UNIVERSITÉ DE SHERBROOKE
Faculté de génie
Département de génie électrique et de génie informatique

Détection et modification des transitoires d'un signal de parole dans le but de rendre un codec plus robuste aux pertes de paquets

Thèse de doctorat
Spécialité : génie électrique

Catherine LEMYRE

Jury : Roch LEFEBVRE (directeur)
Milan JELINEK (co-directeur)
Philippe MABILLEAU
Philippe GOURNAY
Mohammed CHIBANI

Sherbrooke (Québec) Canada Janvier 2012

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-93256-8

Our file Notre référence
ISBN: 978-0-494-93256-8

À Fred

RÉSUMÉ

Résumé

Pour transmettre les signaux de parole de façon efficace, ces derniers sont compressés et transmis en trames typiquement de 10 à 20 ms. Lors de la transmission des trames, il arrive que ces dernières soient perdues. Lors de la reconstruction du signal au décodeur, il est préférable de remplacer les trames perdues par un signal qui se rapproche le plus possible du signal manquant. Le signal perdu est souvent reconstruit en se basant sur l'information des dernières trames reçues, puisque, de façon générale, les propriétés statistiques du signal de parole évoluent relativement lentement d'une trame à la suivante.

Les signaux de parole peuvent être classés en différentes catégories (parole voisée, non-voisée, transitoire, etc.). Afin de mieux exploiter les caractéristiques de chaque catégorie, il est pertinent d'appliquer un classificateur à chaque trame de signal. Cette classification des signaux permet un meilleur camouflage des trames perdues, optimisé pour les différentes classes.

La classification des trames est parfois imprécise lors des transitions entre une trame non-voisée et une trame voisée. Ces erreurs de classification entrainent de mauvaises reconstructions de signal lors des pertes de trames.

Pour pallier ces erreurs, cette thèse propose un nouvel algorithme robuste qui identifie les trames critiques et qui applique la classification appropriée. Pour les trames dont les propriétés ne correspondent pas exactement à l'une des classes disponibles, une modification transparente du signal est appliquée pour rendre ces trames conformes à la classification proposée. Ces modifications permettent d'obtenir une meilleure reconstruction du signal si les trames suivantes sont perdues.

Mots-clés : codeur prédictif, camouflage, opérateur de teager, détection des trames transitoires, modification des trames transitoires

REMERCIEMENTS

REMERCIEMENTS

Un chemin qui n'était pas tracé, quatre années et demie qui se sont prolongées sur plus de 10 ans. Un passage qui devait être plus bref, qui s'est transformé en quelque chose de plus marquant dans ma vie.

Merci vers l'au-delà à Fred qui a su être un vrai ami. Les plus belles années de ma vie, nous les avons partagées ensemble. Que ce soit dans les laboratoires ou autour d'une bonne table, les rires étaient toujours présents. Désormais, il manquera toujours quelqu'un dans les moments importants de ma vie. Des gens comme toi dans une vie, on en rencontre trop peu. Vas, Voleur !

Merci à mon directeur Roch Lefebvre. Merci pour ta grande générosité quand vient le temps de partager tes connaissances. J'ai eu la chance d'avoir un excellent directeur et je te suis très reconnaissante pour toutes ces belles années passées à travailler dans ton laboratoire. Je serai toujours fière d'avoir fait partie de tes étudiantes.

Merci à Milan d'avoir codirigé mon projet de doctorat. Ta force tranquille a été un grand soutien pour moi, particulièrement à travers tes corrections attentives et précieuses. Merci beaucoup pour ta patience et ton écoute et merci aussi pour toutes ces connaissances partagées.

Au labo, il y a un temps pour être sérieux, il y a aussi un temps pour rire et s'amuser. Merci à Jimmy pour toutes ces conversations, les cafés et les moments où j'ai eu besoin de me changer les idées.

Philippe, Bruno, Danielle et les autres étudiants du groupe merci pour les rires, les cafés, le chocolat ;-), vous m'avez donné le goût de venir travailler au labo.

Ceux qui gravitent autour de nous sont aussi importants et contribuent aussi à notre réussite.

Merci à Nicolas, mon conjoint. Merci de m'avoir permis de rester à Sherbrooke pendant presque deux ans. Ton support moral et tes nombreux projets qui m'ont permis de m'évader et de revenir toujours avec des nouvelles idées les lundis matin. Merci beaucoup de ta patience et de tes encouragements.

Un énorme merci à mes parents Robert et Louise pour leur soutien continu et leurs encouragements. Merci aussi aux autres membres de ma famille pour leur motivation.

And in the end, the love you take is equal to the love you make. The Beatles

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

CHAPITRE 1

INTRODUCTION

Le codage de la voix permet la compression d'un signal de parole dans le but de le transmettre ou de le stocker. La grande majorité des codeurs de parole à bas débit traitent le signal de parole en trames de courtes durées (de 10 à 20 ms), période pendant laquelle certains paramètres du signal sont quasi-stationnaires. L'information du signal de parole contenue dans cet intervalle est d'abord analysée et compressée par le codeur pour être ensuite transmise ou stockée. Dans le cas où le signal compressé est transmis à travers un réseau, des trames peuvent être perdues, c'est-à-dire qu'une partie de l'information compressée ne parvient pas au décodeur. Les pertes de trames peuvent être causées par une trame qui arrive en retard (dans une application temps réel), par une trame corrompue ou par une trame qui n'arrive tout simplement pas. Une perte de trame peut entrainer la dégradation du signal au moment du décodage. Le décodeur doit alors remplacer l'information manquante, et ce de façon cohérente pour éviter une perte de qualité du signal de parole décompressé.

Il existe deux catégories de codeurs de parole : les codeurs prédictifs et les codeurs non-prédictifs. Les codeurs prédictifs sont aujourd'hui les plus utilisés. Le PCM (de l'anglais *pulse code modulation*), un codeur non-prédictif, est toutefois encore très populaire lorsque le débit n'est pas une contrainte. Ce standard de l'ITU, le G.711, est utilisé pour la bande téléphonique principalement dans les applications filaires. Il s'agit d'un codeur qui utilise la modulation d'impulsion codée pour représenter les échantillons du signal et pour ensuite les compresser. La représentation PCM se fait sur 8 bits selon une échelle logarithmique. En Amérique du Nord et au Japon, la loi μ est utilisée alors que l'Europe et le reste du monde emploient la loi A. Le signal d'entrée du codeur PCM est échantillonné à 8 kHz et le débit de sortie est de 64 kbps. Le PCM encode chaque échantillon indépendamment des autres, aucune information transmise n'est dépendante des échantillons subséquents ou suivants ce qui en fait un codeur non-prédictif. Les échantillons peuvent être regroupés en trames pour la transmission et le contenu des trames est indépendant d'une trame à la suivante.

Lorsque survient une perte de trames dans le standard G.711, une partie de l'information passée sert à extrapoler l'information perdue [ITU-T G.711, 1999]. Dans un premier temps, une recherche de la fréquence fondamentale du signal (pitch) est faite. La fréquence

fondamentale est l'inverse de la période avec laquelle le signal voisé se répète. Pour trouver le pitch du signal, une recherche par corrélation est faite à travers les dernières bonnes trames reçues. Le camouflage[1] des trames perdues se fait en répétant la dernière période de pitch reçue, jusqu'à ce que la totalité de la trame perdue soit remplie. Pour éviter des artéfacts, une fenêtre de recouvrement (*overlap-add*) permet de lisser la transition entre la dernière bonne trame reçue et la nouvelle trame reconstruite. Le même recouvrement est fait entre la fin de la trame reconstruite et la prochaine bonne trame reçue. Le standard propose des trames de 10 ms [ITU-T G.711, 1999], bien que d'autres longueurs de trames puissent être utilisées.

Un codeur non-prédictif, comme le G.711, ne propage pas d'erreur dans les bonnes trames reçues suite à une trame perdue, puisque le contenu de chaque trame est codé de façon indépendante. Seule la trame suivant une trame perdue est légèrement affectée par le recouvrement entre la trame perdue et la trame suivante bien reçue.

La majorité des codeurs de parole sont des codeurs prédictifs utilisant le modèle CELP. Le modèle CELP (Code-Excited Linear Prediction) a été présenté pour la première fois par Schroeder et Atal [Schroeder et Atal, 1985] et il est illustré à la figure 1.1. Le modèle CELP est une représentation mathématique de la production de la parole chez l'humain. Le prédicteur à court-terme est une approximation du pharynx et de la bouche dans l'anatomie humaine, alors que le prédicteur long-terme représente la vibration des cordes vocales (50-400Hz).

Le prédicteur à court-terme modélise l'enveloppe spectrale du signal. Au codeur, le filtre de prédiction court-terme est obtenu suite à une analyse de prédiction linéaire (analyse LP) et le filtre est composé typiquement de 8 à 10 coefficients pour un signal échantilloné à 8 kHz. Le signal résiduel (ou le résidu) est obtenu suite au filtrage de la parole par le filtre tout-pôle de prédiction court-terme.

Lorsque l'analyse LP est complétée, le signal résiduel obtenu contient encore beaucoup de redondance. Pour l'enlever, le signal résiduel est traité par une analyse de pitch (prédicteur long-terme de la figure 1.1). Le pitch peut aussi être déterminé à l'aide du dictionnaire adaptatif. Ce dictionnaire est construit à l'aide du passé du signal d'excitation du filtre de synthèse. La recherche dans le dictionnaire adaptatif se fait en boucle fermée, c'est-à-dire que la recherche se fait à travers tous les vecteurs formés à partir du passé de l'excitation qui est généralement limité à la longueur maximale considérée pour une période de pitch. Le vecteur qui correspond à l'erreur minimale entre le signal original et le signal synthétisé

[1]Le terme camouflage est utilisé dans le texte pour remplacer le terme anglais *concealment*

dans le domaine perceptuel est choisi. Le principe qui consiste à synthétiser le signal du décodeur à l'encodeur pour optimiser le choix d'un ou de plusieurs paramètres se nomme analyse-par-synthèse.

Le dictionnaire innovateur sert à modéliser l'erreur résiduelle, il contient une quantité définie de vecteurs pouvant représenter ce signal. La recherche du vecteur qui minimise l'erreur entre le signal de synthèse et le signal de parole se fait aussi en boucle fermée en appliquant le principe d'analyse-par-synthèse. L'index du vecteur choisi est transmis au décodeur. Le filtre perceptuel permet de pondérer le signal d'erreur (différence entre le signal de parole original et le signal résiduel), de sorte que le bruit de quantification soit masqué par le signal. Ce filtre permet de contrôler la forme spectrale du bruit de quantification, de sorte qu'il ait une intensité plus faible dans les zones spectrales de faible intensité du signal original, et qu'il ait une intensité plus forte dans les zones spectrales de forte intensité du signal original.

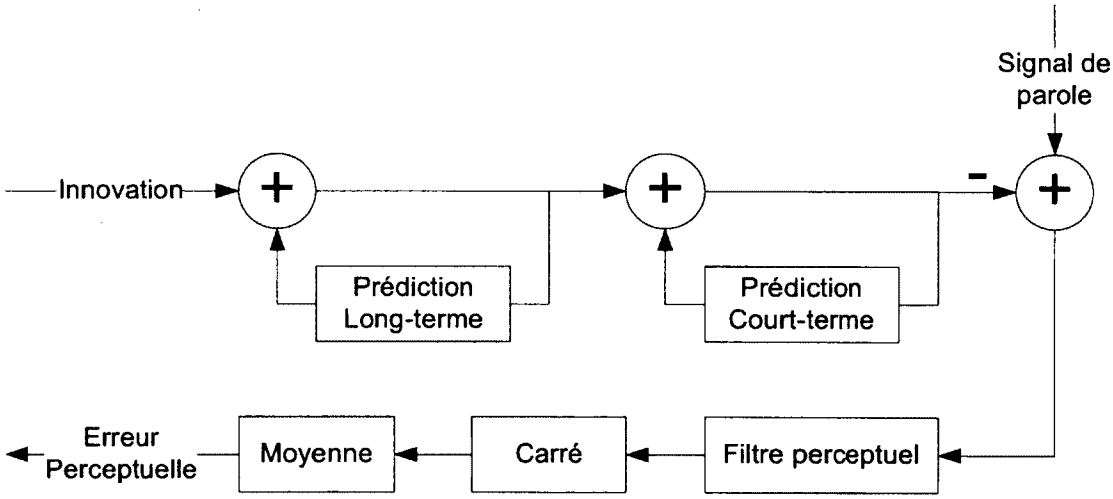


Figure 1.1 Modèle CELP proposé par Schroeder et Atal

En résumé et tel qu'illustré à la figure 1.2, à chaque séquence provenant du dictionnaire innovateur est additionné la contribution du dictionnaire adaptatif. Le signal d'excitation obtenu est filtré par le prédicteur court-terme. Pour alléger les calculs, le filtre perceptuel peut être appliqué directement au signal d'entrée et au signal de sortie du filtre de synthèse. Le signal de parole est comparé au signal de sortie du filtre de synthèse et la moyenne des carrés du signal d'erreur est calculée. L'index du vecteur du dictionnaire innovateur et du dictionnaire adaptatif ayant obtenant l'erreur la plus faible, ainsi que les gains associés à ces deux dictionnaires sont transmis. Le débit des codeurs prédictifs est moindre que celui du PCM ; typiquement 6 à 32 kbps contre 64 kbps.

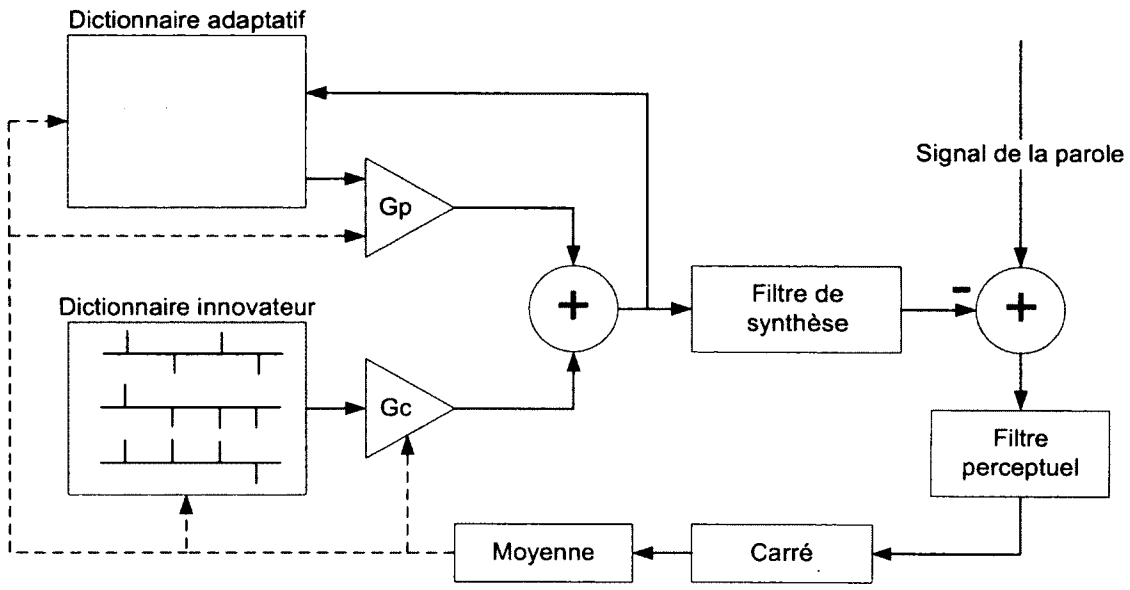


Figure 1.2 Modèle CELP avec dictionnaire adaptatif

Au décodeur, le dictionnaire innovateur modélise le contenu imprédictible (à court ou à long terme) du signal (par exemple tout ce qui est discontinu) alors que le dictionnaire adaptatif modélise la corrélation à long terme du signal (voir figure 1.3). Un gain est appliqué à chacune des sorties des dictionnaires et la somme est filtrée par le filtre de synthèse.

Une technique possible lorsqu'il y a une perte de trame est d'utiliser au décodeur le même filtre de synthèse qu'à la dernière bonne trame reçue et d'atténuer les gains du dictionnaire innovateur et du dictionnaire adaptatif [G.729, 2007]. Dans le cas où la dernière trame reçue est voisée (parole), la contribution du dictionnaire innovateur est mise à zéro, alors que la contribution du dictionnaire adaptatif est construite en utilisant la valeur de pitch de la trame précédente (ou une valeur de pitch extrapolée). Dans le cas où la dernière trame reçue est non-voisée (parole, silence ou bruit), la contribution du dictionnaire adaptatif est mise à zéro et la contribution du dictionnaire innovateur est choisie au hasard.

Comme les différents filtres sont mis à jour en fonction des informations contenues dans le passé du signal, et que le camouflage met à jour ces informations en se basant sur un contenu différent du codeur, cela entraine une désynchronisation entre le codeur et le décodeur. La désynchronisation peut causer une dégradation dans le signal et cette dégradation peut se propager sur quelques trames.

Le camouflage approprié diffère selon le type de trame perdue (voisée, non-voisée, transitoire, etc). Un classificateur qui identifie à quel type une trame appartient peut être

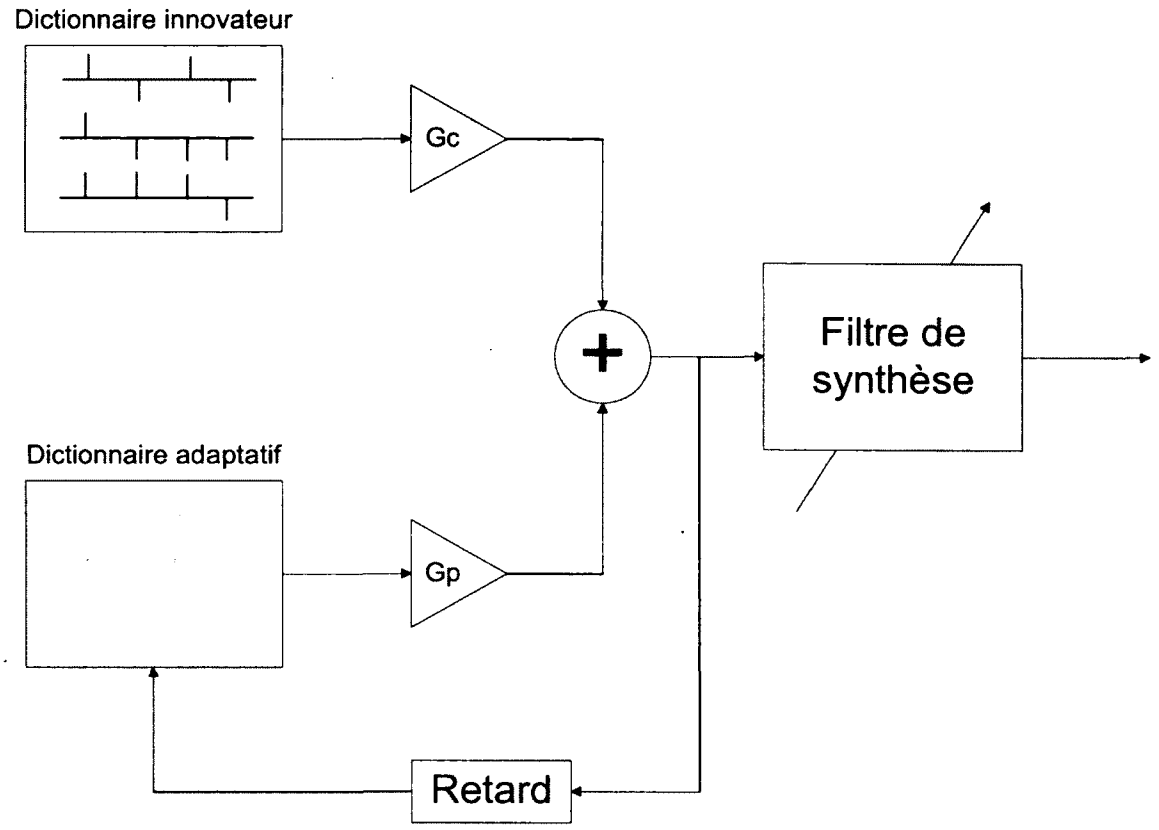


Figure 1.3 Modèle du décodeur CELP

intégré au codeur prédictif. Il est important pour le camouflage de bien classer la trame transitoire entre une trame non-voisée et une trame voisée. Cette trame de transition aussi appelée *onset* marque le début d'un signal voisé et **le terme transitoire, qui est employé dans le reste du document fait référence à cette transition de non-voisé à voisé.** Le bon positionnement de la trame transitoire permet de faire un bon usage des stratégies de camouflage mises de l'avant dans les codeurs prédictifs. Une mauvaise classification dont le résultat serait de classer comme transitoire une trame située avant ou après celle qui contient réellement la transitoire entraine un mauvais camouflage puisque la trame transitoire tranche entre l'utilisation d'un camouflage où la partie périodique du signal n'est pas répétée (trame non-voisée) et le camouflage où le contenu périodique de la trame précédente est répété (trame voisée). Si la trame transitoire est classée trop rapidement et que la trame suivante est perdue, le camouflage optimisé pour les signaux voisés sera appliqué suite à une trame qui en réalité est une trame non-voisée. Typiquement, ce camouflage répète la période de pitch trouvée par le prédicteur long terme, période de pitch qui est basée sur une corrélation fortuite trouvée dans un signal non-voisé plutôt que sur une véritable structure harmonique dans le signal. Cette répétition introduit une

harmonicité artificielle dans le signal. Si la transitoire est classée trop tardivement et que la trame suivant la transitoire réelle est perdue, le camouflage ne reconstruit pas la partie périodique du signal. De plus, l'énergie du signal reconstruit va tendre vers l'énergie de la trame précédente qui peut être beaucoup plus basse, lorsqu'il s'agit d'une trame de silence ou de bruit. Le début du voisement peut aussi être caractérisé par une hausse d'énergie à la fin de la trame et le camouflage qui amène cette énergie rapidement vers zéro crée parfois des artéfacts. Il survient également des cas où le début du voisement ne contient pas une période de pitch complète et bien construite pour assurer un bon camouflage, même si la trame de transition est détectée correctement. Comme la période de pitch n'est pas complète, elle ne peut pas être répétée si la trame suivante est perdue. Lorsqu'un tel cas se présente, il est en général préférable de corriger le résultat du classificateur et de classer la trame suivante comme étant la trame transitoire. Les trames qui contiennent un début de voisement insuffisant pour assurer un bon camouflage pourront être modifiées pour éviter le cas où si la trame suivante est perdue qu'il y ait une hausse d'énergie suivi d'une descente rapide de l'énergie vers zéro. La modification de ces trames sera abordée dans le chapitre 4.

Le travail de recherche de cette thèse porte sur de nouvelles techniques d'amélioration de la robustesse des codeurs de parole aux erreurs de canal, en particulier pour la transmission de la voix sur les réseaux par paquets (VoIP : Voice over IP). Ces mêmes techniques pourront aussi servir pour des transmissions plus traditionnelles telles que les applications de téléphonie cellulaire où il y a aussi des erreurs de canal. Le travail s'intéresse particulièrement à la bonne détection des trames transitoires afin d'assurer un bon camouflage lorsque les trames transitoires ou les trames suivant les trames transitoires sont perdues. La perte de ces trames provoque une importante propagation d'erreurs. Dans certains cas, une mauvaise stratégie de camouflage de ces trames provoque des artéfacts audibles.

Les autres types de trames répondant bien aux techniques de camouflage actuelles, solutionner les cas problèmes devrait améliorer la qualité globale du signal reconstruit en présence des pertes de trames.

Les contributions principales de cette thèse sont les suivantes :

- Nouvel algorithme robuste pour la détection d'une transitoire dans une trame de parole.

- Amélioration de la classification du codeur en fonction du positionnement de la transitoire.

- Modification des trames dans lesquelles il y a un début de voisement insuffisant pour permettre un bon camouflage des trames voisées.
- Correction du suiveur de pitch en boucle ouverte dans les trames dans lesquelles il y a un début de la transitoire.

La thèse est structurée de la façon suivante. Le chapitre 2 explore l'état de l'art qui couvre plusieurs aspects existants de la problématique. Ces aspects sont l'estimation de la fréquence fondamentale, la classification et la modification des signaux de paroles, ainsi que le camouflage appliqué en cas de perte de trame. Pour certains aspects tel que la classification et le camouflage, l'emphase sera mise sur le codeur VMR-WB puisque la majorité des travaux a été effectuée avec ce codeur. La détection de la transitoire y est aussi présentée, mais se rapportant aux signaux musicaux puisque la littérature la présente souvent en faisant référence à la musique. Le chapitre 3 traite de la détection du positionnement précis de la transitoire par le biais de l'opérateur de Teager. Le chapitre 4 explique la modification des transitoires n'ayant pas une période de pitch suffisamment bien construite à la fin de la trame pour l'obtention d'un camouflage sans artéfact. Il est également question de la modification du suiveur de pitch en boucle ouverte. Les résultats des expérimentations sont présentés dans les chapitre 3 et 4. Les conclusions sont exposées dans le chapitre 5.

CHAPITRE 2

État de l'art en lien avec la classification des trames et le camouflage en cas de pertes de trames

La transmission de la voix sur des réseaux par paquets est de plus en plus populaire. Plusieurs applications de transmission de voix par paquets (Voice over IP, téléphonie mobile de 4e génération (4G)) sont disponibles, mais la structure et la congestion du réseau entraînent des pertes de paquets. Lorsqu'il s'agit de données transmises à des applications qui ne sont pas temps réel, les données perdues peuvent être réacheminées à la demande du récepteur. Dans le cas d'une conversation en temps réel, il n'est pas toujours possible de retransmettre les paquets perdus puisqu'ils arriveraient plus tard que les paquets subséquents et ne pourraient pas être traités à temps. La perte des paquets reliée à la transmission de la voix provoque donc une dégradation de la qualité de la voix au décodeur.

Plusieurs méthodes sont utilisées pour pallier les pertes de paquets, soit du coté émetteur, soit du coté récepteur. La correction est plus simple dans les codeurs où chaque trame reçue est indépendante de la précédente ou de la suivante (l'exemple du PCM/G.711 abordé dans l'introduction). Dans les codeurs où les trames sont interdépendantes, la perte d'une trame peut entraîner une dégradation sur quelques trames subséquentes, et ce même si ces trames ont été bien reçues.

Dans les codeurs de parole de type CELP, deux étapes importantes au décodeur utilisent la prédiction. Le prédicteur court-terme modélise l'enveloppe spectrale du signal et le prédicteur long-terme lui ajoute sa périodicité. Lorsqu'une trame est perdue, une façon de la camoufler est de conserver le même filtre pour le prédicteur court-terme et d'estimer la mémoire du prédicteur à long-terme. L'estimation de la mémoire du prédicteur long-terme entraîne une désynchronisation entre le codeur et le décodeur lors de la reconstruction des trames subséquentes. Cette désynchronisation peut propager une erreur sur plusieurs trames suivant la trame perdue. Il existe un cas particulier : la perte de la trame de transition non-voisé à voisé qui suit la trame non-voisée. En effet, lorsque le contenu du prédicteur long-terme déterminé pour une trame non-voisée est extrapolé pour camoufler

une trame voisée, l'erreur sera plus importante puisque la mémoire ainsi que le gain du prédicteur à long terme sont très différents pour ces deux types de signaux (le gain a souvent une valeur supérieure à 1 pour les trames transitoires). Ainsi, à la réception des trames subséquentes, la construction du prédicteur long-terme qui se base sur son passé sera faussée sur plusieurs trames.

Dans le but d'améliorer la robustesse d'un codeur prédictif à la perte de paquets, plusieurs traitements sont à prendre en considération tels que la classification et la modification des signaux, ainsi que le camouflage qui est appliqué au décodeur pour synthétiser un signal qui est le plus près possible du signal original. Avant de modifier, de coder ou de faire quelques autres traitements sur un signal de parole, il faut d'abord être en mesure d'extraire du signal les caractéristiques qui sont pertinentes.

2.1 Codeur VMR-WB

Avant d'aborder chacunes des caractéristiques du signal qui permettent de classifier ou de modifier le signal, un bref survol du codeur VMR-WB qui est utilisé dans cette thèse est présenté. Ce codeur a été choisi pour les travaux de cette thèse puisque la reconstruction au décodeur des trames perdues est basée sur la classification des trames faites par le codeur. Ainsi, la classification de la dernière bonne trame reçue par le décodeur détermine les paramètres de reconstruction de la trame perdue. Comme les améliorations proposées dans cette thèse s'appliquent à la classification des trames en lien avec le camouflage, les modifications proposées pourront être intégrées dans ce codeur et les résultats obtenus seront comparées à ceux obtenus avec le codeur VMR-WB. Le codeur VMR-WB ([Jelinek et Salami, 2007], [Ahmadi, 2005]) est un codeur à débit variable basé sur le CELP. Le débit du codeur varie en fonction du type de trame à coder (voisée, non-voisée, silence, transition). Il est aussi possible d'ajuster le débit en fonction de la bande passante disponible.

La figure 2.1 présente l'organigramme du codeur VMR-WB. Cet organigramme met en évidence quelques points qui seront abordés dans les sections suivantes. Le signal est d'abord analysé dans le domaine des fréquences. Une transformée de Fourier est effectuée deux fois par trame pour déterminer le niveau de bruit du signal, détecter l'activité vocale du signal et réduire le bruit. L'analyse de prédiction linéaire et l'analyse de pitch en boucle ouverte sont faites sur la trame de signal débruitée. L'analyse de pitch en boucle ouverte est expliquée en détail à la section 2.2.3. Par la suite, la classification détermine le type de signal à coder. La figure 2.1 illustre les trois conditions possibles, soit silence, non-

voisé et voisé. Le détail de la classification est exposé à la section 2.3.3. Pour chacune des classes possibles, le type de codage est distinct et optimisé pour ce type de trame. Pour les trames de silence, un bruit de confort est généré pour éviter que l'interlocuteur ne perçoive une coupure dans la communication. Ces trames peuvent être transmises à des intervalles réguliers et de façon discontinues pour réduire le débit du codeur. Pour les trames non-voisées, le filtre de prédiction linéaire est excitée par une source de bruit gaussien. L'outil de prédiction à long terme n'est pas utilisé puisqu'il n'y a pas de périodicité dans ce type de signal. Pour le reste des trames, la stabilité du signal voisé est vérifié. Si le signal voisé est stable, une modification de signal est appliquée (voir les détails à la section 2.4), sinon le signal est codé comme dans le modèle CELP présenté dans l'introduction.

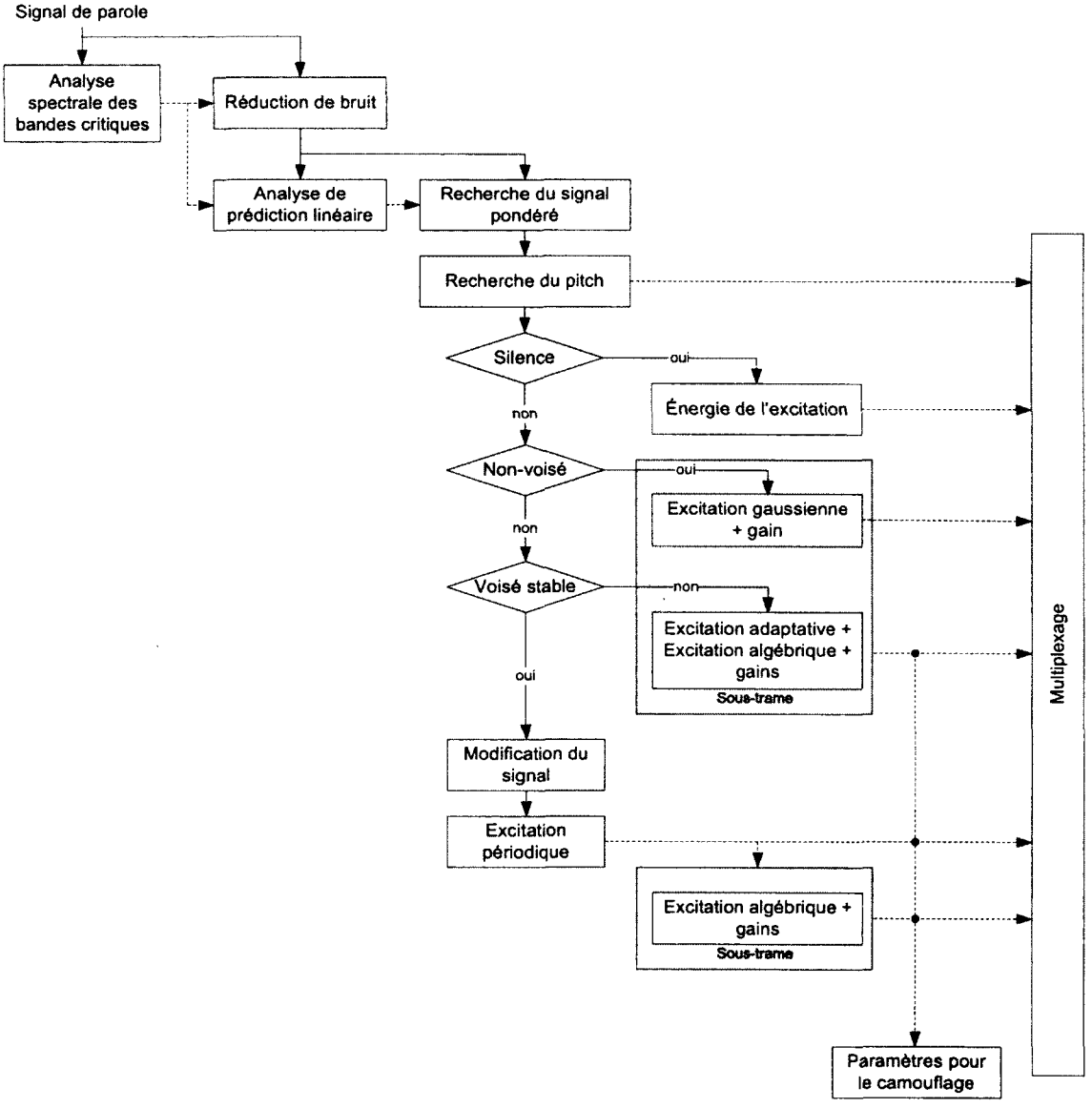


Figure 2.1 Schéma bloc de l'encodeur du VMR-WB

2.2 Techniques d'analyse du pitch

Un des paramètres qui sert à la classification des trames est le taux de voisement. L'estimation du pitch ou l'estimation de la fréquence fondamentale du signal est utilisée pour calculer le taux de voisement. Déterminer la fréquence fondamentale ou le pitch d'un signal de parole est un problème complexe, surtout en présence de bruit ou de distorsion. Plusieurs solutions pour résoudre ce problème ont été proposées à travers les années. Ces solutions se regroupent en deux catégories, soit les approches temporelles et les approches fréquentielles.

La fréquence fondamentale dans le signal de parole est en constante évolution. Les périodes de pitch qui se succèdent varient graduellement en amplitude et en fréquence, d'où la difficulté à bien identifier la valeur du pitch. Puisque la fréquence fondamentale des signaux voisés est variable, l'analyse est faite sur de courts intervalles, typiquement de l'ordre de 5 à 10 ms. Certaines techniques temporelles d'analyse du pitch nécessitent une fenêtre d'analyse qui soit assez grande pour contenir au moins deux périodes de pitch. Comme le pitch varie normalement de 40 à 600 Hz, une fenêtre d'analyse qui correspond à deux fois la longueur du plus long pitch contient nécessairement plusieurs périodes du pitch le plus court et ceci peut aussi dégrader la performance de certains algorithmes de détection de pitch. De plus, la fenêtre d'analyse peut contenir un mélange de signaux voisés et non-voisés, ce qui peut nuire à la détection. Pour concentrer la recherche du pitch sur la plage de fréquence pertinente, certains algorithmes préfiltrent le signal à l'aide d'un filtre passe-bas linéaire ayant une fréquence de coupure autour de 900 Hz.

2.2.1 Techniques temporelles d'analyse du pitch

Les méthodes employant des techniques d'analyses temporelles pour déterminer la fréquence fondamentale du signal sont relativement peu complexes à implémenter et demandent peu de temps de calcul. Lorsque la détection du pitch est faite dans le domaine temporel, il est possible de chercher à quelle fréquence le signal se répète dans un intervalle de temps donné [Gerhard, 2003].

Dans la littérature, une des premières méthodes proposées a été de compter le nombre de passages par zéro et de l'associer à la fréquence fondamentale. Dans le cas d'un sinus pur, le nombre de passages par zéro et la fréquence fondamentale sont effectivement proportionnels. Par contre, pour des signaux plus complexes, ce n'est pas nécessairement le cas. Par exemple, le contenu haute fréquence d'un signal de parole fausse la relation entre

le nombre de passages par zéro et la fréquence fondamentale, puisque le contenu haute fréquence augmente le nombre de passages par zéro.

Dans le cas où le nombre de passages par zéro est utilisé pour faire une approximation de la fréquence fondamentale, les composantes hautes fréquences du signal doivent être éliminées à l'aide d'un filtre passe-bas. La fréquence de coupure du filtre doit être choisie de façon à conserver la fréquence fondamentale du signal tout en éliminant le plus de hautes fréquences possibles. Pour que cette méthode soit valide, il faut que l'amplitude de la fréquence fondamentale soit dominante par rapport aux fréquences résiduelles qui n'ont pas été éliminées par le filtre passe-bas, sinon les autres fréquences présentes dans le signal viendront biaiser le nombre de passage par zéro.

La fonction d'autocorrélation d'un signal peut servir de base aux méthodes temporelles pour déterminer la fréquence fondamentale (voir équation (2.1)), où C_τ est le résultat de la corrélation entre le signal x_j et sa version décalée $x_{j+\tau}$, τ est le décalage entre les deux signaux et W correspond au nombre d'échantillons analysés. Le résultat obtenu est une mesure de ressemblance d'un signal avec une version décalée de lui-même. Alors que le signal est décalé par rapport à lui-même, le niveau de corrélation varie, en étant plus bas que la corrélation obtenue pour un décalage nul. Le résultat de l'autocorrélation atteint un maximum local lorsque le décalage est inversement proportionnel à la fréquence fondamentale du signal.

$$C_\tau = \sum_{j=1}^{W}(x_j x_{j+\tau}) \tag{2.1}$$

En se basant sur l'autocorrélation, Rabiner [Rabiner, 1977] propose d'appliquer diverses modifications non-linéaires au signal avant de calculer la corrélation. Trois types de modifications non-linéaires sont proposées. La première modification consiste à éliminer les échantillons dont la valeur absolue est inférieure à un seuil C_L et à recentrer les échantillons dont la valeur est supérieure au seuil (équation 2.2).

$$y(n) = \begin{cases} x(n) - C_L, & x(n) \geq C_L \\ 0, & |x(n)| < C_L \\ x(n) + C_L, & x(n) \leq -C_L \end{cases} \tag{2.2}$$

La deuxième modification consiste simplement à ramener à zéro les échantillons dont la valeur absolue est inférieure à un seuil C_L (équation 2.3).

$$y(n) = \begin{cases} x(n), & x(n) \geq C_L \\ 0, & |x(n)| < C_L \\ x(n), & x(n) \leq -C_L \end{cases} \tag{2.3}$$

La dernière modification ramène à une valeur constante (1) tous les échantillons dont la valeur est supérieure au seuil C_L et à (-1) tous les échantillons dont la valeur est inférieure au seuil $-C_L$. Les autres échantillons sont mis à zéro (équation 2.4).

$$y(n) = \begin{cases} 1, & x(n) \geq C_L \\ 0, & |x(n)| < C_L \\ -1, & x(n) \leq -C_L \end{cases} \tag{2.4}$$

Grâce à la modification non-linéaire appliquée au signal, le spectre est aplati et la périodicité du signal est augmentée. Pour valider l'approche, dix combinaisons entre les trois modifications proposées et le signal sans modification sont analysées. L'auteur arrive à la conclusion que pour les pitchs courts (voix de femmes et d'enfants), qu'importe la combinaison choisie, l'erreur sur la valeur du pitch est similaire. Pour les pitchs longs, dès que le signal sans modification fait partie de la combinaison, le pourcentage d'erreurs augmente de façon significative. Cette méthode est aussi très sensible quant au choix de la valeur du seuil C_L. Un seuil trop petit pourrait détecter des pics qui sont dus à des harmoniques, alors qu'un seuil trop grand pourrait restreindre le nombre de pics détectés.

Lorsque l'amplitude des harmoniques dans le signal est élevée, le résultat de la corrélation peut être plus élevé pour un des harmoniques que le résultat de la corrélation obtenu pour la fréquence fondamentale. Dans ces cas, la valeur estimée du pitch est erronée. Pour pallier cette difficulté, de Cheveigné [de Cheveigné et Kawahara, 2002] propose d'utiliser une fonction de différence et d'en trouver le minimum. La fonction de différence d_τ est présentée à l'équation (2.5), où x_j est le signal analysé, $x_{j+\tau}$ est le signal analysé décalé, τ est le décalage entre les deux signaux et W est la longueur de la fenêtre d'analyse.

$$d_\tau = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2 \tag{2.5}$$

En développant l'équation (2.5), le résultat suivant est obtenu :

$$d_\tau = C_0 + C_{0,\tau} - 2C_\tau \tag{2.6}$$

où C_0 est la corrélation du signal sans décallage avec lui-même et où $C_{0,\tau}$ est la corrélation du signal avec lui-même pour un décallage égal à τ. Le premier terme de l'équation (2.6) est constant, mais les deux autres varient en fonction de τ. Par contre, le maximum du dernier terme $2C_\tau$ ne coïncidera pas nécessairement au minimum de la fonction d_τ. Ainsi, même si le maximum de la corrélation C_τ est atteint vis-à-vis un harmonique, le maximum des deux autres termes de l'équation (2.6) ne correspond pas nécessairement à cette position. Cette propriété permet d'éviter de fausses détections vis-à-vis les harmoniques du signal et lui confère une certaine robustesse par rapport à la méthode d'autocorrélation (voir équation (2.1)).

La méthode par différence peut encore introduire des erreurs dans l'évaluation de la période de pitch. Si la recherche de pitch se fait à partir du délai nul, le minimum peut correspondre au délai nul ou à un délai très petit. Ce délai peut correspondre au premier formant si ce dernier à une forte amplitude ou à un harmonique qui aurait aussi une forte amplitude. De Cheveigné [de Cheveigné et Kawahara, 2002] propose donc que la décision de la fréquence du pitch ne soit pas prise sur d_τ, mais sur d'_τ (équation (2.7)). Cette fonction de différence cumulative normalisée désaccentue les formants ou les harmoniques qui auraient une valeur avoisinante au délai nul grâce à la valeur initiale égale à 1. Ainsi, le résultat de l'équation (2.7) restera relativement élevé pour les premiers échantillons ce qui permet d'éliminer les premiers harmoniques ou formants sans avoir à déterminer une valeur minimum pour la recherche du pitch.

$$d'_\tau = \begin{cases} 1, & \text{si } \tau = 0 \\ \frac{d_\tau}{\frac{1}{\tau}\sum_{j=1}^{\tau} d_\tau(j)} & \text{sinon} \end{cases} \tag{2.7}$$

Comme alternative à l'autocorrélation, Ross [Ross *et al.*, 1974] propose d'utiliser une fonction de différence appelé AMDF (Average magnitude difference function) et de trouver son minimum pour déterminer le pitch. Cette fonction se définit comme suit :

$$AMDF_j = \sum_{j=1}^{W} |x_j - x_{j+\tau}| \tag{2.8}$$

où x_j est le signal analysé, $x_{j+\tau}$ est le signal analysé décalé, τ est le décalage entre les deux signaux et W est la longueur de la fenêtre d'analyse. Si le signal est parfaitement périodique, le résultat de la fonction est égal à zéro lorsque τ est égal au pitch. Pour les signaux quasi-périodiques, telle que la parole, la valeur atteindra son minimum lorsque le décalage est égal au pitch. Cette méthode a pour avantage d'être plus rapide d'exécution que les méthodes présentées précédemment, puisqu'elle ne nécessite pas de multiplication. Par contre, si le signal est très bruité ou s'il y a de grandes variations dans l'amplitude du signal analysé, il se peut que le minimum calculé ne corresponde pas à la période de pitch. Pour essayer de rendre cette méthode d'analyse du pitch plus robuste au bruit et à la variation d'amplitude du signal qui font en sorte que l'algorithme de détection de pitch trouve parfois un multiple ou un sous-multiple du pitch, Hui [Hui *et al.*, 2006] propose l'analyse en cascade présentée à la figure 2.2.

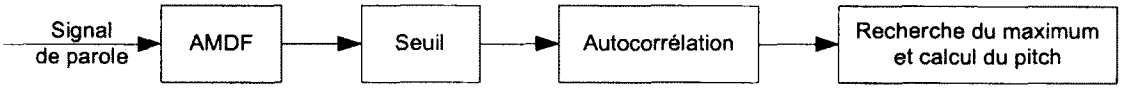


Figure 2.2 Schéma bloc de l'algorithme combiné AMDF et autocorrélation

Dans un premier temps, l'analyse AMDF est faite sur le signal original et un seuil est appliqué au résultat dans le but d'obtenir un signal binaire. Le seuil se calcule selon l'équation suivante : $\alpha(V_{max}+V_{min})$, où α est défini par l'auteur à la valeur fixe de 0,4, V_{max} et V_{min} sont respectivement les valeurs maximales et minimales obtenues après l'analyse AMDF. Puisque l'analyse AMDF tend vers zéro lorsque le décalage correspond au pitch, toutes les valeurs sous le seuil sont arrondies à 1 et toutes les valeurs au-dessus du seuil sont arrondies à 0. L'autocorrélation est donc calculée sur le signal de 1 bit et la valeur du pitch correspond au décalage par rapport au premier pic trouvé (dont le décalage est non nul). Cette méthode est simple à implémenter et demande peu de calculs. L'auteur obtient une meilleure estimation du pitch par cette méthode combinée qu'avec chacune des méthodes séparées.

Pour trouver le pitch, certains auteurs utilisent la méthode de corrélation croisée (équation (2.9)), où x_j est le signal original de longueur N et $\hat{x}_{j+\tau}$ est un signal original d'une longueur inférieure à N ou bien, un signal différent de x_j et de longueur égale ou inférieure à N.

$$CCr_j = \sum_{j=1}^{N}(x_j\hat{x}_{j+\tau}) \tag{2.9}$$

Samad [Samad *et al.*, 2000] calcule les corrélations entre un extrait de signal de parole de longueur N et une version plus courte de ce même extrait N'. L'avantage d'utiliser la corrélation croisée par rapport à l'autocorrélation est que la valeur obtenue par la corrélation croisée est plus grande pour un décalage qui correspond au pitch par rapport à la valeur obtenue avec l'autocorrélation. Les résultats obtenus avec des voix de femmes et d'enfants sont équivalents entre cette méthode et la méthode d'autocorrélation. Pour les voix d'hommes (pitch plus long), la méthode proposée obtient de meilleurs résultats. Pour certains extraits, aucunes des deux méthodes n'est capables d'identifier le bon pitch. Ces régions correspondent à des pitchs très longs qui varient assez rapidement.

Jovicic [Jovicic et Randjolovic, 1987] propose de corréler le signal de parole avec un signal synthétique parfaitement harmonique constitué d'une fréquence fondamentale et de ses harmoniques. La quantité d'harmoniques utilisée pour construire le signal synthétique, ainsi que l'amplitude des harmoniques varient en fonction du rapport signal à bruit. Dans le cas où le signal de parole n'est pas bruité, les harmoniques couvrent toute la bande de fréquences du signal de parole analysé et l'amplitude des harmoniques décroît de façon exponentielle. Dans le cas où il y a beaucoup de bruit dans le signal original, seuls quelques harmoniques (pour couvrir le premier formant) sont ajoutés à la fréquence fondamentale et l'amplitude de ces harmoniques est uniforme. L'intervalle de fréquences utilisé pour construire le signal synthétique varie de 80 à 280 Hz et les corrélations sont effectuées avec une résolution de 10 Hz. Un deuxième balayage est effectué autour de la fréquence dont la corrélation est maximale avec une résolution de 2 Hz. Dans une certaine mesure, cette méthode d'estimation du pitch est relativement insensible à la présence de bruit.

Comme le conduit vocal produit des oscillations (50-80 Hz) qui peuvent interférer avec le pitch qui est produit par la vibration des cordes vocales, Markel [Markel, 1972] propose d'atténuer ces oscillations et de blanchir le spectre du signal de parole pour mieux détecter le pitch. L'algorithme se nomme SIFT (*simplified inverse filter tracking algorithm*) et le schéma bloc présenté à la figure 2.3 illustre le processus pour trouver le pitch et déterminer si la trame analysée est voisée ou non-voisée.

