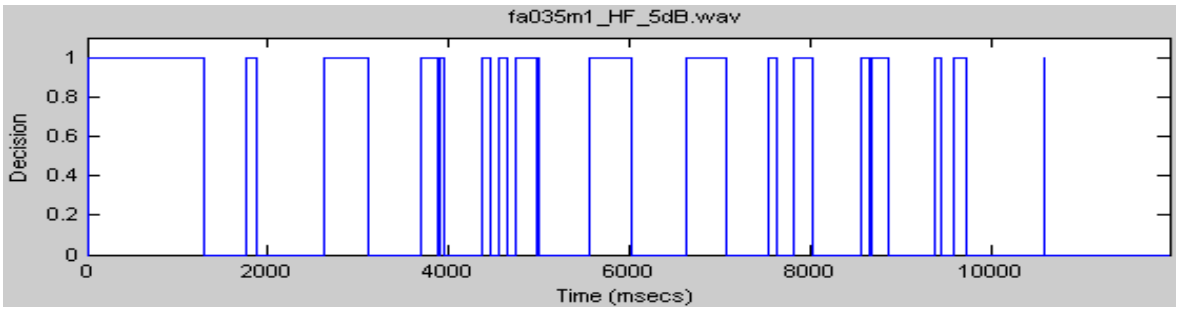

Figure 4.72 : Courbe VAD de l'enregistrement "fa035m1.wav" bruité à 5dB avec du bruit de canal radio HF

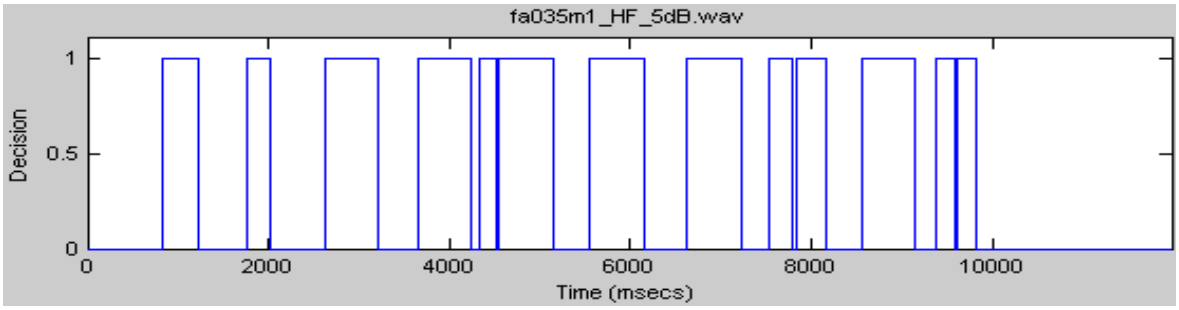

Figure 4.73 : Courbe VAD de l'enregistrement "fa035m1.wav" bruité à 5dB avec du bruit de canal radio HF avec temps de maintien et ajustement du seuil

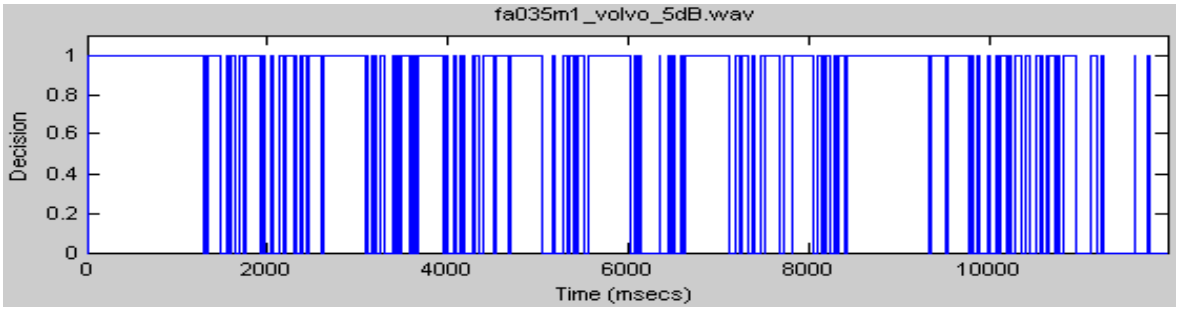

Figure 4.74 : Courbe VAD de l'enregistrement "fa035m1.wav" bruité à 5dB avec du bruit de voiture

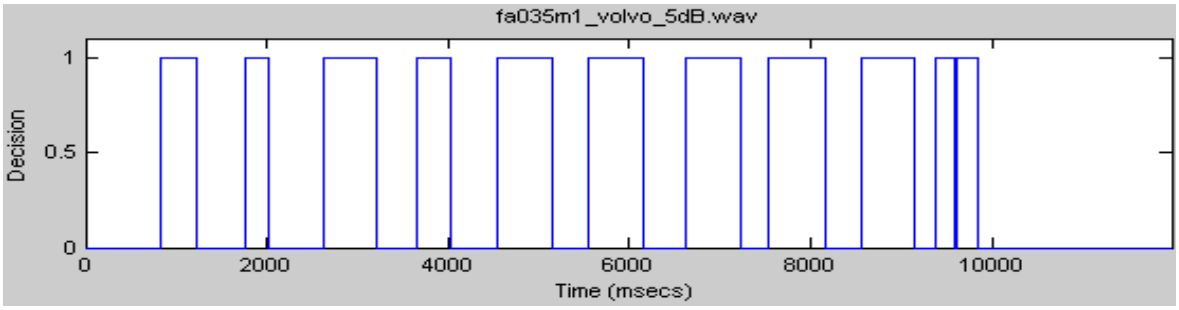

Figure 4.75 : Courbe VAD de l'enregistrement "fa035m1.wav" bruité à 5dB avec du bruit de voiture avec temps de maintien et ajustement du seuil

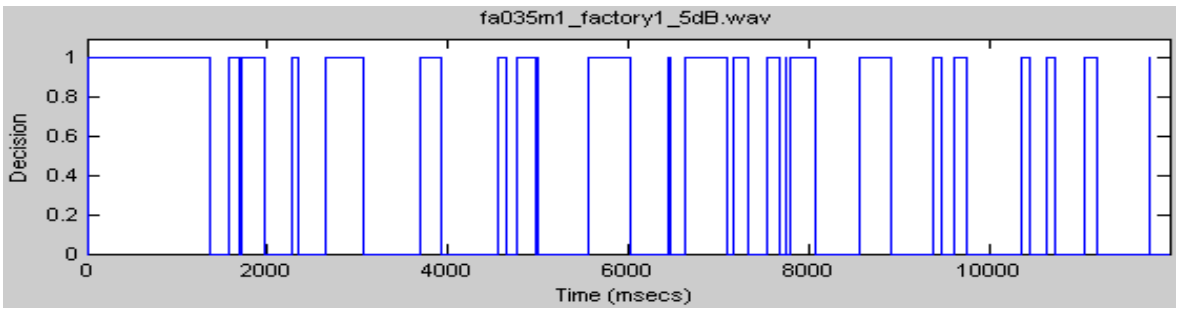

Figure 4.76 : Courbe VAD de l'enregistrement "fa035m1.wav" bruité à 5dB avec du bruit d'usine

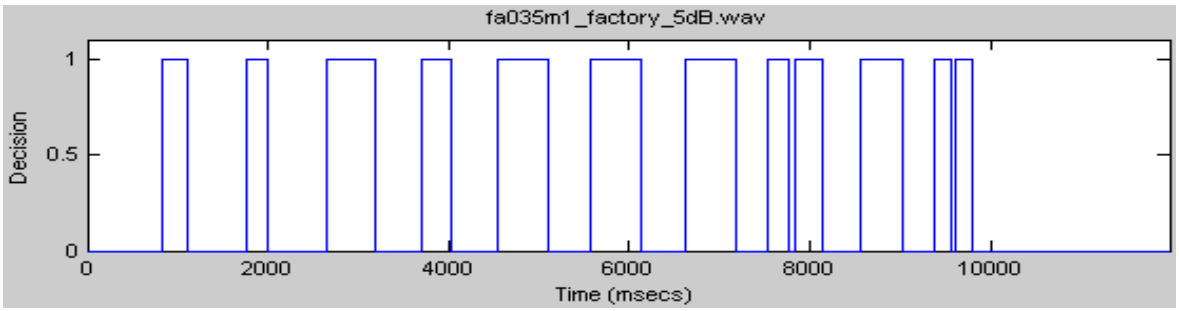

Figure 4.77 : Courbe VAD de l'enregistrement "fa035m1.wav" bruité à 5dB avec du bruit d'usine avec temps de maintien et ajustement du seuil

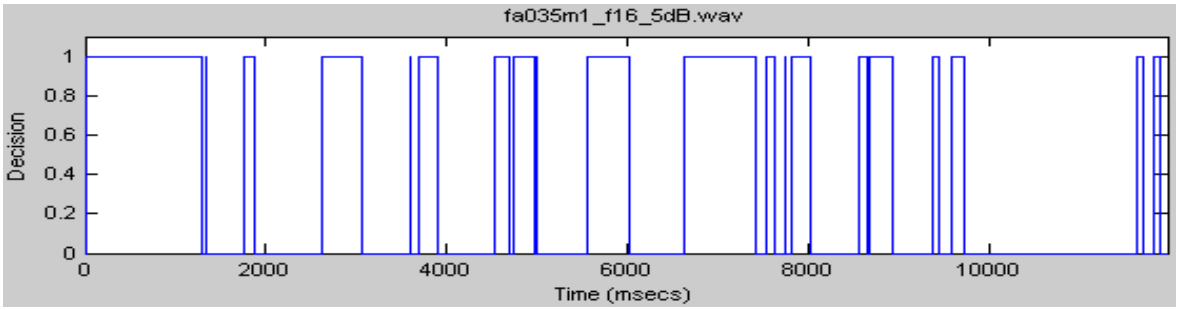

Figure 4.78 : Courbe VAD de l'enregistrement "fa035m1.wav" bruité à 5dB avec du bruit d'avion

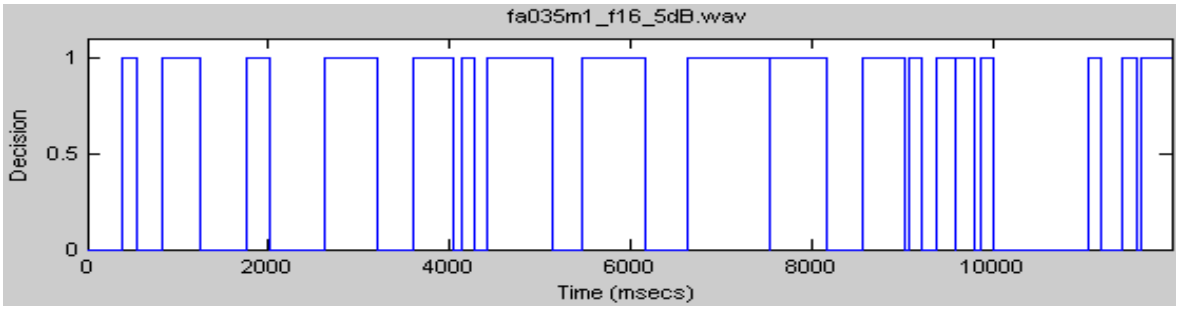

Figure 4.79 : Courbe VAD de l'enregistrement "fa035m1.wav" bruité à 5dB avec du bruit d'avion avec temps de maintien et ajustement du seuil

Enfin, on effectue un bruitage de l'enregistrement ma042m1.wav avec le reste des bruits sélectionnés, à différents SNR. Les Figure 4.80 – 4.87 soulignent l'apport de nos modifications (introduction d'un temps de maintien et ajustement du seuil) sur la prise de décision VAD pour l'enregistrement ma042m1.wav bruité à un SNR = 5 dB.