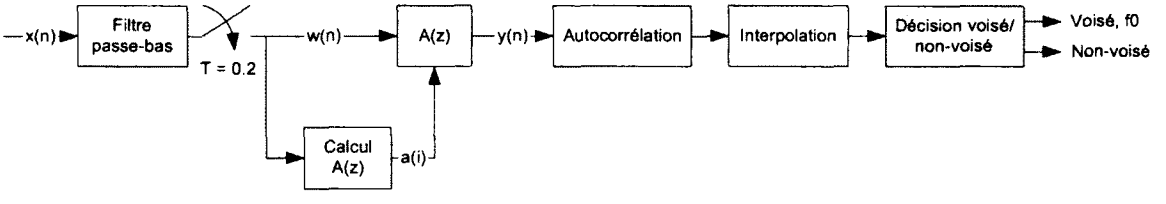


Figure 2.3 Schéma bloc de l'algorithme SIFT

Le signal de parole $x(n)$ est d'abord filtré par un filtre FIR passe-bas dont la fréquence de coupure est de 800 Hz et sous-échantillonné pour obtenir le signal $w(n)$. Ensuite, une analyse de prédiction linéaire est faite sur le signal (a_i) et l'inverse du filtre linéaire obtenu est utilisé pour filtrer le signal de parole. L'ordre du filtre de prédiction linéaire doit être assez long pour bien représenter l'enveloppe spectrale du signal, mais pas trop long pour éviter de modéliser les détails fins du signal. Le signal filtré $y(n)$ a un spectre plus plat, ce qui facilite la détection du pitch. Pour connaître le pitch du signal, l'autocorrélation du signal filtré est calculée. Le pic maximal obtenu dans l'intervalle de recherche (40-500 Hz) correspond au pitch, si cette valeur dépasse un seuil déterminé. Si le pic maximal est plus petit que le seuil, une deuxième recherche à partir de la valeur obtenue à la trame précédente est faite avec un seuil plus bas. Si aucun pic ne dépasse le deuxième seuil, aucun pitch n'est attribué à la trame.

L'utilisation d'un filtre de prédiction ou d'un filtrage perceptuel pour blanchir le spectre du signal, tel que proposé par Markel, est une pratique courante dans les codeurs de parole. Les standards G.729 [G.729, 2007] et le AMR-WB [G.722.2, 2002] utilisent ce principe.

2.2.2 Techniques fréquentielles d'analyse du pitch

Dans le domaine des fréquences, il est possible de trouver la fréquence fondamentale même si cette dernière n'est pas apparente dans le signal. Par exemple, à l'aide d'une détection de pics, les fréquences les plus importantes dans le signal sont identifiées. La fréquence fondamentale est le plus petit dénominateur commun des fréquences partielles identifiées [Schroeder, 1968].

Le cepstre (la transformée de Fourier inverse du logarithme du spectre) du signal peut être utilisé pour trouver le pitch (voir [Noll, 1967] et [Ahmadi et Spanias, 1999]). Le maximum local du cepstre dans l'intervalle de délais possibles est trouvé et cette valeur est comparée à un seuil. Si la valeur est supérieure au seuil, le pitch est défini par le temps correspondant à cette position. Une vérification pour s'assurer que le résultat trouvé n'est pas le double du pitch est faite. Les pics du cepstre à la position correspondant à la moitié du pitch trouvé sont analysés sur un intervalle de +/- 0,5 ms. Si un pic dépasse la valeur du seuil dans cet intervalle, le pitch correspond à cette nouvelle valeur trouvée.

Pour trouver le pitch, Martin [Martin, 1982] propose de calculer la corrélation entre le spectre du signal et la réponse en fréquence d'un filtre en peigne (qui est une série de raies régulièrement espacées). L'amplitude des raies qui constituent la réponse du filtre décroit en fonction de la distance par rapport à la fréquence fondamentale et la distance

entre les raies est égale à la fréquence fondamentale testée. Le pitch correspond à la fréquence ayant obtenue la valeur maximale de corrélation. Cette méthode fréquentielle est équivalente à la méthode temporelle de Jovicic présentée précédemment. Ces deux méthodes sont d'avantage utilisés sur les signaux musicaux que sur les signaux de parole.

2.2.3 Technique d'analyse du pitch dans le codeur de parole VMR-WB

Dans le codeur VMR-WB [Jelinek et Salami, 2007], l'analyse du pitch se fait par une approche temporelle, basée sur l'autocorrélation du signal. Une valeur de pitch est évaluée pour chaque intervalle de 5 ms (donc deux valeurs de pitch par trame puisque la durée d'une trame est de 10 ms). Au traitement de chacune des trames, une troisième valeur de pitch s'ajoute et correspond au pitch de la demi-trame d'avance disponible (appelée look-ahead dans les publications en anglais).

Les trois valeurs de pitch sont évaluées selon la même procédure. Le signal utilisé pour cette évaluation est $s_w(n)$, le signal d'entrée pondéré par un filtre perceptuel. Ce signal est décimé d'un facteur 2 à l'aide d'un filtre FIR d'ordre 5 dont les coefficients sont : $\{0,13; 0,23; 0,28; 0,23; 0,13\}$. Le signal pondéré et décimé est noté $s_{wd}(n)$.

Pour trouver le pitch du signal analysé, l'autocorrélation du signal est calculée pour plusieurs décalages qui varient entre 10 et 115 échantillons. La longueur des signaux corrélés varient aussi en fonction du décalage (voir énumération plus bas). La corrélation est calculée selon l'équation suivante :

$$C(d) = \sum_{n=0}^{L_{sec}} s_{wd}(n) s_{wd}(n-d) \tag{2.10}$$

où $s_{wd}(n)$ est le signal de parole pondéré et décimé tel qu'expliqué précédemment, d représente le nombre d'échantillons de décalage entre les deux signaux et L_{sec} est une valeur qui dépend de d. L_{sec} est déterminé comme suit :

- $L_{sec} = 40$, si $d = 10, ..., 16$
- $L_{sec} = 40$, si $d = 17, ..., 31$
- $L_{sec} = 62$, si $d = 32, ..., 61$
- $L_{sec} = 115$, si $d = 62, ..., 115$

Pour favoriser la continuité du pitch entre les trames, les valeurs de corrélation autour du pitch de la trame précédente sont rehaussées selon une fenêtre triangulaire de longueur 27 (équation 2.11).

$$w_{pn}(13+i) = w_{pn}(13-i) = 1 + \alpha_{pn}(1 - i/14) \tag{2.11}$$

où α_{pn} est un facteur de pondération qui dépend de la valeur normalisée de la corrélation de la trame précédente, ainsi que de la stabilité du pitch. Cette valeur est limitée à 0,7.

La corrélation normalisée est ensuite calculée selon l'équation suivante :

$$C_{norm}(d_{max}) = \frac{C(d_{max})}{\sqrt{\sum_{n=0}^{L_{sec}} s_{wd}^2(n) \sum_{n=0}^{L_{sec}} s_{wd}^2(n - d_{max})}} \tag{2.12}$$

où d_{max} est le nombre d'échantillons de décalage pour laquelle la valeur de corrélation est la plus élevée dans chacune des quatre intervalles définies précédemment. La valeur maximale d'autocorrélation pour chacun des intervalles L_{sec} est sélectionnée. Si dans un intervalle supérieur la valeur maximale de la corrélation normalisée correspond à un multiple de la valeur maximale de la corrélation normalisée sélectionnée dans un intervalle inférieur, cette dernière est pondérée d'un facteur 1,17 (valeur trouvée expérimentalement). En terminant, les trois valeurs de pitch évaluées pour la trame s'influencent pour s'assurer que ces valeurs sont relativement proches entres elles et aussi proche de la trame passée.

2.3 Classification des signaux

Les paramètres présentés à la section précédente peuvent être utilisés pour faire la classification de trames. Entre autres, le nombre de passages par zéro peut être utilisé pour déterminer si un signal de parole est voisé ou non-voisé. Un signal voisé croise l'axe du zéro à une fréquence beaucoup moins importante qu'un signal non-voisé, qui lui possède un contenu haute fréquence plus important.

Greenwood [Greenwood et Kinghorn, 1999] présente une façon simple de classer les trames d'un signal en trois catégories, soit voisée, non-voisée et silence. La classification est faite en fonction du nombre de passages par zéro et de l'énergie du signal à court terme. Les valeurs du nombre de passages par zéro et de l'énergie à court-terme sont calculées avec des fenêtres rectangulaires de 10ms. Ce système permet une classification correcte dans 65% des cas. Deux critères ne semblent pas suffisants pour faire la distinction entre les trames voisées, non-voisées et un silence. De plus, les performances passables sont dues à l'utilisation d'un critère d'énergie en absolu qui doit toujours être utilisé avec précaution

puisque ce critère dépend beaucoup de l'enregistrement du signal, ainsi qu'au nombre de passages par zéro qui peut être faussé par les composantes hautes fréquences du signal et par la présence de bruit.

Ahmadi [Ahmadi et Spanias, 1999] ajoute au nombre de passages par zéro et à l'énergie à court-terme du signal les pics cepstraux pour améliorer la classification voisé / non-voisé. La classification des différentes trames débute par la construction d'un histogramme des trois paramètres choisis (première analyse complète du signal). La médiane de chaque histogramme devient le seuil de décision pour faire le classement des trames. Dans un deuxième temps, le signal de parole est analysé trame par trame selon les seuils de décision trouvés lors de la première analyse. Selon les évaluations, le pourcentage d'erreur de classification de cette méthode tourne autour de 1,6%. Le pourcentage de mauvais classement de trames non-voisées classées voisées est légèrement plus faible (1,49%) que le pourcentage de mauvais classement de trames voisées classées non-voisées (1,73%). Bien que très efficace par rapport à l'algorithme précédent, la nécessité d'analyser le signal à deux reprises (une première fois pour définir les seuils de décision et une deuxième fois pour faire la classification des trames) limite les possibilités de l'algorithme à être utilisé dans un contexte temps-réel.

2.3.1 Discrimination entre la parole et les autres signaux

Dans le même esprit de classification de la voix en signaux voisés et non-voisés, Lu [Lu *et al.*, 2002] propose un classificateur qui fait la distinction entre la parole et d'autres types de signaux. Les signaux qui ne sont pas de la parole sont ensuite séparés en signaux musicaux, en sons ambiants et en silences. Cette méthode utilise les paramètres de classification suivants : le nombre de passages par zéro, le ratio de l'énergie à court-terme, le flux spectral, la divergence entre les LSP (line spectral pairs), la périodicité dans chaque bande de fréquence et le ratio de bruit par trame.

Le NPZE, soit le nombre de passage par zéro élevé, représente le nombre de fois, dans une période d'une seconde, où le NPZ (nombre de passage par zéro) est plus élevé d'un facteur 1,5 par rapport à la moyenne des NPZ. Le NPZE se calcule de la façon suivante :

$$NPZE = \frac{1}{2N}\sum_{n=0}^{N-1}[sgn(NPZ(n) - 1,5avNPZ) + 1] \tag{2.13}$$

$$avNPZ = \frac{1}{N}\sum_{n=0}^{N-1} NPZ(n) \tag{2.14}$$

où n représente l'index dans la trame, N le nombre total de trames dans une période d'une seconde et $avNPZ$ est la moyenne des NPZ sur une période.

De la même façon que le NPZE, le ratio de l'énergie à court-terme (RECT) est calculé en fonction de sa variation dans une période d'une seconde. Le ratio représente le nombre de trames où l'énergie est inférieure à la moitié de l'énergie moyenne de la période.

$$RECT = \frac{1}{2N}\sum_{n=0}^{N-1}[sgn(0,5avECT - ECT(n)) + 1] \tag{2.15}$$

$$avECT = \frac{1}{N}\sum_{n=0}^{N-1} STE(n) \tag{2.16}$$

où $ECT(n)$ représente l'énergie à court-terme de la trame n, N est le nombre de trames traitées dans un intervalle d'une seconde et $avECT$ est la moyenne de l'énergie à court-terme, toujours sur une période d'une seconde.

D'autres auteurs proposent d'utiliser le nombre de passages par zéro pour discriminer la parole et la musique. En plus de ce paramètre, Saunders [Saunders, 1996] utilise la forme de l'enveloppe de l'énergie, c'est-à-dire le fait que l'enveloppe temporelle du signal de parole produit plus de pics et de vallées que les signaux de musique. Panagiotakis [Panagiotakis et Tziritas, 2005] utilise la distribution de l'amplitude du signal en combinaison avec la distribution du nombre de passages par zéro pour faire la discrimination entre les signaux de parole et de musique.

Keum [Keum et Lee, 2006] utilise une approche fréquentielle pour discriminer la parole et la musique. Il calcule les différents pics spectraux contenus dans le signal, ainsi que la durée des différents pics dans le signal. Le nombre de pics spectraux dépassant un seuil fait la différence entre la parole (peu de pics spectraux dépassant le seuil) et la musique (plusieurs pics spectraux dépassant le seuil).

2.3.2 Classification des signaux de parole appliquée aux codeurs de parole

La classification des signaux de parole peut être utilisée dans les codeurs de parole de façon à réduire le débit d'information à transmettre ou pour améliorer les méthodes de camouflage en cas de perte de trames. Les différentes classes présentes sont généralement : les trames voisées, les trames non-voisées, les trames de transition et les trames de silence. Les trames voisées regroupent les signaux qui sont quasi-périodiques. Les trames non-voisées regroupent les signaux où il y a très peu de corrélation à long terme entre les échantillons.

Les trames de transitions sont celles où il y a des changements rapides d'énergie et de composition spectrale. Les trames de silence sont celles où il n'y a pas d'activité vocale.

Pour ce qui est de réduire le débit d'information à transmettre, chaque type de trame est transmis avec un débit distinct et prédéterminé [Wang et Gersho, 1989]. Les trames voisées nécessitent une bonne précision pour tous les paramètres de codage transmis (prédicteur court-terme, prédicteur long-terme et excitation). Par contre, les trames non-voisées n'ont pas besoin du prédicteur long-terme puisqu'elles ne contiennent pas de périodicité. Dans ce cas, le prédicteur long-terme peut ne pas être transmis [Paksoy *et al.*, 1993] afin de réduire le débit de ce type de trame. Enfin, les trames transitoires évoluent rapidement, et comme aucun compromis ne doit être fait sur l'information à transmettre, un débit plus important est alloué aux trames transitoires afin d'augmenter la précision de la partie excitation. Le débit des trames transitoires est donc plus élevé que celui des trames voisées et non-voisées. Pour réduire d'avantage le débit global, il est possible de faire la distinction entre les trames où il y a de l'activité vocale et les trames de silence. Un minimum d'information est transmis dans les trames de silence qui ont ainsi un débit inférieur aux trames non-voisées.

2.3.3 Classification des signaux de parole pour le codeur VMR-WB

Dans le codeur prédictif VMR-WB, la classification du signal sert à réduire le débit moyen du codeur en appliquant un débit précis à chacune des classes du signal de parole (le silence ou les signaux non-voisés ayant un débit plus bas que les signaux voisés). La même classification est aussi utilisée dans le but d'aider le camouflage. Dans certains modes d'opérations, lorsqu'une trame est perdue, l'état (la classe) de la trame précédente est connu et utilisé pour choisir une stratégie de camouflage. Dans le cas des signaux non-stationnaires, les caractéristiques fréquentielles et temporelles changent rapidement. Le camouflage fait alors converger rapidement les paramètres du décodeur vers un état de bruit ambiant. Dans le cas particulier des signaux quasi périodiques et des signaux non-voisés stationnaires, leurs caractéristiques varient peu et les paramètres de codage peuvent donc être maintenus constants même lorsque plusieurs trames sont perdues, sans grande conséquence sur la qualité du signal synthétisé.

Le codeur VMR-WB classe chaque trame dans une des catégories suivantes : *transitoire*, *voisé*, *transition voisée*, *non-voisé* et *transition non-voisée*. En ce qui concerne la classification pour le camouflage des trames perdues dans VMR-WB, les trames de parole non-voisées et les trames où il n'y a pas d'activité vocale sont regroupées dans les trames *non-voisées*. Lorsqu'il y a un offset dans une trame voisée, c'est-à-dire que le signal passe

de voisé à non-voisé au cours de la trame, le classement est *non-voisé*. Le classement non-voisé permet de ne pas introduire de périodicité lors du camouflage si la trame suivante est perdue, car vers la fin de la trame courante le signal a des caractéristiques d'une trame non-voisée. Les trames de *transition non-voisées* regroupent également les trames non-voisées où il y a à la fin un début de transitoire qui ne contient pas au moins une période de pitch bien construite. Il est préférable de ne pas classer la trame comme une *transitoire* (classe expliquée plus loin) puisque le camouflage reconstruirait la partie périodique du signal en se basant sur une information de pitch erroné ; ceci parce que le camouflage présume qu'une trame suivant une trame *transitoire* est une trame *voisée*. Les trames de *transition voisées* regroupent les trames voisées où les caractéristiques ne sont pas stables, c'est-à-dire que l'enveloppe spectrale, le pitch et l'énergie du signal varient assez rapidement (passage entre deux voyelles ou fin d'un voisement). Les trames *voisées* ont des caractéristiques stables et bien définies, c'est-à-dire que l'enveloppe spectrale, le pitch et l'énergie du signal varient peu d'une trame à l'autre. Les trames *transitoires* rassemblent les trames qui suivent des trames non-voisées et qui sont périodiques au moins vers la fin de la trame, c'est-à-dire qu'elles ont au moins une période de pitch bien construite.

Le diagramme d'état présenté à la figure 2.4 donne les contraintes à respecter lors de la classification des signaux dans le codeur VMR-WB. Husain [Husain et Cuperman, 1995] utilise une classification similaire des signaux (voisée, non-voisée, transitoire et silence) avec le standard G.728.

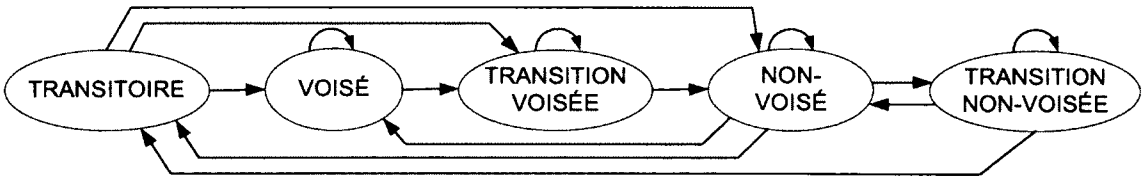


Figure 2.4 Diagramme d'état du classement des trames dans le codeur VMR-WB

Pour faire la classification des trames dans le standard VMR-WB [Ahmadi, 2005], les paramètres suivants sont utilisés : la corrélation normalisée $\overline{R'}_{xy}$, le ratio d'énergie entre les basses et les hautes fréquences e'_t, le ratio signal à bruit snr, la stabilité du pitch pc, l'énergie à la fin de la trame courante par rapport à l'énergie à long-terme E_{rel} et un décompte du nombre de passage par zéro zc. La définition de ces paramètres est rappelée dans les paragraphes suivants.

Chaque paramètre est normalisé selon la fonction générale $p^s = k_p p_x + c_p$, $0 \leq p^s \leq 1$, où p^s est la version normalisée du paramètre p_x. Dans le standard [Ahmadi, 2005], le tableau 5.22-1 donne les valeurs de k_p et c_p pour chacun de ces paramètres. Les valeurs normalisées de chacun des paramètres sont utilisées pour calculer une fonction de coût :

$$f = \frac{1}{7}(2\overline{R'}^s_{xy} + e^s_t + snr^s + E^s_{rel} + zc^s + pc^s) \tag{2.17}$$

Chaque trame est ensuite classée selon la valeur de la fonction de coût obtenue (standard [Ahmadi, 2005], voir tableau 5.22-2).

Le standard VMR-WB utilise une combinaison de deux corrélations normalisées pour évaluer $\overline{R'}_{xy}$, soit la corrélation normalisée de la deuxième moitié de la trame courante $C_{norm}(d1)$ et la corrélation normalisée de la première moitié de la trame suivante $C_{norm}(d2)$:

$$\overline{R'}_{xy} = 0,5(C_{norm}(d1) + C_{norm}(d2)) \tag{2.18}$$

Le calcul des corrélations $C_{norm}(d1)$ et $C_{norm}(d2)$ a été expliqué à la section 2.2.3.

Pour le signal voisé, l'énergie est généralement concentrée dans les basses fréquences, alors que dans le cas d'un signal non-voisé, l'énergie se situe principalement dans les hautes fréquences. Le ratio entre les basses fréquences et les hautes fréquences, $e_{tilt} = \frac{E_{basse}}{E_{haute}}$ contient l'information de la distribution de l'énergie en fonction des fréquences présentes dans le signal. Ce ratio n'est pas calculé sur toute la trame, mais sur chacune des deux moitiés de la trame. Dans le cas présent, le ratio des deux demi-trames courantes est pris en compte sous la forme logarithmique suivante :

$$e_t = 10 \cdot log(e_{tilt}(0)e_{tilt}(1)) \tag{2.19}$$

Le rapport signal à bruit (équation (2.20)) est généralement plus élevé lorsque le signal est voisé. Il est calculé entre l'énergie du signal d'entrée pondéré E_{sw} et l'énergie de l'erreur entre le signal d'entrée pondéré et le signal de synthèse pondéré $e = s_w(n) - s(n)$.

$$snr = \frac{E_{sw}}{E_e} \tag{2.20}$$

Le signal d'entrée pondéré est donné par :

$$s_w(n) = s(n) + \sum_{i=1}^{16} a_i \gamma_1^i s(n-i) + \beta_1 s_w(n-1) \tag{2.21}$$

où $s_w(n)$ est le signal de parole pondéré, $s(n)$ est le signal de parole, a_i sont les coefficients du filtre LP, γ_1^i est le facteur de pondération perceptuel et β_1 est une valeur fixe égale à 0.68.

L'énergie relative est la différence entre l'énergie de la trame courante et l'énergie à long-terme du signal en dB (2.22). Ce paramètre permet d'éviter certaines erreurs de classification en particulier sur des signaux de très faible énergie. L'énergie de la trame courante est la somme de l'énergie de chacune des deux demi-trames courantes (2.24). L'énergie à long-terme est une fraction de l'énergie de la trame courante additionnée à l'énergie à long-terme déterminée précédemment (2.23).

$$E_{rel} = E_t - \bar{E}_f \tag{2.22}$$

$$\bar{E}_f = 0,99 * \bar{E}_f + 0,01 E_t \tag{2.23}$$

$$E_t = 10log(E_{frame}(0) + E_{frame}(1)) \tag{2.24}$$

La stabilité du pitch pc est liée à la fréquence fondamentale du signal qui représente la fréquence de vibration des cordes vocales. Lorsque le pitch est stable, c'est-à-dire que sa valeur varie peu d'une sous-trame à l'autre, il s'agit d'un signal voisé. Dans le cas contraire, il est probable qu'il s'agisse d'un signal non-voisé. Dans le VMR-WB, la stabilité du pitch est basée sur trois valeurs de la période de pitch, soit celle de la première moitié de la trame courante d_0, celle de la deuxième moitié de la trame courante d_1 et celle de la première moitié de la trame suivante d_2. Ces trois valeurs sont comparées entre elles pour obtenir la stabilité du pitch selon :

$$pc = |d_1 - d0| + |d_2 - d1| \tag{2.25}$$

Le taux de passage par zéro ou *zero-crossing rate (ZCR)* donne une indication sur le contenu spectral du signal du fait que plus un signal contient des hautes fréquences, plus le signal va osciller de part et d'autre de l'axe zéro.

2.4 Modification des signaux de parole

La modification des signaux de parole peut consister à altérer le signal de sorte qu'il est perceptuellement identique (ou très peu différent) du signal original mais qu'il s'adapte mieux au modèle de codage ou de camouflage utilisé. Dans cette thèse, une section porte sur la modification des trames transitoires incomplètes. Ces trames contiennent le début d'un signal voisé, mais la partie voisée dans la trame ne contient pas suffisamment d'échantillons pour que le camouflage puisse l'utiliser sans introduire d'artéfacts audibles ; lorsque la fin de la trame ne contient pas au moins une période de pitch bien construite, le camouflage va répéter une fraction de la période de pitch réelle. Cette répétition peut dans certains cas introduire des actéfacts harmoniques audibles. Dans la littérature, il est possible de trouver des informations quant à la modification de signal pour les codeurs CELP ou la modification de la partie transitoire du signal pour les codeurs paramétriques. Cette section en fait le résumé et soulève des points quant à la pertinence d'adapter ces méthodes à la problématique actuelle.

Lorsqu'il faut modifier un signal de parole de façon transparente, un critère très important est de ne pas introduire de distorsion audible dans le signal modifié. Lorsqu'une distorsion audible est introduite dans le signal avant le codeur, il est fort probable que le signal distortionné décodé soit de moins bonne qualité que le signal original codé-décodé.

Kleijn [Kleijn *et al.*, 1992] introduit le codeur RCELP au début des années 90 dans le but de réduire le débit (bits/s) nécessaire pour coder la parole en maintenant une bonne qualité. Pour réduire le débit, le pitch n'est transmis qu'une fois par trame (contre quatre fois par trame dans les codeurs CELP standard). Cette proposition est appropriée puisqu'en général la valeur du pitch de la voix évolue plus lentement que la nouvelle fréquence proposée pour transmettre cette valeur. Le pitch est calculé en fin de trame et son évolution linéaire entre la fin de la trame précédente et la fin de la trame courante est ensuite imposée au signal original. Il est important de faire la modification sur le signal original pour que tous les paramètres du codeur soient calculés à partir du signal modifié. La modification sur le signal original permet de s'assurer que les valeurs de pitch et de gain déterminées par le dictionnaire adaptatif sont cohérentes avec le signal utilisé pour faire l'analyse-par-synthèse. La modification du signal peut entraîner l'accumulation d'échantillons à la fin de la trame (la trame modifiée contient plus ou moins d'échantillons que la trame originale), il faut alors permettre au codeur de se resynchroniser. Comme les modifications de signal ne sont appliquées qu'aux trames voisées (puisque les trames non-voisées ne sont pas périodiques), la resynchronisation se fait dans les trames non-voisées.

La transmission de la valeur du pitch limitée à une fois par trame a aussi des avantages lorsqu'il y a des trames perdues [Nahumi et Kleijn, 1995]. La qualité du signal reconstruit dépend grandement du pitch. En réduisant la quantité de bits accordés au pitch, si des erreurs de canal surviennent, la dégradation du signal reconstruit est moindre par rapport aux codeurs qui transmettent quatre valeurs de pitch par trame. De plus, en limitant la variation du pitch possible et en ajoutant un peu de débit pour transmettre la différence entre le pitch courant et le pitch précédent, il est possible de réduire davantage l'effet des pertes de trames sur le signal reconstruit. Dans le cas où le délai de reconstruction du signal n'est pas critique et qu'une seule trame est perdue, la différence entre le pitch courant et le pitch précédent de la trame suivante peut être utilisé pour reconstruire la trame perdue avec le bon pitch. Si le délai est critique, la trame est reconstruite avec le pitch de la trame précédente. À la réception de la trame suivante, la trame perdue est retraitée avec le bon pitch pour mettre à jour les paramètres du codeur avant de traiter la trame reçue.

Le eX-CELP [Gao *et al.*, 2001a] combine l'approche en boucle fermée plus traditionnelle du CELP à une approche en boucle ouverte sur un signal pondéré par un filtre perceptuel. La modification qui est intégrée au eX-CELP ressemble à celle proposée dans le RCELP, mais en plus de rendre plus linéaire le contour du pitch, le gain du dictionnaire adaptatif est aussi augmenté. La modification du signal s'applique aux trames qui sont voisées et dont le pitch est stable. Pour ces trames, le pitch et le gain du dictionnaire adaptatif sont évalués en boucle ouverte uniquement et un contour de pitch est imposé au signal (comme pour le RCELP) [Gao *et al.*, 2001b]. Tout comme le RCELP, le pitch est transmis qu'une fois par trame. Dans les zones de transitions, le pitch du signal change rapidement. Le eX-CELP met en oeuvre deux options, soit un lissage des harmoniques ou un interpolation du signal pour uniformiser l'évolution du pitch dans le but d'accélérer la construction du dictionnaire adaptatif. Cette méthode a aussi pour effet d'augmenter le gain du dictionnaire adaptatif.

Tammi ([Tammi *et al.*, 2005] et [Tammi et Jelinek, 2002b]) applique au codeur VMR-WB une modification qui ressemble à celle proposée par Kleijn. Cette modification s'applique uniquement aux trames parfaitement voisées et a pour but de réduire le débit du codeur. Contrairement au RCELP, la trame de signal n'est pas modifiée en totalité et la synchronisation à la fin des trames est gardée intacte. La synchronisation en fin de trame est plus simple à implémenter puisqu'il n'y a pas de délai supplémentaire à gérer. De plus, la synchronisation en fin de trame fait en sorte que la modification de signal proposée peut être intégrée facilement dans un codeur à débit variable où les conditions de codage varie d'une trame à l'autre.

La modification impose une évolution linéaire du pitch qui est bornée par le dernier pitch de la trame précédente et le dernier pitch de la trame courante de façon à respecter la synchronisation en fin de trame. Pour effectuer la modification du signal, certaines conditions doivent être rencontrées pour assurer la transparence de la modification proposée. Pour appliquer la modification à une trame, il faut que celle-ci soit classée comme étant voisée, de plus il faut une évolution lente du pitch à l'intérieur de la trame. Si la trame rencontre cette condition, les trois étapes de la modification débutent.

La première étape consiste à identifier tous les cycles de pitch présents dans la trame. Pour ce faire, une estimation du pitch est effectuée en boucle ouverte. Ensuite, le dernier cycle de pitch de la trame précédente est corrélé avec le signal de la trame courante pour positionner chaque cycle de pitch. La trame est segmentée pour que chacun des segments contienne un cycle de pitch. La frontière entre les deux cycles est positionnée à mi-chemin entre les maximums de chacun des cycles. La zone qui correspond au milieu entre les cycles de pitch correspond aussi à la jonction entre les signaux modifiés. Cette partie du signal a une énergie plus faible ce qui réduit le risque d'introduire des artéfacts audibles dans le signal modifié.

La deuxième étape consiste à déterminer l'évolution linéaire du pitch qui sera imposée au signal modifié, ainsi que la valeur du pitch qui sera transmise au décodeur. L'évolution linéaire est déterminée en fonction du dernier pitch de la trame courante et du dernier pitch de la trame précédente. Pour éviter les problèmes d'oscillations d'une trame à l'autre, il est proposé de faire une évolution linéaire de la valeur du délai et ensuite, de stabiliser cette valeur [Tammi et Jelinek, 2002a].

La dernière étape consiste à appliquer la modification. La premier cycle de pitch modifié est copié de la trame précédente pour suivre l'évolution du pitch qui a été déterminée précédemment. Les cycles subséquents sont aussi une copie du cycle qui les précède. Au cours des trois étapes de la modification, si les caractéristiques calculées ne sont pas valides et qu'ainsi la modification du signal pourrait entraîner des artéfacts audibles, il y a un arrêt de la procédure de modification et la trame est codée avec un débit régulier.

Vafin [Vafin *et al.*, 2001] propose de modifier l'emplacement des parties transitoires d'un signal de telle sorte que les parties transitoires se retrouvent à un endroit prédéfini dans une trame donnée. La modification de l'emplacement de la partie transitoire du signal sert à diminuer le nombre de sinusoïdes amorties nécessaires à décrire cette partie du signal. Il est rapporté que ces modifications n'entraînent pas de changement audible dans le signal.

Le cadre du travail de Vafin est un codeur paramétrique où le signal est décomposé en trois parties : la partie transitoire, la partie stationnaire (représentée par une série de sinusoïdes) et le bruit. Le problème que rencontre ce genre de codeur quant à la partie transitoire du signal est que si cette partie se situe dans le milieu d'un segment analysé, le nombre de sinusoïdes amorties nécessaire est considérable lorsqu'on le compare au nombre de sinusoïdes amorties requises lorsque la partie transitoire se situe au début de la trame.

Pour faire la modification de localisation des parties transitoires, Vafin propose de détecter les transitoires à l'aide de fenêtres glissantes rectangulaires. La détection des parties transitoires est faite grâce à l'énergie du signal et permet de cadrer la partie transitoire en déterminant où elle commence et où elle se termine. Par la suite, les échantillons qui représentent la partie transitoire sont simplement décalés (coupés-collés) au début de la trame. Le signal entre les deux parties transitoires est ensuite reconstruit pour ne pas laisser un vide.

Pour reconstruire l'intervalle entre les deux transitoires, il est possible de modifier l'échelle temporelle (*time-warping*) si la fréquence fondamentale f_0 n'est pas affectée de plus de 0.2%. Dans le cas contraire, il est possible de séparer l'intervalle en deux parties, une première partie qui contient les premières 10 ms du signal et une deuxième partie contenant le reste de l'intervalle traité. Si cette procédure ne comble pas le vide entre les deux intervalles, il faut le combler en appliquant une méthode de recouvrement (*overlap-add*). Il est à noter que l'effet psychoacoustique de masquage qui suit une transitoire permet une plus grande tolérance quant à la variation de la fréquence fondamentale. La variation de la fréquence fondamentale ne doit jamais être plus grande que 2% pour respecter la condition de masquage psychoacoustique.

Il est aussi possible de modifier la totalité du signal, sans avoir besoin de classifier les trames. Jensen [Jensen *et al.*, 1999] propose de modifier le signal original pour le rendre plus facile à coder tout en faisant bien attention de ne pas y introduire de distorsion. La modification est faite en boucle fermée à l'aide de l'algorithme du moindre carré et est optimisée de façon à augmenter le gain de prédiction. Dans le cadre du présent travail, augmenter le gain de prédiction n'est pas favorable puisque la propagation de l'erreur dans le cas où la dernière trame bien reçue est une transitoire ou une trame voisée ne sera que plus longue.

Avendano [Avendano et Goodwin, 2004] propose d'améliorer la qualité d'un signal audio en modifiant l'importance de la partie transitoire du signal. Pour ce faire, il utilise un détecteur de transitoire graduel, c'est-à-dire qui ne prend pas de décision binaire (oui/non).

Ainsi, une réponse graduelle est appliquée au flux spectral du signal pour construire une fonction caractéristique de transition continue. Cette fonction est expliquée dans les paragraphes suivants.

Premièrement, l'amplitude du spectre du signal est calculée et la différence de premier ordre $\Delta[n,k]$ est calculée ($|X(n,k)| - |X(n-1,k)|$), où n est la trame de signal traitée et k est la bande de fréquence. Le flux spectral non-normalisé ρ est ensuite calculé :

$$\rho[n] = \sum_k |\Delta[n,k]|^{1/2} \tag{2.26}$$

Ce flux met en évidence les endroits où il y a des transitoires dans le signal, mais est dépendant des caractéristiques de traitement (longueur des trames et taille de la fft) et de l'amplitude du signal. Pour rendre le processus plus robuste, un lissage est appliqué :

$$\begin{aligned} si(\rho[n] > \beta_{n-1}) \\ \beta_n = \rho[n] \\ sinon \\ \beta_n = \gamma\beta_{n-1} \end{aligned} \tag{2.27}$$

où β_0 est initialisé à une très grande valeur (ici, l'auteur propose 2000) et γ est fixé à une valeur proche de 1 (l'auteur propose 0.99). Pour terminer, le flux est normalisé à partir du facteur β_n trouvé précédemment :

$$\Phi[n] = \frac{\rho[n]}{\beta_n} \tag{2.28}$$

Dans le but d'obtenir une modification de la transitoire qui varie en fonction du signal d'entrée, la réponse paramétrisée gradée du flux spectral est :

$$\alpha[n] = G_{\{\alpha_0,\lambda,\Phi_0\}}(\Phi[n]) \tag{2.29}$$

$$= (\frac{\alpha_0+1}{2}) + (\frac{\alpha_0-1}{2})tanh[\pi\lambda(\Phi[n]-\Phi_0)] \tag{2.30}$$

où λ représente la pente de la courbe, Φ_0 en est le point d'inflexion et α_0 est l'amplitude de la réponse paramétrisée. La valeur de α_0 est plus grande que 1 si une amplification des zones transitoires est désirée, dans le cas contraire α_0 est plus petite que 1 d'afin d'en diminuer l'importance.

La modification des zones transitoires peut se faire selon deux approches : linéaire ou non-linéaire. Le signal $\hat{X}[n,k]$ est obtenu à l'aide d'une transformation du type :

$$\hat{X}[n,k] = H_\alpha[n,k]X[n,k] \tag{2.31}$$

L'approche linéaire la plus simple est l'application d'un gain :

$$H_\alpha[n,k] = f\{\alpha[n]\} = \alpha[n] \tag{2.32}$$

Le gain appliqué aux parties stationnaires est approximativement égal à 1. Les parties transitoires sont augmentées si $\alpha_0 > 1$ et sont diminuées si $\alpha_0 < 1$. La modification linéaire proposée introduit des artéfacts et de la distorsion est perçue dans le résultat final.

La modification non-linéaire est du type :

$$|\hat{X}[n,k]| = (|X[n,k]| + 1)^{\alpha[n]} - 1 \tag{2.33}$$

Selon l'auteur, cette modification n'entraîne pas de distorsion et a un son beaucoup plus naturel que la modification linéaire. Il est à noter que la modification est presque linéaire à l'échelle logarithmique et que cette courbe suit un peu la courbe de perception de l'oreille.

2.5 Le camouflage

Lorsqu'un signal de parole est transmis par paquets, il arrive que des paquets soient perdus ou qu'ils arrivent avec un trop grand délai pour être traités. Il en résulte un manque d'informations pour faire la reconstruction ce qui entraîne une dégradation du signal. Cette dégradation peut être locale, c'est-à-dire n'affecter qu'une seule trame, ce qui est le cas pour un codeur non-prédictif. Par contre, dans le cas où le codeur est prédictif, la dégradation affecte la trame perdue ainsi que quelques trames subséquentes (en fonction de l'importance de la trame perdue). Pour augmenter la qualité du signal au décodeur, il est possible de combler les trames manquantes de diverses façons. Les différents types de camouflage se classent parmi deux grandes catégories, soit le camouflage fait au codeur et celui fait au décodeur. Plusieurs techniques sont présentées en revue par Wah [Wah *et al.*, 2000] et Perkins [Perkins *et al.*, 1998].

Le camouflage au codeur peut être du type actif ou passif. Le type actif est simplement une retransmission. La retransmission permet souvent de recevoir le paquet perdu, mais nécessite un long temps de traitement (s'apercevoir que le paquet est perdu, envoyer une requête de retransmission puis recevoir et traiter le paquet qui arrive en retard). Ce long délai est prohibitif pour les applications de voix en temps réel et restreint l'utilisation de cette alternative. Une méthode efficace pour limiter les dégâts inhérents aux paquets perdus est de mélanger les trames et de former des paquets hétérogènes. Une séquence de 16 trames réparties en 4 paquets peut être envoyée selon un ordre chronologique ([1-2-3-4],[5-6-7-8],[9-10-11-12],[13-14-15-16]), mais aussi selon un ordre entrelacé d'une trame sur 4 ([1-5-9-13],[2-6-10-14],[3-7-11-15],[4-8-12-16]). Ainsi, si le troisième paquet est perdu, les trames perdues ne seront pas consécutives et il sera plus facile de les camoufler au décodeur. Cette méthode a par contre le désavantage d'imposer un certain délai entre l'émetteur et le récepteur puisque pour envoyer le premier paquet, la treizième trame est nécessaire.

Toujours selon Perkins [Perkins *et al.*, 1998], une autre grande classe de camouflage faite au codeur est celle du FEC (Forward Error Correction). Ces méthodes servent à réparer le signal lorsque des trames sont perdues. Le FEC se subdivise en deux méthodes en vertu de la dépendance ou l'indépendance entre le camouflage et les caractéristiques du signal. La méthode indépendante du signal utilise n paquets pour créer un paquet supplémentaire qui sera lui aussi envoyé au décodeur. Parmi les différentes techniques rencontrées, l'opération ou-exclusif (XOR) est fréquemment employée. L'opérateur est appliqué à plusieurs paquets pour générer un paquet de parité. Lorsqu'un seul des paquets n est perdu, il est possible de reconstruire le paquet perdu grâce au paquet de parité. La méthode dépendante du signal quant à elle, consiste à envoyer certaines caractéristiques du signal avec le paquet suivant (par exemple, un codage secondaire avec une largeur de bande réduite et une moins bonne qualité).

Bien que ces camouflages permettent de récupérer le signal, ils entraînent des délais additionnels et une augmentation de la bande passante nécessaire selon la quantité d'informations supplémentaires transmises.

D'autres techniques peuvent être appliquées au codeur pour réduire l'impact d'une trame perdue. Kabal [Tosun et Kabal, 2005] propose d'ajouter de la redondance dans l'information transmise. La redondance est concentrée sur les paramètres de l'excitation de la trame suivante selon l'importance de la trame. Cette importance est caractérisée par une différence importante de l'énergie entre les paramètres d'excitation des deux trames (courante et suivante). La redondance est incluse dans environ 11% des trames envoyées. Le

fait de ne pas envoyer de redondance à toutes les trames permet de limiter l'augmentation de débit introduit par cette méthode.

Andersen [Andersen *et al.*, 2002] suggère un prédicteur long-terme indépendant d'une trame à l'autre. Ainsi, suite à une trame perdue, il n'y a pas de désynchronisation entre le codeur et le décodeur ce qui limite la propagation d'erreurs. Eksler [Eskler et Jelinek, 2008] remplace le prédicteur long-terme par un dictionnaire d'impulsions glottales non-prédictif pour les trames qui suivent les transitoires. Les trames transitoires sont plus sensibles à la perte de trames parce que la construction de la partie périodique du signal ne peut être basée sur la dernière bonne trame reçue qui est non-voisée. Le remplacement du prédicteur long-terme par un dictionnaire d'impulsions glottales permet de limiter plus rapidement la propagation d'erreurs lorsque la trame perdue est une transitoire.

Du coté décodeur, le camouflage se fait par insertion. Perkins [Perkins *et al.*, 1998] et Bhute [Bhute et Shrawankar, 2008] résument plusieurs approches possibles. Il est possible d'insérer des zéros, du bruit ou répéter la trame précédente. Ces méthodes sont très simples à implémenter, mais sont très peu efficaces. Il est aussi possible de remplacer la trame perdue par une combinaison de la trame précédente et de la trame suivante. La première partie de la trame perdue est remplacée par la dernière moitié de la trame précédente et la deuxième moitié par la première partie de la trame suivante. Lorsqu'une trame perdue est située entre deux trames voisées, la reconstruction peut se faire en fonction du pitch de la trame précédente et celui de la trame suivante. De cette façon, la reconstruction produit un pitch cohérent. Lorsque la trame perdue est entre deux trames non-voisées, il est possible de simplement répéter la trame précédente.

Lorsqu'une trame est perdue, la méthode de camouflage utilisée estime la trame manquante. Avec un codeur prédictif, une estimation du filtre de prédiction à long-terme et du filtre de synthèse est faite par le camouflage. Cette estimation peut faire en sorte qu'il y a propagation de l'erreur sur plusieurs trames puisque l'état des filtres du codeur et du décodeur n'est plus identique. Pour réduire la propagation de l'erreur lorsqu'une trame est en retard, Gournay [Gournay *et al.*, 2003] propose d'utiliser les trames qui arrivent en retard pour mettre à jour l'état des filtres de prédiction. Lorsqu'une trame est reçue trop tard pour pouvoir être synthétisée, mais quand même assez rapidement pour faire la mise à jour du filtre prédictif et du filtre de synthèse pour resynchroniser le codeur et le décodeur avant la synthèse de la trame suivante, l'erreur est alors circonscrite à deux trames (ce qui permet d'éviter les discontinuités entre la trame perdue et la trame suivante).

Dans le codeur VMR-WB, le camouflage fait en sorte qu'il y a une convergence des paramètres du signal vers les paramètres correspondants au bruit de fond. La vitesse de convergence est fonction de la classification de la dernière bonne trame reçue. Lorsque la dernière trame reçue est non-voisée, la vitesse de convergence dépend aussi de la stabilité du filtre de synthèse. Dans le cas où le filtre est stable, le signal converge plus lentement. Si le filtre est moins stable, le signal converge plus rapidement.

L'excitation du filtre de synthèse de la trame perdue est construite à partir d'une partie aléatoire et d'une partie périodique. La partie aléatoire est toujours présente pour tous les types de trames dans l'excitation. La partie périodique n'est pas utilisée pour la reconstruction des trames non-voisées, la classification se rapportant toujours à la dernière bonne trame reçue.

La partie aléatoire de l'excitation est obtenue en générant un signal aléatoire de distribution uniforme et pondérée en fonction de l'énergie du signal de la dernière bonne trame reçue. Le gain de la partie aléatoire g_s est initialisé en fonction du gain obtenu dans la dernière bonne trame reçue g dans chacune des quatre sous-trames ($g(0)$, $g(1)$, $g(2)$ et $g(3)$) selon l'équation suivante :

$$g_s = \gamma_0 g(0) + \gamma_1 g(1) + \gamma_2 g(2) + \gamma_3 g(3) \tag{2.34}$$

où $\gamma_0 = 0,1$, $\gamma_1 = 0,2$, $\gamma_2 = 0,3$ et $\gamma_3 = 0,4$.

Le gain évolue selon l'équation suivante :

$$g_s^1 = \alpha g_s^0 + (1 - \alpha) g_n \tag{2.35}$$

où α est la vitesse de convergence qui dépend de la classification de la dernière bonne trame reçue, g_s^1 est le gain au début de la trame suivante, g_s^0 est le gain au début de la trame courante et g_n est le gain de l'excitation lors des trames de silence. Le gain évolue de façon linéaire (échantillon par échantillon) jusqu'à atteindre la valeur g_s^1 à la fin de la trame.

La partie périodique de l'excitation est reconstruite soit à partir de la dernière période de pitch de la dernière bonne trame reçue, soit à partir de la dernière période de pitch de la dernière trame voisée stable reçue. Les trames considérées comme étant voisées stables sont : trame voisée précédée d'une trame voisée, transitoire voisée ou transitoire. Les deux

périodes de pitch sont comparées entre elles à l'aide de (2.36) :

$$\text{si } ((T_3 < 1,8T_s)\text{ET}(T_3 > 0,6T_s))\text{OU}(T_{cnt} \geq 30), \text{ donc } T_c = T_3, \text{ sinon } T_c = T_s \quad (2.36)$$

T_3 est la période de pitch de la 4e sous-trame de la dernière bonne trame reçue, T_s est la dernière période de pitch de la dernière trame voisée stable, T_{cnt} représente le nombre d'échantillons maximums qu'il peut y avoir entre T_3 et T_s. T_c est le pitch appliqué à la trame reconstruite.

2.6 Détection des transitoires

Dans cette thèse, la transitoire est définie comme étant la transition entre le signal non-voisé et le signal voisé. Il s'agit d'un événement bien défini dans le temps, soit quelques échantillons de signal.

Dans la littérature, la détection de la transitoire est généralement abordée dans un contexte musical. Pour bien mettre la détection de transitoires en contexte, il est important de distinguer le contexte musical du contexte vocal. Au niveau du vocabulaire, les termes onset, attaque et transitoire sont ici utilisés dans un contexte musical contrairement au reste du document. La figure 2.5 illustre bien les trois concepts différents. Lorsqu'il est question d'onset, il s'agit du début du signal actif et cet événement est très ponctuel. L'attaque définit la montée d'énergie/amplitude du signal, alors que la transitoire est la partie du signal qui évolue de façon imprévisible ou non-prédictible.

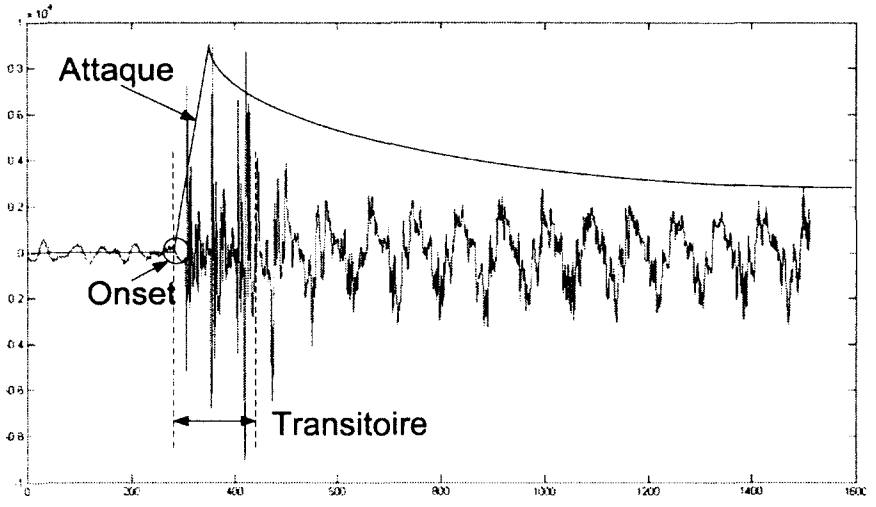


Figure 2.5 Définition de l'onset, de l'attaque et de la transitoire

Dans le domaine musical, trois types d'onset distincts peuvent être détectés. Il s'agit d'onsets provenant de percussions qui ne produisent qu'une seule valeur de pitch (tambour, cymbale, etc), d'onsets provenant de percussions qui peuvent produire plusieurs valeurs de pitch (piano, etc) et d'onsets provenant d'instruments qui peuvent produire plusieurs valeurs de pitch (violon, clarinette). Généralement, l'approche pour détecter les onsets de ces trois familles d'instruments diffère. Dans la première famille (percussion avec un seul pitch), il y a une montée nette d'énergie associée à chaque note jouée. Pour la famille des percussions dont le pitch peut varier, il y a aussi une montée d'énergie liée à chaque note, ainsi qu'une variation en fréquence lorsqu'il y a une variation dans le pitch. Pour ce qui est des instruments qui ne sont pas des percussions, la variation d'énergie entre les notes peut être presque nulle, il faut donc d'avantage se fier à la variation de fréquence pour détecter les onsets.