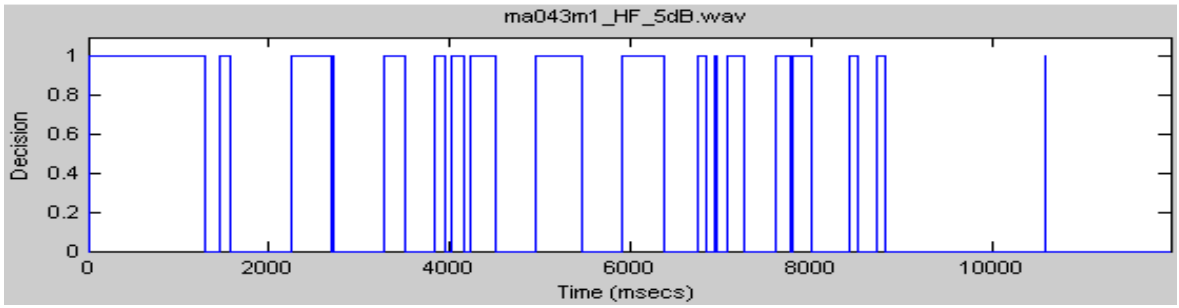

Figure 4.80 : Courbe VAD de l'enregistrement "ma042m1.wav" bruité à 5dB avec du bruit de canal radio HF

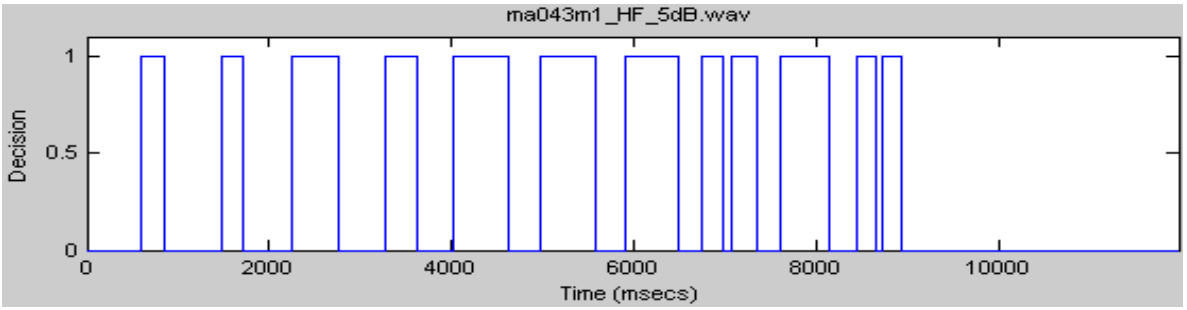

Figure 4.81 : Courbe VAD de l'enregistrement "ma042m1.wav" bruité à 5dB avec du bruit de canal radio HF avec temps de maintien et ajustement du seuil

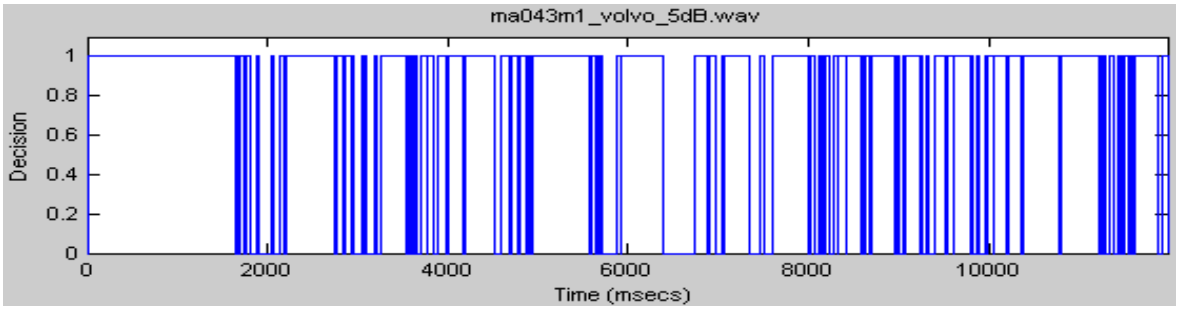

Figure 4.82 : Courbe VAD de l'enregistrement "ma042m1.wav" bruité à 5dB avec du bruit de voiture

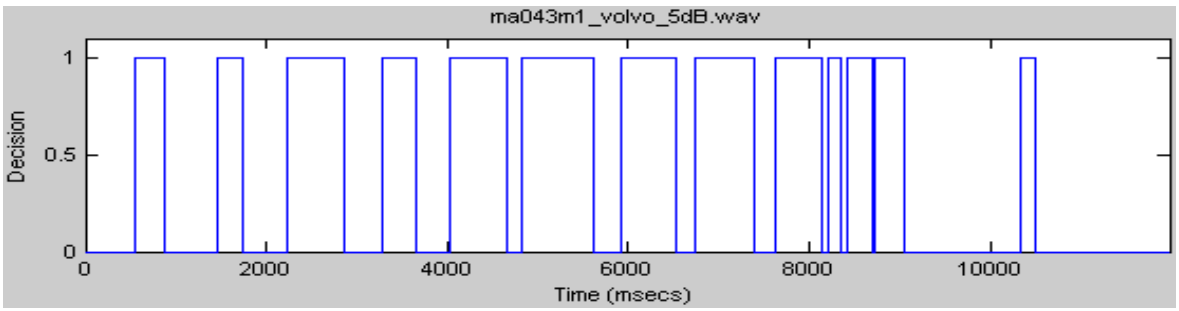

Figure 4.83 : Courbe VAD de l'enregistrement "ma042m1.wav" bruité à 5dB avec du bruit de voiture avec temps de maintien et ajustement du seuil

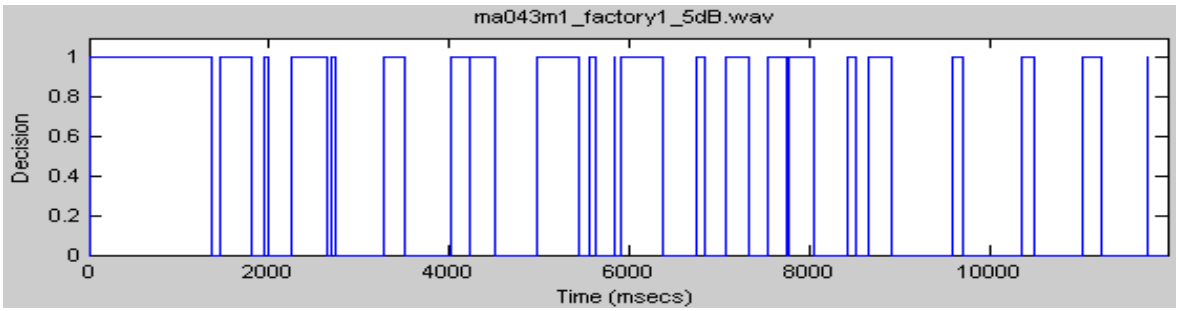

Figure 4.84 : Courbe VAD de l'enregistrement "ma042m1.wav" bruité à 5dB avec du bruit d'usine

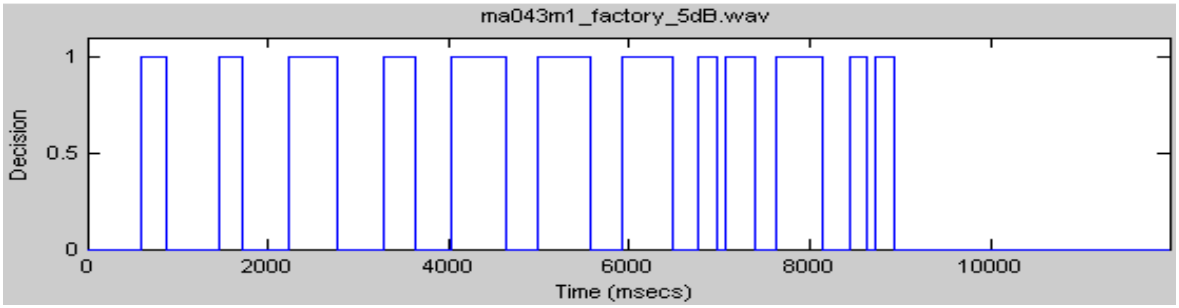

Figure 4.85 : Courbe VAD de l'enregistrement "ma042m1.wav" bruité à 5dB avec du bruit d'usine avec temps de maintien et ajustement du seuil

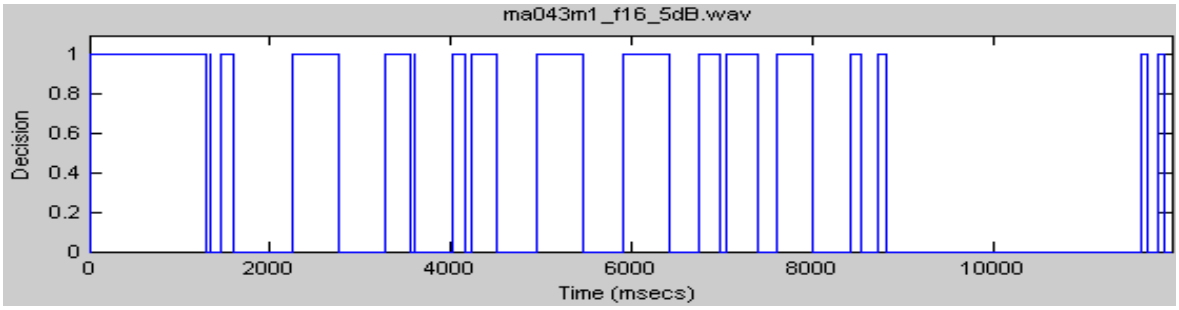

Figure 4.86 : Courbe VAD de l'enregistrement "ma042m1.wav" bruité à 5dB avec du bruit d'avion

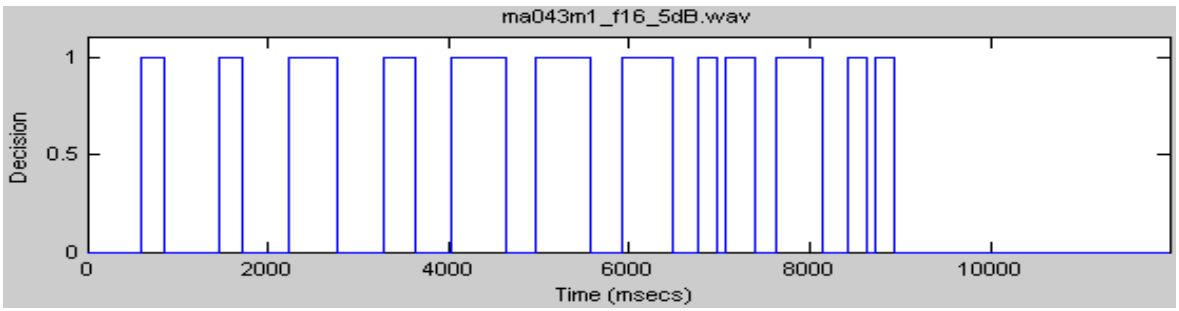

Figure 4.87 : Courbe VAD de l'enregistrement "ma042m1.wav" bruité à 5dB avec du bruit d'avion avec temps de maintien et ajustement du seuil

Discussion

La méthode de lissage de la recommandation G.729 annexe B présente quelques manquements notamment dans la quatrième étape de lissage qui définit un essai qui s'appuie uniquement sur la différence d'énergie entre la trame actuelle et l'énergie du bruit de fond pour statuer sur la prise de décision. Ainsi, elle ne tient pas compte des décisions antérieures alors que le but de la fonction de lissage est de tenir compte du caractère stationnaire à long terme du signal vocal. Ceci peut introduire le classement erroné de la parole en tant que bruit lorsque l'énergie du bruit de fond devient trop élevée.

Parallèlement à cela, on remarque que l'essai qui permet de décider en faveur du mode BRUIT, lorsque la différence entre le signal du moment et le bruit qui précède est trop petite peut produire un mauvais classement de la parole en tant que bruit. C'est le cas souvent en présence de sons dont les caractéristiques approchent celles du bruit comme les fricatives. Cette erreur de classification est accentuée en milieu fortement bruité.