La détection de transitoires dans les signaux musicaux peut se faire dans le domaine temporel, mais en général, elle ne se fait pas directement sur le signal original. Dans les divers cas étudiés, il y a un pré-traitement du signal, de telle sorte que certaines caractéristiques sont amplifiées ou atténuées. Cette étape a pour but de préconditionner le signal pour améliorer les performances des étapes subséquentes. Par la suite, le signal, souvent sous-échantillonné, est traité par une fonction de détection. Pour terminer, un algorithme de détection de pics localise les transitoires. Pour mieux comprendre le processus de détection, les étapes énumérées précédemment seront reprises et explicitées.

2.6.1 L'étape de prétraitement

L'étape de prétraitement sert à mettre en évidence certaines caractéristiques du signal dans le but de faciliter les étapes subséquentes. Pour contribuer efficacement à la détection des onsets, le prétraitement doit faire ressortir les transitoires dans le signal. Deux approches sont fréquemment suggérées ; la première est la séparation du signal en différentes bandes de fréquences et la deuxième est la séparation du signal en parties transitoires et stationnaires.

La séparation du signal en bandes de fréquences précède une analyse distincte sur chacune des bandes. La combinaison des résultats obtenus est ensuite utilisée pour la détection des onsets. Par exemple, il est possible de suivre l'apparition et la disparition de notes à travers les différentes bandes de fréquences (de façon individuelle) et de recombiner le tout dans une application de détection de tempo (voir Scheirer [Scheirer, 1998]). D'autres approches combinent le résultat de chacune des bandes pour déterminer la position de la transitoire (voir Klapuri [Klapuri, 1999]).

2.6.2 La fonction de détection

La fonction de détection met en évidence les transitoires qui composent le signal. Le signal obtenu grâce à la fonction de détection est souvent sous-échantillonné. Deux grands groupes de fonctions de détection sont généralement utilisés ; ceux basés sur les caractéristiques du signal et ceux basés sur des modèles statistiques.

Les caractéristiques du signal peuvent se trouver dans le domaine temporel ou dans le domaine fréquentiel. Dans le domaine temporel, une transitoire s'accompagne d'une augmentation d'amplitude dans l'enveloppe du signal. Un simple détecteur d'enveloppe peut être construit à l'aide d'un redresseur et d'un filtre passe-bas. Ce genre de détecteur fonctionne bien avec des événements très distincts du bruit de fond (ayant une forte amplitude), mais trouve difficilement les événements plus subtils (ayant une faible amplitude). En raffinant cette approche, il est possible de travailler avec la dérivée de l'énergie du signal en fonction du temps ou d'utiliser l'échelle logarithmique de la pression acoustique perçue par l'oreille humaine, soit $logP(n)$. Puisque l'information de la structure temporelle du signal se trouve dans la phase, cette dernière peut aussi être utilisée. Lorsque le signal est stationnaire, la fréquence instantanée du signal est à peu près constante. La fréquence instantanée se calcule comme suit :

$$f_k(n) = (\frac{\varphi_k(n) - \varphi_k(n-1)}{2\pi h})f_s \tag{2.37}$$

où f_s est la fréquence d'échantillonnage, $\varphi_k(n)$ est la phase à l'instant n et $\varphi_k(n-1)$ est la phase à l'instant $n-1$ pour la bande de fréquence k et h est la distance entre les fenêtres d'analyses.

Donc, lorsque la fréquence instantanée est à peu près constante, on peut écrire :

$$(\varphi_k(n) - \varphi_k(n-1)) \cong (\varphi_k(n-1) - \varphi_k(n-2)) \tag{2.38}$$

et donc :

$$\Delta\varphi_k(n) = \varphi_k(n) - 2\varphi_k(n-1)) + \varphi_k(n-2) \cong 0 \tag{2.39}$$

En zone de transition, puisque la fréquence instantanée n'est pas constante, $\Delta\varphi_k(n) >> 0$. Comparer $\Delta\varphi_k(n)$ à un seuil de détection judicieux permet alors de détecter une transitoire.

Dans le domaine fréquentiel, si le signal est analysé du point de vue spectral, les transitoires sont détectées par la présence d'une grande quantité d'énergie spectrale dans les hautes

fréquences. Il est possible de faire la somme de l'énergie des différentes bandes de fréquences en appliquant une pondération différente à chacune d'elles et d'établir un seuil au-dessus duquel une transitoire est déclarée.

Différentes méthodes statistiques peuvent être utilisées pour décrire le signal analysé. En fonction de la justesse des modèles utilisés, les résultats obtenus pour la détection de la transitoire seront plus ou moins concluants. L'une des méthodes proposées se nomme *sequential probability ratio test* [Bello *et al.*, 2005].

Une autre approche proposée par Abdallah [Abdallah et Plumbley, 2003] est basée sur la «surprise» de l'événement. Un signal est normalement prédictible et prévisible. Par contre, lorsqu'une transitoire survient le signal change de telle sorte que son évolution devient difficile à prévoir. L'effet de surprise étant localisé vis-à-vis les zones de transition, sa mesure peut servir de fonction de détection. L'équation 2.40 définit l'effet de surprise.

$$S(n) \equiv S(x(n)) = -logP(x(n)) \tag{2.40}$$

où P est la probabilité du signal $x(n)$.

Gainza [Gainza *et al.*, 2005] propose d'utiliser un filtre en peigne, combiné à l'augmentation de l'énergie dans le signal pour détecter les onsets. Cette technique a été testée sur des onsets dont l'évolution est plus lente, comme ceux produits par les instruments à vent. Le filtre en peigne est une sommation entre le signal original et une version décallée de lui-même, comme présenté à l'équation 2.41, où g est le gain du filtre en peigne et D est le délai appliqué au filtre.

$$y[n] = x[n] + g \cdot x[n - D] \tag{2.41}$$

La détection des onsets se fait dans le domaine fréquentiel. Dans un premier temps, une transformée de Fourier est appliquée sur le signal. Le résultat de la transformée de Fourier est ensuite filtré par le filtre en peigne selon l'équation 2.42, où m représente le numéro de trame traitée, k représente la bande de fréquence et D_i est le délai du filtre en peigne qui varie entre D_{min} et D_{max}.

$$Y_{D_i}(m, k) = X(m, k) \times H(D_i, k) \tag{2.42}$$

Ensuite, l'énergie du résultat du filtrage par le filtre en peigne $Y_{D_i}(m, k)$ est calculé selon l'équation 2.43, où M est le nombre de bandes de fréquence obtenu par la transformée de

Fourier.

$$E(m, D_i) = \sum_{k=1}^{M} \left| Y_{D_i}(m, k)^2 \right| \tag{2.43}$$

Selon l'équation 2.41, la valeur maximale de $y[n]$ est de $2 \cdot x[n]$, si g vaut 1 et que $x[n] = x[n+D]$. Ainsi, l'énergie maximale de $y[n]$, soit $E(y_{max})$ est égale à $4 \cdot x[n]^2$. En normalisant l'équation 2.43 par $E(y_{max})$, une mesure de ressemblance entre le filtre en peigne utilisé et le filtre en peigne idéal est obtenu. Dans le cas où le filtre en peigne suit parfaitement les harmoniques d'un signal, le ratio de 1 sera obtenu. Dans le cas où le signal est composé de plusieurs fréquences qui ne sont pas harmoniques entre elles, un ratio près de 0 sera obtenu. C'est le cas lorsqu'il y a présence d'un onset dans le signal analysé.

Pour faciliter l'analyse, la logique de l'équation 2.43 est modifiée selon l'équation 2.44 pour obtenir la valeur 0 pour un signal harmonique et la valeur 1 pour les onsets.

$$E'(m, D_i) = abs(E(m, D_i) - 1) \tag{2.44}$$

Pour détecter les onsets, une dérivée est appliquée sur le résultat (équation 2.45).

$$dE(m) = \sum_{i=D_{min}}^{D_{max}} [E'(m, D_i) - (E'(m-1, D_i) - 1)]^2 \tag{2.45}$$

Collins [Collins, 2005] détermine la fréquence fondamentale du signal afin de mieux cibler la recherche des onsets. La fréquence fondamentale du signal est trouvée grâce à l'analyse du spectre selon une résolution *constant-Q*. Comme l'intervalle entre les fréquences fondamentales des différentes notes de musique, ainsi que leurs harmoniques suivent une échelle logarithmique, l'analyse d'un signal de musique peut être faite avec une résolution constant-Q (résolution qui suit une échelle logarithmique). De cette façon, l'énergie d'une note et ses harmoniques est distribuée à travers les différentes bandes de fréquences.

Suite à l'analyse, le résultat du suiveur de pitch est converti dans une échelle de demi-ton (plus petite variation de fréquence entre deux notes consécutives). Par la suite, une analyse des vibratos est faite sur le signal. Les vibratos sont des variations de faibles amplitudes qui modulent la fréquence du signal. Lorsque des vibratos sont identifiés, ils sont retirés du signal. L'étape de détection des onsets est ensuite faite en cherchant les diférences significatives entre les valeurs de pitchs trouvées.

Les méthodes de détection d'onsets sont souvent conçues pour détecter soit les variations d'énergie marquées, soit les variations de fréquences qui sont facilement détectables, plu-

sieurs auteurs proposent des approches combinées (deux méthodes ou plus) de façon à détecter les onsets de nature distincte.

Zhou [Zhou *et al.*, 2008] classifie les signaux à analyser et détermine la meilleure approche à utiliser pour détecter les onsets, soit une approche basée sur la détection du pitch, soit une approche basée sur l'énergie pour détecter la position des onsets. Une analyse en fréquence est faite sur le signal. Ensuite, une différence de premier ordre est calculée sur le spectre du signal et une moyenne est faite sur la totalité du signal. Si cette moyenne est plus élevée qu'un seuil prédéfini, l'approche basée sur l'énergie est utilisée. Dans le cas contraire, l'approche basée sur le pitch est utilisée pour trouver les onsets.

L'approche basée sur l'énergie utilise la même différence de premier ordre appliquée à chacune des bandes spectrales analysées. La différence est comparée à un seuil, et seulement les bandes dont la différence est plus élevée que le seuil seront considérées dans le calcul de la moyenne. La moyenne de toutes les bandes retenues est faite et cette valeur est comparée à un deuxième seuil. Dans le cas où la moyenne est supérieure au seuil, la position est retenue comme un candidat potentiel pour un onset. Si deux onsets sont détectés dans un intervalle de moins de 50 ms, seulement celui dont la moyenne est la plus élevée est retenu.

L'approche basée sur le pitch sépare le signal en parties transitoires et en parties stationnaires à partir d'un détecteur de pitch. À partir des parties stationnaires du signal, une recherche est faite dans le passé pour trouver l'endroit où il y a une transition d'énergie marquée qui correspond au début de l'onset. Cette transition doit être plus grande qu'un seuil prédéterminé. Cette recherche est faite dans les bandes de fréquences qui correspondent au pitch du signal et à ses harmoniques (généralement l'énergie est concentrée dans les dix premières harmoniques du signal). L'intervalle de temps où la recherche est effectuée est de 300 ms avant le début de la partie stationnaire du signal. Encore une fois, si l'espacement entre plusieurs candidats est inférieur à 50 ms, celui dont la transition d'énergie est la plus marquée est retenu.

Tan [Tan *et al.*, 2010] combine aussi deux approches pour détecter les onsets soit par une variation d'énergie, soit par une variation de fréquences. Pour la détection des onsets où il y a présence d'une variation rapide d'énergie, le résultat de la transformée de Fourier rapide est utilisé. Une variation de l'énergie en fonction des bandes de fréquences est recherchée et chacun des onsets potentiels est gardé en mémoire. Une détection de pitch avec la méthode du *constant-Q* est utilisée pour les onsets qui sont détectés grâce à la variation de fréquence. En fonction du pitch identifié et de la présence de variations d'énergie, les onsets potentiels sont catégorisés selon les trois catégories décrites au début de la section,

soit les onsets percussifs sans pitch associé, les onsets percussifs avec pitch associé et les onsets non-percussifs avec pitch associé. Pour les onsets percussifs, l'énergie du signal sur toute la bande de fréquence est comparée à un seuil pour confirmer la présence d'un onset. Dans le cas des onsets percussifs avec pitch associé, la détection se fera à la fois par une détection de variation de fréquence et d'énergie. Il faut savoir par contre que la variation de fréquence a priorité dans ces détections d'onsets et que la variation d'énergie n'est pas essentielle pour que l'onset soit détecté. Si la variation d'énergie est prise en compte, c'est seulement l'énergie de la bande de fréquence associée au pitch qui est utilisé pour faire la validation. Quant aux onsets non-percussifs, seule la variation de pitch est prise en compte pour faire la détection de l'onset.

2.6.3 Localisation de pics

La dernière étape de la détection de transitoire musicale se trouve dans la localisation des pics. Si la fonction de détection est efficace, les transitoires seront facilement localisés par des maximums locaux du résultat de la fonction de détection. Ces maximums peuvent par contre être plus ou moins francs et être noyés dans du bruit. Pour trancher entre les pics qui sont réellement des transitoires et les autres pics, il faut établir un seuil au-dessus duquel le maximum du pic sera déclaré comme une transitoire. Le seuil peut être fixe ou adaptatif.

Levine [Levine, 1998] propose d'utiliser l'approche de décomposition du signal en sinusoïdes et en bruit pour trouver le positionnement des transitoires. Une augmentation significative de la différence entre le modèle ainsi obtenu et le signal original correspond à une transitoire. En période de transition, le modèle sinusoïde et bruit ne modélise pas bien le signal, d'où l'augmentation d'énergie dans la différence entre le modèle et le signal original.

2.7 Conclusion

Différentes approches ont été présentées pour l'estimation de la fréquence fondamentale, la classification des trames, la modification du signal, la détection des trames transitoires pour l'amélioration du codage et les différentes méthodes utilisées pour le camouflage. Chaque étape représente un aspect de la thèse proposée et servira d'inspiration pour la mise en oeuvre de solutions aux différents aspects du problème à résoudre.

CHAPITRE 3

Application de l'opérateur de Teager à la détection de transitoires

Hypothèse #1 : Il est possible d'améliorer la qualité du signal de parole lorsqu'il y a pertes de trames en modifiant la classification des trames à l'encodeur.

Ce chapitre présente une première contribution de la thèse qui vise l'amélioration de la détection des transitoires dans un signal de parole. L'objectif est d'améliorer la qualité du signal de synthèse dans le cas où un codeur CELP est utilisé sur un canal avec pertes de paquets. Plus spécifiquement, le standard VMR-WB [Ahmadi, 2005] sera utilisé comme exemple d'un codeur CELP récent. Le standard VMR-WB étant basé sur un modèle de codage multi-mode pour le signal d'excitation, la classification (et notamment la détection de transitoires) a une incidence à la fois sur la qualité de l'encodage et sur l'algorithme de camouflage des trames perdues au décodeur. Plusieurs classes sont possibles dans le standard VMR-WB et sont énumérées à la section 2.3.3.

Il est démontré [Jelinek et Salami, 2007] que le camouflage de trames perdues est plus efficace lorsqu'il peut être adapté en fonction du type de signal perdu, et donc en fonction d'une information de classification. De plus, considérant des cas particuliers, il est encore possible d'apporter des améliorations. Spécifiquement, trois cas de figure entourant la période transitoire entre les trames non-voisées et les trames voisées peuvent être identifiés. Ces trois cas sont exposés en détail dans la section suivante.

3.1 Détection de la transition non-voisé $\rightarrow$ voisé

Comme introduit dans le chapitre 2, le standard VMR-WB utilise une fonction de coût définie comme une somme pondérée de six variables pour effectuer la classification des trames. Cette fonction permet, en général, de bien discriminer les différentes classes possibles pour améliorer le camouflage des trames perdues. Par contre, il arrive au classificateur de se tromper. Une mauvaise classification peut avoir un impact sur le camouflage puisque celui-ci est basé sur la classification de la dernière trame reçue. L'impact négatif d'une mauvaise classification est particulièrement problématique pour les transitions non-

voisées à voisées appelées simplement «transitoires» dans le texte suivant. Cette section traite plus particulièrement de cette problématique.

Un premier cas problème survient lorsque le classificateur détecte la transitoire trop rapidement. Une trame totalement non-voisée ou inactive est alors déclarée comme étant une transitoire. Si la trame suivant la transitoire déclarée est perdue, la dernière période de pitch évaluée par le codec est répétée (le camouflage suivant une trame transitoire est le même que celui suivant une trame voisée étant donné qu'une trame transitoire est définie comme une trame ayant une structure périodique bien définie vers sa fin). Cette période de pitch répétée par le processus de camouflage ne correspond pas à la structure réelle du signal manquant puisque la période de pitch n'est pas associée à une réelle structure harmonique dans le signal. Il peut en résulter différents artéfacts audibles. Ce cas est illustré à la figure 3.1. Comme il est possible de voir à la sous-figure 3.1(b), lorsque le camouflage répète une période de pitch sans fondement à cause d'une mauvaise classification, la désynchronisation entre le codeur et le décodeur causée par la trame perdue peut s'échelonner sur plusieurs trames.

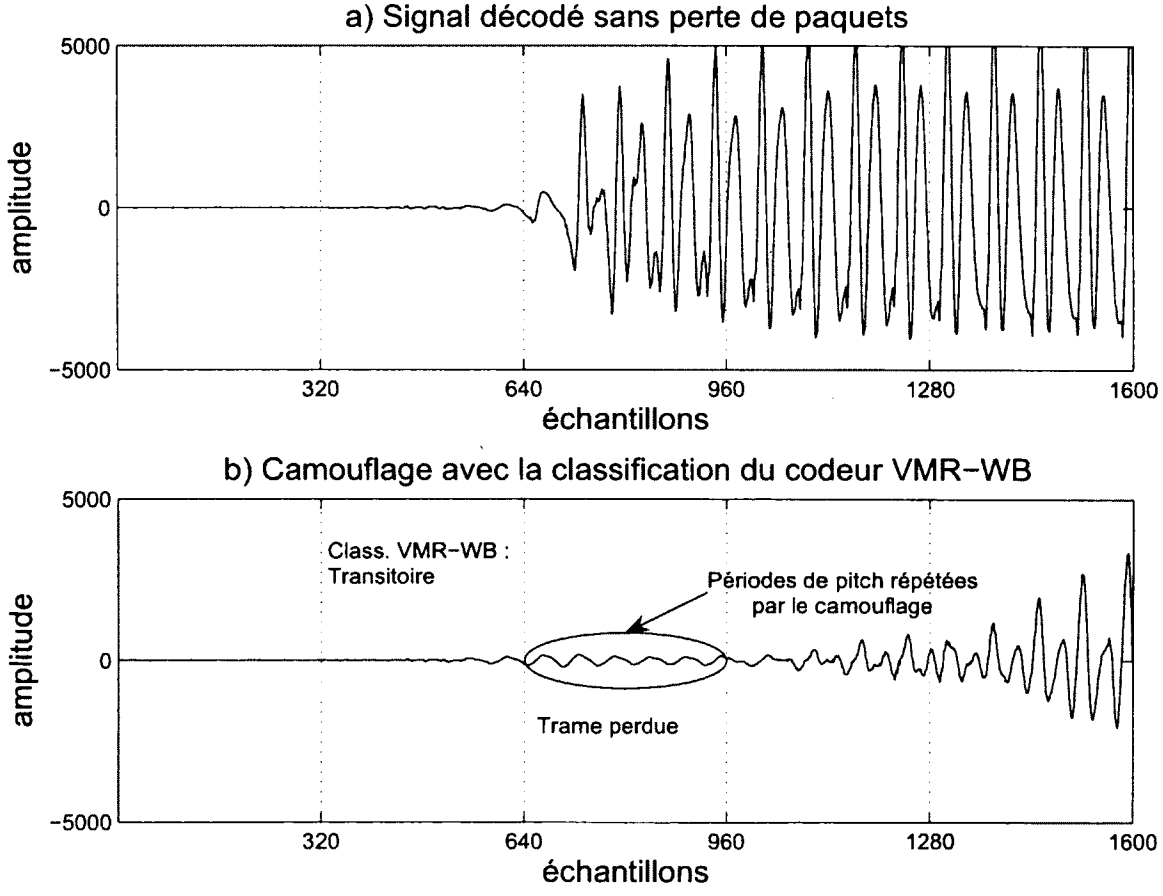


Figure 3.1 Camouflage lorsque la transitoire est déclarée trop rapidement

Un deuxième cas problème est très similaire au cas décrit ci-dessus ; seule une fraction de la bonne période de pitch est présente à la toute fin de la trame, mais pas suffisamment pour que le camouflage s'effectue correctement si la trame suivante est perdue. Le camouflage répétera la période de pitch évaluée à la fin de la dernière bonne trame reçue. Comme cette période est incomplète, elle ne représente qu'une fraction de la véritable période de pitch. Cette mauvaise période de pitch aura des répercussions sur plusieurs trames puisqu'il y aura désynchronisation entre le codeur et le décodeur. La figure 3.2 illustre ce cas.

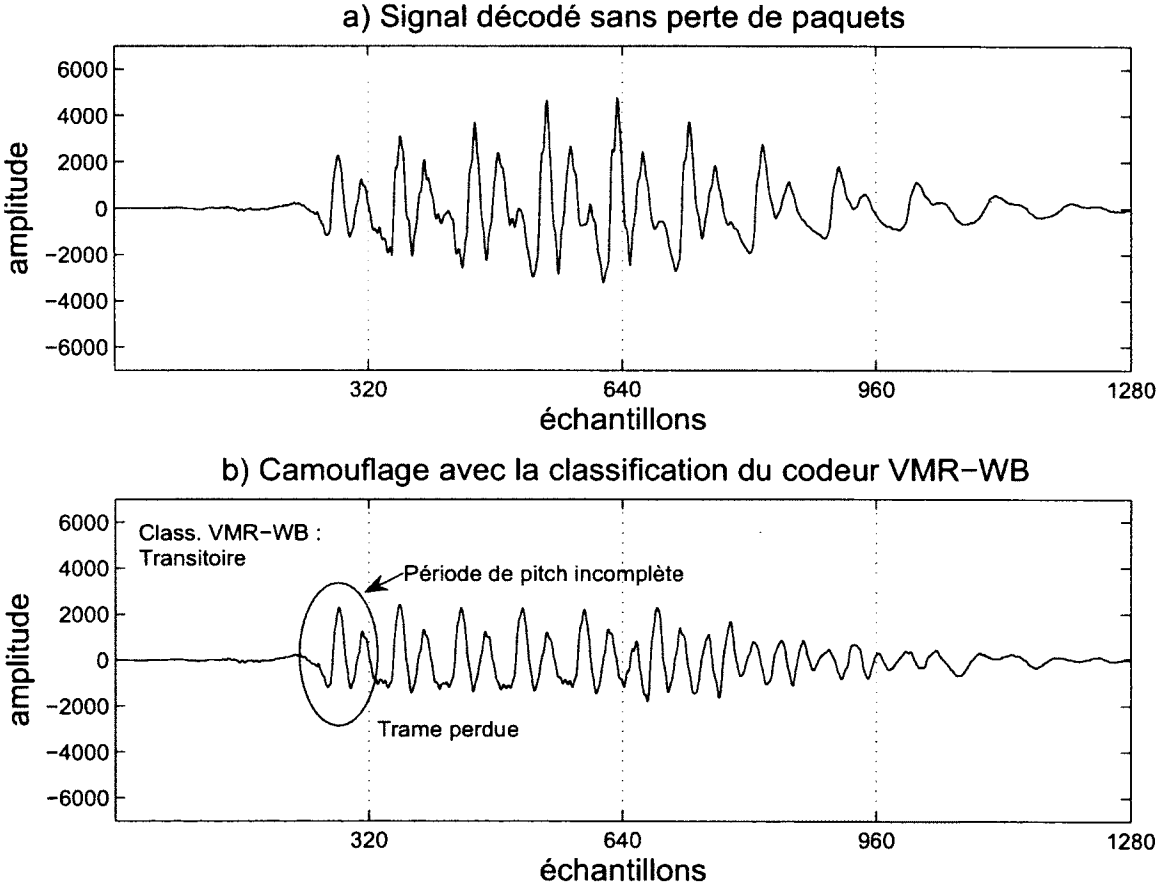


Figure 3.2 Camouflage lorsque la trame précédant la trame manquante comprend une période de pitch incomplète

Le troisième cas est différent des deux précédents. Il se produit lorsque la trame transitoire est classée comme étant une transition non-voisée et non stationnaire. Dans cette situation, si la trame suivante est perdue, le camouflage réagit en amenant rapidement l'énergie à zéro. Il survient alors un artéfact bien audible puisqu'il y a une montée d'énergie suivie d'une descente rapide. Ce cas est illustré à la figure 3.3. Le cas où une trame transitoire est classée comme étant une trame non-voisée donne un résultat similaire même si le signal de la trame est stationnaire. Dans le cas d'un signal non-voisé stationnaire, l'énergie du signal

reconstruit sera plus stable. Néanmoins un artéfact audible pourrait survenir puisque le camouflage n'inclut pas la partie périodique dans le signal reconstruit.

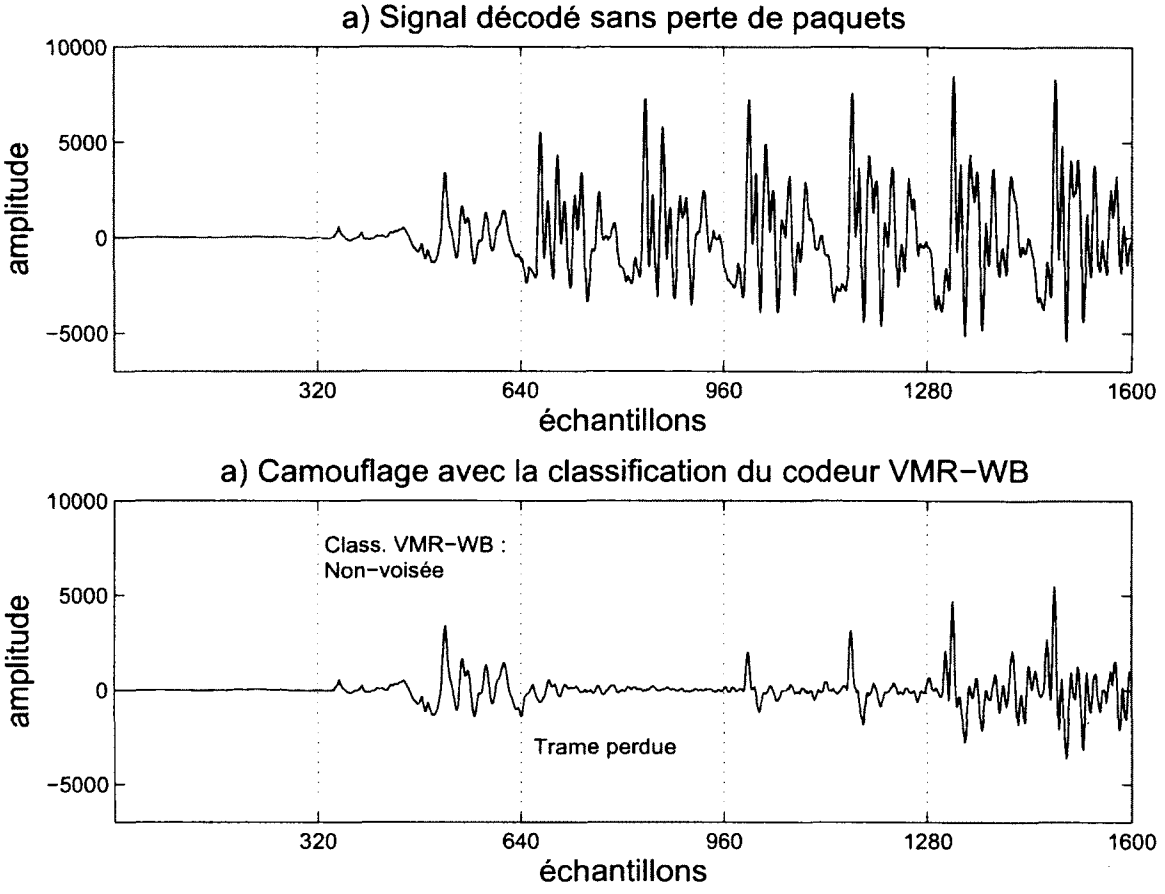


Figure 3.3 Camouflage lorsque la transitoire est déclarée comme étant une trame *non-voisée*

Ces trois cas, même s'ils sont relativement peu fréquents dans un signal de parole typique, peuvent altérer la qualité du signal reconstruit lorsqu'il y a des pertes de trames. Pour éviter ce problème, il est nécessaire d'identifier avec précision le début des transitions non-voisées à voisées. Nous allons montrer dans ce qui suit qu'une technique basée sur une analyse d'énergie instantannée avec l'opérateur de Teager présente des avantages. Cet opérateur calcule l'énergie instantanée qui est proportionnelle à la fois à l'amplitude du signal et aux fréquences qu'il contient. Pour justifier l'utilisation de cet opérateur, la définition traditionnelle de l'énergie ainsi que la détection d'enveloppe sont d'abord présentées.

3.2 L'énergie et l'enveloppe d'un signal de parole

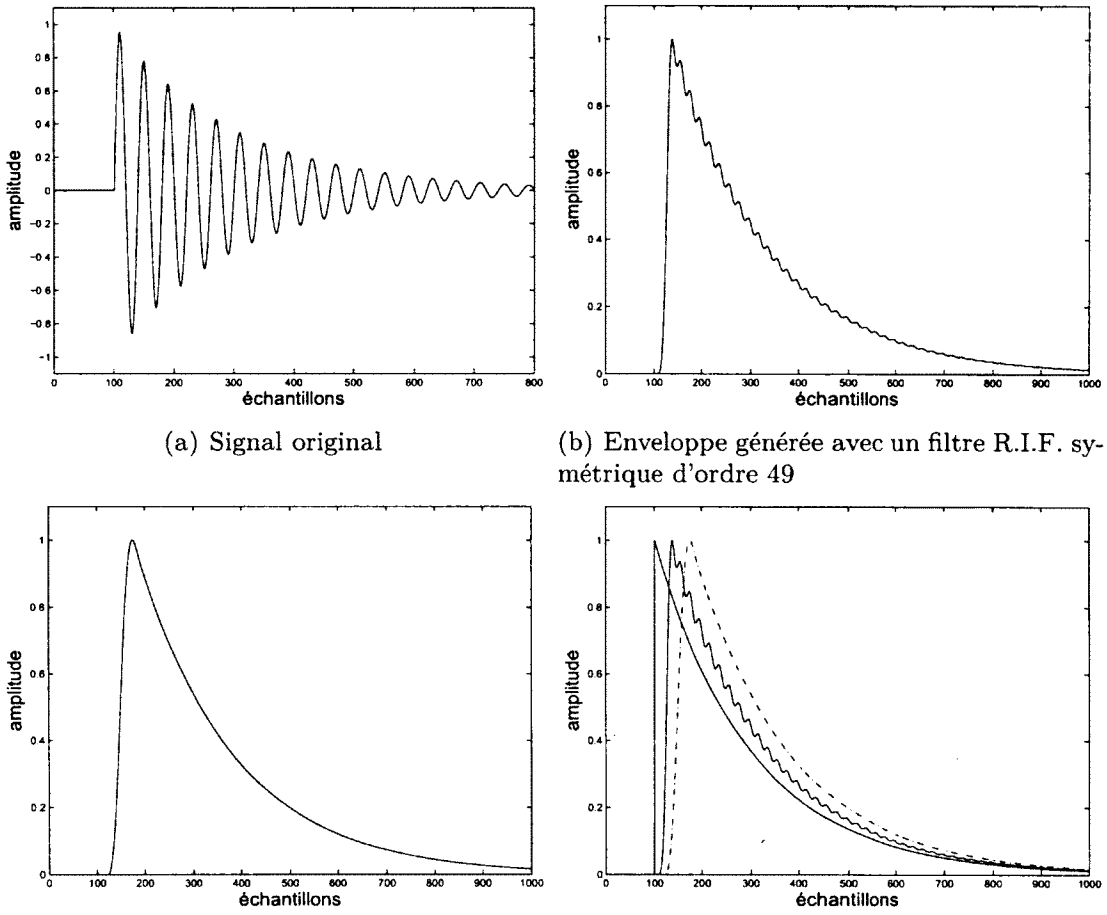
L'énergie contenue dans un signal numérique $x(n)$ de durée N échantillons peut être calculée à partir de l'équation $E = \sum_{n=0}^{N-1} x^2(n)$.

L'enveloppe du signal $x(n)$ en donne l'allure générale c'est-à-dire qu'elle permet de voir l'évolution instantannée de son énergie dans le temps. Une technique souvent employée pour extraire l'enveloppe consiste à appliquer un filtre passe-bas à la valeur absolue du signal.

La figure 3.4 présente un premier exemple d'extraction d'enveloppe pour un signal sinusoïdal d'amplitude décroissante. Deux filtres R.I.F. différents sont utilisés : le premier est un filtre R.I.F. passe-bas d'ordre 49 et le second un filtre R.I.F. passe-bas d'ordre 99. Le filtre d'ordre 99 est le plus petit ordre possible dans le cas démontré pour obtenir une enveloppe suffisamment lisse. Si l'ordre du filtre est diminué de moitié (ici, 49), l'enveloppe obtenue n'est pas lisse et présente encore des oscillations à court-terme du signal. Dans les deux cas, la fréquence de coupure du filtre (point à -3 dB) est de 50 Hz et le signal est échantillonné à 16000 Hz. La figure 3.4 (d) compare l'enveloppe réelle du signal ainsi que les enveloppes obtenues par la technique de redressement et filtrage. Plus le filtre est d'ordre élevé, plus l'enveloppe tend à être lisse. Par contre un ordre élevé implique un retard qui augmente de façon proportionnelle à la longueur du filtre. Ce retard peut être trop grand pour des applications telle que la détection des transitoires pour le camouflage des trames perdues dans un système de communication. Notamment, le retard du filtre ne doit pas être supérieur au nombre d'échantillons d'avance (en anglais, *look-ahead*) disponible dans le traitement à l'encodeur, sinon il n'est pas possible de filtrer tous les échantillons à traiter.

La figure 3.5 illustre la détection d'enveloppe appliquée à un signal de parole échantillonné à 16000 Hz. Les deux filtres R.I.F. utilisés sont respectivement d'ordre 49 et d'ordre 499 (ordre minimal dans le cas de ce signal pour obtenir une enveloppe lisse). La fréquence de coupure des filtres est de 80 Hz. Les résultats montrent que le filtre R.I.F. d'ordre 49 est insuffisant pour bien extraire l'enveloppe du signal. Le filtre R.I.F. d'ordre 499 donne une meilleure approximation, mais son délai de traitement de 250 échantillons (presqu'une trame complète) rend prohibitive son utilisation dans le contexte du standard VMR-WB à moins de permettre un délai de traitement supplémentaire.

La sous-figure 3.5 d) montre le même exemple de signal mais cette fois, il est filtré avec un filtre R.I.F. d'ordre 255 qui a un délai de traitement de 128 échantillons. Ce délai correspond au nombre d'échantillons disponibles de la demi-trame d'avance du standard



(a) Signal original

(b) Enveloppe générée avec un filtre R.I.F. symétrique d'ordre 49

(c) Enveloppe générée avec un filtre R.I.F. symétrique d'ordre 99

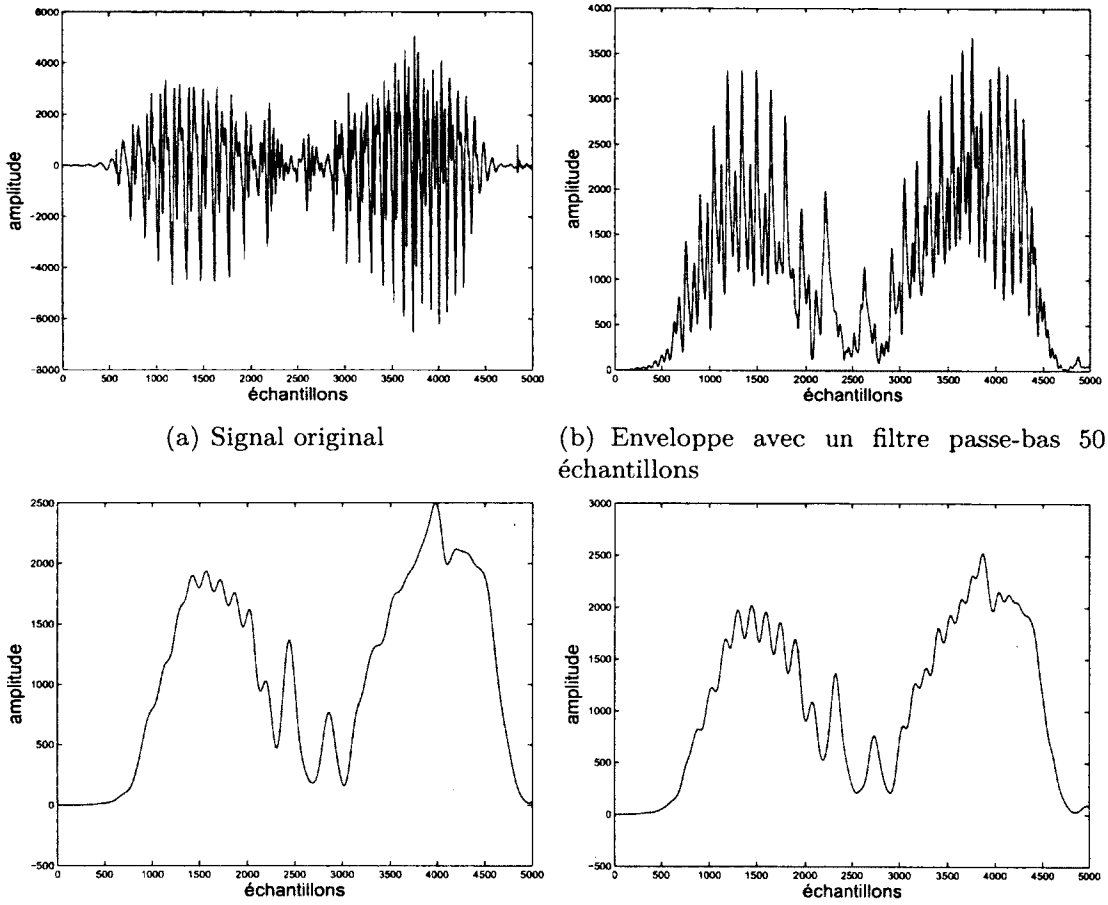
(d) Enveloppe du signal original, plus les deux enveloppes estimées avec des filtres R.I.F. symétriques d'ordre 49 et d'ordre 99

Figure 3.4 Démonstration de détection d'enveloppe sur un signal sinusoïdal décroissant

VMR-WB. C'est à partir de ce signal d'enveloppe qu'il faut maintenant déterminer s'il est possible de bien positionner le début des transitoires.

La figure 3.6 montre un exemple de signal où la position du début de la transition non-voisée à voisée peut être déterminée à l'aide de l'enveloppe du signal à l'aide d'un filtre R.I.F d'ordre 255 et 499. La sous-figure a) montre le signal original et les sous-figures b) et c) montrent respectivement les enveloppes obtenues avec les filtre R.I.F d'ordre 499 et 255. Dans les deux cas, il est possible de bien positionner le début de la transitoire si le seuil d'énergie choisi varie environ entre 200 et 800. L'ordre du filtre n'a pas d'influence sur les résultats obtenus.

Par contre, il y a des cas de signaux où la partie non-voisée du signal de parole contient beaucoup d'énergie relativement à la partie voisée. Dans le cas illustré à la figure 3.7, le



(a) Signal original

(b) Enveloppe avec un filtre passe-bas 50 échantillons

(c) Enveloppe avec un filtre passe-bas 500 échantillons

(d) Enveloppe avec un filtre passe-bas 256 échantillons

Figure 3.5 Démonstration de détection d'enveloppe avec un signal réel

détecteur d'enveloppe positionne le début de la transitoire beaucoup trop rapidement si un seuil d'énergie équivalent à l'exemple précédent est choisi (soit entre 200 et 800), et ce, peu importe l'ordre du filtre choisi. Le seuil de détection de l'énergie devrait être augmenté au dessus de 1100 pour éviter cette fausse détection et pour que le début de la transitoire soit bien positionné.

Une augmentation du seuil à une plus grande valeur n'est pas nécessairement souhaitable puisqu'il arrive que l'amplitude du signal recherché (i.e. la transitoire) soit relativement faible. La figure 3.8 illustre un cas où la détection par l'enveloppe nécessiterait un ajustement du seuil à la baisse. Le seuil de détection minimum devrait être situé autour de 450-500 pour pouvoir détecter le début de la transitoire. Si un seuil autour de 1100 est choisi, le début de la transitoire sera détecté avec trois trames de retard. Encore une fois, l'ordre du filtre n'a pas d'influence sur les résultats obtenus.

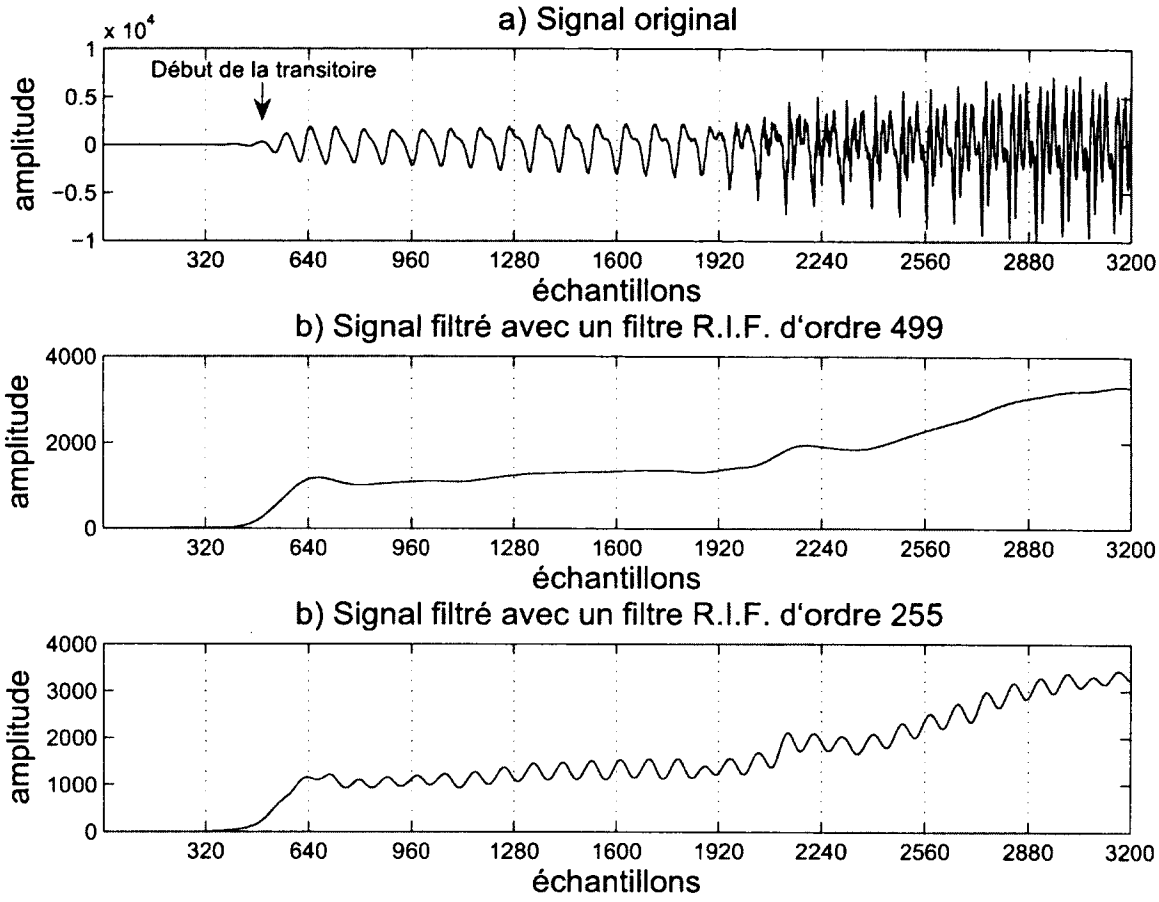


Figure 3.6 Utilisation de l'enveloppe pour déterminer la position de la transitoire avec un filtre R.I.F d'ordre 499 et un filtre R.I.F d'ordre 255

La prochaine section montre qu'en modifiant la définition de l'énergie instantannée, on obtient une mesure présentant un délai presque nul (approche basée sur l'opérateur de Teager) et que le délai global de l'approche proposée incluant le pré-filtrage du signal respecte le délai disponible à l'encodeur. Ainsi, aucun délai supplémentaire n'est introduit par cette approche.