Afin de palier à ces inconvénients, une substitution de la méthode de lissage définie dans la recommandation G.729 annexe B par une méthode de temps de maintien de la décision VOIX durant N décisions BRUIT successives à été faite. L'apport de cette modification a été prouvé, il en résulte une diminution de l'oscillation de la décision entre parole et bruit observée sur les nouvelles courbes de détection d'activité vocale extraites après introduction de cette modification. Des notes d'opinion objective (PESQ) viennent appuyer ces résultats confortant l'approche que nous avons proposée.

Néanmoins l'introduction de ce temps de maintien affecte l'optimisation de la bande passante, qui est un aspect très important dans les applications de voix sur IP. Cet impact est cependant assez minime dans un environnement calme mais tend à s'accroitre dans un environnement fortement bruité, où le bruit est assimilé assez souvent à de la parole.

Afin de préserver les avantages en termes d'optimisation de la bande passante du module VAD, un ajustement du seuil d'énergie (fixé à 15 dB dans l'actuelle recommandation), a été réalisé. Ce seuil a un poids important dans l'algorithme de détection d'activité vocale. Selon que l'énergie de la trame actuelle soit supérieure ou inferieure à ce seuil, une décision VOIX ou BRUIT définitive est prise et une initialisation des caractéristiques du bruit de fond est effectuée durant les 32 premières trames. Passé les 32 premières trames la décision de détection d'activité vocale initiale est forcée à 0 ou passe par des régions de décision à frontières multiples suivant que l'énergie moyenne de la trame actuelle soit supérieure ou inferieure à ce seuil.

La méthode que nous proposons effectue une adaptation de ce seuil en se basant sur la valeur moyenne de l'énergie de trames dans la pleine bande de fréquences calculée durant les 32 premières trames (temps d'acquisitions des caractéristiques du bruit de fond). Cette adaptation doit s'accompagner d'un calibrage du temps de maintien pour préserver la qualité du signal vocal perçue.

La méthode d'ajustement du seuil que nous avons proposée vise à séparer les bruits ambiants de niveau élevé de la parole proprement dite. Donc, son objectif premier est d'optimiser la capacité. La méthode du temps de maintien vise quand à elle l'amélioration de la qualité de la parole transmise. La combinaison de ces deux méthodes que nous avons appliquées à nos signaux vocaux en milieu bruité a pour but d'optimiser la capacité tout en préservant la qualité de la parole.

Les résultats expérimentaux obtenus avec des vecteurs de tests de la recommandation G.729 annexe B de l'ITU ainsi que la base de données Arabes ARADIGIT en environnement calme et bruité montre l'efficacité de l'amélioration du VAD que nos avons proposée. Les mesures objectives effectuées ont mis en évidence une amélioration de la qualité de la parole transmise. L'évaluation statistique sur les trames de paroles, trames SID et les trames non transmises a débouché sur une préservation, sinon une légère augmentation de la capacité (consommation de la bande passante) compensée par une amélioration de la qualité de la parole transmise.

Conclusion Générale

Conclusion Générale

Afin d'optimiser la capacité du canal de transmission, sachant qu'une partie non négligeable de l'échange verbale entre deux personnes communicantes est constituée de silence, l'adjonction d'un module de détection de l'activité vocale (VAD) au codec s'est récemment imposée dans les réseaux de communications, en particulier les télécommunications mobiles.

Dans ce travail, nous avons d'abord évalué l'efficacité de la VAD, après avoir implémenté le codec G729B. Cette évaluation a été effectuée au moyen de vecteurs de test fournis par la recommandation ITU G729 annexe B, ainsi qu'avec de la parole issue de la base de données arabe ARADIGIT.

En environnement calme, la VAD répond favorablement à toute apparition de zones de silence dans le flot de parole. Elle est seulement prise en défaut lorsqu'il y'a présence de longue occlusives non voisées dans le segment de parole. Ce qui occasionne une instabilité de la fonction VAD préjudiciable à la bonne qualité d'écoute, et même à l'intelligibilité de la parole reçue.

Par ailleurs, l'investigation que nous avons menée, notamment en milieu bruité, fait apparaitre certains dysfonctionnements de la détection d'activité vocale. En effet, suivant la nature du bruit et suivant le SNR, la courbe VAD assimile assez souvent des portions de bruits ambiant à de la parole faisant perdre au codec son efficacité quand à l'optimisation de l'utilisation du canal. Dans certains cas, il y a même une perte d'information quand la parole est interprétée comme du bruit. Ces observations ont été mises en évidence en utilisant les vecteurs de test de la recommandation ITU G729 annexe B ainsi qu'avec des segments de parole de la base ARADIGIT, le tout bruité avec la base NOISEX-92.

Afin de remédier à ces défaillances nous avons proposé dans ce mémoire des modifications dans l'algorithme VAD, afin d'améliorer la discrimination silence / parole. Les modifications introduites dans l'algorithme VAD concernent principalement :

- Une substitution de la méthode de lissage actuelle par un retard calculé introduit avant le passage du mode PAROLE au mode BRUIT, qui correspond à N décisions BRUIT détectées consécutivement.
- Ajustement d'un seuil d'énergie de détection en se basant sur la valeur moyenne de l'énergie de trame dans la pleine bande de fréquences calculée durant les 32 premières trames.

Les modifications intégrées à l'algorithme VAD de l'annexe B de la recommandation G729B de l'ITU, offrent une meilleure qualité vocale, tout en maintenant une efficacité élevée d'utilisation de la largeur de bande, adaptée aux applications de transmission de la voix par paquets.

En perspective, nous pensons opérer des changements plus importants dans l'implémentation du codec G729B. Ainsi, nous estimons dans une première étape qu'une analyse acoustique des segments phonémiques, et leurs durées, peuvent être mises à profit pour calculer au plus près les zones de silence.

Nous estimons également qu'une bufferisation des trames successives de silence et l'application d'un seuil de durée, tiré de l'analyse acoustique des phonèmes, peuvent aider à une meilleure séparation silence / parole.

Dans une perspective à moyen terme, l'intégration de nouveaux modules spécifiques à la VAD pour la réalisation de ces taches, peut accroitre davantage l'efficacité des codecs vocaux et optimiser la capacité des canaux dans les réseaux de communication.