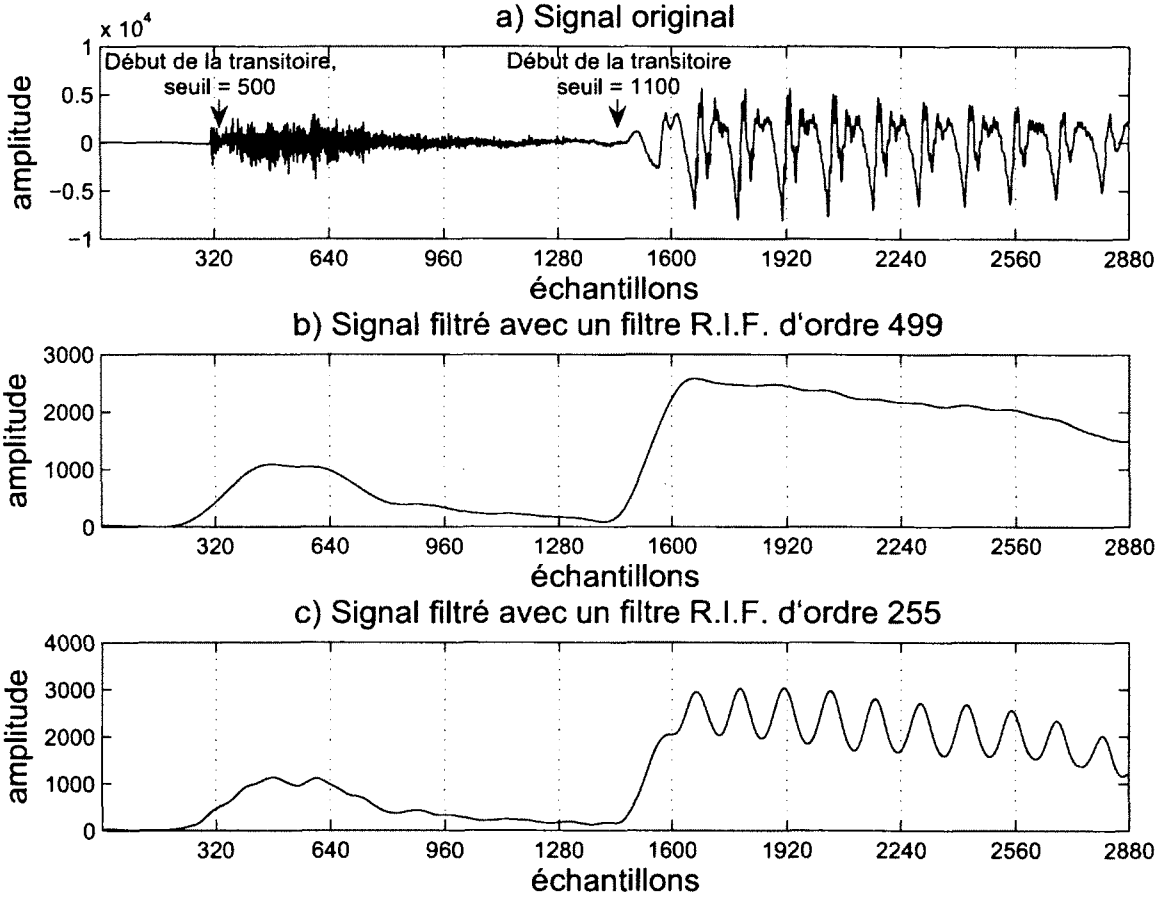


Figure 3.7 Détection du signal de parole non-voisé avec le détecteur d'enveloppe avec un filtre R.I.F d'ordre 499 et un filtre R.I.F d'ordre 255

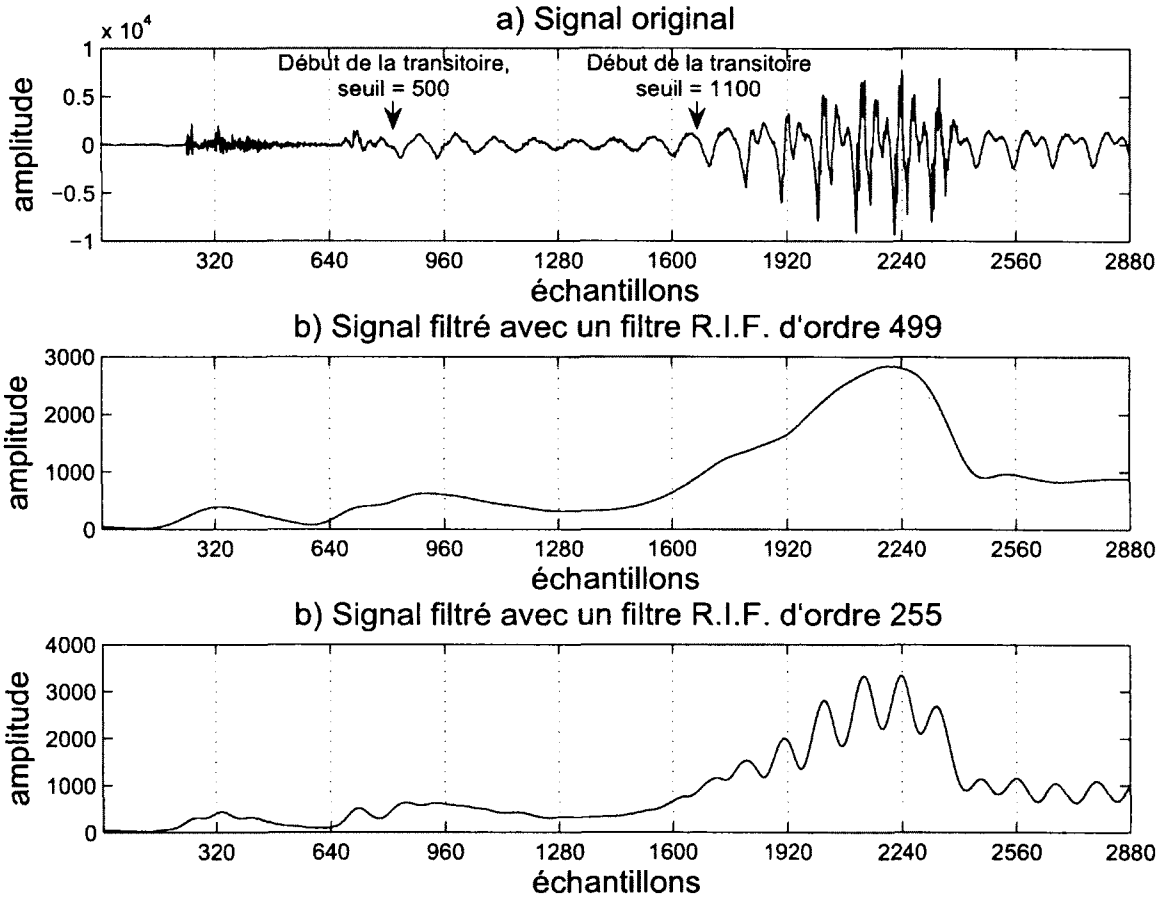


Figure 3.8 Absence de détection de la transitoire avec le détecteur d'enveloppe avec un filtre R.I.F d'ordre 499 et un filtre R.I.F d'ordre 255

3.3 L'opérateur de Teager, son origine

L'opérateur de Teager permet de représenter l'énergie nécessaire à la génération d'un signal harmonique tel que présenté dans l'article de Kaiser [Kaiser, 1990]. Il peut être appliqué par exemple au calcul de l'énergie instantannée dans un système masse-ressort, un système harmonique simple. L'équation de mouvement d'un tel système est $\frac{d^2x}{dt^2} - \frac{k}{m}x = 0$, où x représente le déplacement d'une masse m suspendue au ressort dont la constante élastique est k et t représente le temps. La solution de cette équation du mouvement est une sinusoïde unique, de la forme $x(t) = Acos(\Omega t + \phi)$. L'énergie totale du système est présentée à l'équation (3.1). Elle est la somme de l'énergie potentielle du ressort et de l'énergie cinétique de la masse.

$$E = \frac{1}{2}kx^2 + \frac{1}{2}m\dot{x}^2 \tag{3.1}$$

En introduisant la valeur de $x(t) = Acos(\Omega t + \phi)$ qui représente le mouvement d'oscillation du système masse-ressort où A est l'amplitude du mouvement et $\Omega = \sqrt{k/m}$ représente la fréquence de l'oscillation du système masse-ressort ($k = \Omega^2 m$) dans l'équation 3.1, on montre que :

$$E = \frac{1}{2}k(Acos(\Omega t + \phi))^2 + \frac{1}{2}m(-A\Omega sin(\Omega t + \phi))^2 \tag{3.2}$$

$$E = \frac{1}{2}\Omega^2 m(Acos(\Omega t + \phi))^2 + \frac{1}{2}m(-A\Omega sin(\Omega t + \phi))^2 \tag{3.3}$$

$$E = \frac{1}{2}\Omega^2 mA^2(cos(\Omega t + \phi)^2 + sin(\Omega t + \phi)^2) \tag{3.4}$$

$$E = \frac{1}{2}\Omega^2 mA^2 \tag{3.5}$$

$$E \propto \Omega^2 A^2 \tag{3.6}$$

Ainsi, l'énergie est non seulement proportionnelle au carré de l'amplitude, mais aussi au carré de la fréquence.

On peut également obtenir l'information d'énergie instantannée d'un signal sinusoïdal $x(t)$ sans connaître son amplitude et sa fréquence. On va prendre ici l'exemple d'un signal échantillonné $x(n)$, de fréquence normalisée ω et d'amplitude A (inconnus).

On pose d'abord :

$$x(n) = Acos(\omega n + \phi) \tag{3.7}$$
$$\text{d'où } x(n-1) = Acos(\omega(n-1) + \phi) \tag{3.8}$$
$$\text{et } x(n+1) = Acos(\omega(n+1) + \phi) \tag{3.9}$$

En utilisant la propriété trigonométrique suivante :

$$cos(\alpha + \beta)cos(\alpha - \beta) = \frac{1}{2}[cos(2\alpha) + cos(2\beta)] \tag{3.10}$$

On obtient :

$$x(n+1)x(n-1) = \frac{A^2}{2}[cos(2\omega n + 2\phi) + cos(2\omega)] \tag{3.11}$$

Appliquant l'identité :

$$cos2\alpha = 2cos^2\alpha - 1 = 1 - 2sin^2\alpha \tag{3.12}$$

On obtient ensuite :

$$x(n+1)x(n-1) = A^2cos^2(\omega n + \phi) - A^2sin^2(\omega) \tag{3.13}$$
$$x(n+1)x(n-1) = x^2(n) - A^2sin^2(\omega) \tag{3.14}$$

En retravaillant l'équation (3.14), on obtient :

$$A^2sin^2(\omega) = x^2(n) - x(n+1)x(n-1) \tag{3.15}$$

Or, on peut faire l'approximation $sin(\omega) \approx \omega$ si une petite valeur est attribuée à ω. Kaiser [Kaiser, 1990] propose de limiter $\omega < \frac{\pi}{4}$, ainsi l'erreur sur l'approximation de $sin(\omega)$ par ω est de moins de 11%. Dans ce cas, $f/f_s < 1/8$ où f_s est la fréquence d'échantillonnage du signal. En appliquant cette approximation à (3.15), on obtient :

$$T(n) = A^2\omega^2 \approx x^2(n) - x(n+1)x(n-1) \tag{3.16}$$

Pour démontrer la validité de l'approximation de l'équation (3.16), prenons un signal sinusoïdal d'amplitude 100, de fréquence 50Hz et échantillonné à 16000Hz. En utilisant l'équation 3.16 comme approximation, on obtient 3.8548. Le signal et le résultat obtenu sont illustrés à la figure 3.9, le calcul est fait sur une durée de 0.5 secondes.

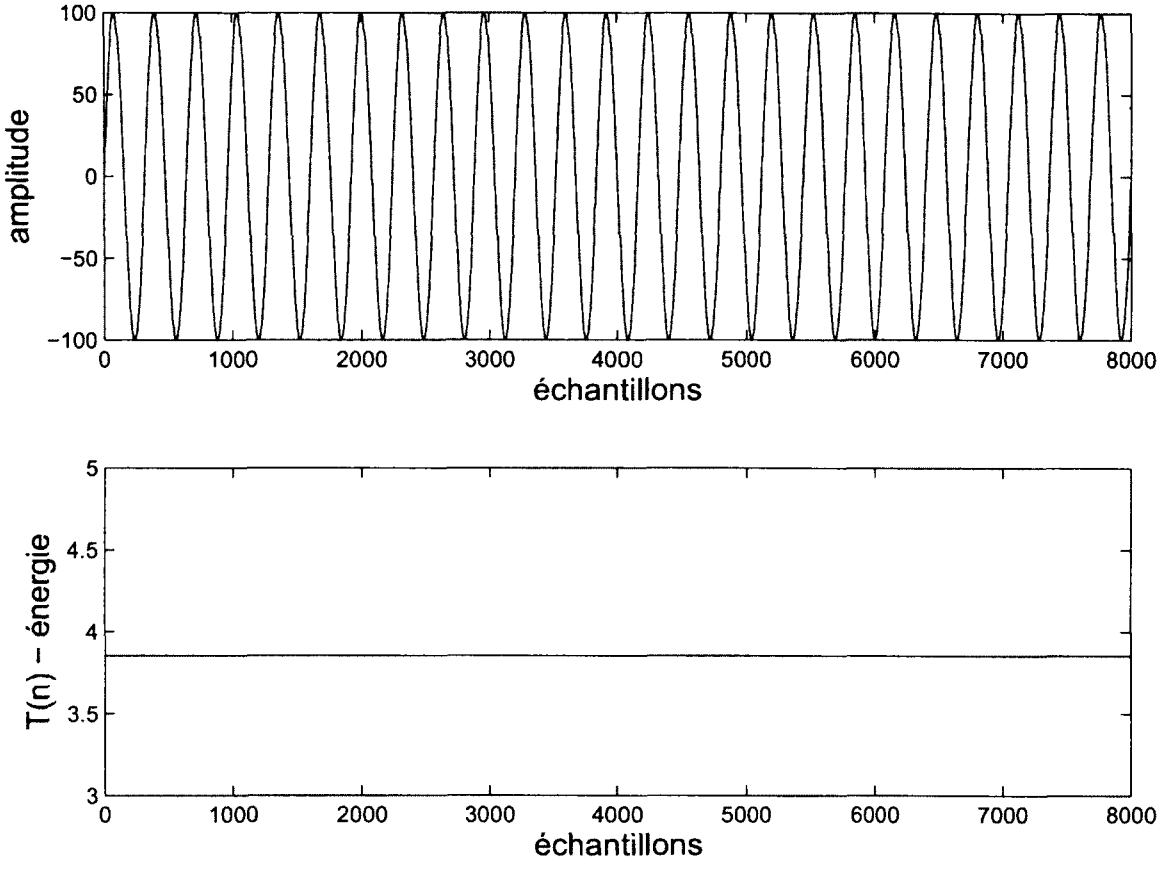


Figure 3.9 Signal sinusoïdal et son énergie avec l'approximation de Teager

En comparaison, l'énergie obtenue à l'aide de l'équation (3.6) est 3.8553. La différence séparant les résultats des deux méthodes est négligeable et s'explique par le fait que l'équation (3.16) est une bonne approximation de la définition originale de l'opérateur de Teager, surtout pour les faibles fréquences, ce qui est le cas dans cet exemple avec $f = 50$ Hz.

L'opérateur de Teager présente plusieurs avantages. Il est indépendant de la phase du signal et est très robuste même lorsqu'un échantillon du signal est égal à zéro puisque l'opérateur ne comporte aucune division. Une division par zéro entraine un résultat invalide, il est donc nécessaire de s'assurer qu'un tel cas ne se produise pas lorsqu'un opérateur comportant une opération de division est utilisé. Seulement deux multiplications et une soustraction pour chaque échantillon sont nécessaires en appliquant l'équation (3.16). Ceci en fait un

opérateur simple d'implémentation et nécessitant peu de calculs. L'opérateur de Teager répond aussi très rapidement (quelques échantillons) à tous changements d'amplitude ou de fréquence.

Les exemples illustrés aux figures 3.10 et 3.11 démontrent la rapidité de réaction de l'opérateur de Teager. Dans le premier cas, l'opérateur de Teager est calculé sur un signal sinusoïdal dont l'amplitude varie dans le temps, mais dont la fréquence est constante (figure 3.10). Dans le second cas l'amplitude demeure fixe alors que la fréquence subit des changements ponctuels (figure 3.11). Dans les deux cas, l'opérateur de Teager réagit instantanément aux variations du signal.

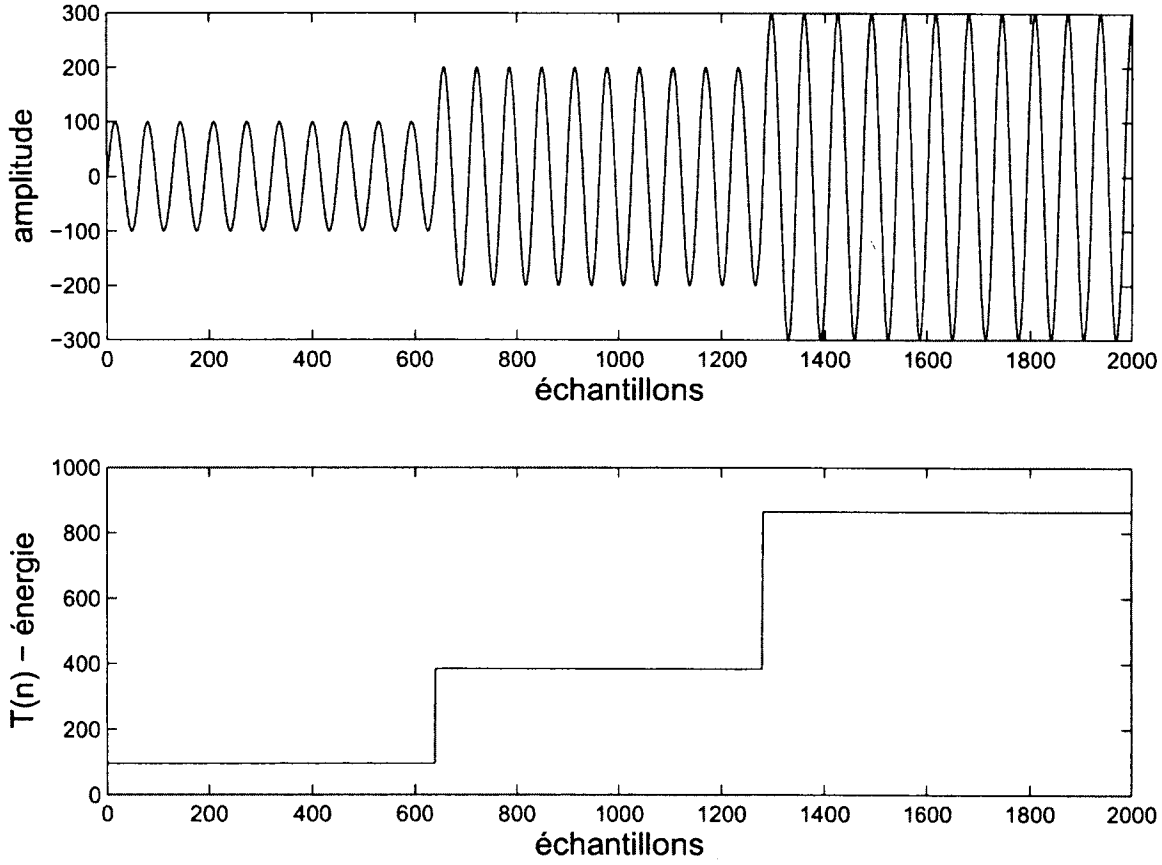


Figure 3.10 Effet de la variation d'amplitude sur le résultat de l'opérateur de Teager

Parmi les propriétés de l'opérateur de Teager, on note également qu'il donne un résultat nul pour tout signal constant de fréquence nulle (DC), quelle qu'en soit l'amplitude. Cet opérateur ne comporte pas seulement des avantages, il présente aussi quelques inconvénients. L'opérateur de Teager n'est applicable qu'à des signaux à bande étroite, idéalement des sinusoïdes pures. Il est sensible au bruit puisque qu'en présence de bruit, le signal tend

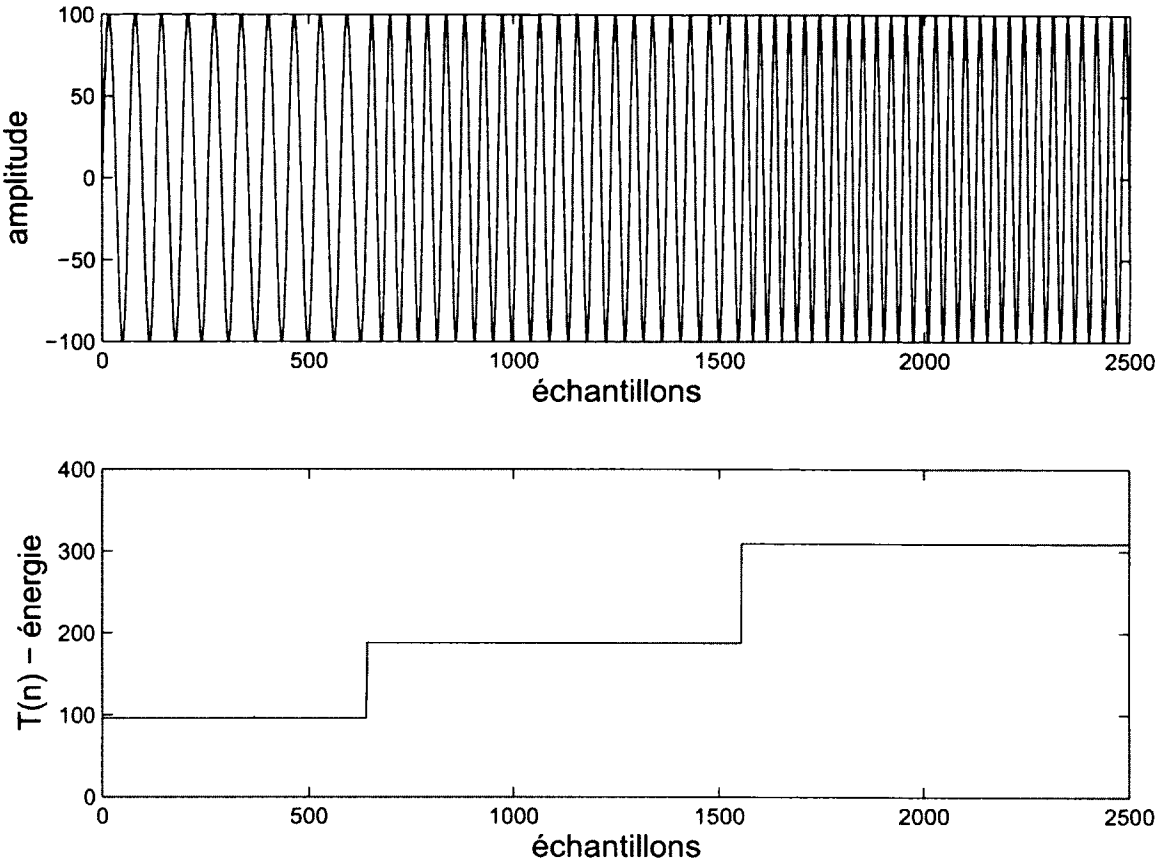


Figure 3.11 Effet de la variation de la fréquence sur le résultat de l'opérateur de Teager

à s'éloigner de la sinusoïde pure idéale et le résultat de l'opérateur de Teager n'est plus constant d'un échantillon à l'autre. Pour illustrer le fait que les signaux doivent être le plus près possible d'une sinusoïde pure, prenons l'exemple $x(n) = 100 \cdot sin(2\pi \cdot 250n/16000)$ et $x(n) = 50 \cdot sin(2\pi \cdot 500n/16000)$.

On constate que lorsque le signal est composé de plus d'une sinusoïde, la mesure donnée par l'opérateur de Teager n'est plus constante même si l'amplitude et la fréquence de ses composantes ne changent pas. Dans un tel cas, la valeur instantannée de l'opérateur de Teager ne peut être utilisée pour discriminer un changement d'amplitude ou de fréquence dans le signal. Il est alors préférable d'utiliser une moyenne du résultat de l'opérateur de Teager en fonction du temps. Cet opérateur s'utilise donc sur des signaux à bande étroite, en respectant l'hypothèse que $sin(\omega) \approx \omega$ et que la fréquence maximale du signal est plus petite que $f_s/8$. Ces signaux doivent se rapprocher le plus possible d'une sinusoïde simple pour que le résultat obtenu avec l'opérateur de Teager soit cohérent avec le signal analysé et pour que l'approximation de l'équation (3.16) soit valide.

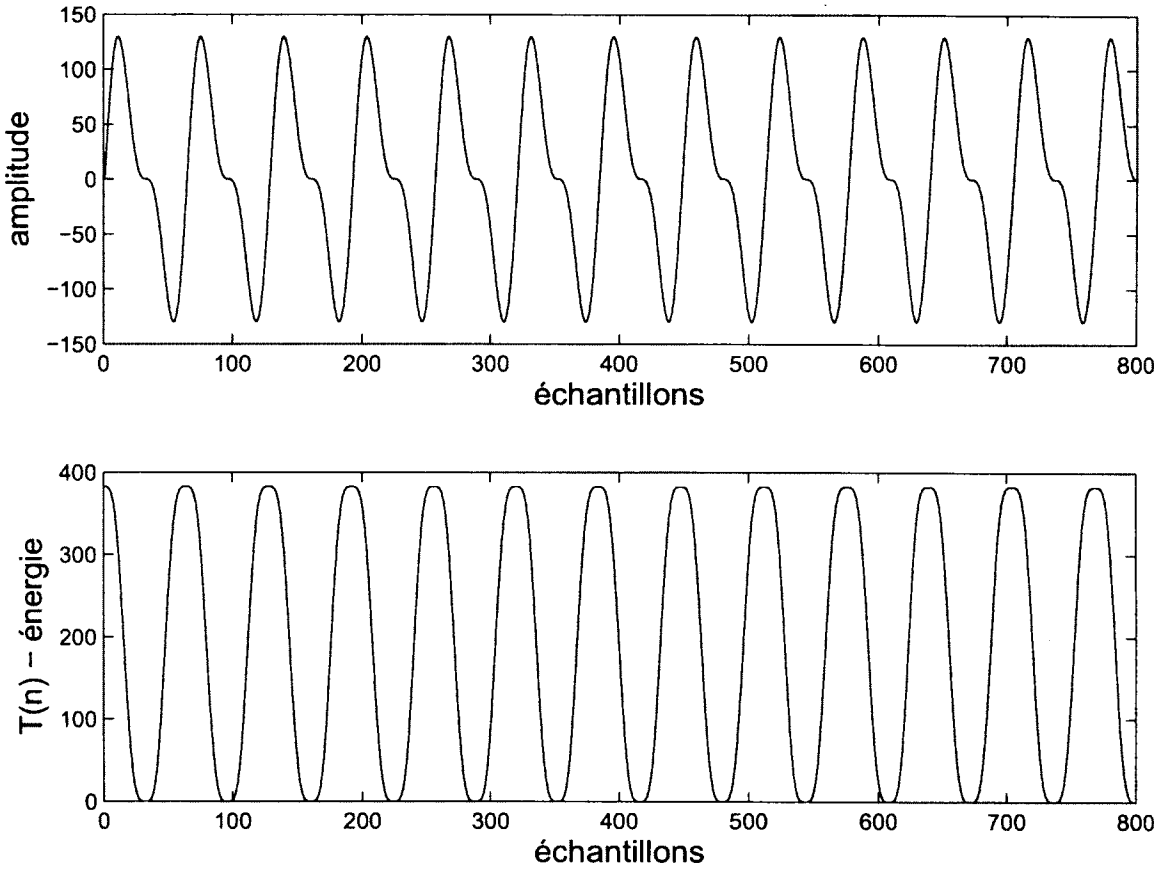


Figure 3.12 Combinaison de deux sinusoïdes de fréquence et d'amplitude différentes

3.4 Détection des transitoires

Maintenant que les caractéristiques et avantages de l'opérateur de Teager ont été décrits, il est proposé d'appliquer l'opérateur de Teager à la détection de transitions non-voisées à voisées dans un signal de parole. L'opérateur de Teager a été choisi pour deux de ses propriétés : la possibilité de détecter les variations de fréquences autant que les variations d'amplitudes ainsi que la rapidité de réaction de l'opérateur à tout changement de fréquence ou d'amplitude dans le signal analysé.

Une analyse fréquentielle effectuée sur des signaux de parole permet de voir les variations de la composition en fréquences du signal lors des transitoires et de montrer qu'il est pertinant d'utiliser l'opérateur de Teager pour détecter ces signaux. La figure 3.13 illustre le spectrogramme (3.13 b)) d'un signal de parole. Le spectrogramme montre l'évolution des fréquences dans le signal, où les zones foncées représentent les composantes fréquentielles les plus énergétiques dans le signal. À l'endroit où les transitoires débutent (définies par les lignes verticales pointillées dans la figure), il est possible de voir une augmentation des basses fréquences présentes dans le signal. L'opérateur de Teager devrait donc permettre d'identifier rapidement ces zones dans le signal puisqu'il réagit rapidement à toutes variations d'amplitude et de fréquence.

La figure 3.14 montre un segment élargi de la figure 3.13. Il est possible de noter que le signal non-voisé précédant le début de la transitoire a des composantes beaucoup plus hautes en fréquences (5000 à 7000 Hz), alors que les composantes de la transitoire sont très basses fréquences (moins de 2000 Hz).

Le schéma bloc de la figure 3.15 illustre les différentes étapes pour permettre la détection des transitoires selon l'algorithme proposé. Cette section explique étape par étape la détection des transitoires avec l'opérateur de Teager en se référant au schéma bloc de la figure 3.15. Pour déterminer si la trame courante contient une transitoire, il faut que la ou les trames précédentes soient classées non-voisées ou transition non-voisées, puisque l'objectif est d'identifier les transitions non-voisés à voisés. Le schéma s'applique donc uniquement lorsque cette condition est rencontrée.

Dans le contexte où l'opérateur de Teager est utilisé pour la détection des transitoires, le signal de parole doit être séparé en bandes de fréquences étroites, tel qu'illustré à la figure 3.15. La bande de fréquences qui correspond aux pitchs possibles dans le codeur VMR-WB varie de 55 Hz à 640 Hz, ainsi les périodes de pitch varient de 10 à 115 échantillons pour une fréquence d'échantillonnage de 6400 Hz (fréquence d'échantillonnage utilisée par le suiveur de pitch du VMR-WB). Pour s'assurer que toutes les transitoires sont détectées,

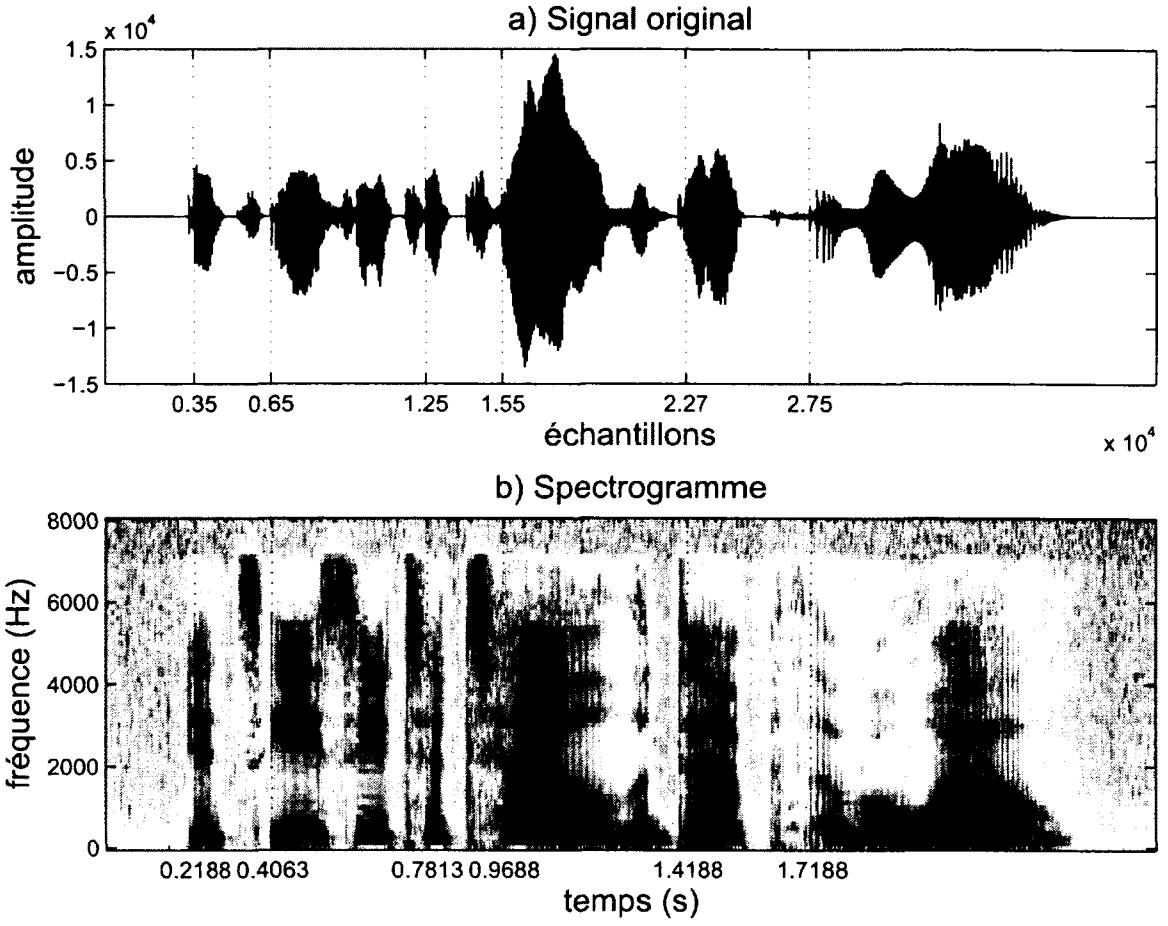


Figure 3.13 Variations des fréquences dans un signal de parole

la plage de fréquences analysées s'étend de 0 Hz à 800 Hz, puisque si on se limitait aux bandes de fréquences du pitch (55 Hz à 640 Hz), certaines détections seraient manquées. De plus, en limitant la fréquence maximale analysée à 800 Hz, la condition $\omega < \frac{\pi}{4}$, c'est-à-dire que $f/f_s < 1/8$ est respectée (dans ce cas précis $f/f_s <= 1/16$, car les calculs seront effectués sur le signal échantillonné à 12800 Hz, soit la fréquence d'échantillonnage utilisée par le codeur VMR-WB). Le fait de limiter la bande de fréquence analysée à 800 Hz permet aussi de réduire l'erreur sur l'approximation $sin(\omega) \approx \omega$ et dans notre cas, l'erreur maximale est de 2.6% (lorsque $\omega = 0.3927$). Cette plage est découpée en bandes de 50 Hz avec un recouvrement de 50%, soit 0-50 Hz, 25-75Hz, 50-100 Hz etc. Une largeur de bande restreinte (50 Hz) diminue les possibilités d'occurence de plusieurs harmoniques dans une même sous-bande, se rapprochant ainsi de l'hypothèse requise pour l'opérateur de Teager, soit le suivi d'une seule sinusoïde. Par exemple, lorsque la fréquence fondamentale d'un signal est de 55 Hz, la deuxième harmonique est située à 110 Hz, donc ces deux harmoniques sont situées dans des bandes d'analyse distinctes. Le recouvrement

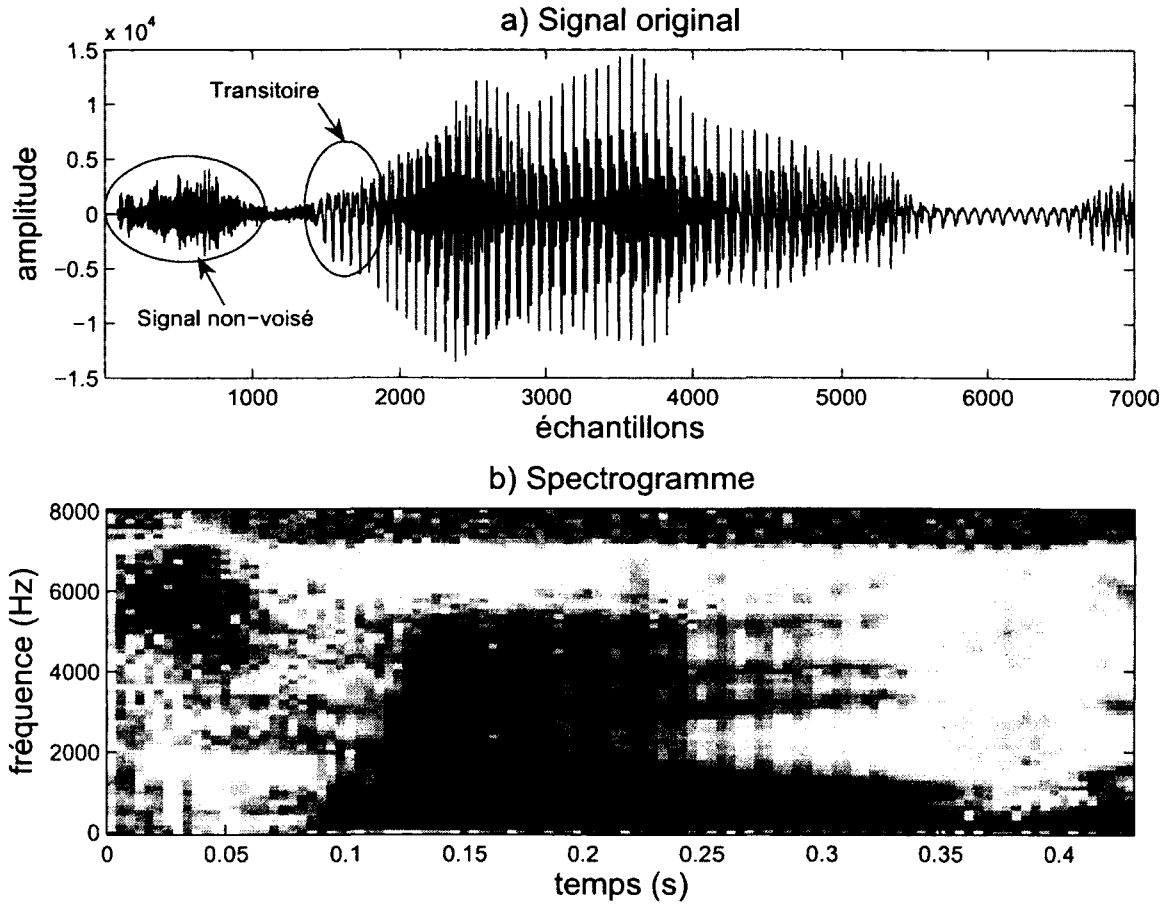


Figure 3.14 Variations des fréquences dans un signal de parole

des bandes de fréquences permet d'inclure les fréquences fondamentales qui seraient à la jonction des bandes passantes de deux filtres, par exemple une fréquence fondamentale de 48 Hz qui se trouve à la jonction des filtres 0-50 Hz et 50-100 Hz). Sans le recouvrement, une fréquence située à la jonction de deux filtres (donc atténuée partiellement par les deux filtres) pourrait ne pas être détectée. Ce cas est illustré à la figure (3.16) où une sinusoïde de 48 Hz est filtrée par les filtres 0-50 Hz, 25-75 Hz et 50-100 Hz. L'amplitude de la sinusoïde filtrée est maximale pour le filtre 25-75 Hz, mais atténuée par les deux autres. Ce phénomène est causé par la réponse en fréquence non rectangulaire des filtres (voir annexe). Les filtres R.I.F. passe-bas et passe-bande utilisés sont d'ordre 255 et ont un délai de 127 échantillons. Le délai des filtres correspond à la demi-trame d'avance (en anglais, *look-ahead*) du signal disponible, soit 128 échantillons. En respectant le délai offert par la demi-trame d'avance disponible dans le codeur VMR-WB, les filtres en sous-bande n'introduisent pas de délai supplémentaire au codeur. À chaque trame un total de 256 échantillons sont analysés pour trouver la position possible d'une transitoire et ces échantillons sont ceux qui correspondent à la trame courante.

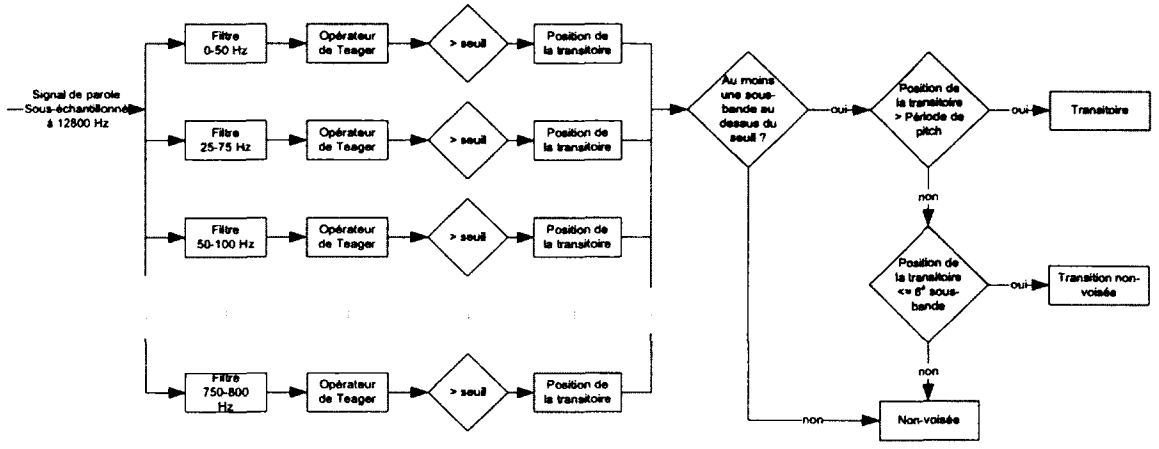


Figure 3.15 Schéma bloc de la détection de transitoires avec l'opérateur de Teager

Un exemple de signal filtré pour une sous-bande est illustré à la figure 3.17. Le signal de la figure 3.17 b) qui est la sortie du filtre en sous-bande 100-150Hz montre un signal à peu près sinusoïdal sur lequel l'opérateur de Teager pourra être appliqué (la sous-bande 100-150 Hz a été choisie puisque la fréquence fondamentale de cet exemple de signal varie entre 110 et 135 Hz).

Suite au banc de filtres, l'opérateur de Teager est appliqué à chacune des sous-bandes (voir figure 3.15). Ensuite, le résultat de l'opérateur de Teager est comparé à un seuil pour chacune des sous-bandes. Ce seuil doit tenir compte du fait que l'opérateur de Teager varie en fonction de l'amplitude et de la fréquence du signal. Pour utiliser un seuil unique pour toutes les bandes de fréquences, il faut tenir compte du fait que la contribution des fréquences dans le résultat de l'opérateur de Teager est différente pour chacune des bandes de signal analysées. Les paragraphes suivants expliquent la normalisation appliquée au seuil pour tenir compte de cette réalité.

Pour évaluer la contribution de chacune des bandes de fréquences, une sinusoïde pure d'amplitude constante égale à 1 a été analysée pour toutes les bandes de fréquence possibles (0 à 800 Hz). Le résultat de l'opérateur de Teager appliqué à ces différentes sinusoïdes obtenu est donc uniquement proportionnel au carré de la fréquence. Le résultat obtenu est illustré à la figure 3.18.

On observe à la figure 3.18 que pour chaque bande de fréquences à traiter, la contribution de la partie fréquentielle dans l'équation $E \approx A^2\omega^2$ diffère. La contribution de la partie fréquentielle (ω^2) de l'équation $E \approx A^2\omega^2$ pour la bande de fréquences 0-50 Hz est en moyenne de 1,3e-4 alors que pour la bande de fréquences 750-800 Hz, cette valeur est en

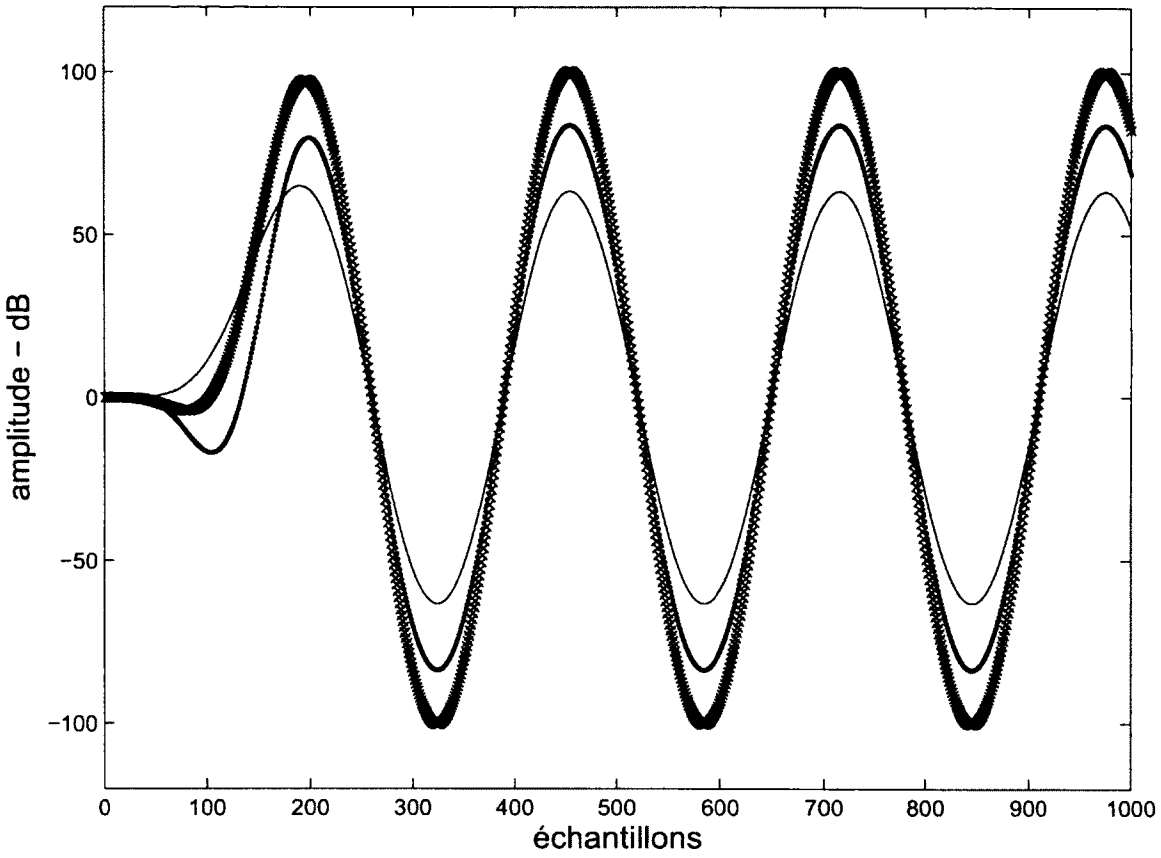


Figure 3.16 Sinusoïde de 48 Hz filtré par le filtre 0-50 Hz (ligne continue), par le filtre 25-75 Hz ("x") et par le filtre 50-100 Hz (".")

moyenne de 0,08. La contribution de la partie fréquentielle ω^2 dans chaque bande de fréquence doit donc être prise en compte pour la détermination d'un seuil. Une possibilité est d'appliquer un seuil qui est proportionnel à la fréquence centrale de la bande de fréquence analysée. Dans le cas présent une autre approche est préconisée, soit l'application d'un unique seuil pour toutes les bandes de fréquence. Pour y arriver, le résultat de l'opérateur de Teager est normalisé par rapport à la fréquence centrale de chaque bande (figure 3.19). La normalisation est calculée selon l'équation (3.17)

$$T_N(n) = \frac{(x^2(n) - x(n+1)x(n-1))}{\omega_c} \tag{3.17}$$

où $T_N(n)$ est le résultat de l'opérateur de Teager normalisé, $x(n)$ est le signal de parole filtrée et ω_c est proportionnelle à la valeur centrale de la bande de fréquence analysée (ex : $\omega_c = 2\pi \cdot 50/12800$, pour la bande de fréquence 0-100 Hz à une fréquence d'échantillonnage

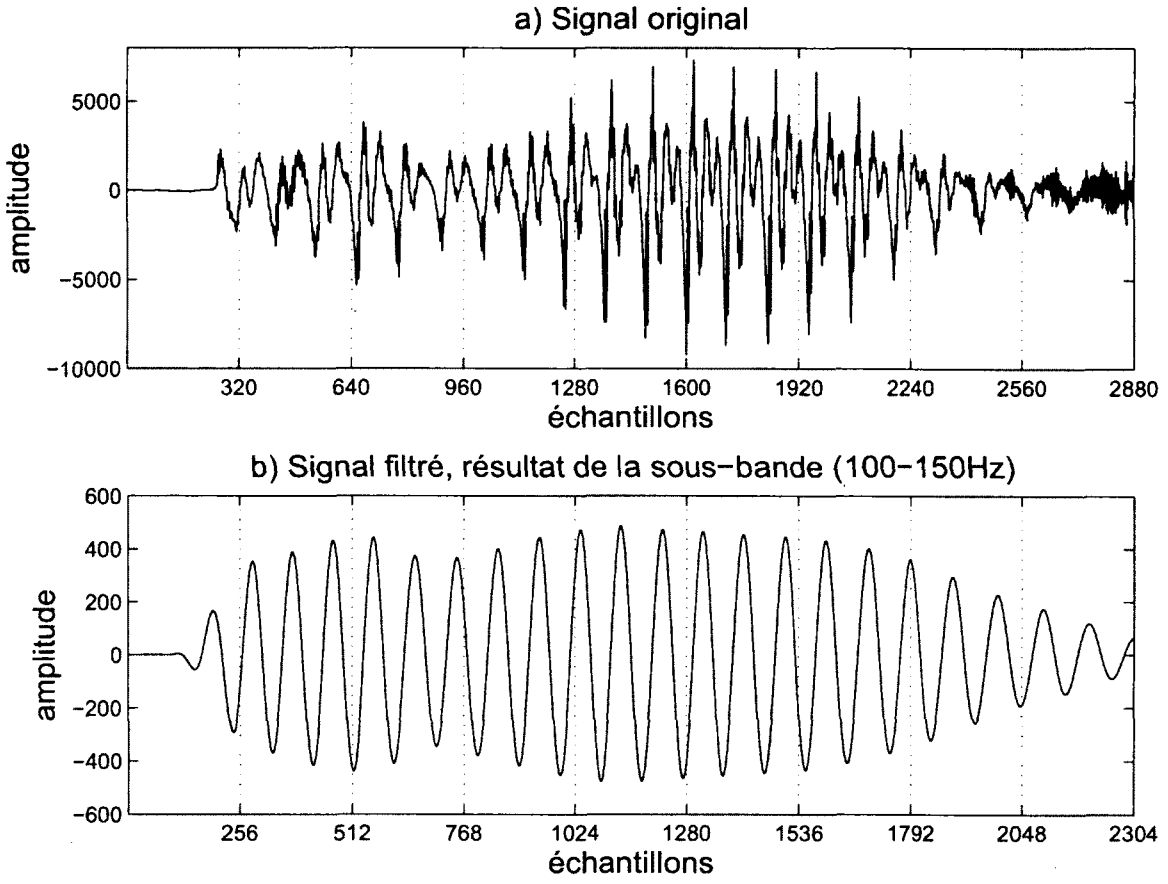


Figure 3.17 Signal original et signal filtré résultant dans la bande de fréquence 100-150Hz

de 12800 Hz). La normalisation permet l'utilisation d'un seuil unique pour toutes les bandes de fréquences analysées.

Afin que la méthode soit indépendante du niveau du signal d'entrée, les variations d'amplitude du signal doivent aussi être considérées. Ces variations sont causées par les changements d'intensité de la voix de l'interlocuteur. Pour en tenir compte, le seuil employé est proportionnel à la mesure d'énergie à long terme du signal. Cette valeur se calcule en effectuant un lissage de l'énergie d'une trame E_{tot}. L'énergie moyenne E_{moy} du signal est calculée à l'équation (3.18), telle que déjà définie et calculée dans le codeur VMR-WB.

$$E_{moy} = 0,99 \cdot E_{moy} + 0,01 \cdot E_{tot} \tag{3.18}$$

E_{moy} est mise à jour à toutes les trames et est calculée une seule fois par trame.

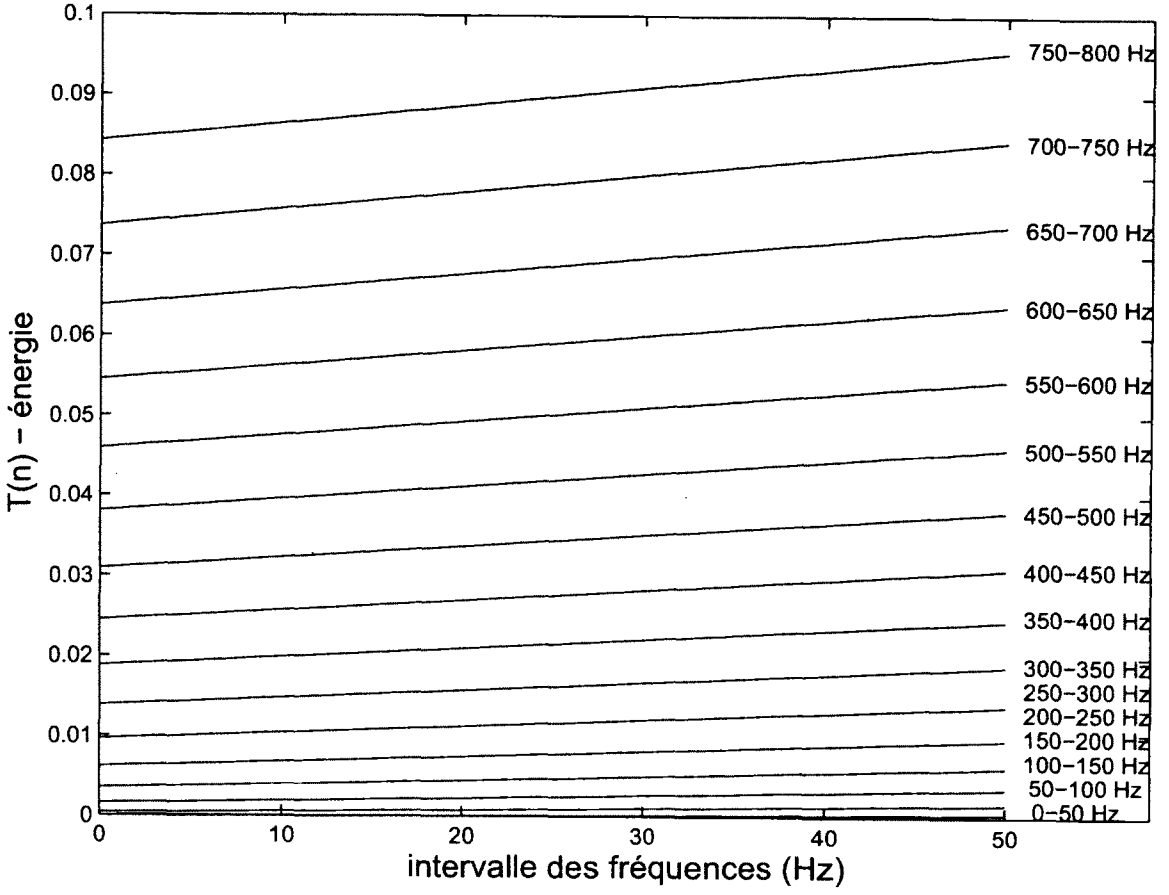


Figure 3.18 L'énergie obtenue avec l'opérateur de Teager, balayage des différentes bandes de fréquences pour un signal d'amplitude égale à 1

Le lissage de l'énergie totale est calculée en dB. L'équation (3.19) détermine le seuil T_{teager}.

$$T_{teager} = \tau \cdot exp^{(0,1 \cdot (Emoy \cdot log(10))) \cdot (1+(\frac{Nmoy}{Emoy}))} \tag{3.19}$$

où N_{moy} est la variation de l'énergie du bruit à long-terme telle que calculée dans le codeur VMR-WB. La valeur $\tau = 0,95$ a été trouvée de façon expérimentale. Le choix de la valeur de τ repose sur la justesse avec laquelle l'opérateur de Teager doit être en mesure de positionner le début de la transitoire. Des expérimentations ont été conduites en variant la valeur de τ et la position du début de la transitoire a été vérifiée sur un échantillonnage de 16 phrases, soit 110 transitions non-voisées à voisées. La valeur de τ où le positionnement du début de la transitoire paraissait le plus juste par rapport à la position théorique de la transitoire a été retenue.

Pour déterminer si une trame est une transitoire, il a déjà été énoncé que la trame précédente doit être classée comme étant non-voisée. Le deuxième critère est que pour au moins

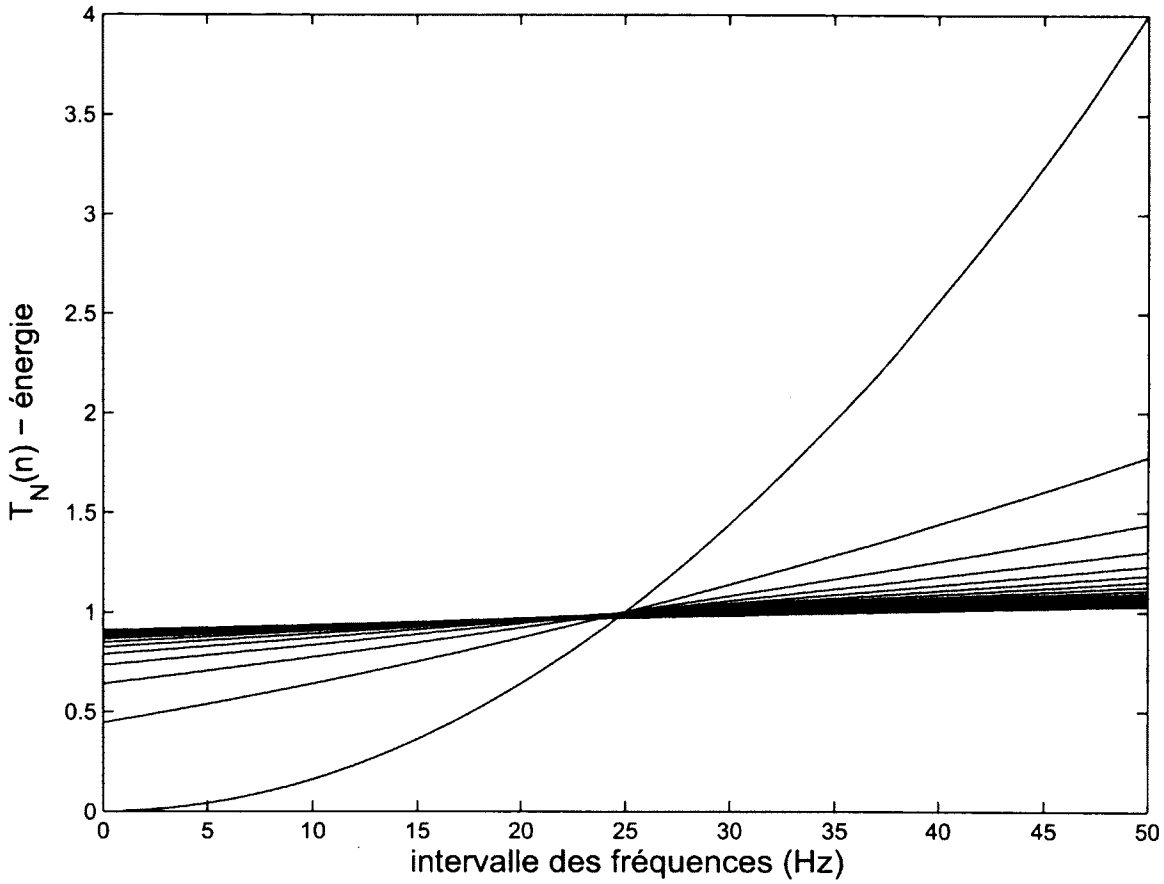


Figure 3.19 L'opérateur de Teager normalisé par rapport à la fréquence centrale, balayage des différentes bandes de fréquences

une sous-bande, le résultat de l'opérateur de Teager soit au-dessus du seuil calculé pour la trame courante (voir figure 3.15). La recherche se fait à partir de la sous-bande la plus basse (0-50 Hz) jusqu'à la sous-bande la plus élevée (750-800 Hz). Parmi les sous-trames ayant atteint le seuil, celle qui correspond à la sous-bande de plus basses fréquences est conservée.

Ainsi lorsque dans une trame, l'opérateur de Teager calculé pour une sous-bande passe au-dessus du seuil de détection, cette sous-bande contient un signal dont le changement de l'amplitude est significative. Ce critère n'est pas suffisant pour déterminer si une trame est une transitoire ; trois choix de classement sont encore possibles. Premièrement, la trame courante peut être non-voisée et contenir un événement ponctuel qui est par exemple une plosive [1] ou tout autre événement avec quelques échantillons d'amplitude plus élevée. Cet événement ne correspond généralement pas à un début de voisement. Lorsqu'il s'agit d'un

[1]En phonétique articulatoire, une plosive fait intervenir un blocage complet de l'écoulement de l'air au niveau de la bouche, du pharynx ou de la glotte, et le relâchement soudain de ce blocage. *source : Wiktionnaire*

début de voisement, il ne se compose généralement que de quelques échantillons (voir figure 3.20 a). Deuxièmement, le début du voisement peut se trouver en fin de trame et même s'il est bien formé, il n'est pas suffisamment long pour contenir au moins une période complète de pitch (voir figure 3.20 b). Dans ce cas, la trame transitoire sera classée comme étant trame de *transition non-voisée* pour que le camouflage non-voisé s'applique si la trame suivante est perdue. Dans le dernier cas, lorsqu'il y a au moins une période de pitch dans la trame courante, la trame est déclarée comme étant une *transitoire* (voir figure 3.20 c)).

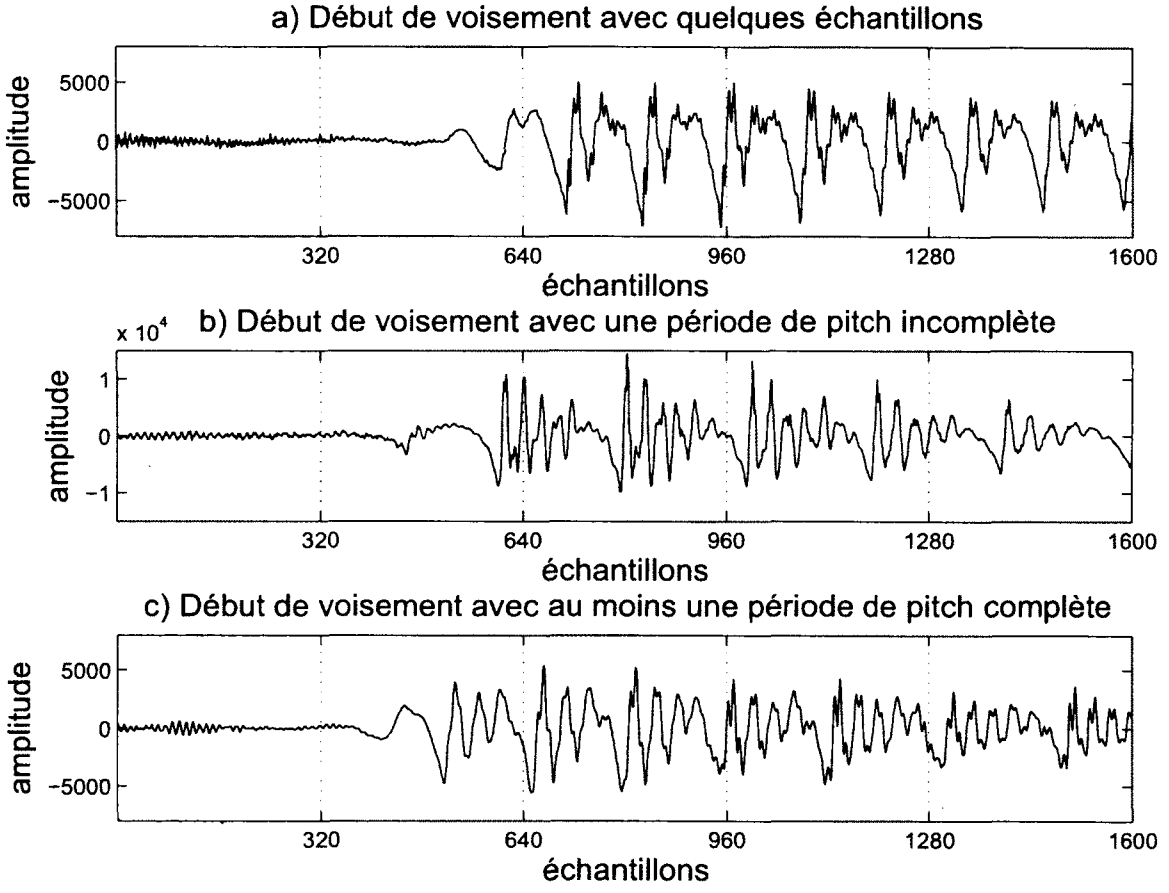


Figure 3.20 Différents positionnements de la transitoire par rapport à la trame

La vérification de la présence d'une période de pitch à la fin de la trame analysée est faite de la façon suivante (voir figure 3.15). La trame est découpée en huit sous-trames pour permettre de bien positionner le début de la transitoire à l'intérieur de la trame. Le début de la transitoire est localisé dans l'une des huit sous-trames possibles (soit 32 échantillons par sous-trame, pour un signal échantillonné à 12800 Hz). Ensuite, la position du début de la transitoire à l'intérieur de la trame est comparée avec la dernière période de pitch évaluée dans la trame courante, soit $d2$ (voir section 2.3.3). Si le nombre d'échantillons compris entre le début de la transitoire trouvée avec l'opérateur de Teager et la fin de la trame

est plus grand que la longueur de la période de pitch déterminée pour cette trame, il y a présence d'au moins une période de pitch dans la trame et celle-ci est donc déclarée comme une trame *transitoire*. Si le nombre d'échantillons compris entre le début de la transitoire et la fin de la trame est plus petit que la période de pitch évaluée, la trame ne contient qu'une fraction de la période de pitch. Si la position de la transitoire déterminée par Teager se trouve dans les six premières sous-trames, la trame est classée comme étant une *transition non-voisée*. Si la position de la transitoire se trouve dans les deux dernières sous-trames, la trame est déclarée comme étant *non-voisée*. La terminologie de la classification des trames fait toujours référence à la classification utilisée dans le VMR-WB tel qu'exposée dans la section 2.3.3. La combinaison entre la position de l'opérateur de Teager et la valeur de la période de pitch $d2$ détermine ainsi la classe de la trame, soit *transitoire*, *transition non-voisée* ou *non-voisée*.

3.5 Résultats des expérimentations

Cette section présente un comparatif entre la détection des transitoires avec l'opérateur de Teager et la détection de transitoires dans le codeur VMR-WB. Cette section est séparée de la façon suivante. Dans un premier temps, une analyse des erreurs de positionnement de la trame transitoire dans le codeur VMR-WB est faite. Ensuite, une analyse comparative entre la classification originale du VMR-WB et la détection des trames transitoires par l'opérateur de Teager est exposée. Pour terminer, les erreurs de classification de l'opérateur de Teager sont analysées, ainsi que la précision de la détection obtenue grâce à l'opérateur de Teager.

La classification de trames originale du VMR-WB combine sept paramètres distincts. La valeur calculée par la combinaison des sept paramètres et la classe attribuée à la trame précédente déterminent la classe qui est attribuée à la trame courante. Les détails du calcul ont été exposés à la section 2.3.3. Lors du début des travaux pour cette thèse, la possibilité de modifier la classification existante a été étudiée. Que ce soit pour les trames où la transitoire est en retard ou en avance, aucune modification ne diminuait de façon significative les erreurs de classification.

Pour qu'une trame transitoire soit classée comme telle, il faut que le résultat de la fonction de coût (voir équation 2.17) soit supérieur à 0,63, combiné avec le fait que la trame précédente doit être classée comme étant *non-voisée* ou bien *transition non-voisée*. Chacun des paramètres qui compose la fonction de coût est normalisé de sorte que plus la trame présente des caractéristiques d'un signal voisé, plus les paramètres tendent vers la valeur 1 et inversement, plus la trame présente des caractéristiques d'un signal non-voisé, plus les paramètres tendent vers 0. Ainsi, la contribution totale des paramètres doit atteindre la valeur de 4,41 (avant la division par 7) pour que la trame soit classée comme une *transitoire*. Dans un tel cas, il est intéressant d'avoir un maximum de paramètres donc la valeur est supérieure à 0,5 et qui idéalement tend vers 1. Tous les paramètres qui sont inférieurs à 0,5 et qui tendent vers 0 nuisent à la possibilité de bien détecter les trames transitoires.

Lorsque les mauvaises classifications de la trame transitoire sont analysées en détail, il est possible de constater que les erreurs de classification ne sont pas dues à un ou quelques-uns des paramètres utilisés. Selon le cas, chacun des paramètres contribue à une mauvaise classification de la trame transitoire. Dans un fichier contenant 840 trames *transitoires*, il a été compté que 83 *transitoires* ont été positionnées en avance, alors que 68 *transitoires* ont été positionnées en retard.

Lorsque la trame transitoire est classée en retard, le résultat de la fonction de coût est inférieur à 0,63 pour la trame transitoire réelle. Les 68 trames où la classification de la transitoire est en retard ont été analysées et les conclusions suivantes sont obtenues. Dans la majorité des cas, les valeurs de ratio entre les basses fréquences et les hautes fréquences (e'_t) et le nombre de passage par zéro (zc) sont saturées (donc tendent à faire augmenter la valeur de la fonction de coût). Il y a seulement et respectivement 9 et 12 occurences où ces valeurs ne sont pas saturées à 1 (leurs valeurs minimales sont respectivement de 0,25 et de 0,6). L'énergie relative (E_{rel}) est souvent supérieure à 0,5 (10 occurences inférieures à 0,5), mais n'est jamais égale à 0. Pour la corrélation normalisée ($\overline{R'}_{xy}$), dans 14 cas cette valeur est égale à 0, donc nuit à la fonction de coût. Dans 39 cas, elle donne une valeur inférieure à 0,5. Comme la corrélation normalisée compte double dans l'équation de la fonction de coût, son influence est non-négligeable. La stabilité du pitch (pc) est égale à 0 dans 30 cas et le ratio signal à bruit (snr) est égale à 0 dans 9 cas. Donc pour 46 cas, les mauvaises détections sont attribuables (seules ou en combinaison) à la corrélation normalisée, de la stabilité du pitch et du ratio signal à bruit. Pour 22 cas, aucun paramètre n'est significativement plus bas que les autres, c'est donc la totalité des paramètres qui contribuent à la fausse détection.

Lorsque la trame transitoire est classée en avance, le résultat de la fonction de coût est supérieure à 0,63 alors que cette trame doit être classée non-voisée. La saturation d'un ou plusieurs paramètres expliquent ces résultats. Comme dans le cas précédent, les valeurs de ratio entre les basses fréquences et les hautes fréquences (e'_t) et le nombre de passage par zéro (zc) sont majoritairement saturées. Il y a 4 cas où le ratio entre les basses fréquences et les hautes fréquences n'est pas saturé et la valeur minimale est de 0,6. Pour les 11 cas où le nombre de passage par zéro n'est pas saturé, la valeur minimale est de 0,4. Il y a 26 cas (corrélation normalisée), 44 cas (stabilité du pitch), 5 cas (ratio signal à bruit) et 2 cas (énergie relative) où ces paramètres sont saturés. Ces cas représentent 60 cas sur les 83 cas répertoriés (pour chacun des cas, au moins un paramètre nommé est saturé). Sinon, comme dans le cas précédent, il y a 23 cas où tous les paramètres sont assez élevés sans être saturés pour faire en sorte que la fonction de coût pour cette trame soit assez élevée (0,63) pour que la trame soit classée *transitoire*.

Pour améliorer la classification existante, une nouvelle mesure avec l'opérateur de Teager et indépendante de la classification du codeur VMR-WB est proposée. La nouvelle classification avec l'opérateur de Teager a été comparée à la classification existante du VMR-WB. Pour faire la comparaison, les transitoires ont d'abord été marquées manuellement pour chacune des phrases qui composent le test. Les trames transitoires correspondent à la

première trame au moins partiellement voisée suivant une ou des trames non-voisées qui contient au moins une période de pitch bien construite. Les trames en faute sont classées selon deux catégories, soit les trames où le classement de transitoire est attribué trop rapidement, soit les trames où le classement de transitoires est attribué en retard. Le pourcentage de trames où le classement est erronné est compté pour chacun des deux cas et les résultats sont présentés au tableau 3.1.

	Classification du codeur VMR-WB	Classification avec l'opérateur de Teager
Transitoires en avance	9,9%	3,8%
Transitoires en retard	8,1%	4,8%
Total	18%	8,6%

Tableau 3.1 Pourcentage de trames où le classement est erroné

L'opérateur de Teager permet de diminuer de 9,4% le taux d'erreur sur la classification des trames transitoires. Le deux tiers de cette diminution est en lien avec les trames transitoires qui sont détectées trop rapidement (6,1%), alors que la balance est attribuable aux trames transitoires détectées en retard. L'opérateur de Teager permet donc de diminuer de moitié les erreurs de classification autour des trames transitoires.

Il est assez difficile d'évaluer la position exacte du début du segment voisé (visuellement ou avec une mesure quelconque). Pour valider la performance de l'opérateur de Teager, toutes les trames transitoires ont été passées en revue (visuellement). La position du début de la transitoire correspond au début d'un changement dans l'allure du signal, changement qui correspont à un début de signal de forme plus sinusoïdale. L'analyse visuelle de la position de l'opérateur de Teager par rapport au début estimé de la transitoire donne les résultats suivants. Il y a un total de 6,6% (56 cas sur 840) des trames où la position de l'opérateur de Teager est jugée fausse par rapport au début estimé de la transitoire. De ce pourcentage, la position de l'opérateur de Teager est en retard par rapport à la position estimée de la transitoire dans 5,2% (44 cas sur 840) des cas. Dans 1,4% (12 cas sur 840) des cas, la position de l'opérateur de Teager est en avance sur la position estimée de la transitoire.

Pour les trames où la position de l'opérateur de Teager est en retard par rapport à la transitoire réelle, seulement 2,3% des cas (19 cas sur 840) posent problèmes quant à la classification de la trame (classification de la trame transitoire en retard par rapport à la trame transitoire réelle). Pour les trames classées transitoire en avance, c'est 0,8% (7 cas sur 840) qui sont dus à une mauvaise estimation du début de la transitoire par l'opérateur de Teager. Du tableau 3.1 qui présentent les résultats obtenus avec l'opérateur de Teager

pour la classification des transitoires, il y a 32 cas où la méthode proposée classe la trame *transitoire* en avance. De ces cas, 22% sont dus à un mauvais positionnement de l'opérateur de Teager, le reste (88%) est dû à une erreur quant à la combinaison de l'opérateur de Teager et la valeur calculée du pitch. Pour les 40 cas où la trame *transitoire* est en retard, 19 cas sont dus à une mauvaise position de l'opérateur de Teager. Les 21 cas restants sont dus à une mauvaise combinaison entre la position de l'opérateur de Teager et la valeur du pitch.

La figure 3.21 présente un cas où le début de la transitoire est pointée trop tardivement par l'opérateur de Teager, ce qui occasionne un retard de la trame transitoire.

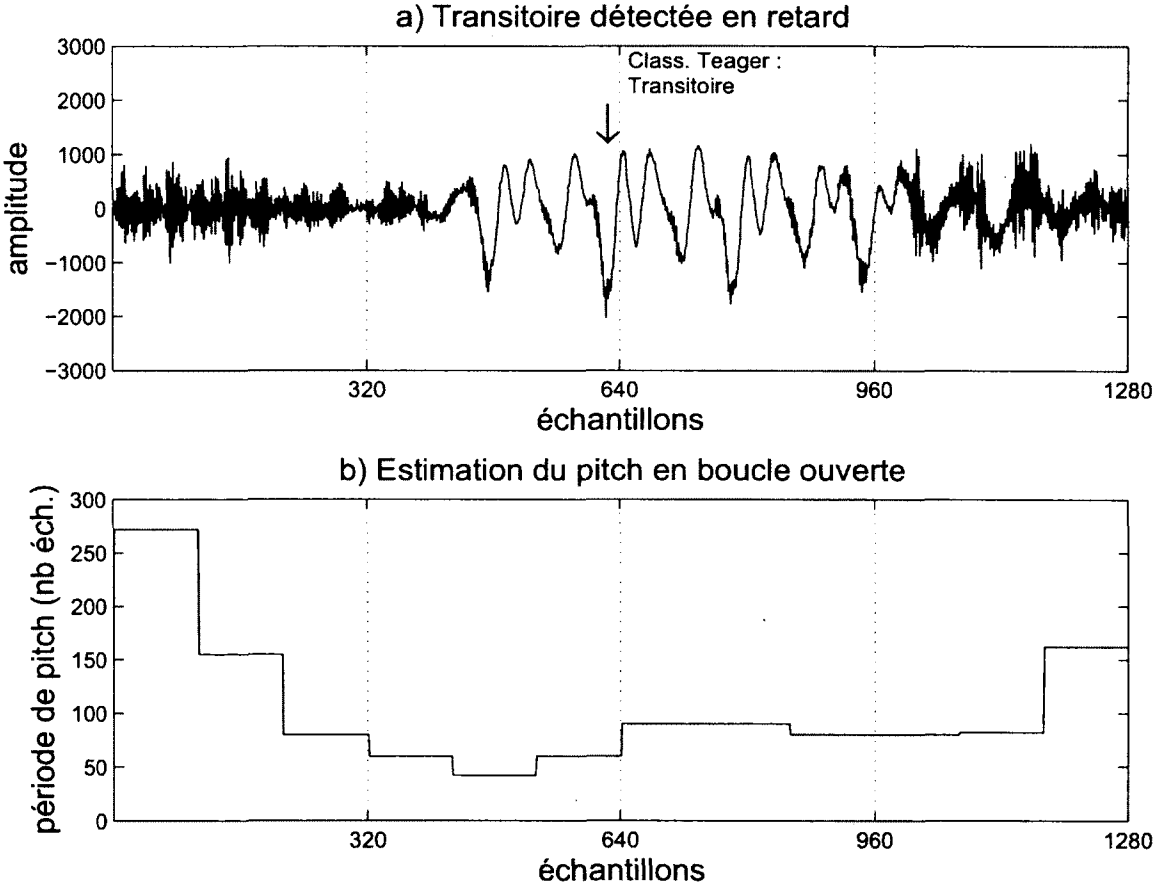


Figure 3.21 Illustration d'un positionnement de la trame transitoire en retard à cause du positionnement de l'opérateur de Teager

La figure 3.22 présente un cas où le début de la transitoire est pointée trop rapidement par l'opérateur de Teager, ce qui occasionne une avance de la trame transitoire.

Pour ce qui est des classifications qui sont trop en avance, dans la majorité des cas l'opérateur de Teager trouve le début de la transitoire à la bonne position. Par contre, la

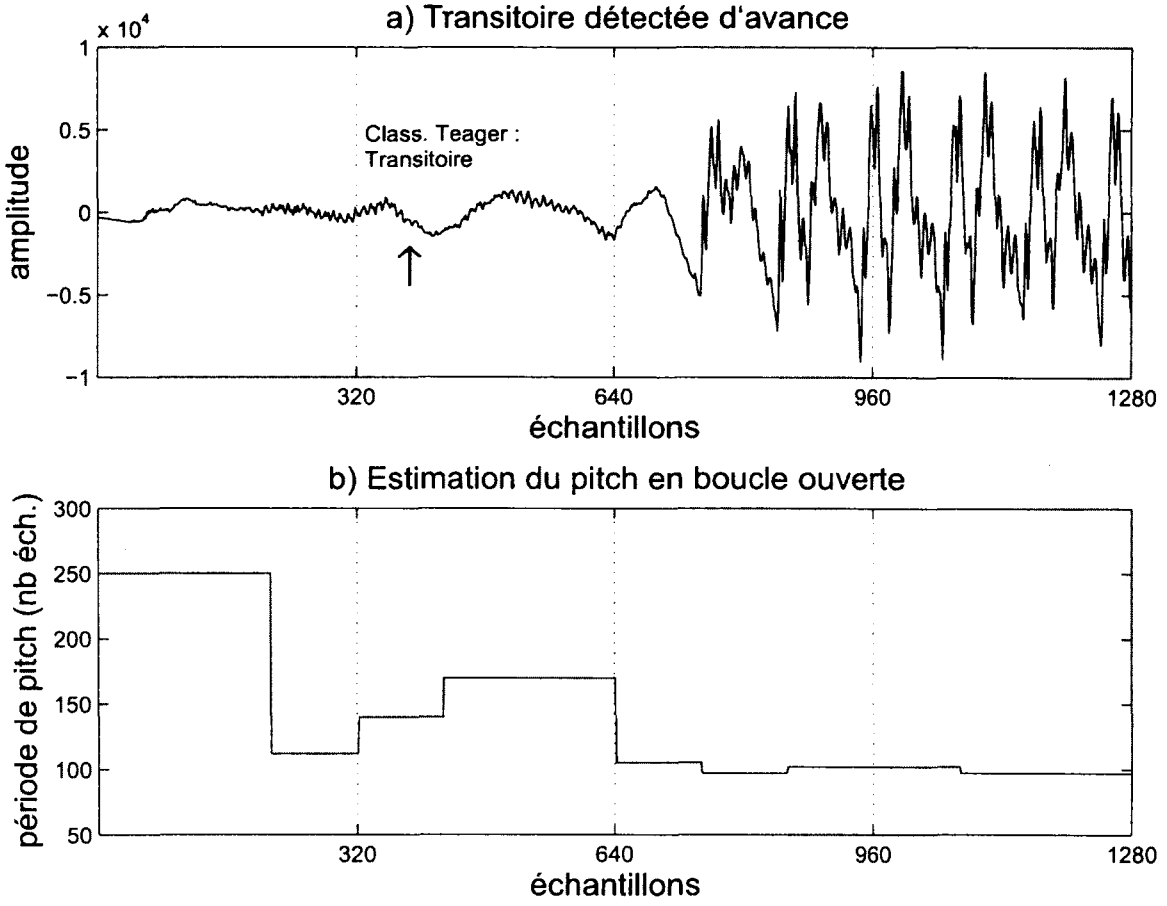


Figure 3.22 Illustration d'un positionnement de la trame transitoire en avance à cause du positionnement de l'opérateur de Teager

combinaison entre la position du début de la transitoire pointée par l'opérateur de Teager et de la longueur du pitch donne une fausse information quant à la validité d'une période de pitch bien construite dans la trame. Un exemple est illustré à la figure 3.23. La flèche pointe sur la zone où l'opérateur de Teager indique le début de la transitoire.

Les mauvaises classifications obtenues par la combinaison du résultat de l'opérateur de Teager et de la longueur de pitch sont attribuables à deux raisons distinctes. La première raison est la détection de pitchs longs, qui sont parfois moins énergétiques au début. Dans ce cas, l'opérateur de Teager se trouve décallé par rapport au début de la transitoire et que le résultat de l'opérateur de Teager est en fin de trame, la combinaison d'un pitch long avec la position du résultat de l'opérateur de Teager invalide la trame pour la classification de transitoire. Un exemple est présenté à la figure 3.24, où la fléche indique la position où l'opérateur de Teager positionne le début de la transitoire.

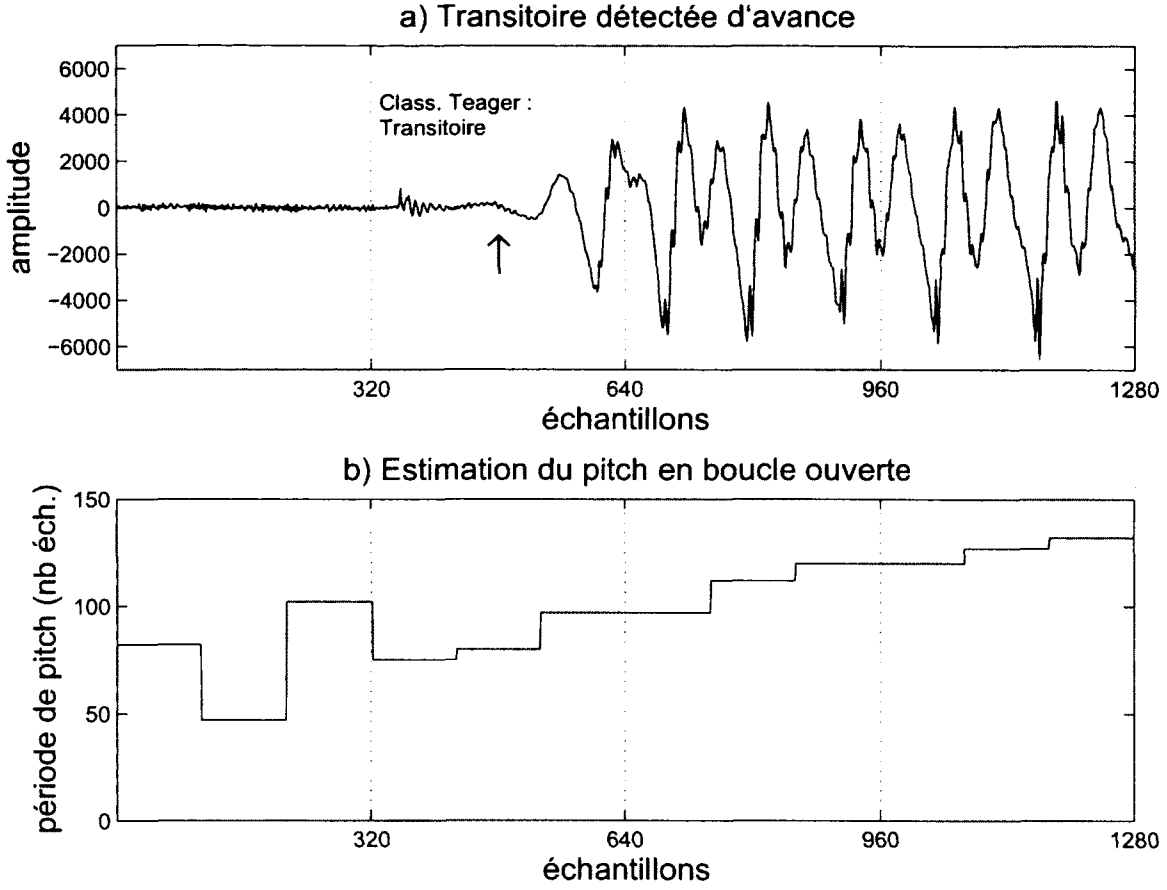


Figure 3.23 Illustration d'un positionnement de la trame transitoire en avance à cause d'une valeur de pitch erronée

Dans le deuxième cas, malgré que le pitch soit court, la position de l'opérateur de Teager est en fin de trame. Encore une fois, la combinaison entre la longueur de pitch et la position de l'opérateur de Teager invalide la trame pour la classification de transitoire. Un exemple est donné à la figure 3.25, où la flèche indique la position où l'opérateur positionne le début de la transitoire. Ici, la longueur de pitch est de 97 échantillons et la position de l'opérateur de Teager se trouve dans la sixième sous-trame. À cette position, la période de pitch maximale permise est de 96 échantillons.

3.5.1 Détection des transitoires avec le détecteur d'enveloppe

Il est intéressant d'analyser les résultats obtenus si les mêmes étapes appliquées à l'opérateur de Teager sont reprises avec un détecteur d'enveloppe. La seule différence entre les deux approches se situe au niveau du choix du seuil qui doit être ajusté à la baisse puisqu'avec le détecteur d'enveloppe, seule la composante d'amplitude en valeur absolue

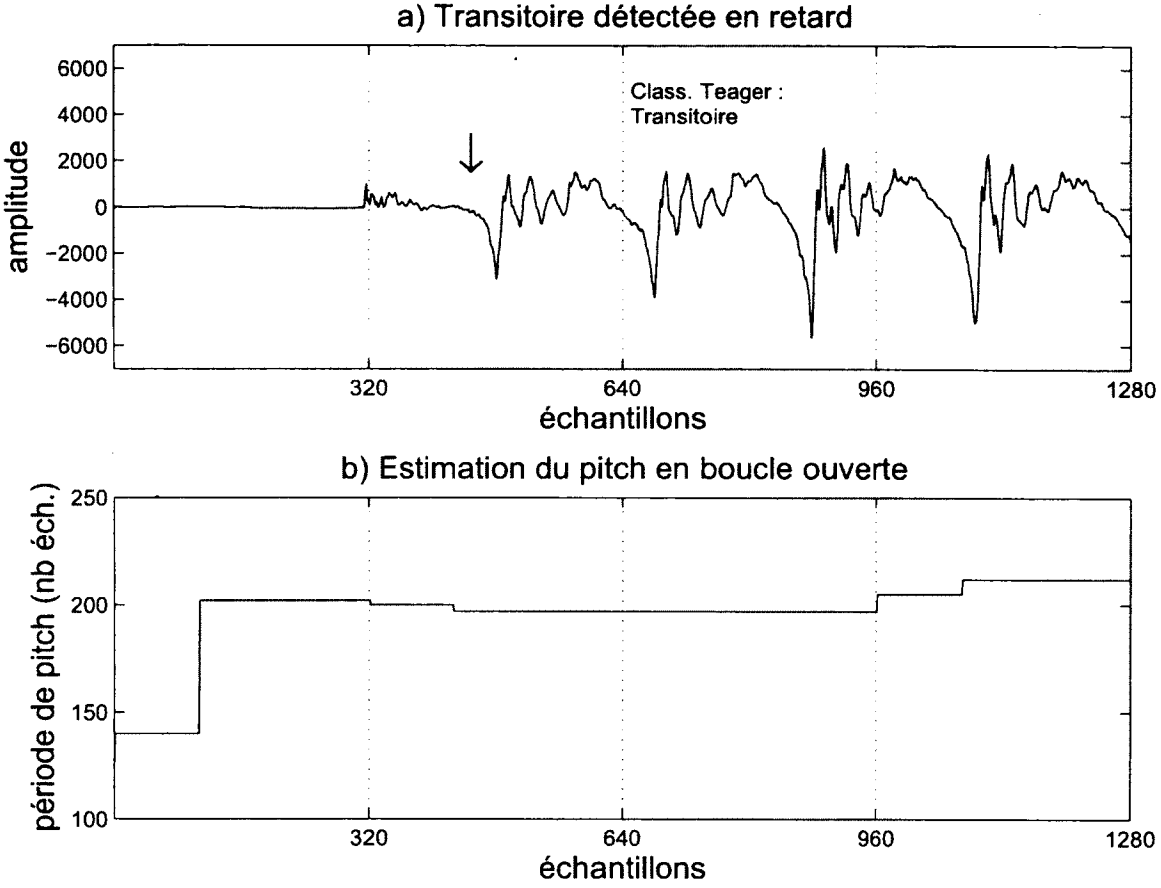


Figure 3.24 Illustration d'un positionnement de la trame transitoire en retard avec l'opérateur de Teager, avec un pitch long

est considérée, contrairement à l'opérateur de Teager qui varie en fonction du carré de l'amplitude et du carré de la fréquence.

En fonction des valeurs de seuils choisies, on se trouve devant les mêmes problèmes exposés à la section 3.2. Dans un premier temps, si le seuil choisi est trop élevé, le début de la transitoire sera manqué. À la figure 3.26, c'est toute une syllabe d'un mot qui n'est pas détectée. Pour chacune des figures 3.26-3.28, les sous-figures a) représentent le signal original, alors que les sous-figures b) montrent la sous-bande dans laquelle le début de la transitoire a été détecté. Il y a détection de la transitoire seulement lorsque la valeur de la sous-bande est plus grande que zéro.

Même si on réduit la valeur du seuil de moitié, certains cas problématiques sont encore trouvés (voir figure 3.27 où la sous-figure a) représente le signal original, alors que la sous-figure b) montre la sous-bande dans laquelle le début de la transitoire a été détecté).

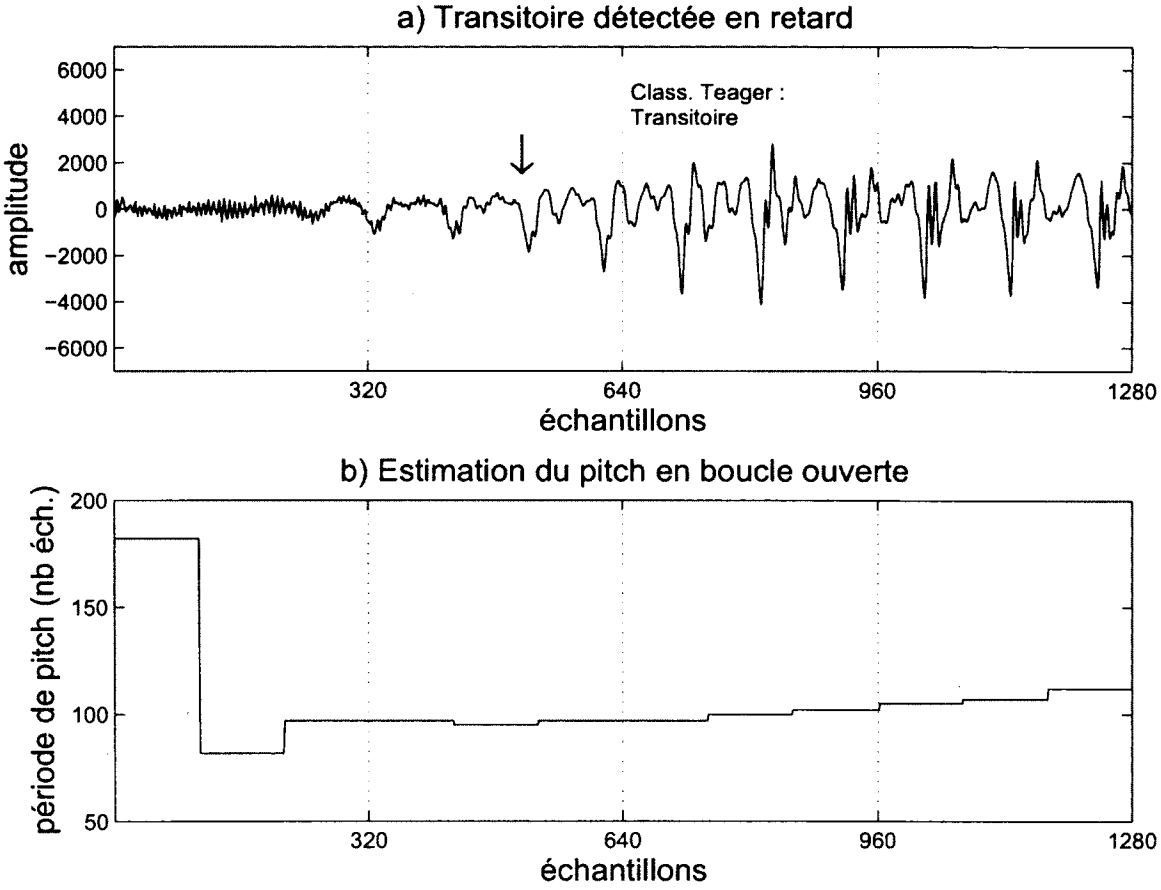


Figure 3.25 Illustration d'un positionnement de la trame transitoire en retard avec l'opérateur de Teager, avec un pitch court

Si le seuil est encore réduit de moitié, le cas problème inverse se produit. Dans ce cas, plusieurs détections sont trop rapides et sont associées à une partie non-voisée du signal, tel qu'illustré à la figure 3.28, où la sous-figure a) représente le signal original, alors que la sous-figure b) montre la sous-bande dans laquelle le début de la transitoire a été détecté.

Le détecteur d'enveloppe, même si appliqué sur les sous-bandes de signal, ne permet pas une détection meilleure ou équivalente à l'opérateur de Teager puisque l'amplitude des zones non-voisées a une influence sur le choix du seuil de détection et empêche une détection robuste du début des transitoires.