Annexe A :

Schémas de principe du codec CS-ACELP G.729

Annexe A :

Schémas de principe du codec CS-ACELP G.729

La recommandation G.729 de l'ITU normalise un codeur de parole CS-ACELP à 8 kbits/s [17]. Les deux figures suivantes donnent respectivement les schémas de principe du codeur et du décodeur de cette recommandation.

Le codeur G.729, dont la circulation des signaux est donnée en figure A.1, opère sur des trames de parole de 10 millisecondes, qui correspondent à 80 échantillons numérisés sur 16 bits pour une fréquence d'échantillonnage de 8 kHz [12]. Le signal de parole est analysé à chaque trame pour extraire les coefficients du filtre de Prédiction Linéaire (LP) du $10^{éme}$ ordre, qui sont convertis en lignes de raies spectrales (LSP) et numérisés par quantification vectorielle prédictive à deux étapes. Par la suite, les paramètres d'excitation tels que la période de pitch, les index ainsi que les gains des dictionnaires, fixe et adaptatif, sont estimés sur la base de sous-trames de 40 échantillons, soit 5 millisecondes. Le signal d'excitation est choisi au moyen d'une procédure de recherche par analyse et synthèse dans laquelle l'erreur, entre le signal vocal original et celui reconstruit, est minimisée en fonction d'une mesure de distorsion pondérée.

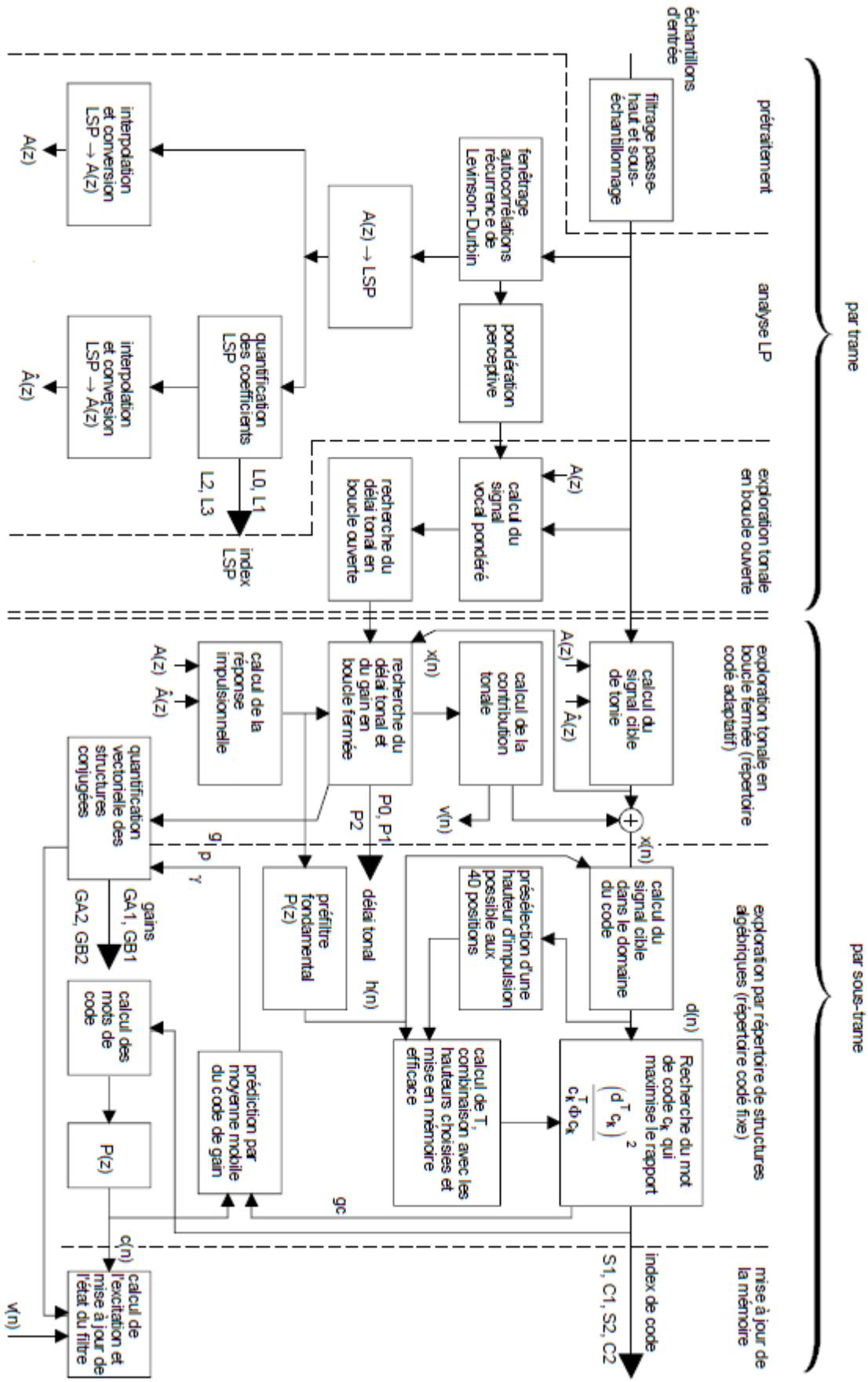

Figure A.1 : Codeur CS-ACELP G.729 (Recommandation G729 ITU)

Les paramètres de codage, correspondant à une trame vocale de 10 millisecondes, sont ensuite traités par le décodeur G.729 décrit par la figure A.2. Ainsi, Les coefficients LSP sont interpolés et reconvertis en coefficients de filtre de prédiction linéaire pour chaque sous-trame de 5 millisecondes; l'excitation est construite par combinaison des codes vectoriels adaptatif et fixe, multipliés par leur gain respectif; enfin, le signal vocal est reconstitué par le filtre de synthèse LP et amélioré à l'aide d'un bloc de post-traitement suivi d'un filtre passe-haut et d'un échantillonneur-normalisateur.