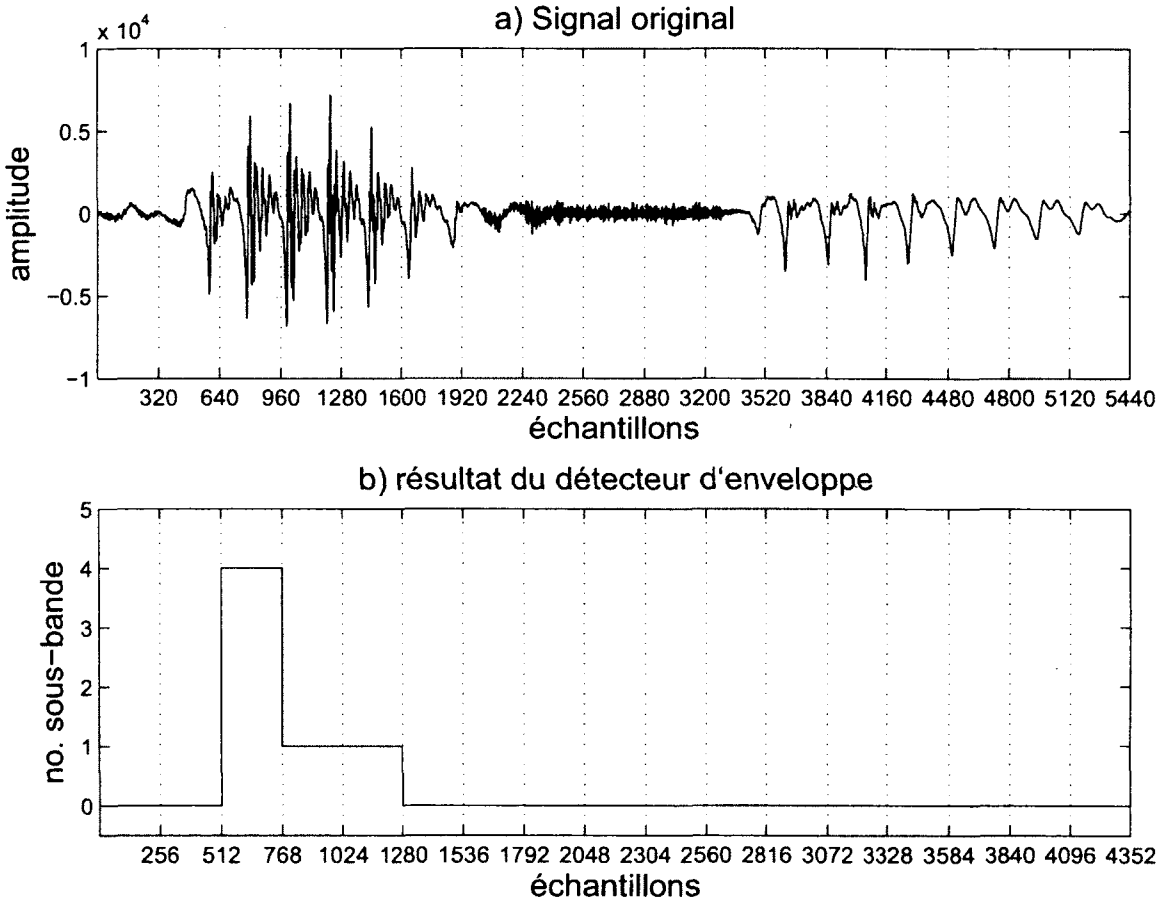


Figure 3.26 Détection avec un détecteur d'enveloppe appliqué en sous-bande (seuil égal à 0.01), détection manquée

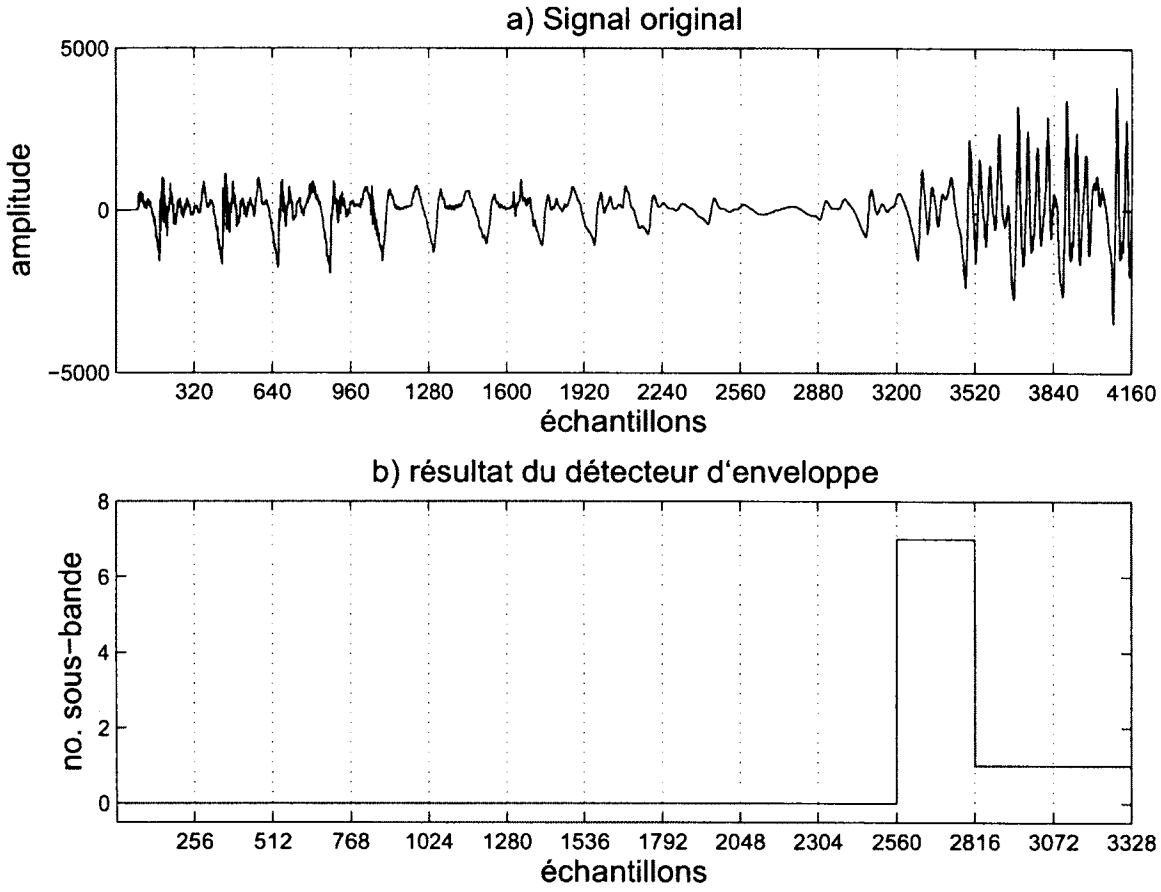


Figure 3.27 Détection avec un détecteur d'enveloppe appliqué en sous-bande (seuil égal à 0.005), détection manquée

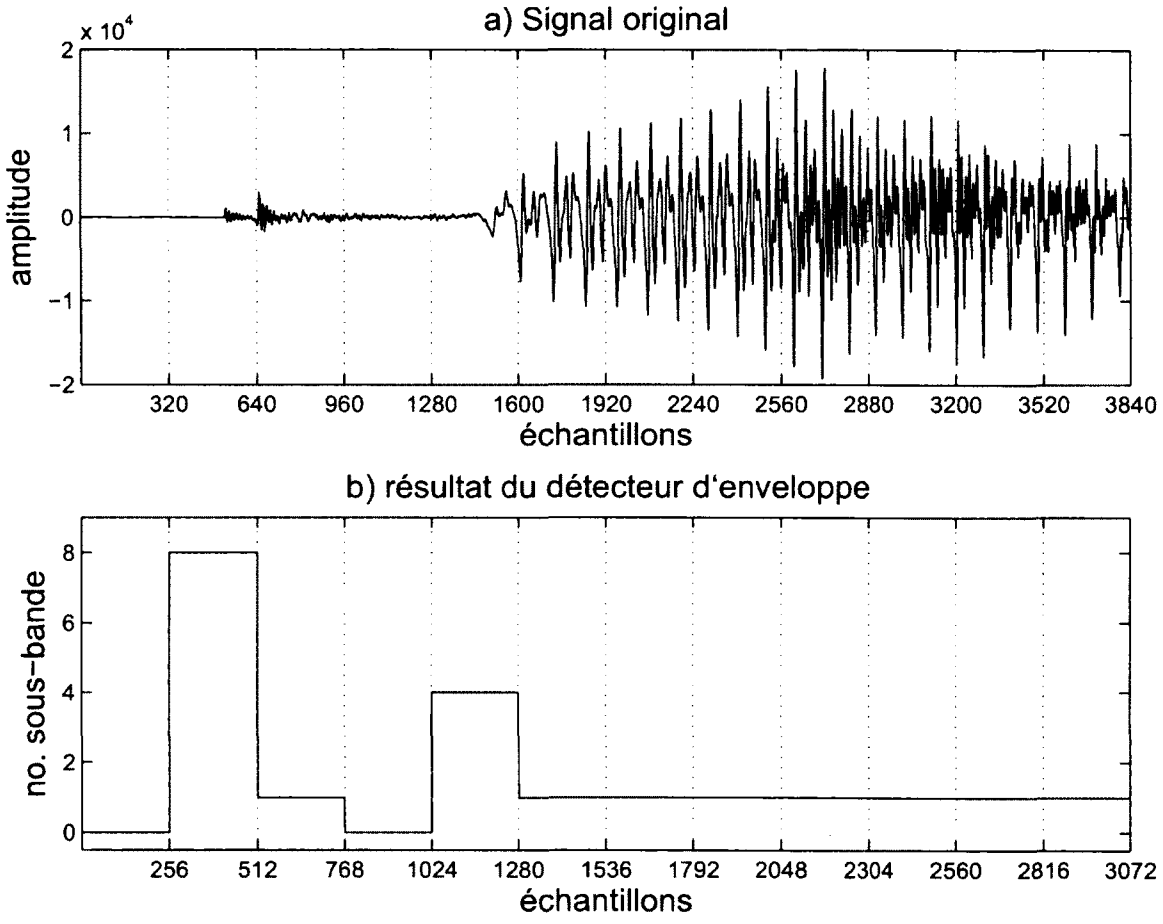


Figure 3.28 Détection avec un détecteur d'enveloppe appliqué en sous-bande (seuil égal à 0.0025), détection trop rapide

3.6 Évaluation subjective

Pour valider la pertinence d'appliquer une nouvelle classification qui améliore la détection des trames transitoires, des tests d'écoute ont été faits. Les résultats de tests d'écoute exposés dans cette section ont été présentés dans les articles suivants ([Jelinek *et al.*, 2007] et [Lemyre *et al.*, 2008]). Les tests ont pour but de valider l'amélioration de la performance du codec VMR-WB en cas de pertes de trames et ils ont été faits en deux volets. Dans le premier volet, la trame perdue est toujours celle qui contient la transitoire réelle. Les transitoires réelles ont été déterminées manuellement pour toutes les phrases contenues dans le fichier. Ce test évalue l'impact d'une mauvaise classification de trames dans le cas où la trame précédant la transitoire réelle, généralement non-voisée, est classée comme *transitoire* par le classificateur original du VMR-WB. La figure 3.29 illustre ce cas. Lorsqu'une trame non-voisée est classée comme une *transitoire*, le camouflage répète une période de pitch fictive dans la trame reconstruite, ce qui impose une structure harmonique artificielle dans le signal. Dans le cas où la trame non-voisée est bien classée, le camouflage n'extrapole pas de partie périodique dans le signal et évite toute forme d'artéfact. Si la trame est considérée comme stable, l'énergie est maintenue, sinon elle tend rapidement vers zéro.

Le test a été construit de la façon suivante. Deux fichiers ont été créés, un à partir du codeur original et l'autre à partir du codeur original VMR-WB avec la classification modifiée (avec l'opérateur de Teager). La classification des trames des deux fichiers a été comparée et lorsque l'une ou l'autre des classifications (ou les 2 en même temps) classait la transitoire trop rapidement (la trame précédant la transitoire réelle), cette phrase était retenue pour le test. Pour simuler les trames perdues, toutes les transitoires réelles où la classification diffère entre les deux modes de classification ont ensuite été effacées. Ce test mesure l'impact d'une mauvaise classification qui applique un camouflage voisé sur une trame non-voisée. Les fichiers sont comparés selon un test Mushra [BS.1534, 2001] et les résultats obtenus sont présentés à la figure 3.30.

En tout, neuf auditeurs ont écouté chacun 18 groupes de quatre phrases. Lors de chaque écoute, l'auditeur doit identifier la phrase originale (normalement notée 100%), et donne une note variant entre 1 et 100% à chacune des trois autres phrases écoutées. Les auditeurs ont identifiés correctement la version originale sans perte de trames dans 99,6% des cas (version originale du signal codé/décodé). La version originale de la classification du VMR-WB, qui contenait des pertes de trames, a obtenu une note de 89,8%, comparativement à la version parfaite où la classification est "forcée" manuellement aux endroits désirés qui

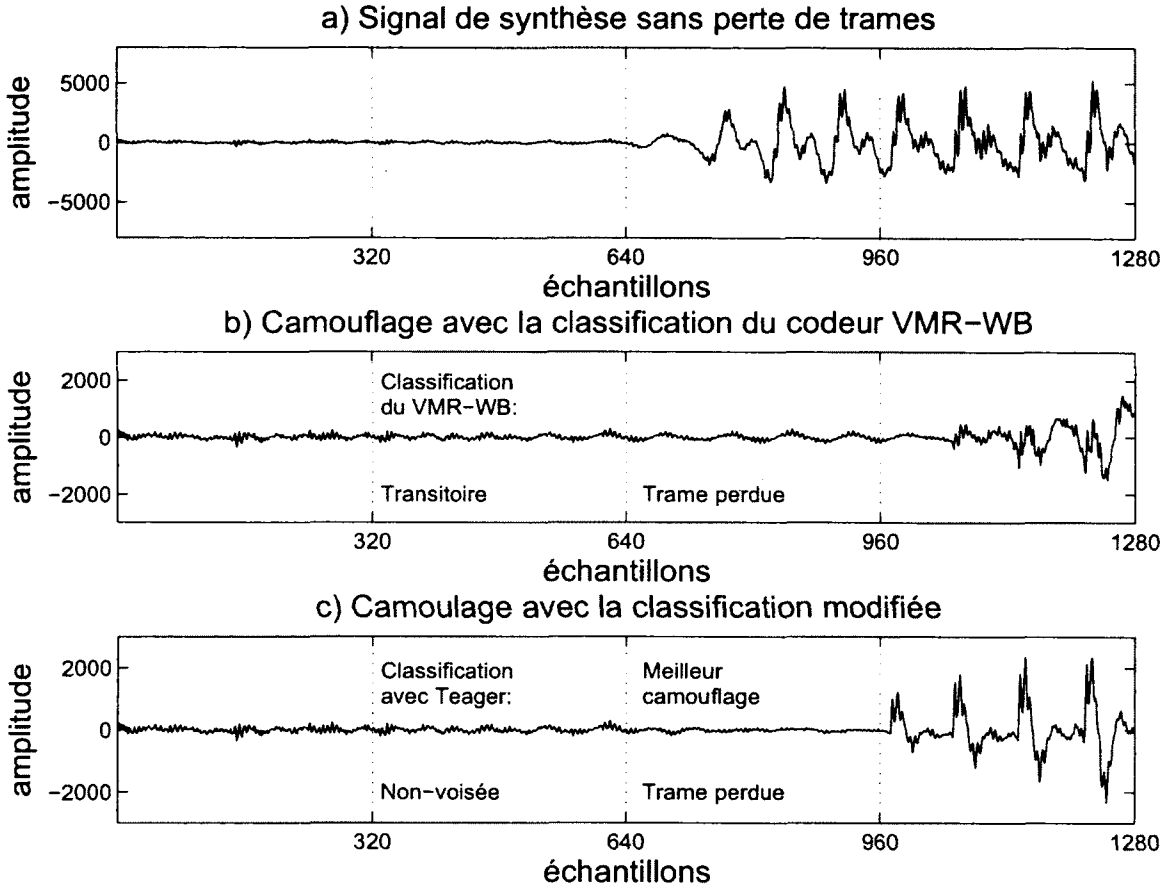


Figure 3.29 Illustration de camouflage si la trame de la transitoire est perdue

a obtenu 92,1% avec les mêmes trames perdues. La version classifiée par l'opérateur de Teager a quant à elle obtenue une note de 91,4%, toujours selon le même patron de trames perdues.

L'analyse des résultats avec l'intervalle de confiance montre que les trois classifications sont statistiquement équivalents. Aucune des classifications ne se démarque des autres sur le plan subjectif. Deux éléments expliquent l'écart restreint qui les séparent et l'équivalence statistique. Premièrement, peu de trames sont perdues à chaque phrase (en moyenne une seule trame perdue par phrase). Deuxièmement, l'énergie du signal avant la transitoire est assez basse, ce qui a pour effet de limiter les artéfacts audibles.

Dans le deuxième volet, la trame perdue est celle qui suit la trame transitoire réelle. Selon l'exemple de la figure 3.31, il arrive parfois que le classificateur original du codeur VMR-WB classe la trame transitoire réelle comme une trame non-voisée. Dans ces cas, si la trame suivant la transitoire réelle est perdue, le camouflage se fait selon la dernière trame bien reçue (ici classée *transitoire non-voisée*). Deux cas de camouflage sont alors possible,

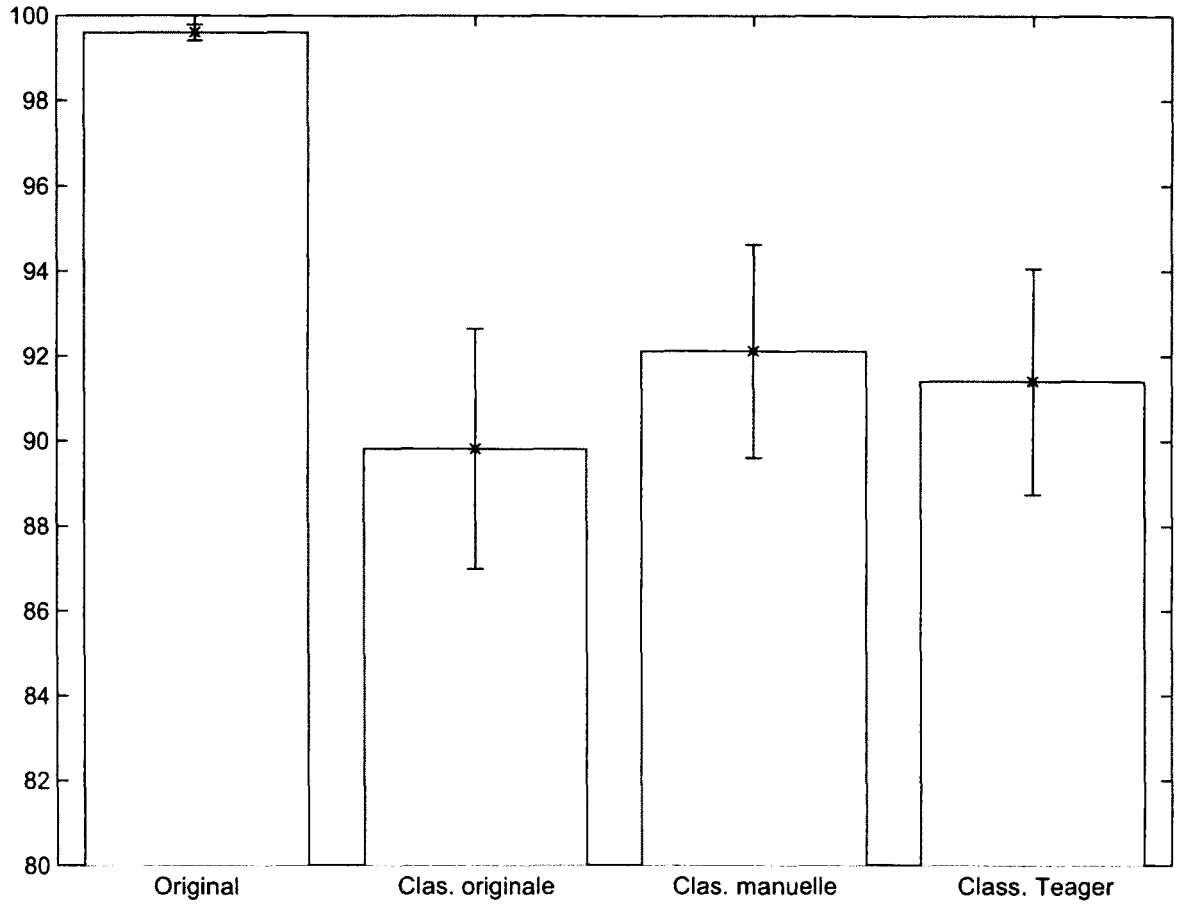


Figure 3.30 Résultats du test Mushra si la trame transitoire est perdue

soit il y aura une diminution rapide de l'énergie, soit il y a une génération d'une excitation aléatoire sans périodicité et la diminution de l'énergie est plus lente. L'artéfact créé par le premier camouflage est très audible puisqu'il y a une montée d'énergie dans la dernière bonne trame reçue suivi d'une descente rapide de l'énergie dans la trame créée par le camouflage.

Un test de format identique au précédent a été utilisé pour mesurer la performance de l'opérateur de Teager dans ce cas de figure. Dans ce test, la trame perdue est celle qui suit la transitoire et les phrases sélectionnées sont toujours celles où la classification diffère entre la classification parfaite, la classification du VMR-WB et la classification avec l'opérateur de Teager.

Dans ce cas précis, les résultats sont plus éloquents et sont présentés à la figure 3.32. Dans 98,9% des cas, les auditeurs ont été en mesure de bien identifier la phrase sans perte de trames. La note de 67,3% a été accordée à la classification originale du codeur VMR-WB, avec un intervalle de confiance qui varie de 64,4 à 71%. La note 77,3% a été

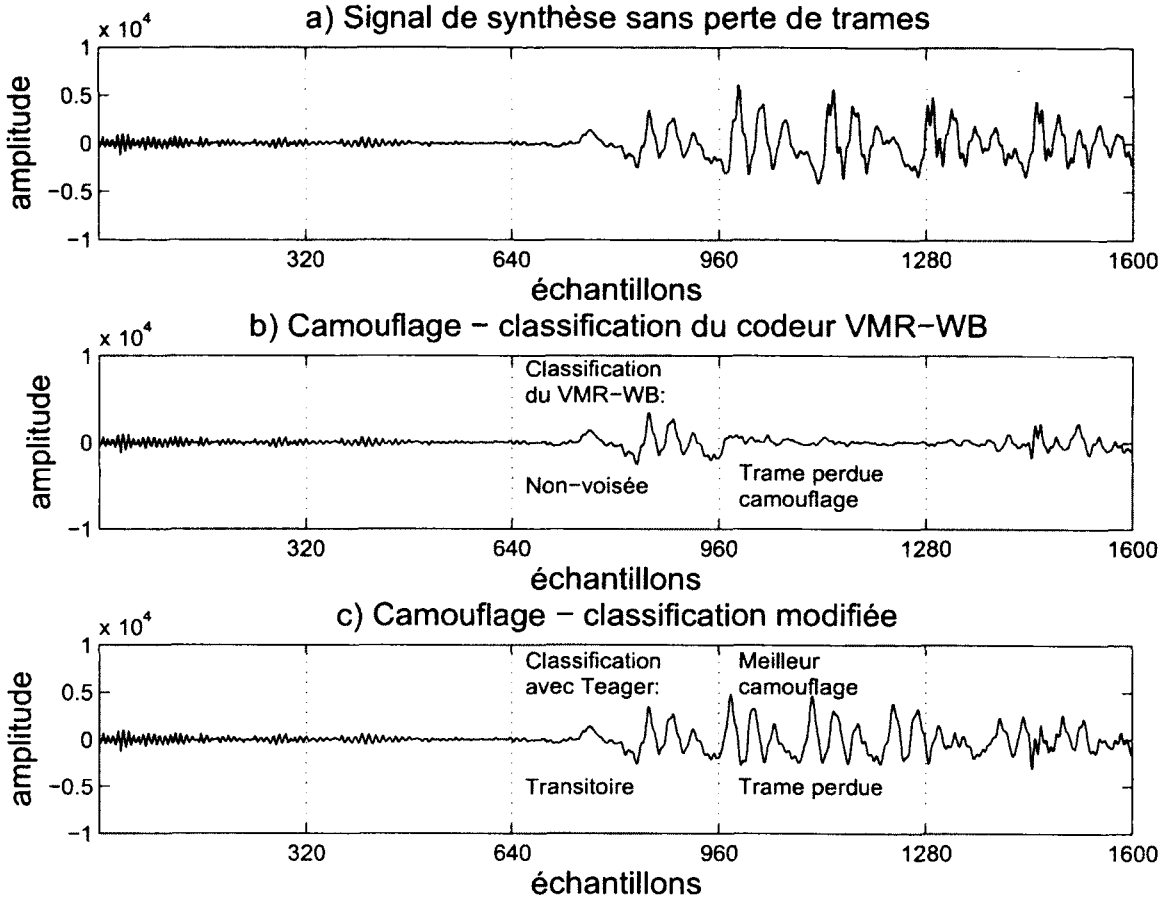


Figure 3.31 Illustration du camouflage si la trame après la transitoire réelle est perdue

accordée à la classification parfaite, avec un intervalle de confiance qui varie de 74,7 à 80,6%. La classification avec l'opérateur de Teager obtient quant à elle une note de 74,6%, avec un intervalle de confiance qui varie de 71,5 à 78%. Les classifications (parfaite et avec l'opérateur de Teager) sont statistiquement équivalentes selon l'analyse de l'intervalle de confiance à 95%. Quant à la classification originale du codeur VMR-WB, l'intervalle de confiance confirme que ces artéfacts sont plus dérangeants pour l'auditeur que ceux obtenus avec les deux autres classifications.

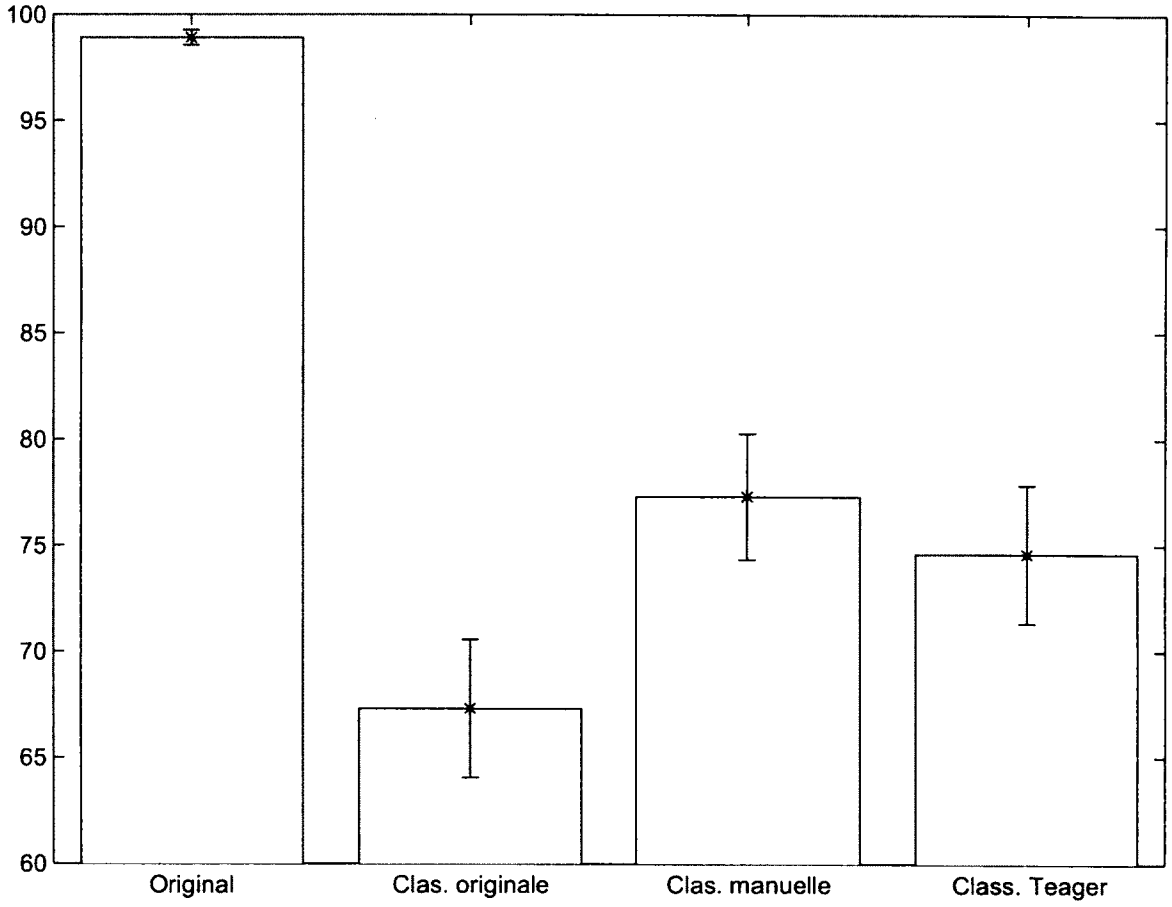


Figure 3.32 Résultats du test Mushra si la trame après la transitoire réelle est perdue

3.7 Conclusions sur l'utilisation de l'opérateur de Teager pour la détection de transitoires

En somme, cette thèse propose d'utiliser l'opérateur de Teager dans le but de faire une meilleure identification des transitions non-voisées à voisées. L'opérateur de Teager est efficace pour suivre les variations d'amplitude et de fréquence de signaux peu complexes (composantes fréquentielles restreintes). Un grand avantage de l'opérateur de Teager est sa réaction presque instantanée (en quelques échantillons) à tout changement d'amplitude ou de fréquence. De plus, cet opérateur est simple d'utilisation, puisque seulement une soustraction et deux multiplications sont suffisantes pour l'obtention d'un résultat. L'opérateur de Teager doit par contre être utilisé sur des bandes de fréquences restreintes ce qui nécessite un filtrage en bande du signal. Malgré ce fait, il est avantageux d'utiliser l'opérateur de Teager en remplacement d'un simple filtre passe-bas car ce dernier doit être beaucoup plus long pour obtenir une enveloppe lisse que les filtres utilisés pour obtenir les bandes de fréquences restreintes nécessaires à l'opérateur de Teager. Le filtre passe-bas introduisant un délai équivalent à la moitié de sa longueur, le retard introduit par le filtre est plus long avec cette méthode. De plus, l'opérateur de Teager réagit à la variation de fréquence contenue dans le signal, ce qui lui permet de détecter le début d'une transitoire même si l'augmentation de l'amplitude se fait très graduellement, ce que ne permet pas de détecter un filtre passe-bas qui extrait l'enveloppe du signal.

En focalisant l'analyse du signal dans les bandes de fréquences fondamentales de la voix (55-640 Hz pour la voix, étendu à 0-800 Hz pour la détection avec l'opérateur de Teager), l'opérateur de Teager permet de détecter le début des segments voisés. Ces débuts de voisements se caractérisent par une montée de la valeur de l'opérateur de Teager au-dessus d'un seuil. Ce seuil n'est pas fixe, il est variable en fonction de l'énergie à long-terme du signal. Ce seuil adaptatif permet également de détecter le début du signal voisé peu importe la dynamique du signal (faible (chuchotement), forte (cri) ou variante). De plus, comme il a été mentionné précédemment, l'opérateur de Teager réagit rapidement à tout changement d'amplitude ou de fréquence (quelques échantillons). Cette caractéristique de l'opérateur de Teager améliore la résolution temporelle du début du signal voisé. Ainsi, avec la méthode proposée, la résolution obtenue est de 1/8 de trame. Tel qu'exposé dans les résultats, l'opérateur de Teager permet d'atteindre cet objectif dans 93,4% des transitoires détectées. Le mauvais positionnement de l'opérateur de Teager n'est responsable que du tiers des mauvaises détections, le restant étant dû à une mauvaise combinaison de la position de l'opérateur de Teager avec la valeur du pitch. De plus, une amélioration substantielle des

fausses détections (autant en avance qu'en retard) est obtenue avec l'opérateur de Teager (diminution de plus de la moitié des mauvaises détections).

Des tests d'écoute de type Mushra démontrent que l'opérateur de Teager est efficace pour bien détecter les trames transitoires. Ces tests ont révélé que les auditeurs préférent la nouvelle classification (sans changement à la méthode de camouflage) à la classification originale du codeur VMR-WB dans le cas de perte de trames suivant une trame transitoire. Si la trame transitoire est perdue, les auditeurs ne font pas la distinction entre les méthodes de détection.

L'hypothèse qu'il est possible d'améliorer la qualité du signal de parole lorsqu'il y a pertes de trames, en ayant un positionnement plus précis de la trame transitoire, est vérifiée.

CHAPITRE 4

Deuxième contribution : Modification des transitoires partielles

L'hypothèse #2 : il est possible d'améliorer la qualité du signal de parole lorsqu'il y a pertes de trames, en modifiant les trames où il y a des transitoires partielles avant de les coder

En présence d'une détection précise de la position des trames transitoires, certains cas de figure démontrent une amélioration notable du camouflage en cas de pertes de trames. Par contre, il subsiste encore des cas où des artéfacts sont audibles en cas de pertes de trames. Cela survient en particulier lorsqu'une transitoire partielle se trouve en fin de trame (soit un début de segment voisé dont la durée dans une trame est plus courte que la période de pitch). Rappelons qu'avec la stratégie de classification des trames proposée au chapitre 3, ces trames étaient classées comme étant *non-voisées* (si la position du début de la transitoire est dans les deux dernières sous-trames, soit les sous-trames 7 et 8) ou comme étant une *transition non-voisée* (si le début de la transitoire se trouve dans les six premières sous-trames, soit les sous-trames 1 à 6). Dans les deux cas, lorsque la trame qui suit ce début de transitoire est perdue, le camouflage réagit en atténuant l'énergie de la trame qu'il doit générer. Dans l'éventualité où le début de la transitoire contenait beaucoup d'énergie, un artéfact audible sera créé à la frontière entre la trame passée et la trame courante générée par camouflage.

Il est souhaitable d'éliminer cet artéfact, qui se trouve à être le début de la transitoire voisée. Avec une bonne précision de détection du début de la transitoire à l'intérieur même de la trame, il est possible de remplacer dans le signal original ce début d'un signal voisé par un signal non-voisé moins énergique. Les transitions incomplètes à la fin des trames seront éliminées, ne laissant que des trames où les transitoires auront au moins une période de pitch bien construite précédée de trames complètement non-voisées. Cette technique requiert par contre une certaine prudence afin d'éviter l'introduction de nouveaux artéfacts qui pourraient dégrader le signal lorsqu'il n'y a pas de pertes de paquets.

Ce chapitre propose une méthode de détection des trames à modifier, soit les trames transitoires qui ne contiennent pas au moins une période de pitch complète, ainsi qu'une méthode pour modifier ces trames. La modification du signal est implémentée dans le codeur EV-VBR [Jelinek *et al.*, 2008]. Ce codeur est basé sur le VMR-WB, mais il offre

plus de versatilité quant au débit et à la largeur de bande du signal transmis. Les bases utilisées pour la classification du signal et expliquées à la section 2.3.3 restent les mêmes.

4.1 Conditions de modifications des trames transitoires

Le critère de performance le plus important pour la modification du signal original est que la modification soit transparente pour l'auditeur. La première étape consiste à déterminer les critères de sélection qui permettent la détection d'un maximum de trames à modifier tout en éliminant celles qui doivent rester intactes. Les critères menant à la modification du signal sont illustrés à la figure 4.1 et sont expliqués en détail dans les paragraphes suivants.

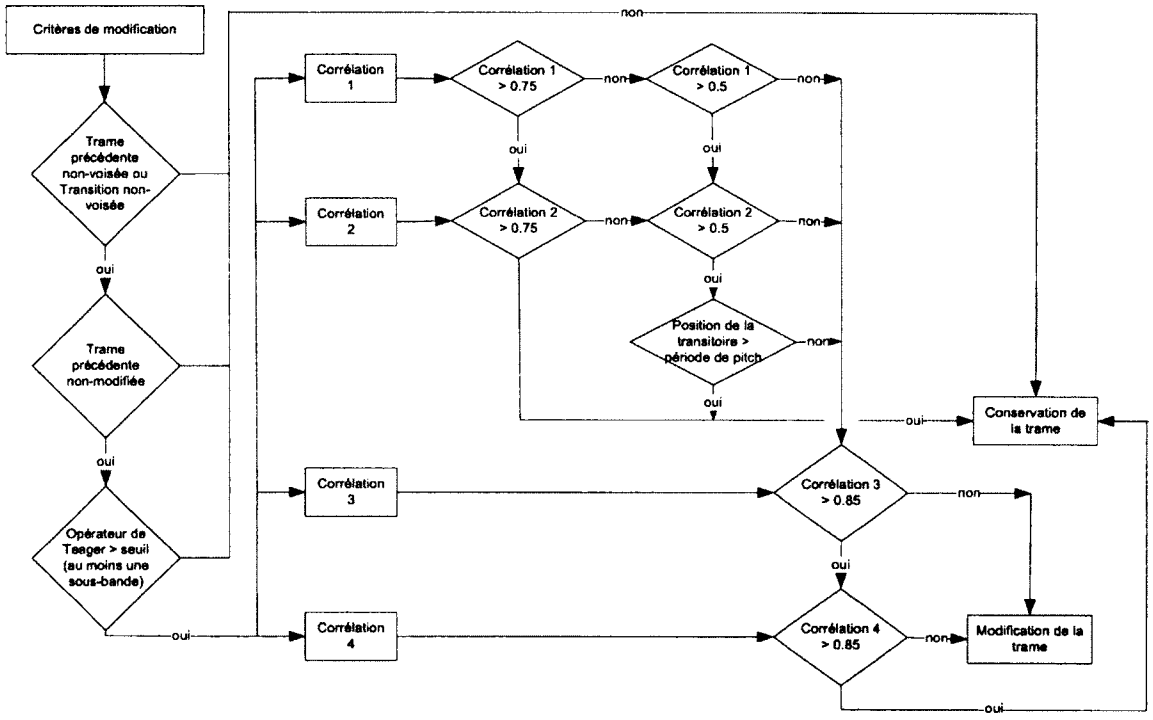


Figure 4.1 Critères qui déterminent les trames à modifier

Trois conditions initiales doivent être respectées pour la sélection d'une trame à modifier. Premièrement, pour la trame actuelle, le résultat du détecteur de Teager doit être supérieur au seuil pour au moins une sous-trame, ce qui indique un début de transition dans le signal. Deuxièmement, la trame précédente doit être classée comme étant *non-voisée* ou *transition non-voisée* par le classificateur original du EV-VBR. Troisièmement, la trame précédente ne doit pas avoir été modifiée, car la modification de deux trames consécutives provoque dans certains cas des artéfacts audibles.

Dans la section 3.5, il a été démontré qu'utiliser l'opérateur de Teager pour détecter les transitoires était profitable puisque le taux de mauvaises détections diminuait de plus de moitié. Pour ce qui est des mauvaises détections restantes, plus de la moitié étaient dues à une mauvaise combinaison entre la position de l'opérateur de Teager et la valeur de pitch. Pour éviter ces problèmes, l'approche proposée dans le chapitre présent est différente de l'approche proposée au chapitre 3 et quatre calculs de corrélation seront utilisés pour valider la présence d'au moins une période de pitch complète dans la trame.

Pour chaque trame où les trois conditions initiales sont rencontrées, le calcul de quatre corrélations doit déterminer si au moins une période de pitch est présente dans la trame courante. Les résultats de corrélation sont tous normalisés et calculés selon l'équation (4.1) :

$$R_{x(n),x(n-j)} = \frac{\sum x(n)x(n-j)}{\sqrt{\sum x(n)^2 \sum x(n-j)^2}} \tag{4.1}$$

où $x(n)$ et $x(n-j)$ sont respectivement les vecteurs de signaux de parole à l'instant n et $n-j$.

La longueur des vecteurs à corréler est égale à la longueur du pitch. Pour chacune des quatre corrélations définies dans cette section, le maximum parmi cinq valeurs de corrélations calculées autour d'un décallage égal à la valeur estimée de la période du pitch est choisi.

$$R = argmax \left\{ \frac{\sum x(n)x(n-J)}{\sqrt{\sum x(n)^2 \sum x(n-J)^2}} \right\} \tag{4.2}$$

où J varie de $j-2$ à $j+2$ et J est la longueur du pitch.

Les quatre corrélations calculées sont définies comme suit :

- Corrélation 1 - corrélation entre la dernière période de pitch située à la fin de la trame courante et la première période de pitch de la trame d'avance (1-A :1-B).

- Corrélation 2 - corrélation entre la dernière période de pitch à la fin de la demi-trame d'avance et la période de pitch précédente (2-A :2-B).

- Corrélation 3 - corrélation entre la période de pitch à la fin de la trame courante et la période de pitch précédente (3-A :3-B).

- Corrélation 4 - corrélation entre l'avant dernière période de pitch de la trame courante et la période de pitch précédente (4-A :4-B).

La figure 4.2 montre le positionnement des quatre corrélations pour une période de pitch inférieure à 128 échantillons (soit plus petite qu'une demi-trame). La corrélation (1-A :1-B) et la corrélation (2-A :2-B) ont pour objectif de confirmer la présence d'un début de voisement et le fait que ce voisement est bien contruit dans la trame future. Les corrélations (3-A :3-B) et (4-A :4-B) sont concentrées dans la trame courante et ont pour but de déterminer si le voisement est bien construit dans la trame courante. Elles aident également à trouver les cas où le pitch évolue entre la trame courante et la trame future. La figure 4.3 montre le positionnement des quatre corrélations pour une période de pitch supérieure à 128 échantillons. Comme la demi-trame d'avance ne contient que 128 échantillons, il n'est pas possible de calculer la première corrélation (1-A :1-B) avec le total des échantillons nécessaires si le pitch est plus grand que 128. La corrélation est alors calculée avec seulement une fraction des échantillons (128 échantillons, partie en gris sur la figure).

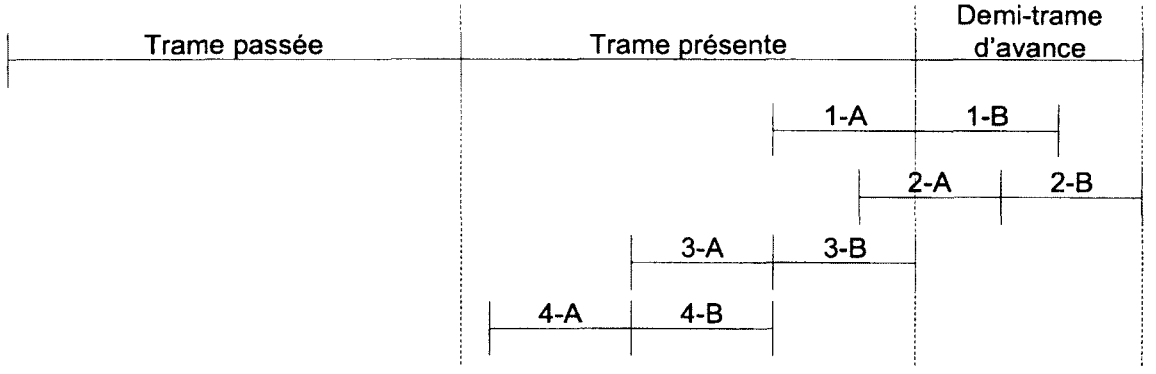


Figure 4.2 Corrélations calculées avec une période de pitch inférieure ou égale à 128 échantillons.

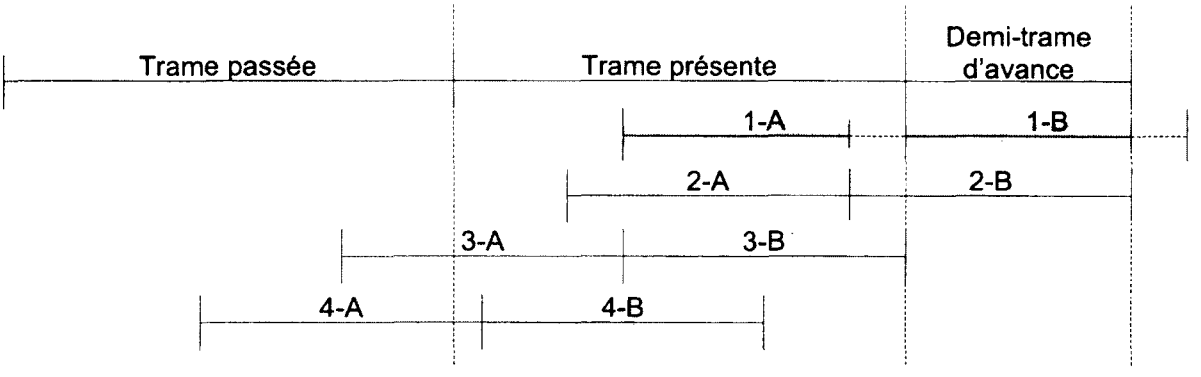


Figure 4.3 Corrélations calculées avec une période de pitch supérieure à 128 échantillons.

Une fois les corrélations calculées, l'étape suivante consiste à déterminer les caractéristiques des trames à modifier. Ces dernières ne contiennent que des fractions de période de pitch et présentent en conséquence un faible résultat pour les corrélations 1 et 2. Par contre, l'évolution rapide de la période de pitch en zone de transition pourrait nuire à la détection ce qui justifie le calcul des corrélations 3 et 4. Ainsi, s'il y a une bonne corrélation à l'intérieur de la trame courante, cette trame ne sera pas modifiée et les moins bonnes corrélations 1 et 2 seront attribuées au fait qu'il y a une évolution rapide de la période de pitch. Des corrélations élevées peuvent aussi être obtenues dans un signal complètement aléatoire. Ces corrélations sont tout à fait fortuites et il n'est pas possible de prédire à quelle intervalle dans le signal elles se produisent. Dans un signal de parole, ces corrélations fortuites peuvent masquer une trame qui devrait être modifiée. Pour éviter de trouver de bonnes corrélations fortuites, deux corrélations décalées (corrélation 1 et 2, par exemple) sont calculées.

Les seuils des valeurs de corrélation ont été déterminés et optimisés de façon expérimentale. Les valeurs des quatre corrélations, pour chacune des transitoires, ont été analysées et les valeurs pivot ont été choisies en fonction du comportement désiré de l'algorithme présenté dans cette section.

Si la corrélation 1 et la corrélation 2 sont plus grandes que 0,75 (bonne corrélation entre la dernière période de pitch de la trame courante et la période de pitch suivante, et bonne corrélation entre les deux dernières périodes de pitch qui chevauchent la trame courante et la demi-trame d'avance) ou bien, si la corrélation 3 et la corrélation 4 sont plus grandes que 0,85 (très bonne corrélation vers la fin de la trame courante), la modification de signal est écartée. Les trames répondant à ces critères sont classées comme étant des transitoires puisqu'elles contiennent au moins une période de pitch bien construite et que si la trame voisée suivante est perdue, le signal pourra être reconstruit sans artéfact.

Si la corrélation 1 et la corrélation 2 sont plus grandes que 0,5 et qu'au moins une des deux corrélations est plus petite que 0,75 (corrélation basse mais non nulle entre la dernière période de pitch de la trame courante et la demi-trame d'avance ainsi que dans les deux dernières périodes de pitch qui chevauchent la trame courante et la période de pitch suivante), un autre critère doit être analysé pour déterminer si la trame doit être modifiée. Si la distance entre la position du début du voisement et la fin de la trame est plus grande qu'une période de pitch, la trame n'est pas modifiée. La plus basse corrélation dans ce cas est probablement due au fait que les débuts de transitoires sont parfois moins bien formés que les périodes de pitch subséquentes, mais l'information disponible est assez complète pour l'obtention d'un bon camouflage en cas de perte de trame. Si la distance

entre la position du début du voisement et la fin de la trame est plus petite qu'une période de pitch, la trame est alors modifiée puisqu'il n'y a pas une période de pitch complète disponible pour faire le camouflage en cas de perte de trame.

Si la corrélation 1 et la corrélation 2 sont plus petites que 0,5, la trame est alors modifiée puisque le signal de la trame courante est trop différent du signal de la demi-trame d'avance, et ce peu importe la position du début de la transitoire et de la période de pitch.

4.2 Modification du suiveur de pitch

La valeur de pitch qui est utilisée pour le camouflage est très fiable dans les zones voisées du signal, mais est parfois inexacte dans les zones de transitions. Ces inexactitudes posent un problème dans les étapes subséquentes des modifications proposées, puisque la corrélation est utilisée pour détecter la présence d'une période de pitch complète en fin de trame. Lorsque la valeur du pitch est inexacte, il peut y avoir détection de faux positifs (détection d'une période de pitch complète, alors qu'il n'y en a pas) ou de faux négatifs (détection d'un signal incomplet, alors qu'il y a au moins une période de pitch complète à la fin de la trame).

Dans cette section, une méthode pour contraindre la recherche du pitch est proposée. Cette méthode est basée sur l'opérateur de Teager. Ce dernier étant déjà utilisé pour déterminer l'emplacement du début de la transitoire, une partie des calculs peut être récupérée. Cette méthode ne nécessite pas la présence d'une période de pitch complète pour être en mesure d'en évaluer la fréquence ce qui lui donne avantage sur une méthode par corrélation où au moins deux périodes de pitch sont nécessaires pour parvenir au même résultat.

4.2.1 Séparation de la composante fréquentielle du signal

Par définition, pour un signal discret, l'opérateur de Teager est défini par (4.3).

$$T(n) = A^2\omega^2 \tag{4.3}$$

L'amplitude et la fréquence peuvent être obtenues de façon indépendante par (4.4) et (4.5).

$$|A| = \frac{T[x(n)]}{\sqrt{T[\dot{x}(n)]}} \tag{4.4}$$

$$\omega = \sqrt{\frac{T[\dot{x}(n)]}{T[x(n)]}} \tag{4.5}$$

où $\dot{x}(n)$ est la dérivée du signal $x(n)$.

Maragos [Maragos *et al.*, 1992], [Maragos *et al.*, 1993] utilise l'opérateur de Teager pour séparer les composantes d'amplitude et de fréquence d'un signal discret. Il obtient ainsi (4.6) et (4.7).

$$\Omega(n) \approx arccos(1 - \frac{T[y(n)] + T[y(n+1)]}{4 \cdot T[x(n)]}) \tag{4.6}$$

$$|a(n)| \approx \sqrt{\frac{T[x(n)]}{1 - (1 - \frac{T[y(n)]+T[y(n+1)]}{4T[x(n)]})^2}} \tag{4.7}$$

où $\Omega(n)$ est la composante fréquentielle du signal $x(n)$, $|a(n)|$ la composante d'amplitude et $y(n)$ est définie par (4.8).

$$y(n) = x(n) - x(n-1) \tag{4.8}$$

L'estimation de la fréquence assume que $0 < \Omega(n) < \pi$, ainsi en théorie l'estimateur de fréquence fonctionne pour des fréquences allant jusqu'à la moitié de la fréquence d'échantillonnage. L'estimateur assume également que la composition fréquentielle du signal est à bande étroite. Les trois exemples théoriques qui suivent expliquent les limites de la séparation en fréquences avec l'opérateur de Teager.

On peut observer à la figure 4.4 a) un signal d'amplitude constante (100) échantillonné à 16 kHz dont la fréquence évolue dans le temps (250 Hz - 350 Hz - 450 Hz) et à la figure 4.4 b) l'évolution du résultat de l'opérateur de Teager qui lui est associé. La figure 4.4 c) montre l'évolution de la composante des fréquences, calculée à l'aide de l'équation (4.6). On remarque que cette valeur est égale à la valeur théorique de la fréquence. La figure 4.4 d) montre l'amplitude qui reste constante (comme celle du signal), calculée avec l'équation (4.7). Les résultats obtenus démontrent que (4.6) et (4.7) donnent le résultat théorique attendu avec un signal harmonique simple.

La figure 4.5 présente le résultat des mêmes opérateurs que ceux de la figure précédente. Ici par contre, le signal traité est issu du mélange de deux signaux n'appartenant pas à la même bande de fréquence et dont les amplitudes diffèrent. Rappelons que l'opérateur de Teager est employé sur des bandes de fréquences limitées à une largeur de 50 Hz ; pour cet exemple les deux fréquences choisies se trouvent alors dans des bandes distinctes. Le signal principal est défini par (4.9) et le signal secondaire par (4.10).