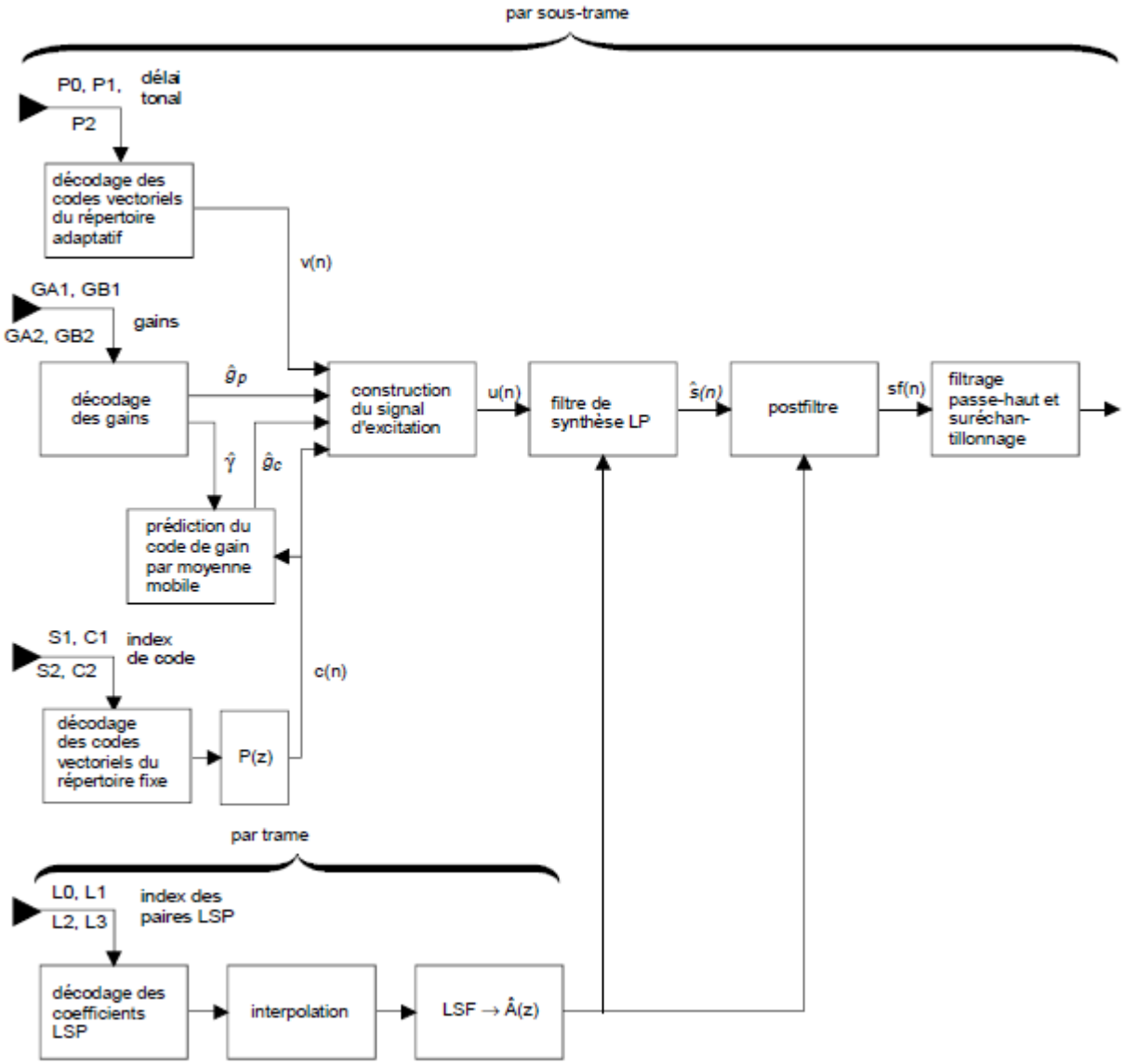

Figure A.2 : Décodeur CS-ACELP G.729 (Recommandation G729 ITU)

Bibliographie

Bibliographie

[1] D. O'Shaughnessey, ''Speech Communication - Human and Machine'', Addison-Wesley Publishing Company, 1987.

[2] D. B. Pisoni, and R. E. Remez, ''The Handbook of Speech Perception'', Blackwell Publishing Company, 2005.

[3] P. Roach, ''Phonetics'', Oxford University Press, 2001.

[4] R. Boite, H. Bourlard, T. Dutoit, J. Hancq, and H. Leich, ''Traitement de la parole'', Edition Presses Polytechniques Universitaire Romande, 2000.

[5] A. M. Kontoz ''Digital speech : coding for low bit rate communication systems '', Wiley Publishing Company, 2004.

[6] G. Baudoin, J. Cernocky, P. Gournay, and G. Chollet, ''Codage de la Parole à Bas et Très Bas Débits'', Annale des Télécommunications, No 9-10, Vol. 55, pp. 462-482, 2000.

[7] ITU-T Recommandation P.800, ''Methods for subjective determination of transmission quality'', August 1996.

[8] ITU-T Recommandation P.862, ''Perceptual evaluation of speech quality (PESQ) : An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs'', February 2001.

[9] M. R. Schroeder and B. S. Atal, ''Code-Excited Linear Predictive (CELP) : High Quality Speech at Very Low Bit Rates'', Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing, Vol. 10, pp. 937-940, March 1985.

[10] ITU-T Recommandation G.729, ''Coding of Speech at 8 kbits/s using Conjugate Algebraic Code-Excited Linear Prediction (CS-ACELP)'', June 1995.

[11] ITU-T Recommandation G.729 - Annex A, ''Reduced complexity 8 kbit/s CS-ACELP speech codec'', Novembre 1996.

[12] G. Madre, ”Application de la transformée en nombres entiers à l’étude et au développement d’un codeur de parole pour transmission sur réseaux IP”, Doctorat de l’Université de Bretagne Occidentale, Mention Electronique, Octobre. 2004.

[13] R. Salami, C. Laflamme, J. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham ‘’Design and Description of CS-ACELP: A Toll Quality 8 kb/s Speech Coder‘’, IEEE Transactions on Speech and Audio Processing, Vol. 6, No. 2, pp. 116-130, March 1998.