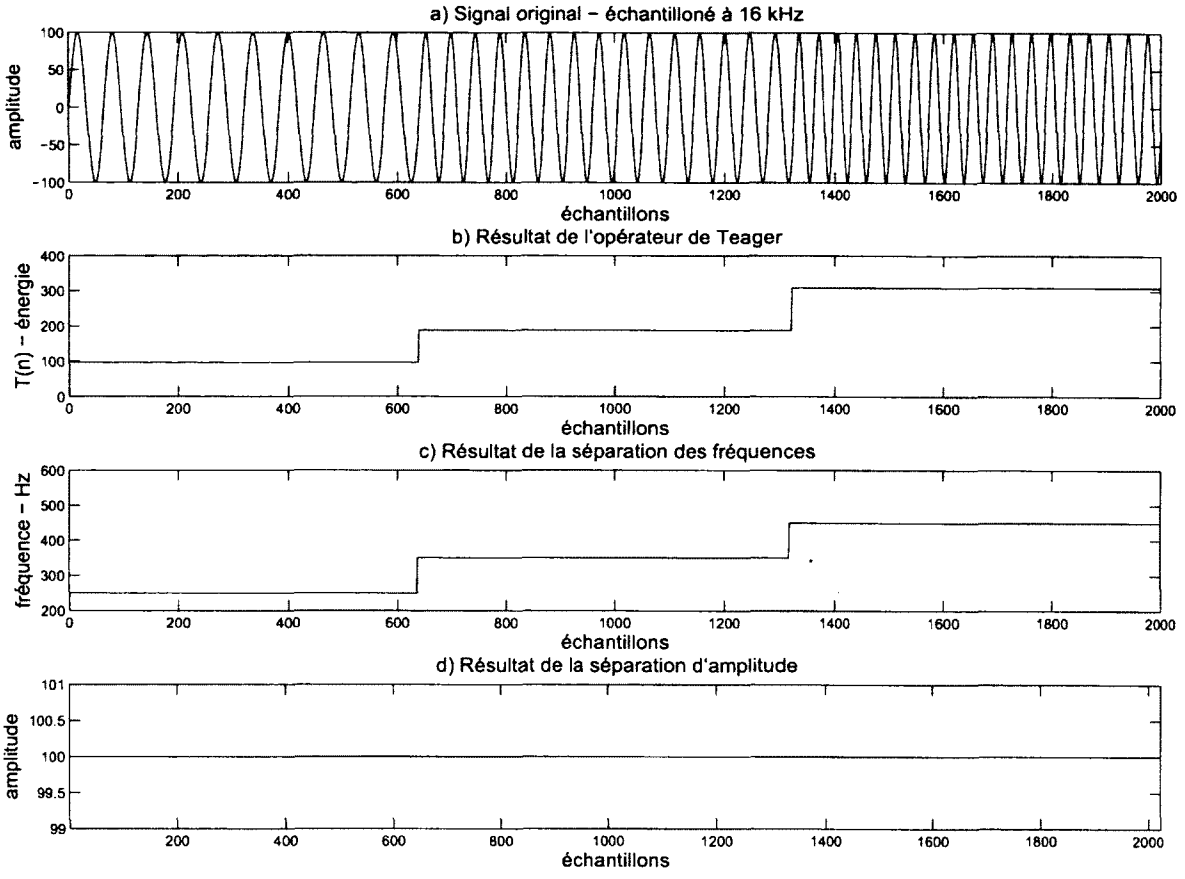


Figure 4.4 Résultats de l'opérateur de Teager du calcul de ses composantes d'amplitude et de fréquence pour un signal harmonique simple

$$100 \cdot sin(2\pi \cdot 250 \cdot n/fe) \tag{4.9}$$

$$10 \cdot sin(2\pi \cdot 50 \cdot n/fe) \tag{4.10}$$

La fréquence d'échantillonnage fe des deux signaux est de 16 kHz. La composante à 50 Hz influence le résultat de l'opérateur de Teager et cause une variation de la fréquence instantannée calculée par l'équation (4.6). Le résultat varie ainsi entre 238 et 260 Hz. Le calcul de la moyenne du résultat sur un total de 1600 échantillons donne 250 Hz. Le calcul de l'amplitude, quant à lui, donne un résultat variant entre 92 et 110 et ayant une moyenne de 100. À long terme, les valeurs d'amplitudes et de fréquences trouvées correspondent au signal de plus forte amplitude. En calculant les moyennes sur un intervalle plus court (2,5 ms), figure 4.5 e) et f), on peut valider que le résultat donné par l'équation (4.6) reste

cohérent sur une plus petite période. On obtient ainsi une fréquence variant entre 248 Hz et 252 Hz et une amplitude se situant entre 98 et 101. Bien que la valeur exacte ne soit pas trouvée, une bonne estimation est obtenue.

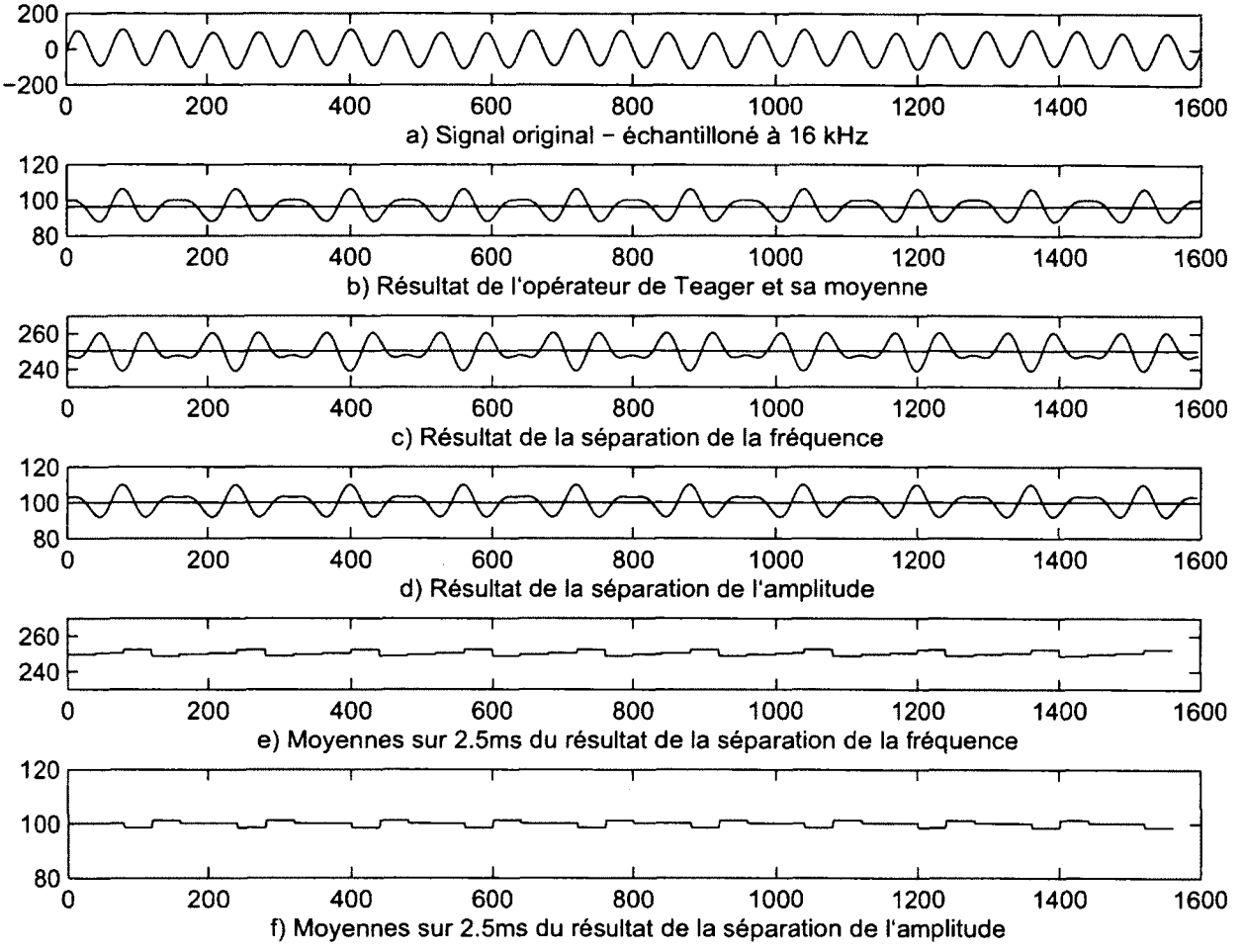


Figure 4.5 Résultats de l'opérateur de Teager et ses composantes d'amplitude et de fréquence pour un signal composé de deux sinusoïdes

On reprend le calcul du résultat des mêmes opérateurs, mais cette fois sur un signal composé de deux sinusoïdes, d'amplitudes similaires et de fréquences appartenant à la même sous-bande, voir équations (4.11) et (4.12).

$$100 \cdot sin(2\pi \cdot 250 \cdot t) \tag{4.11}$$

$$75 \cdot sin(2\pi \cdot 275 \cdot t) \tag{4.12}$$

La fréquence d'échantillonnage est toujours de 16 kHz. Ici, les variations obtenues avec (4.6) pour les fréquences instantanées sont plus grandes, soit de 28 Hz à 261 Hz. La fréquence moyenne sur 2,5 ms est plus stable, 175 Hz à 261 Hz, alors que la fréquence

moyenne sur tout le signal est de 250 Hz. La moyenne à long terme tend vers la fréquence du signal ayant l'amplitude la plus élevée. Si l'amplitude était supérieure pour le signal de 275 Hz, la moyenne à long terme serait de 275 Hz. En ce qui concerne l'amplitude instantanée, cette dernière varie de 24 à 175. L'amplitude moyenne calculée par intervalles de 2,5 ms varie de 35 à 173 mais donne 115 lorsqu'on la calcule sur la totalité du signal. La moyenne absolue réelle du signal est de 72, donc la valeur obtenue par (4.7) n'est pas exacte.

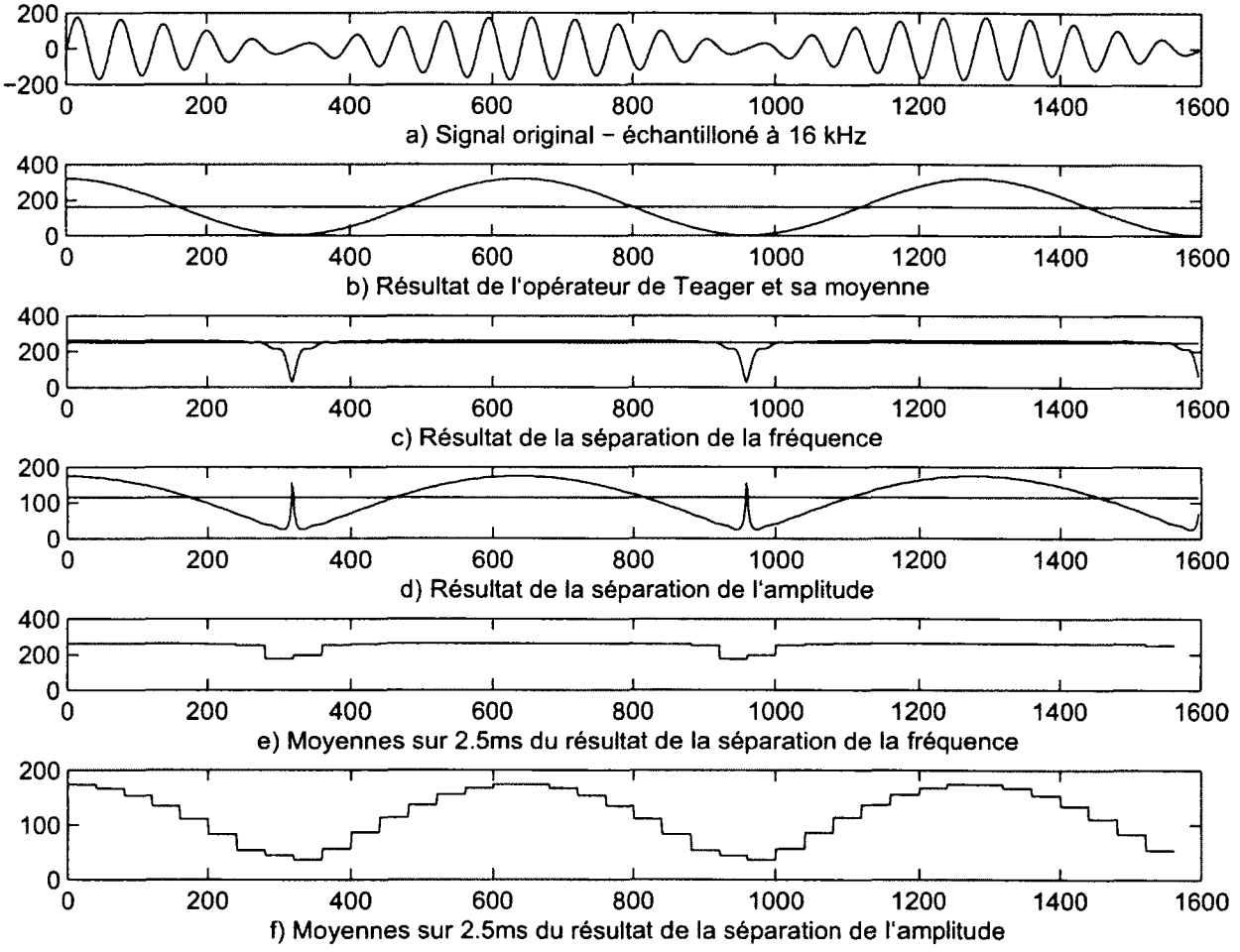


Figure 4.6 Résultats de l'opérateur de Teager et du calcul de ses composantes d'amplitude et de fréquence pour un signal composé de deux sinusoïdes limité à une sous-bande

Les résultats sur des signaux synthétiques des trois précédents exemples donnent une indication sur la composante fréquentielle ayant l'amplitude la plus élevée dans le signal analysé. En pratique, si le signal est composé de plus d'une sinusoïde, la valeur obtenue pour la fréquence instantanée ne correspond pas à l'une des fréquences qui composent le signal. Le résultat peut s'en approcher si l'une des composantes fréquentielles du signal

a une amplitude très élevée par rapport aux autres composantes fréquentielles du signal. Comme il se doit avec l'opérateur de Teager, le signal doit donc être à bande étroite et d'une composition assez simple.

4.2.2 Application de la séparation de fréquences

Il est proposé d'utiliser la séparation de fréquences pour diriger le suiveur de pitch afin de lui imposer une plage de fréquences où chercher. Pour ce faire, l'estimateur de fréquence de Maragos (4.6) est appliquée à toutes les sous-bandes déjà traitées pour la détermination du positionnement de la transitoire. La moyenne des fréquences instantanées est calculée sur des périodes de 2,5 ms et le résultat obtenu pour chaque sous-bande est borné de telle sorte que les fréquences permises sont de plus ou moins 100 Hz par rapport à la bande de fréquences analysée. Par exemple, la bande 200-250 Hz permet une fréquence instantanée de 100-350 Hz. Cet élargissement de la fréquence instantanée permise est nécessaire car le filtre passe-bande n'est pas très sélectif et laisse passer des fréquences avoisinantes.

Pour déterminer la plage de fréquences possibles pour le pitch, la fréquence instantanée trouvée dans la plus basse bande de fréquence où l'opérateur de Teager est passé au-dessus du seuil est gardée en mémoire pour chacune des sous-trames de 2.5 ms. La recherche du pitch en boucle ouverte par le suiveur de pitch du EV-VBR est bornée par les valeurs minimales et maximales trouvées parmi les huit valeurs de fréquences instantanées calculées dans la trame courante (une valeur pour chaque sous-trame).

Le prochain exemple démontre l'utilité de corriger le suiveur de pitch. Rappelons que le pitch mesuré entre la trame courante et la trame suivante est utilisé pour vérifier si la trame courante contient au moins une période de pitch bien construite, et ce dans le but de déterminer si une modification de la trame doit être faite (voir section 4.1). Dans le cas où la trame ne contient pas au moins une période de pitch correctement formée, elle sera classée comme étant une *transition non-voisée* (si la corrélation normalisée varie entre 0,5 et 0,75 et que la distance entre le début de la transitoire et la fin de la trame est plus petite que la période de pitch) ou comme étant une trame à modifier. Dans ces conditions, lorsque l'estimation du pitch est fausse, deux choses peuvent se produire. La première est que de bonnes corrélations fortuites soient trouvées (corrélations élevées dans un signal non-périodique) et que la trame soit classée comme étant une trame *transitoire.* La deuxième est que la trame contenant au moins une bonne période de pitch bien formée soit modifiée parce que les corrélations calculées auront obtenues un résultat concluant qu'il n'y a pas au moins une période de pitch bien formée dans la trame. La figure 4.7 b) illustre la mauvaise estimation du pitch pour la demi-trame d'avance.

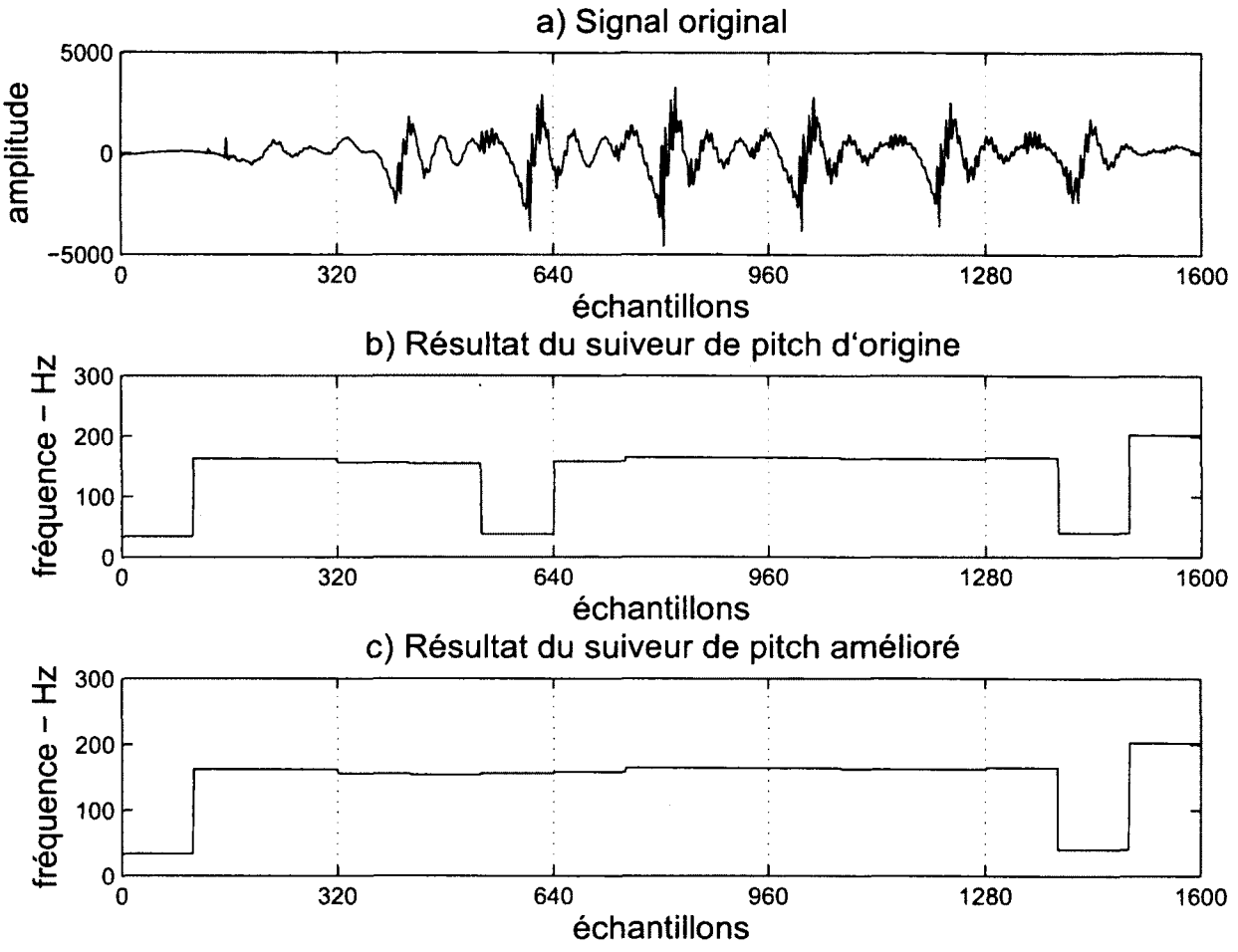


Figure 4.7 Modification à cause d'une estimation de pitch erronée, et estimation corrigée du pitch

La première trame est *non-voisée*, la deuxième trame devrait être la trame *transitoire*, mais la valeur du pitch est erronée (38 au lieu de 156). Ainsi la corrélation trouvée est très basse (plus petite que 0,5 au lieu de plus grand que 0,75) et la trame est identifiée comme une trame à modifier (les détails de la modification sont abordés dans la section suivante). Cette décision n'est pas la bonne et une modification du signal serait faite là où ce n'était pas nécessaire. La figure 4.7 c) montre la bonne estimation du pitch grâce à la correction apportée par l'opérateur de Teager et son estimation de la fréquence instantanée du signal. Ici, la deuxième trame serait bien classée comme étant la *transitoire* et le signal ne serait pas modifié.

4.3 Modification des trames transitoires

La modification des trames transitoires partielles consiste à retirer la partie voisée localisée en fin de trame. Cette partie est remplacée par un signal non-voisé semblable au signal passé. La modification du signal se fait dans le domaine du résidu pour lisser les discontinuités qu'il pourrait y avoir lors de la construction du nouveau signal. Le signal original est traité à l'aide d'un filtre à prédiction linéaire. La prédiction linéaire modélise le signal par une combinaison linéaire des échantillons passés, et ce en minimisant l'erreur obtenue entre le signal original et le signal modélisé. L'équation (4.13) résume la prédiction linéaire, où $e(n)$ est l'erreur, $x(n)$ est le signal original, et a_i sont les coefficients du filtre en nombre égal à P. Le filtre de prédiction linéaire modélise l'enveloppe spectrale du signal et le résidu $e(n)$ est le résultat du filtrage.

$$e(n) = x(n) + \sum_{i=1}^{P} a_i x(n-i) \tag{4.13}$$

Le signal est modifié en fonction de la position du début de la transitoire dans la trame. La position du début de la transitoire est déterminée avec une précision de 1/8 de trame par l'opérateur de Teager. Dans la section 3.5, il a été démontré que l'opérateur de Teager est fiable quant à la position du début de la transitoire dans une trame, puisque dans 93,4% des trames la position de l'opérateur de Teager est exacte.

Malgré le fait que l'opérateur de Teager permet de détecter le début de la transitoire avec une précision de 1/8 de trame, les modifications proposées s'effectuent avec une précision de 1/4 de trame. Cette précision a été choisie en fonction des filtres de prédiction qui sont calculés quatre fois par trame. Après l'essai de plusieurs schémas de modifications possibles, le signal reconstruit contient moins d'artéfacts si la modification du résidu est faite de façon synchrone avec les filtres de prédiction. De plus, pour ajouter de la robustesse à la détection, la modification du signal débute deux ou trois sous-trames (précision de 1/8 de trame) avant la position où l'opérateur de Teager a positionné le début de la transitoire. Toutes les sous-trames suivantes sont aussi modifiées et ce, jusqu'à la fin de la trame.

4.3.1 Modification du résidu

Le début de la transitoire se caractérise par un début de périodicité dans le signal, combiné à une augmentation de l'énergie plus ou moins rapide. La montée d'énergie est visible dans le signal résiduel dont l'énergie augmente aussi en début de transitoire. Pour éliminer le début de la transitoire, il faut donc réduire cette énergie. Plusieurs modifications du résidu

ont été essayées. Par exemple, le résidu a été remplacé par un signal aléatoire dont l'énergie était équivalente au signal résiduel qui précédait le début de la transitoire. Une simple copie du résidu passé, sans modifier l'énergie de ce dernier a aussi été testé. Pour toutes les combinaisons essayées, il y avait toujours des artéfacts audibles même sur le signal sans perte de trames. Les paragraphes suivants expliquent la méthode où les artéfacts sont les moins audibles et donc l'approche retenue pour cette thèse.

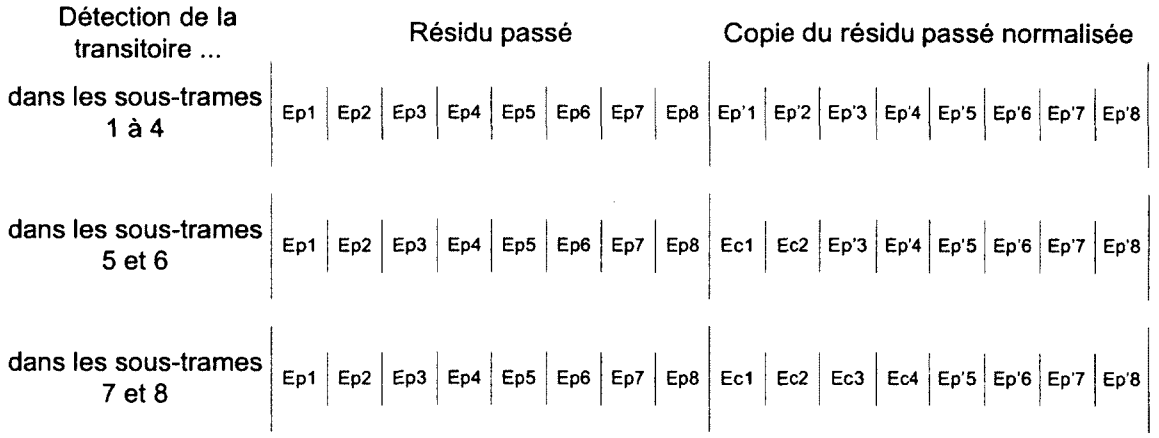


Figure 4.8 Patron de modification d'un signal résiduel

La modification consiste à remplacer les sous-trames contenant le début de la transitoire par une copie altérée des sous-trames qui les précèdent. Ce processus est illustré à la figure 4.8. Les sous-trames Ep(1 à 8) représentent le signal résiduel de la trame précédente, les sous-trames Ep'(1 à 8) représentent le signal résiduel de la trame précédente qui est copié et modifié dans la trame courante et les sous-trames Ec(1 à 8) représentent le signal résiduel de la trame courante. Par exemple, lorsque le début de la transitoire y est détecté à la sous-trame 5, le signal résiduel Ec3 à Ec8 est remplacé par le signal résiduel Ep3 à Ep8 auquel une pondération est appliquée dans le but d'éliminer le potentiel d'un artéfact audible (Ep'3 à Ep'8).

Le signal résiduel pondéré (Ep') est le signal résiduel passé auquel est appliqué le ratio d'énergie entre l'écart-type de l'énergie du signal résiduel des huit dernières sous-trames qui précèdent le début de la modification (Ep et Ec) et l'énergie de l'échantillon du signal résiduel passé (Ep) qui est copié à la place du signal résiduel courant. Le ratio s'applique uniquement lorsque l'énergie du signal résiduel passé copié dans la trame courante est plus grand que l'écart-type du signal résiduel passé. De cette façon, seuls les échantillons dont l'amplitude est plus grande que l'écart-type sont normalisés. Ceci permet de limiter l'énergie dans la trame modifiée tout en assurant un minimum d'énergie pour limiter les

artéfacts audibles. L'équation(4.14) résume le calcul de l'écart-type pour les huit sous-trames qui précèdent la modification du signal résiduel.

$$S = \sqrt{\frac{1}{N}(\sum_{i=p}^{p+N} \epsilon(i)^2) - \bar{\epsilon}^2} \tag{4.14}$$

où S est l'écart-type de l'énergie du signal résiduel, N est le nombre d'échantillons considérés dans le calcul de la moyenne et p est la position du premier échantillon des huit dernières sous-trames. Pour les échantillons qui précèdent le début de la modification du signal résiduel (l'équivalent de huit sous-trames), $\bar{\epsilon}^2$ est le carré de la moyenne de l'énergie de ce signal et $\frac{1}{N}(\sum_{i=p}^{p+N} \epsilon(i)^2)$ est la moyenne des carrés de l'énergie de ce même signal.

Lorsque l'énergie d'un échantillon du signal résiduel de remplacement Ep' est plus grande que l'écart-type de l'énergie du signal résiduel passé alors l'échantillon est normalisé selon l'équation suivante :

$$e'(n) = e(n) \cdot \frac{S}{\epsilon(n)} \tag{4.15}$$

où $e(n)$ est l'échantillon du signal résiduel, S est l'écart-type de l'énergie du signal résiduel des sous-trames Ep et Ec et $\epsilon(n)$ est l'énergie de l'échantillon $e(n)$.

Ainsi, l'énergie du signal résiduel est limitée pour les sous-trames remplacées afin d'éviter qu'il y ait une hausse d'énergie en fin de trame. Cette normalisation s'applique à toute la trame modifiée, même dans les cas où le signal résiduel original est conservé en début de trame. La figure (4.9) illustre un exemple de signal résiduel modifié où l'on voit que les deux pics plus énergiques du signal original ne sont plus présents dans le signal et que l'énergie du signal résiduel passé a été normalisée dans le signal modifié.

4.3.2 Modification du filtre de synthèse

Les filtres de synthèse utilisés pour passer du domaine du résidu au domaine du signal doivent être modifiés pour les rendre cohérents avec les modifications apportées au résidu. Quatre filtres de synthèse sont calculés pour chaque trame de 20 ms (256 échantillons à 12,8 kHz). Chaque filtre de synthèse couvre deux sous-trames de 2,5 ms (64 échantillons à 12,8 kHz), tel que calculé dans le EV-VBR.

Il existe trois cas possibles de modifications des filtres de synthèse en fonction de la position de la première sous-trame modifiée dans le résidu. Ces trois cas sont présentés à la figure 4.10, où T1 à T8 représentent la sous-trame où l'opérateur de Teager a identifié la

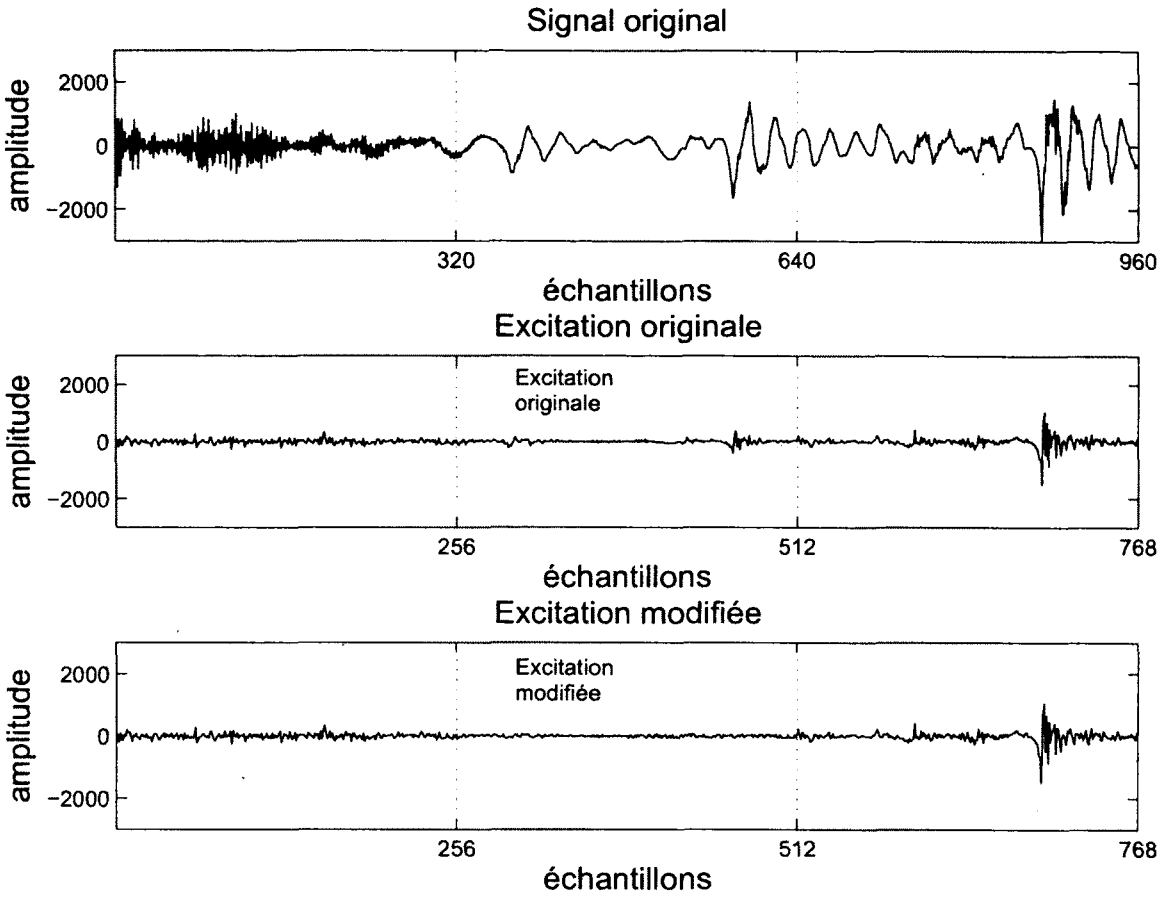


Figure 4.9 Exemple de modification d'un signal résiduel

transitoire. Dans le premier cas, la détection de la transitoire se fait dans les sous-trames T1 à T4. Les filtres utilisés pour la synthèse de toute la trame sont alors les filtres A1 à A3 (voir figure 4.10) calculés à la trame précédente. Le filtre s'appliquant à la sous-trame A4 n'est pas utilisé puisque lors du calcul des filtres dans le codeur EV-VBR, une fenêtre d'analyse de 30 ms est utilisée et déborde sur la sous-trame suivante (5 ms avant la trame et 5 ms après la trame courante). Ainsi, pour s'assurer que le dernier filtre calculé n'ait pas pris en compte le début du voisement (situé dans la trame suivante), il est remplacé par le filtre précédent A3. Dans le second cas, la détection de la transitoire se fait dans les sous-trames T5 ou T6 et les filtres utilisés pour la synthèse de la trame sont les filtres A1 à A4. Finalement, lorsque la détection de la transitoire se fait dans les sous-trames T7 ou T8 : les filtres utilisés pour la synthèse sont les filtres A2 à B1. Le type de modification à apporter aux filtres de synthèse a été déterminé par expérimentation lors de tests d'écoute de type A-B.

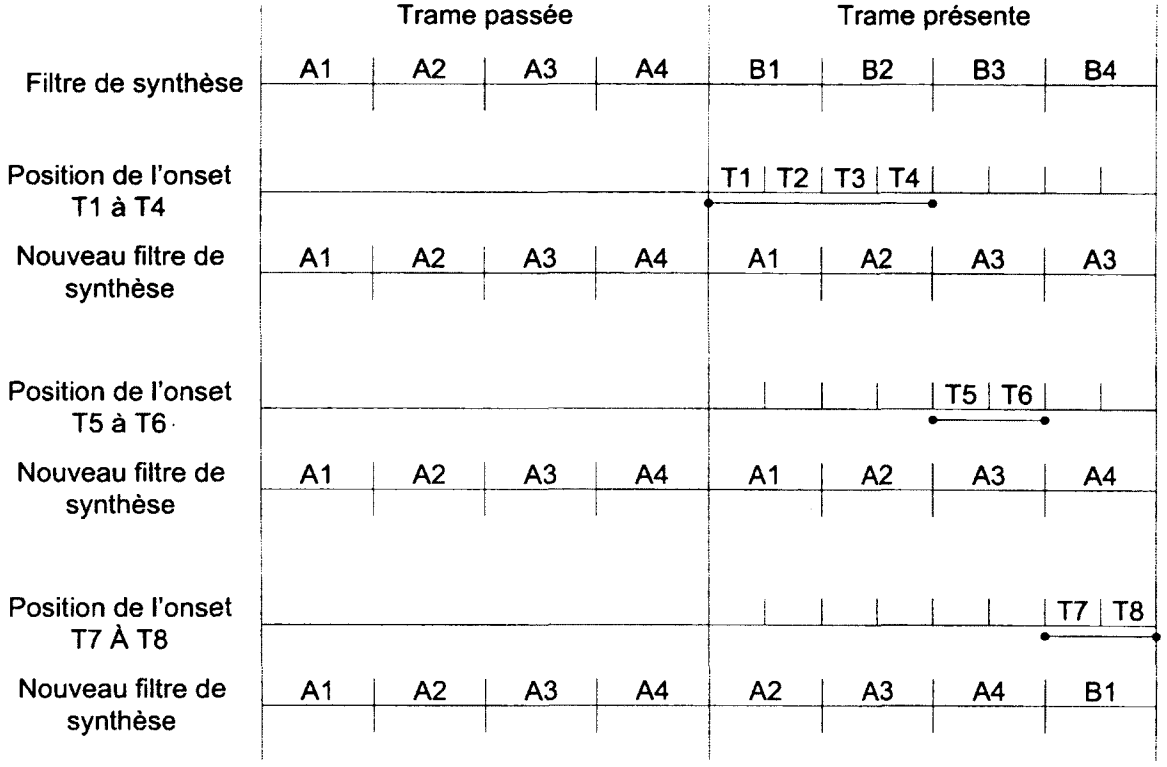


Figure 4.10 Patrons de modification du signal, pour les filtres de synthèse et les résidus

Les figures suivantes (4.11 à 4.14) illustrent les filtres de synthèse d'ordre 16 calculés pour la trame passée (sous-figure du haut) et la trame courante (sous-figure du bas). Comme dit précédemment, il y a quatre filtres de calculés pour chacune des trames et les filtres sont dans l'ordre représentés par les courbes suivantes, soit le trait continu, le trait en tiret, le trait en tiret pointillé et le trait en pointillé. Ces figures illustrent les choix faits pour la modification des filtres de prédiction linéaire.

La figure 4.11 illustre un exemple de réponse en fréquence des filtres de synthèse de la trame transitoire, calculés lorsque la position de l'opérateur de Teager est dans la deuxième sous-trame, ainsi que de la trame non-voisée précédente. Les trois premières réponses en fréquence de la sous-figure du haut sont représentatifs du signal non-voisé qui précède le début de la transitoire (le trait continu, le trait en tiret et le trait en tiret pointillé). La réponse en fréquence du filtre de synthèse qui couvre la dernière partie de la trame non-voisée chevauche aussi le début de la transitoire de la trame suivante. La réponse en fréquence (trait en pointillé de la sous-figure du haut) s'approche des réponses en fréquence des filtres de la trame de transition. Comme le signal résiduel est modifié pour toutes les sous-trames de la trame de transition, il en est de même pour les filtres de synthèse dont

la réponse en fréquence doit être similaire au résultat obtenu dans l'analyse de la trame précédente. Pour la modification des filtres de synthèse, les trois premiers filtres de la trame non-voisée sont repris et le troisième filtre est répété pour la reconstruction du signal, ces filtres caractérisent bien le signal non-voisé de la trame précédant la trame transitoire.

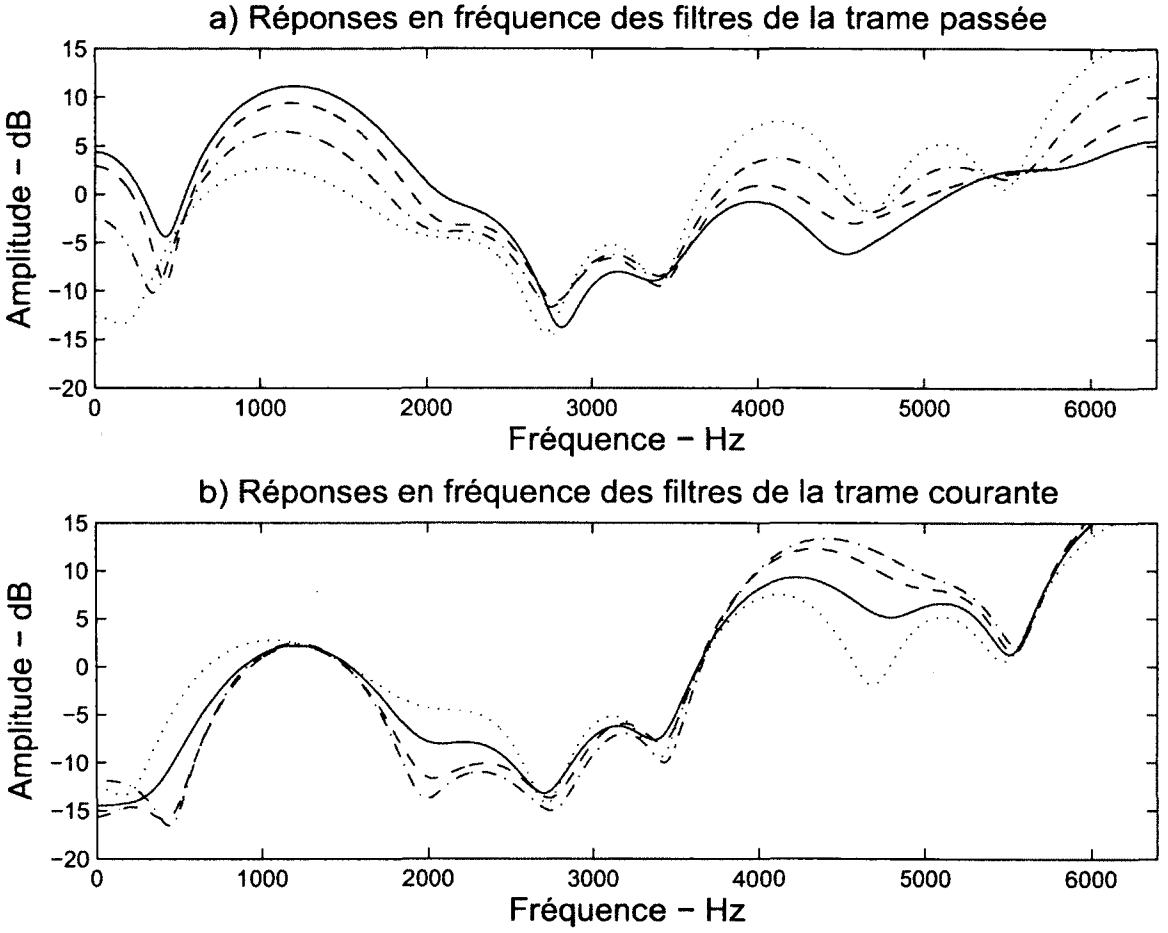


Figure 4.11 Exemple de réponse en fréquence des filtres de synthèse pour la trame passée et la trame courante lorsque l'opérateur de Teager est positionné dans la sous-trame 2

La figure 4.12 illustre un exemple de réponse en fréquence des filtres de synthèse de la trame transitoire, calculés lorsque la position de l'opérateur de Teager est dans la quatrième sous-trame, ainsi que de la trame non-voisée précédente. Dans ce cas, les quatre réponses en fréquence de la trame non-voisée sont très similaires. Les réponses en fréquences de la trame courante sont distinctes des réponses en fréquences de la trame passée non-voisée. Pour garder le même patron de modification que le cas où l'opérateur de Teager est positionné dans les deux premières sous-trames puisque le résidu est modifié de la même façon, le patron de modification des filtres de synthèse est gardé identique au cas précédent.

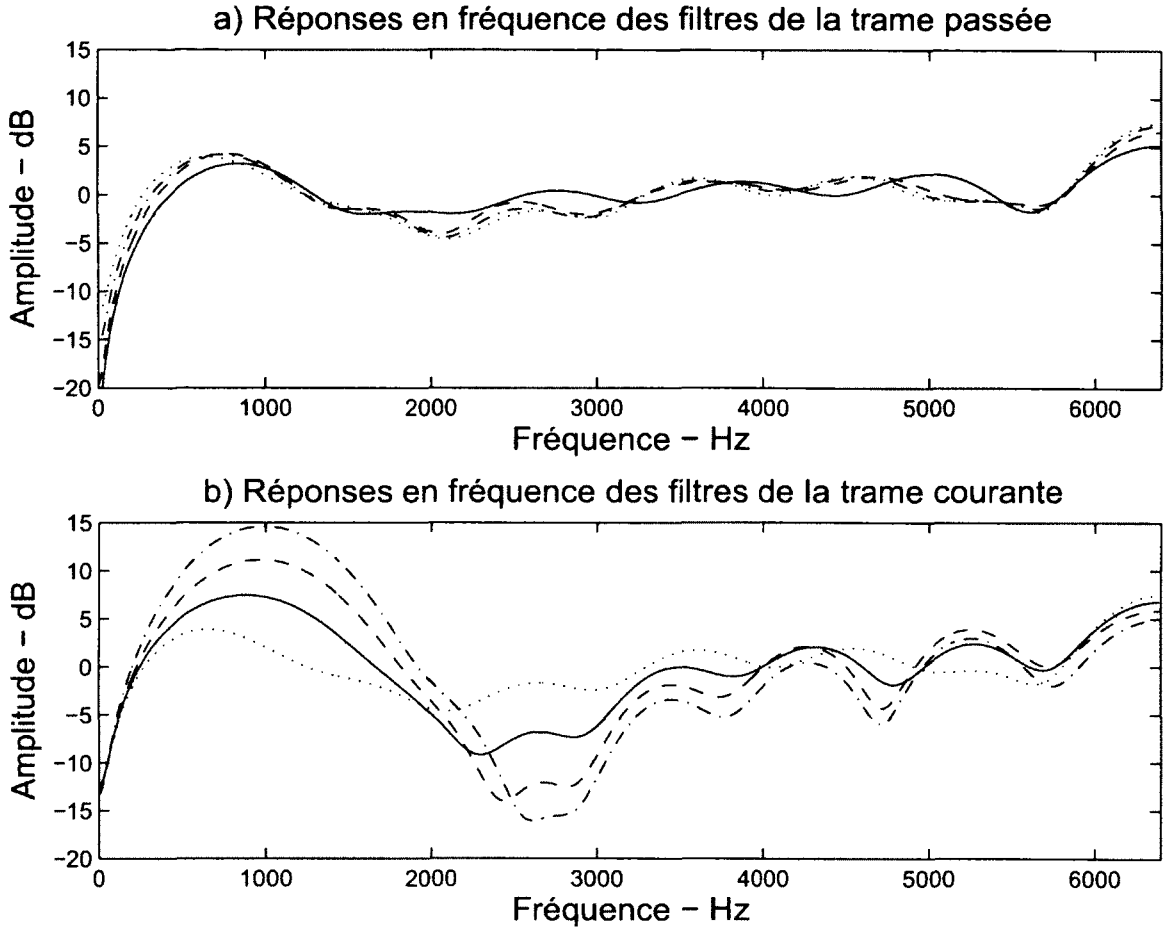


Figure 4.12 Exemple de réponse en fréquence des filtres de synthèse pour la trame passée et la trame courante lorsque l'opérateur de Teager est positionné dans la sous-trame 4

La figure 4.13 illustre un exemple de réponse en fréquence des filtres de synthèse de la trame transitoire, calculés lorsque la position de l'opérateur de Teager est dans la sixième sous-trame, ainsi que de la trame non-voisée précédente. Dans ce cas, les quatre réponses en fréquence de la trame non-voisée sont très similaires. Les réponses en fréquence des deux premiers filtres de la trame courante (le trait continu et le trait en tiret de la sous-figure du bas) ont l'allure des réponses en fréquence de la trame non-voisée, par contre les basses fréquences (0-1000 Hz) des deux derniers filtres se rapprochent des réponses en fréquence du signal voisé (le trait en tiret pointillé et le trait en pointillé de la sous-figure du bas). Comme le signal résiduel est modifié à partir de la troisième sous-trame et que dès le premier filtre de la trame courante les caractéristiques de la réponse en fréquence tendent vers la réponse en fréquence du signal voisé, tous les filtres sont modifiés. Ainsi, les filtres de la trame précédente sont utilisés pour reconstruire la trame courante.

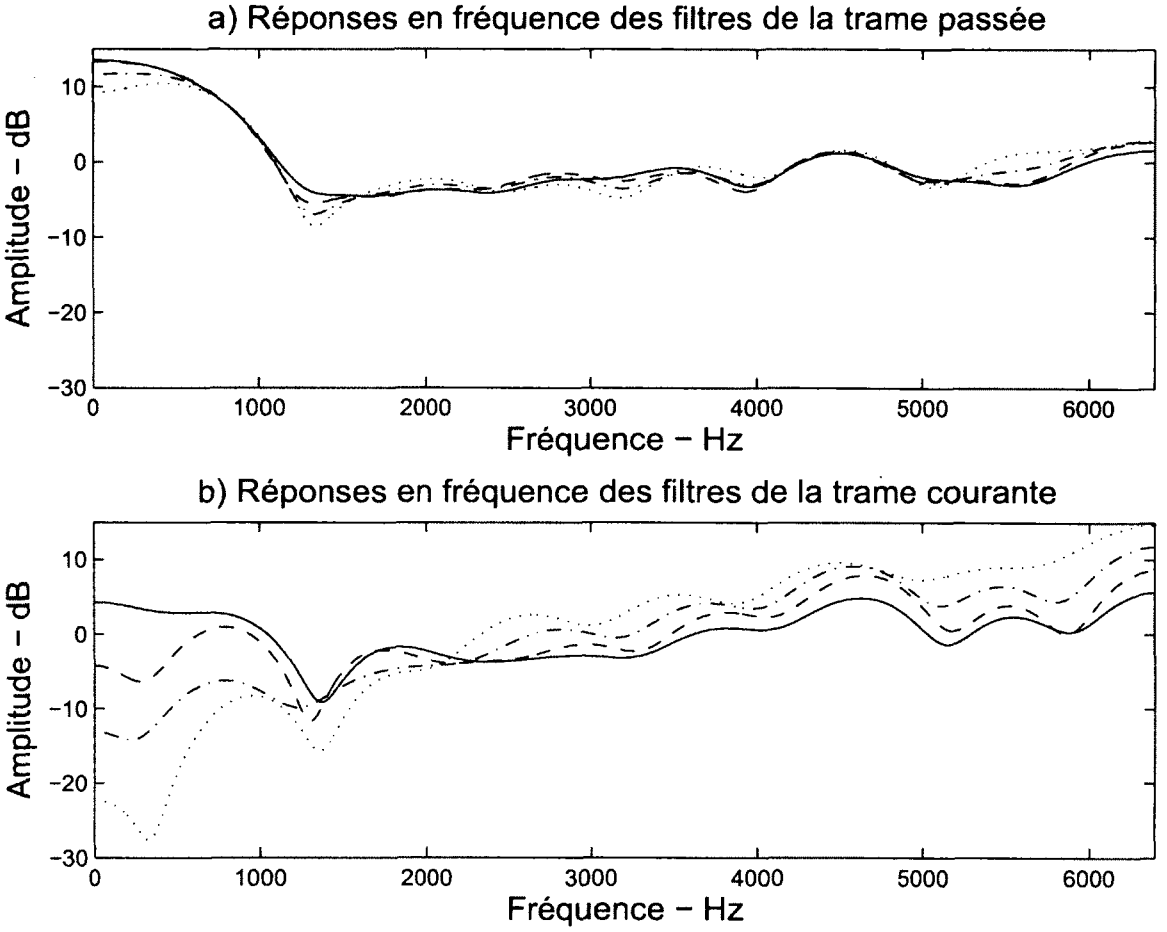


Figure 4.13 Exemple de réponse en fréquence des filtres de synthèse pour la trame passée et la trame courante lorsque l'opérateur de Teager est positionné dans la sous-trame 6

La figure 4.14 illustre un exemple de réponse en fréquence des filtres de synthèse de la trame transitoire calculés lorsque la position de l'opérateur de Teager est dans la dernière sous-trame, ainsi que de la trame non-voisée précédente. Dans ce cas, les quatre réponses en fréquence de la trame transitoire sont très similaires. La réponse en fréquence du premier filtre de la trame transitoire (le trait continu de la sous-figure du bas) est très similaire de la réponse en fréquence du dernier filtre de la trame non-voisée (le trait en pointillé de la sous-figure du haut). Pour les autres réponses en fréquence, elles tendent vers celle du signal voisé (le trait en pointillé de la sous-figure du bas). Comme le signal résiduel est modifié à partir de la cinquième sous-trame et que dès le deuxième filtre de la trame courante les caractéristiques de la réponse en fréquence tend vers la réponse en fréquence du signal voisé, les trois derniers filtres sont modifiés. Ainsi, les trois derniers filtres de la trame précédente sont utilisés pour reconstruire la trame courante.