[14] V. Kapur, M. M. Raghuvanshi, and A. B. Maidamwar, ‘’Real Time Implementation of Speech codec G.729 Using CS-ACELP on TM 1000 VLIW DSP processor‘’, International Journal of Soft Computing and Engineering, Vol. 1, No. 5, pp. 46-49, November 2011.

[15] R. V. Cox and P. Kroon, ‘’Low Bit-Rate Speech Coders for Multimedia Communications‘’, IEEE Commun. Mag., Vol. 34, No. 12, pp. 34–41, December 1996.

[16] R. V. Prasad, A. Sangwan, H. S. Jamadagni, M. C. Chiranth, R. Sah, and V. Gaurav, ‘’Comparison of Voice Activity Detection Algorithms for VoIP‘’, 7th International Symposium on Computers and Communications, 2002, Italy.

[17] ITU-T Recommandation G.729 - Annex B, ‘’Silence Compression Scheme For Optimized For Termnals Conforming To Recommendation‘’, November 1996.

[18] A. Benyassine, E. Shlomot, and H. Y. Su, ‘’ITU-T recommendation G.729 Annex B : A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications‘’, IEEE Commun. Mag., vol. 35, pp. 64-73, September 1997.

[19] P. Setiawan, S. Schandl, H. Taddei, H. Wan, J. Dai, L. Zhang, D. Zhang, J. Zhang, and E. Shlomot, ‘’On the ITU-T G.729.1 silence compression scheme‘’, 16th European Signal Processing Conference, 2008, Lausanne.

[20] ETSI, GSM 06.94, ‘’Digital cellular telecommunication system (Phase 2+); voice activity detector VAD for adaptive multi rate (AMR) speech traffic channels; general description‘’, Tech. Rep. V.7.0.0, February 1999.

[21] J. Stadermann, V. Stahl, and G. Rose, ‘’Voice Activity Detection in Noisy Environments‘’, Eurospeech 2001, Scandinavia.

[22] T. Ravichandran, and K. D. Samy, ‘’Performance Enhancement on Voice Using VAD Algorithm and Cepstral Analysis‘’, Journal of Computer Science, Vol. 2, N° 11, pp. 835-840, 2006.

[23] H. Farsi, M.A. Mozaffarian, and H. Rahmani ‘’A Novel Method to Modify VAD Used in ITU-T G.729B for Low SNRs‘’, International Journal of Computers and Communications, Vol. 2, N° 1, pp. 20-29, 2008.

[24] S. I. Kang, Q. H. Jo, and J. H. Chang,‘’Discriminative Weight Training for a Statistical Model-Based Voice Activity Detection‘’, IEEE Signal Processing Letters, Vol. 15, pp. 170–173, 2008.

[25] S. S. Ahn and Y. C. Lee ‘’ Improved Statistical Model-Based VAD Algorithm With an Adaptive Threshold‘’, Journal of the Chinese Institute of Engineers, Vol. 29, No. 5, pp. 783-789, 2006.

[26] Z. Tuske, P. Mihajlik, Z. Tobler and T. Fegyo, ‘’ Robust Voice Activity Detection Based on the Entropy of Noise-Suppressed Spectrum‘’, Interspeech, 2005, Portugal**.**

[27] C. G. Babu, P.T. Vanathi, R. Ramachandran, M. Senthil Rajaa, and R. Vengatesh ‘’A Comprehensive Analysis of Voice Activity Detection Algorithms for Robust Speech Recognition System Under Different Noisy Environment‘’, (IJCNS) International Journal of Computer and Network Security, Vol. 1, No. 2, November 2009.

[28] E. Nemer, R. Goubran, and S. Mahmoud, ‘’Robust Voice Activity Detection Using Higher-Order Statistics in the LPC Residual Domain‘’, IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 3, pp. 217–231, march 2001.

[29] V. Gilg, C. Beaugeant, M. Schonle, and B. Andrassy, ‘’Methodology for The Design of a Robust Voice Activity Detector for Speech Enhancement‘’, International Workshop on Acoustic Echo and Noise Control (IWAENC2003), September 2003, Kyoto, Japan.

[30] A. Torre, J. Ramirez, C. Benitez, J. C. Segura, L. Garcia, and A. J. Rubio, ‘’Noise Robust Model-Based Voice Activity Detection‘’, Interspeech 2006, ICSLP.

[31] J. Sohn, N. S. Kim, and W. Sung, ‘’A Statistical Model-Based Voice Activity Detection‘’, IEEE Signal Processing Letters, Vol. 6, No. 1, pp 1-3, January 1999.

[32] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano, ‘’Performance Evaluation and Comparison of G.729/AMR/Fuzzy Voice Activity Detectors‘’, IEEE Signal Processing Letters, Vol. 9, No. 3, pp 85-88, March 2002.

[33] S. G. Tanyer and H. Özer, ‘’Voice Activity Detection in Nonstationary Noise’’, IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 4, pp 479-482, July 2000.

[34] R. V. Prasad, R. Muralishankart, S. Vijay, H. N. Shankar, P. Pawelczak, and I. Niemegeers, ‘’Voice Activity Detection for VoIP-An Information Theoretic Approach‘’, Global Telecommunications Conference, 2006. GLOBECOM '06. IEEE.

[35] J. H. Chang, N. S. Kim, and S. K. Mitra, ‘’Voice Activity Detection Based on Multiple Statistical Models‘’, IEEE Transactions on Signal Processing., Vol. 54, No. 6, pp.1965-1976, June 2006.

[36] F. Beritelli, S. Casale, and A. Cavallaro, ‘’A Robust Voice Activity Detector for Wireless Communications Using Soft Computing‘’, IEEE J. Select. Areas Commun., Vol. 16, No. 9, pp. 1818–1829, December 1998.

[37] G. Madre, ”Application de la transformée en nombres entiers à l’étude et au développement d’un codeur de parole pour transmission sur réseaux IP”, Doctorat de l’Université de Bretagne Occidentale, Mention Electronique, Octobre. 2004.