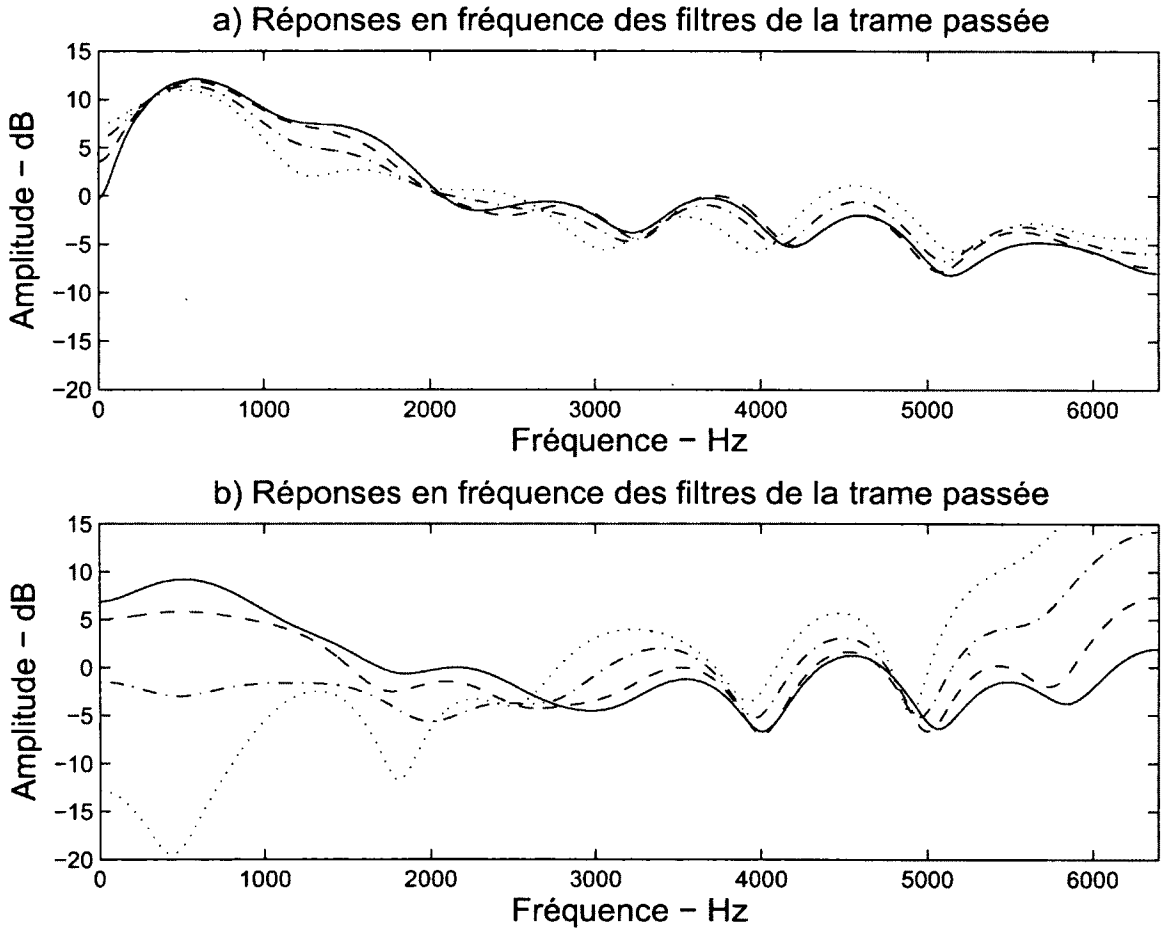


Figure 4.14 Exemple de réponse en fréquence des filtres de synthèse pour la trame passée et la trame courante lorsque l'opérateur de Teager est positionné dans la sous-trame 8

4.3.3 Exemples de modification

Maintenant que les critères de modification sont connus, ainsi que le mécanisme de modification des trames, il est intéressant d'étudier la différence entre le signal original (sans modification) et le signal modifié. Rappelons que l'idée de modifier le signal est d'améliorer les performances lorsqu'il y a perte de trames tout en restant transparent lorsque le signal est bien reçu. Les deux exemples suivants démontrent bien l'impact de la perte de trame lorsqu'il y a une transition incomplète dans la trame précédant la trame perdue (voir figure 4.15 et 4.16, a) et b)). Les deux cas présentés sont des cas où les artéfacts résultants du camouflage sont audibles. Il y a donc un début de transitoire incomplète, suivi d'une trame perdue. Comme la transitoire incomplète est classée comme étant une trame *transitoire non-voisée*, la trame de camouflage diminue rapidement l'énergie.

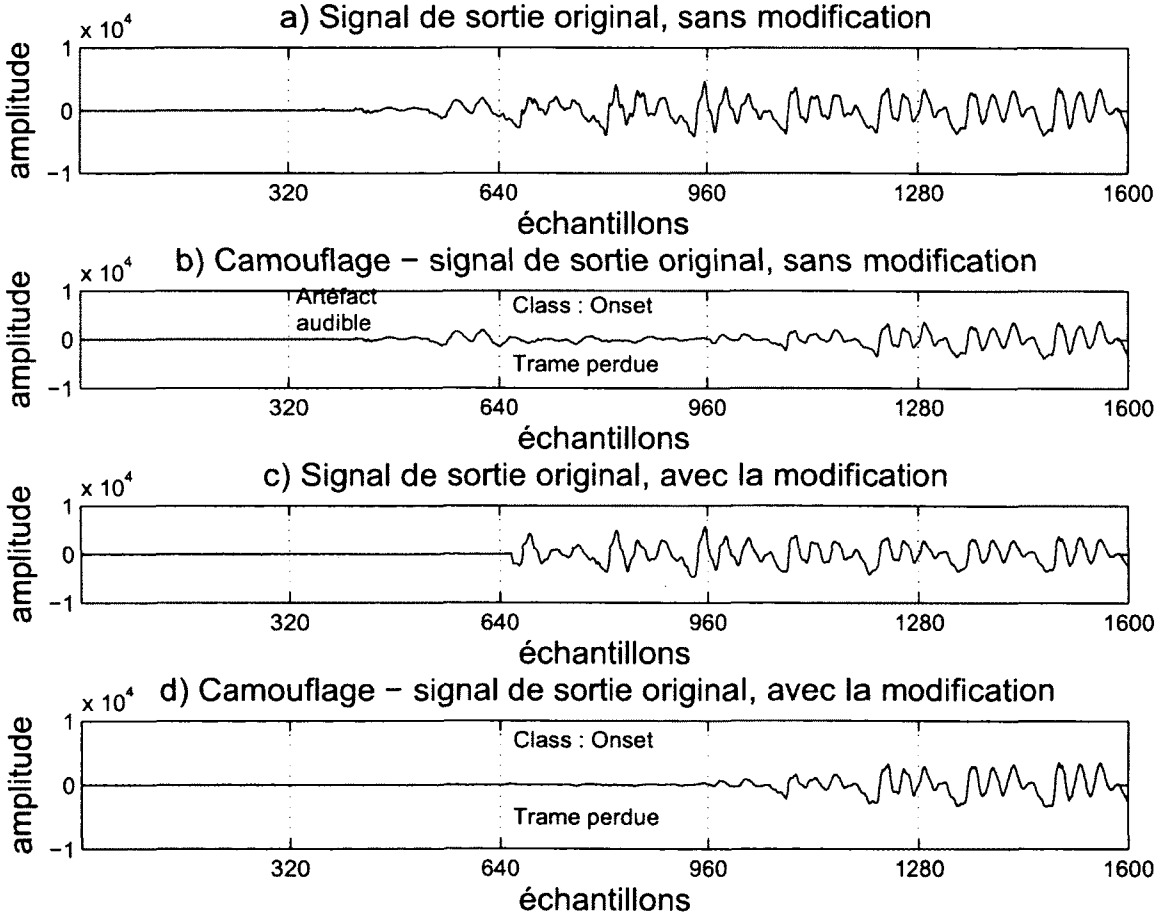


Figure 4.15 Différence entre le signal original et le signal modifié lorsque la trame transitoire est perdue

Avec la modification du signal, la dernière partie de la trame précedent la trame perdue, qui contenaient le début d'un segment voisé, a été remplacée par un signal non-voisé de

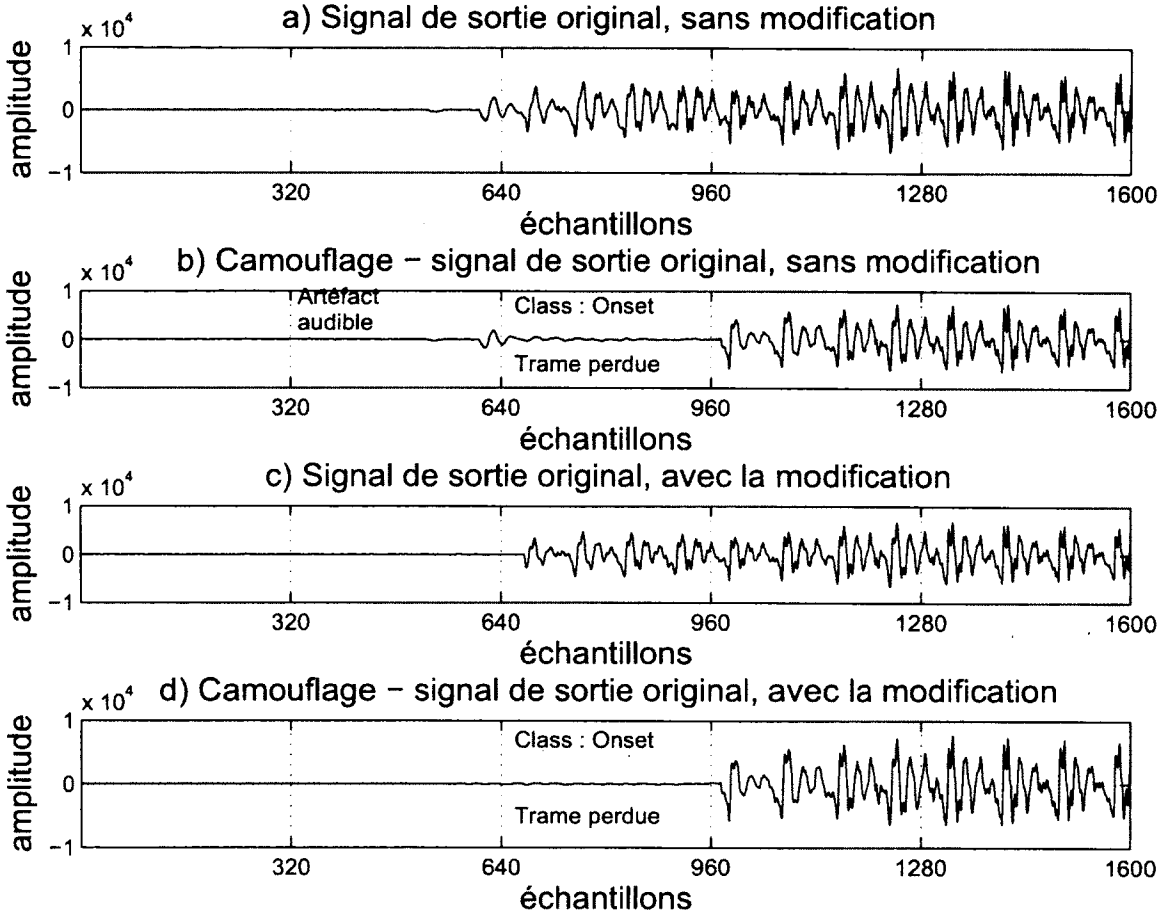


Figure 4.16 Différence entre le signal original et le signal modifié lorsque la trame transitoire est perdue

même énergie que le signal précédent. Ainsi lorsque la trame transitoire est perdue, le camouflage poursuit le signal non-voisé en diminuant l'énergie, mais sans la diminution rapide d'énergie. Alors qu'à la sous-figure 4.16 b), le problème était causé par la montée importante de l'énergie causé par le début de la transitoire, suivi par la diminution rapide d'énergie dû au camouflage.

4.4 Validation de la modification des trames transitoires partielles

Afin de prouver que les modifications des trames transitoires incomplètes apportent vraiment une amélioration au camouflage en cas de perte de trames, des échantillons sonores ont été soumis à un groupe de personnes chargées d'en évaluer la qualité. Au cours des tests, la classification des signaux est demeurée la même pour les signaux modifiés et pour les signaux non-modifiés. Conserver la classification permet de comparer uniquement l'effet de la modification du signal et ce sans aucun biais. La méthode utilisée et celle des tests A-B, c'est-à-dire que l'auditeur a à choisir entre deux candidats celui qu'il préfère en votant ± 1. Lorsque l'auditeur ne perçoit pas de différence entre les deux candidats, il peut voter "0" et dans le cas où sa préférence est marquée, il peut voter ± 2. Pour ce test, les votes négatifs sont pour le signal de référence et les votes positifs sont en faveur du signal modifié. Il est à noter que la majorité des phrases contiennent au moins une trame modifiée. Le fichier de test contient 128 phrases, subdivisées en 22804 trames. De ces trames, un total de 279 ont été modifiées, soit 1.22% des trames de parole active.

Les tests ont été faits en trois étapes. Dans un premier temps, le signal de référence (codé/décodé), ainsi que le signal modifié (codé/décodé) sont comparés. Cette comparaison a pour but de vérifier la transparence des modifications, à savoir si elles sont ou non perçues par les auditeurs. Les résultats obtenus sont affichés à la figure 4.17. Sur un total de 160 votes, les résultats sont répartis de la façon suivante, soit (score -2 : 2 votes), (score -1 : 37 votes), (score 0 : 96 votes), (score 1 : 25 votes) et (score 2 : 0 vote). Dans 24% des cas, les auditeurs ont préféré l'original, contrairement à 16% pour la version modifiée. Un total de 16 votes séparent la version originale et la version modifiée, et ce en faveur de la version originale. Dans 60% des cas, les auditeurs n'ont pas perçu de différence entre les deux versions. Selon l'analyse statistique, la moyenne des votes est égale à 0,1 et l'intervalle de confiance à 95% est de 0,102 (voir encadré de la figure 4.17, où le 'x' désigne la moyenne et les barres verticales délimitent l'intervalle de confiance). La plage de l'intervalle de confiance varie de -0,202 à 0,002. Lorsque l'intervalle de confiance est distribuée de chaque coté du zéro, il en résulte que la différence entre les deux signaux présentés n'est pas statistiquement significative. Ici, bien que ce soit le cas, le partie de l'intervalle qui est à droite du zéro n'est pas très prononcée. Dans ce cas, les conclusions sont que la différence est en faveur du signal original.

Dans un deuxième temps, des pertes de trames ont été introduites dans le signal. Ces pertes de trames sont positionnées de façon stratégique. Ainsi, la trame qui suit la trame

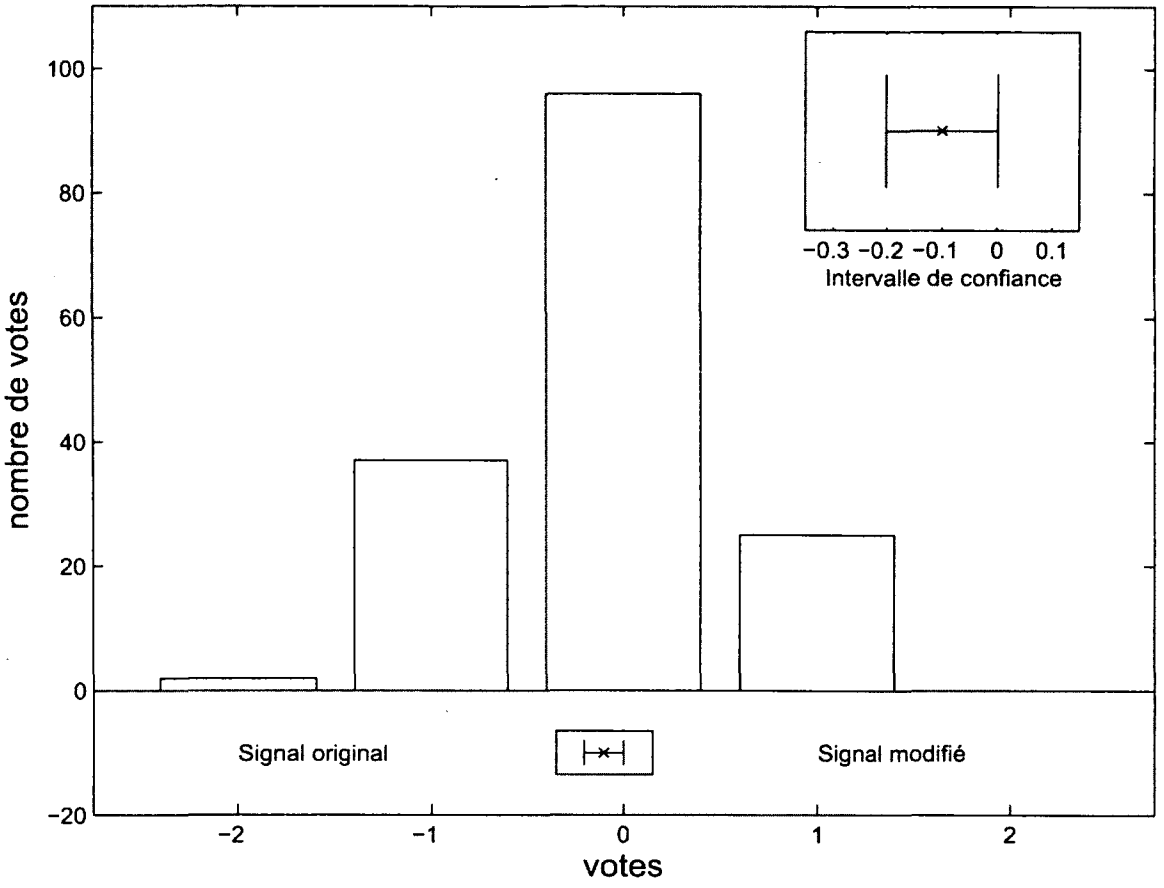


Figure 4.17 Résultats du test AB et son intervalle de confiance, comparaison entre le signal original et le signal modifié, sans erreur de canal

modifiée est systématiquement perdue. Dans le cas du signal de référence, la trame à modifier est conservée alors que dans l'autre, cette trame est modifiée de façon à enlever le début du voisé incomplet. Cette procédure introduit des artéfacts dans le signal sans modification qui contient des pertes de trames et ce, en raison du camouflage non-voisé qui fait tendre le signal de remplacement rapidement vers zéro à la suite du début de la transitoire partielle.

Les résultats obtenus sont distribués comme suit :(score -2 : 3 votes), (score -1 : 37 votes), (score 0 : 92 votes), (score 1 : 40 votes), (score 2 : 20 votes) et sont illustrés à la figure 4.18. Sur un total de 192 votes, près de la moitié (47,9%) ne sont pas en faveur d'une approche en particulier. Un total de 40 votes (20,8%) est favorable à la version originale, alors que 60 votes (31,.2%) sont en faveur du signal modifié. Un tiers de ces derniers votes sont de (+2), ce qui indique que les auditeurs ont entendu une différence frappante entre les deux versions. L'intervalle de confiance est illustré dans l'encadré de la figure 4.18 (où la moyenne est indiquée par un 'x' et les barres verticales délimitent l'intervalle de

confiance) et montre que la moyenne des votes pour ce test est de 0,193 avec un intervalle de confiance variant entre 0,063 et 0,323. L'analyse statistique de l'intervalle de confiance à 95% démontre que les auditeurs préfèrent le signal modifié au signal original.

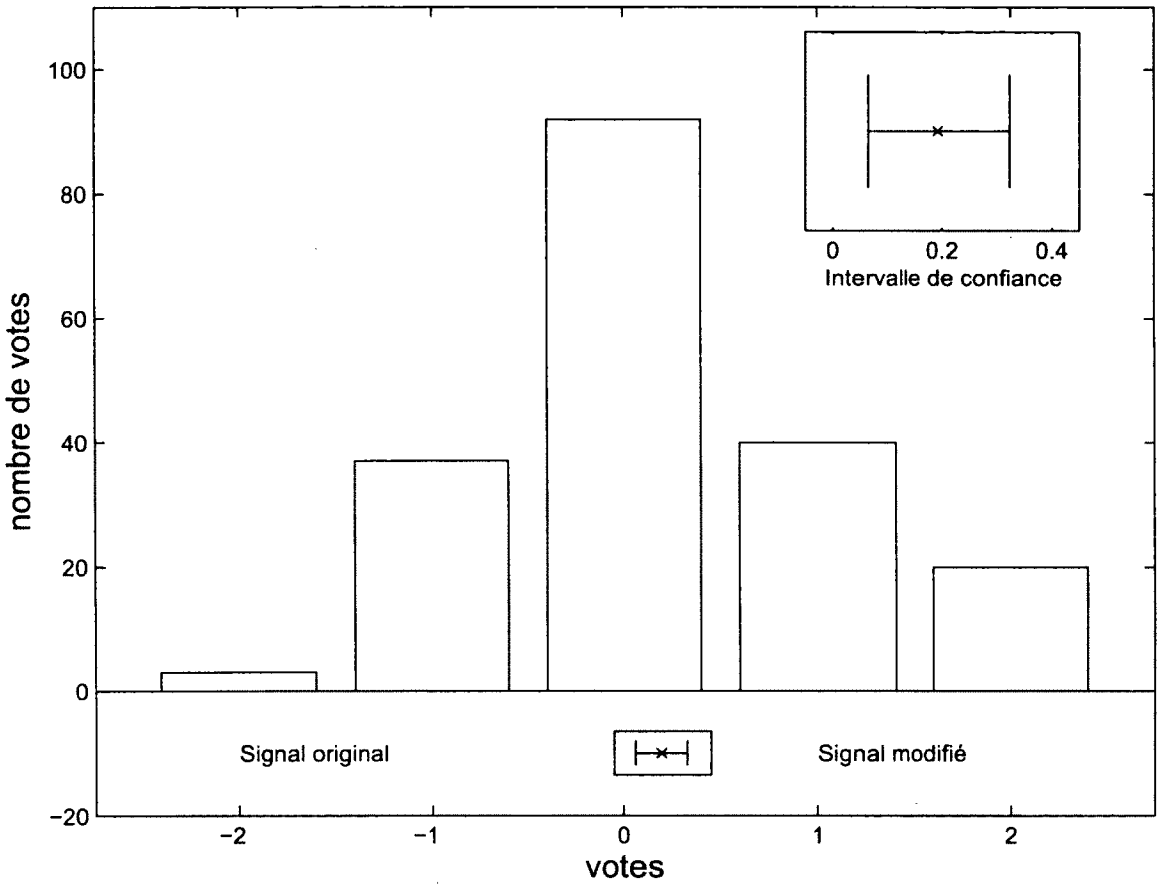


Figure 4.18 Résultats du test AB et son intervalle de confiance, comparaison entre le signal original et le signal modifié, avec erreurs après les trames modifiées

Dans un troisième et dernier temps, les pertes de trames ont été placées de façon aléatoires à travers toutes les phrases. Le taux de trames perdues est de 8% et un total de 279 trames ont été modifiées. De ces trames modifiées, 22 ont été perdues. Les résultats sont présentés à la figure 4.19 et sont distribués comme suit : (score -2 : 7 votes), (score -1 : 35 votes), (score 0 : 103 votes), (score 1 : 38 votes) et (score 2 : 8 votes). Un total de 42 votes (22%) sont en faveur du signal original contre 46 votes (24%) pour le signal modifié. Comme mentionné précédemment, seule une petite partie des erreurs concerne le cas étudié (cas où les pertes de trames surviennent sur la transitoire et où la trame précédente a été modifiée). L'intervalle de confiance illustré dans l'encadré de la figure 4.19 (où la moyenne est indiquée par un 'x' et les barres verticales délimitent l'intervalle de confiance) montre que malgré le fait que le cas étudié ne se produise pas très souvent, les résultats démontrent

une performance similaire pour les deux versions (moyenne égale à 0,026 avec un intervalle allant de -0,092 à 0,145. Ainsi, la dégradation vue pour le canal sans erreur (premier volet du test) semble compensée par l'amélioration de la performance en présence des trames effacées.

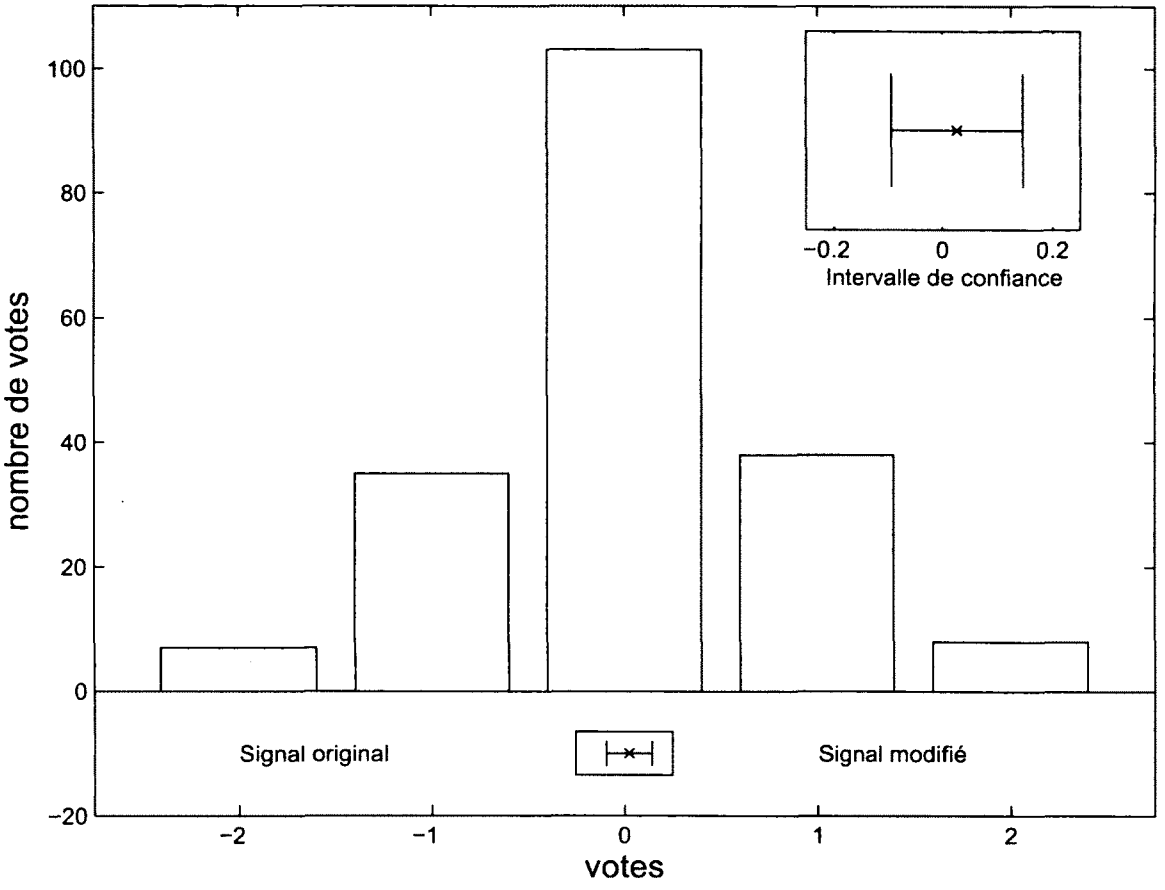


Figure 4.19 Résultats du test AB et son intervalle de confiance, comparaison entre le signal original et le signal modifié, avec un taux d'erreurs aléatoires de 8%

4.5 Conclusion sur la modification du signal

En conclusion, pour qu'une trame soit candidate à la modification de signal, elle doit se terminer par une transitoire qui ne contient pas au moins une période de pitch complète et doit être précédée par une trame *non-voisée*.

Comme la décision de modifier une trame ou non repose sur la corrélation, il est important que la valeur du pitch soit fiable. Le suiveur de pitch actuel est très fiable sur les signaux complètement voisés, mais peut se tromper lorsqu'il y a une transition non-voisé à voisé. Afin d'améliorer la robustesse du suiveur de pitch, une amélioration basée sur l'opérateur de Teager a été proposée. L'opérateur de Teager est utilisé pour faire une approximation du contenu fréquentiel du signal. Les résultats obtenus sont utilisés dans le suiveur de pitch pour en borner l'étendu des recherches et limiter les erreurs en zone de transition.

Les performances de la modification ont été évaluées à l'aide de trois tests d'écoute. Le premier des tests a servi à valider la transparence de la modification. Les résultats montrent que les auditeurs préfèrent légèrement le signal original, mais ils ne perçoivent pas une différence significative entre les deux signaux. L'intervalle de confiance confirme cette conclusion par le fait qu'il croise légèrement le zéro. Le deuxième test évalue l'effet de la modification de trames lorsque la trame suivant la modification est systématiquement perdue. Les auditeurs ont voté dans ce cas de figure plus souvent en faveur de la modification. Encore une fois, près de la moitié des phrases sont perçues comme étant équivalentes. Ce résultat n'est pas surprenant puisque le cas traité ne cause pas d'artéfacts audibles à tout coup dans le signal de référence. L'intervalle de confiance confirme que les auditeurs préfèrent le signal modifié. Le dernier test est celui où les erreurs sont distribuées au hasard sur 8% des trames. Le pourcentage de cas problèmes n'est pas significatif et les auditeurs ne percoivent pas de différence entre le signal original et le signal modifié.

La modification, bien que transparente dans la grande majorité des cas, ne l'est pas toujours et une légère dégradation est perçue par les auditeurs. La technique de modification des trames incomplètes permet une amélioration substantielle de performance lors d'un effacement systématique des trames suivant les trames modifiées ce qui démontre le potentiel de la technique. Dans le cas où les effacements de trames sont aléatoires, les résultats obtenus ne prouvent pas une amélioration générale de la performance. Afin que la technique soit utile en pratique, un des objectifs suivants devrait être atteint : soit éliminer la dégradation en transmission sans erreurs ou augmenter la performance dans le cas des effacements aléatoires.

CHAPITRE 5

Conclusion

La transmission efficace d'un signal de parole numérique à travers un réseau nécessite son découpage en trames. La fiabilité imparfaite des réseaux (ou leurs limites) provoque la perte de trames ou retarde de façon inacceptable certaines d'entre elles. L'information manquante doit alors être reconstruite au décodeur. Diverses techniques de camouflage existent pour pallier ce problème et éviter la présence d'artéfacts audibles. Ces techniques différent selon le type de codeur utilisé.

Dans le cas de codeurs de parole non-prédictifs, la trame précédente est utilisée pour reconstruire la trame perdue. Comme chacune des trames est codée de façon indépendante, l'impact de la perte d'une trame est localisé et très court. Le débit de ce type de codeur est toutefois suffisamment élevé pour justifier l'utilisation de codeurs prédictifs. Le camouflage des codeurs prédictifs utilise également l'information contenue dans la dernière bonne trame reçue pour reconstruire la trame manquante. Toutefois, la perte d'une trame cause une désynchronisation entre le codeur et le décodeur ce qui entraine une dégradation du signal même sur les trames bien reçues suivant la trame perdue. L'impact de la perte d'une trame pour un codeur prédictif est ainsi plus important.

La propagation d'erreurs peut être observée particulièrement lors de la perte des trames de transition entre le non-voisé (parole non-voisée, silence ou bruit) et le voisé (parole voisée). La propagation est d'autant plus étendue que la perte est combinée à une mauvaise classification de trames ; les pires cas étant une trame non-voisée classée voisée et trame voisée classée non-voisée. Dans le premier cas (une perte de trame suivant une trame non-voisée classée voisée), le camouflage répète la dernière période de pitch de la dernière trame reçue. Comme cette trame était non-voisée, la période de pitch n'est pas représentative et le signal reconstruit contient des artéfacts (introduction de périodicité dans le signal). Dans le second cas (une perte de trame suivant une trame voisée, classée comme étant non-voisée), le camouflage diminue parfois rapidement l'énergie du signal reconstruit. Ainsi le signal reconstruit contient une descente rapide d'énergie par rapport à la dernière bonne trame reçue, le tout provoquant un artéfact audible.

La première hypothèse de cette thèse est que le fait d'améliorer la robustesse de la classification en ayant une meilleure détection du début des segments voisés dans les périodes

de transition améliorerait la qualité du signal reconstruit en cas de pertes de trames. La seconde hypothèse proposée est de transformer les transitoires partielles en trames complètement non-voisées afin d'améliorer d'avantage la performance du camouflage lorsque la trame suivant la transitoire est perdue.

L'opérateur de Teager a été choisi pour améliorer la détection des transitions non-voisées à voisées puisque sa variation est fonction de l'amplitude et de la fréquence du signal. L'opérateur de Teager est un opérateur non-linéaire, dont le résultat varie en fonction de l'énergie nécessaire à la production d'un signal harmonique simple. La comparaison du résultat de Teager calculé sur chaque sous-bande du signal permet de détecter avec une précision de 1/8 de trame le début de la transition non-voisée à voisée. Lorsque le résultat de Teager passe au dessus d'un seuil pour au moins une sous-bande, la trame est possiblement une trame de transition (onset). Le seuil utilisé est fonction de l'énergie à long terme du signal. Pour permettre au camouflage d'utiliser une trame en répétant sa partie voisée, il faut que cette dernière contienne au moins une période de pitch complète. Cette information est validée par deux conditions : premièrement la corrélation entre la fin de la trame présente et le début de la trame suivante doit être élevée et deuxièmement la distance entre le début de la transitoire (détecté au moyen de l'opérateur de Teager) et la fin de la trame doit être plus grande que la période de pitch.

Une analyse statistique sur les résultats de classification obtenus avec l'opérateur de Teager en comparaison avec la classification du VMR-WB a été faite. L'opérateur de Teager permet de réduire de plus de la moitié les fausses détections autour de la transition non-voisé à voisé. De plus, la précision de l'opérateur de Teager permet d'identifier le début de la transition avec une précision de 1/8 de trame dans 93,4% des cas.

Des tests d'écoutes ont été faits pour valider les résultats. Les tests ont été répartis en deux volets. Dans le premier test, la trame transitoire est systématiquement perdue. Dans le deuxième test, la trame suivant la trame transitoire est perdue. Dans le premier cas, les résultats obtenus sont similaire entre la classification du codeur VMR-WB et celle basée sur l'opérateur de Teager. Lorsque les trames perdues sont celles qui suivent la transitoire, la différence entre les deux approches est plus significative et en faveur de la nouvelle classification.

La deuxième hypothèse proposée est d'appliquer une modification du signal afin d'éliminer les transitoires partielles. Cette modification permet l'obtention d'une trame totalement non-voisée et ainsi l'utilisation d'un camouflage non-voisé lorsque la trame suivant la trame modifiée est perdue. La modification se fait sur le résidu du signal obtenu lorsque celui-ci

est traité par un filtre de prédiction linéaire. Le résidu du signal correspondant à la partie transitoire est remplacé par une partie du résidu passé. De même, les différents filtres prédictifs correspondants à la partie transitoire du signal sont remplacés par les filtres prédictifs passés correspondants à la partie non-voisée du signal.

Des tests d'écoute de type A-B ont été choisis pour démontrer la transparence de la modification et son utilité. Les tests évaluant la qualité perceptuelle entre le signal modifié codé et le signal de référence codé lorsqu'il n'y a pas de pertes de trames montrent que la modification est transparente dans une grande majorité des cas. Aussi, lorsque l'on efface systématiquement la trame suivant la trame modifiée (perte de la trame transitoire), les auditeurs préfèrent le signal modifié. Afin d'évaluer la performance de la méthode pour un canal simulé, 8% des trames ont été aléatoirement retirées du signal. En présence de ce type de dégradation, les auditeurs n'ont pas de préférence entre le signal modifié et le signal de référence codés. Ces résultats s'expliquent par le fait que la présence de la transitoire partielle n'entraine pas toujours une dégradation du signal si la trame suivante est perdue. Bien que la méthode soit prometteuse (ce qui est indiqué par la tendance positive pour l'effacement systématique de trames suivant les modifications), elle doit être poussée plus d'avant afin d'être en mesure d'augmenter les performances des systèmes réels.

Globalement, une meilleure classification des transitoires et une élimination des transitoires partielles précédant une transitoire permet une amélioration de la qualité du signal en cas de pertes de trames. Le problème adressé survient relativement très rarement, mais il peut causer une dégradation significative. On a ainsi pu observer une amélioration dans les tests subjectifs lors de l'effacement systématique, malgré le fait que les trames affectées ne constituaient qu'une très faible partie du signal testé, soit environ 5% des trames actives, ou 2 à 3% des trames totales du signal. De plus, les problèmes évoqués par une mauvaise classification ou par la présence de transitoires partielles ne causeraient pas nécessairement d'artéfacts audibles lors d'une perte de trame.

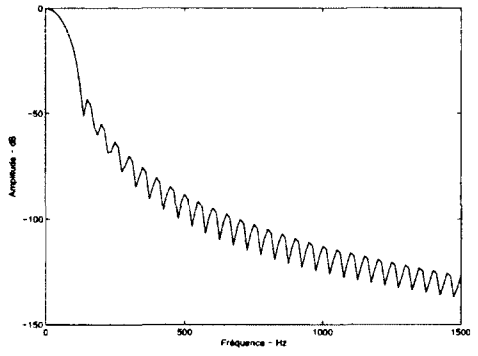
Ces travaux ont démontré une nouvelle utilité de l'opérateur de Teager. L'opérateur améliore la performance et la précision de la détection de la transition non-voisée à voisée à l'intérieur même d'une trame. La modification subséquente du signal, quoique pas totalement transparente, a pour conséquence une amélioration notable dans le cas ou la trame suivant la modification est effacée. Les modifications s'appliquent parfois à une trame entière et dans la grande majorité des cas, la modification est complètement transparente bien que le signal modifié soit totalement différent à l'oeil. Il serait intéressant de pousser d'avantage les recherches du côté de la modification de signal afin d'augmenter la trans-

parence du signal modifié et obtenir un avantage systématique en cas de trames perdues lorsque la modification est appliquée.

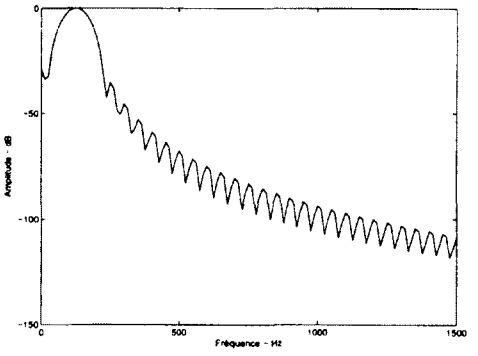
ANNEXE A

Filtres passe-bas et passe-bande

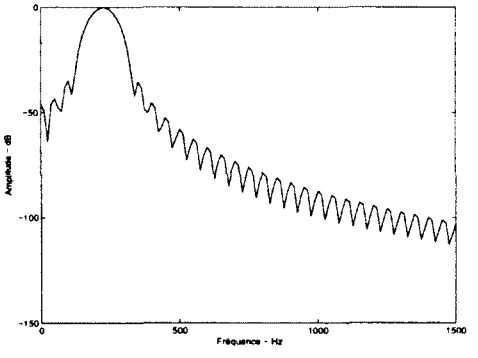
Aperçu des réponses en fréquence de quelques filtres passe-bas et passe-bande, utilisés pour la séparation en fréquence du signal analysé par l'opérateur de teager.



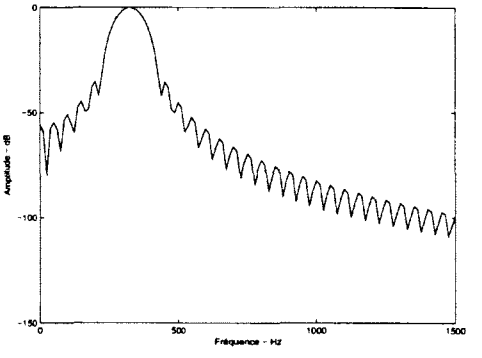
(a) Réponse en fréquence du filtre passe-bas 0-50 Hz



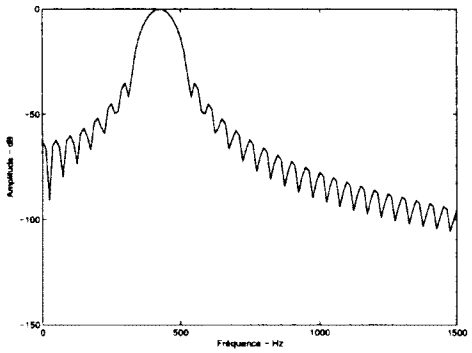
(b) Réponse en fréquence du filtre passe-bande 100-150 Hz



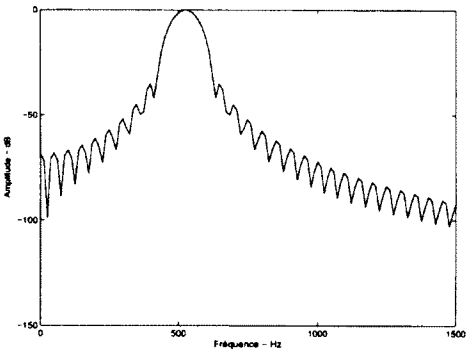
(c) Réponse en fréquence du filtre passe-bande 200-250 Hz



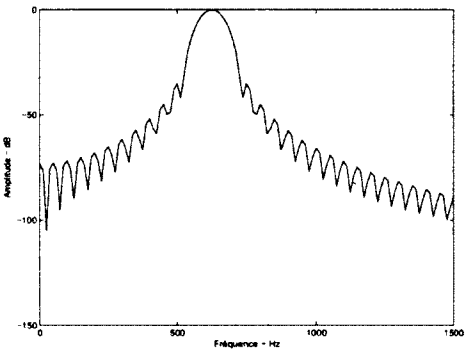
(d) Réponse en fréquence du filtre passe-bande 300-350 Hz



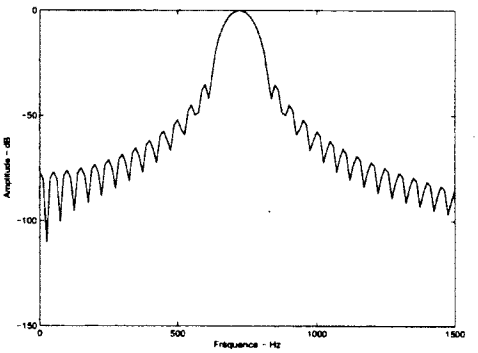
(e) Réponse en fréquence du filtre passe-bande 400-450 Hz



(f) Réponse en fréquence du filtre passe-bande 500-550 Hz



(g) Réponse en fréquence du filtre passe-bande 600-650 Hz



(h) Réponse en fréquence du filtre passe-bande 700-750 Hz

LISTE DES RÉFÉRENCES

Abdallah, S. et Plumbley, M. (mars 2003) Unsupervised onset detection : A probabilistic approach using ica and a hidden markov classifier. *Proceedings of the Cambridge Music Processing Colloquium.*

Ahmadi, S. (mars 2005) *Source-Controlled Variable-Rate multimode wideband speech codec (VMR-WB), Version 1.0.* 3GPP2.

Ahmadi, S. et Spanias, A. S. (mai 1999) Cepstrum-based pitch detection using a new statistical v/uv classification algorithm. *IEEE transactions on speech and audio processing*, p. 333–338.

Andersen, S. V., Kleijn, W. B., Hangen, R., Linden, J., Murthi, M. N. et Skoglund, J. (octobre 2002) Ilbc - a linear predictive coder with robustness to packet losses. *IEEE Workshop Proceedings Speech Coding*, p. 23–25.

Avendano, C. et Goodwin, M. (octobre 2004) Enhancement of audio signals based on modultaion spectrum processing. *Proceeding of 117th AES Convention.*

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M. et Sandler, M. (septembre 2005) A tutorial on onset detection in music signals. *IEEE transactions on speech and audio processing*, volume 09, n° 5, p. 1035–1047.

Bhute, V. et Shrawankar, U. (juin 2008) Error concealment schemes for speech packet transmission over ip network. *15th International Conference on Systems, Signals and Image Processing*, p. 185–188.

BS.1534, I.-R. (2001) *Method for the subjective assessment of intermediate quality level of coding systems.* ITU-R.

Collins, N. (septembre 2005) Using a pitch detector for onset detection. *Proceedings of 6th International conference on music information retrieval*, p. 100–106.

de Cheveigné, A. et Kawahara, H. (avril 2002) Yin, a fundamental frequency estimatior for speech and music. *J. Acoustic. Soc. Am.*, volume 111, n° 4, p. 1917–1930.

Eskler, V. et Jelinek, M. (avril 2008) Transition mode coding for source controlled celp coders. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, p. 4001–4004.

G.722.2, I.-T. (janvier 2002) *Codage vocal en large bande à 16 kbit/s environ par codage adaptatif multidébit à large bande (AMR-WB).* ITU-T.

G.729, I.-T. (janvier 2007) *Coding of speech at 8kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP).* ITU-T.

Gainza, M., Coyle, E. et Lawlor, B. (octobre 2005) Onset detection using comb filter. *IEEE Workshop on applications of signal processing to audio and acoustics*, p. 263–266.

Gao, Y., Benyassine, A., Thyssen, J., Huan-yu, S. et Shlomot, E. (mai 2001a) ex-celp : a speech coding paradigm. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, p. 689–692.

Gao, Y., Shlomot, E., Benyassine, A., Thyssen, J., Huan-yu, S. et Murgia, C. (mai 2001b) he smv algorithm selected by tia and 3gpp2 for cdma applications. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, p. 709–712.

Gerhard, D. (novembre 2003) Pitch extraction and fundamental frequency : history and current techniques. *Technical report TR-CS 2003-06.*

Gournay, P., Rousseau, F. et Lefebvre, R. (avril 2003) Improved packet loss recovery using late frames for prediction-based speech coders. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, p. 108–103.

Greenwood, M. et Kinghorn, A. (1999) Suving : automatic silence/unvoiced/voiced classification of speech.

Hui, L., qian Dai, B. et Wei, L. (mai 2006) A pitch detection algorithm based on amdf and acf. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, p. 377–380.

Husain, A. et Cuperman, V. (mai 1995) Reconstruction of missing packets for celp-based speech coders. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, p. 245–248.

ITU-T G.711, A. . (septembre 1999) *Pulse code modulation (PCM) of voice frequencies, Appendix I : A high quality low-complexity algorithm for packet loss concealment with G.711.* ITU-T.

Jelinek, M., Eksler, V., Lemyre, C. et Lefebvre, R. (novembre 2007) Classification-based techniques for improving the robustness of celp coders. *Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*, p. 1480–1484.

Jelinek, M. et Salami, R. (mai 2007) Wideband speech coding advances in vmr-wb standard. *IEEE transactions on audio, speech, and language processing*, volume 15, n° 4, p. 1167–1179.

Jelinek, M., Vaillancourt, T., Ertan, A., Stachurski, J., Ramo, A., Laaksonen, L., Gibbs, J. et Bruhn, S. (avril 2008) Itu-t g.ev-vbr baseline codec. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, p. 4749–4752.

Jensen, J., Jensen, S. H. et Hansen, E. (juin 1999) A perturbation-based pre-processing algorithm for celp-coders. *1999 IEEE Workshop on Speech Coding Proceedings*, p. 153–155.

Jovicic, S. et Randjolovic, P. Z. (novembre 1987) Crosscorrelation pitch extraction method. *Electronics letters*, volume 23, n° 24, p. 1317–1318.

Kaiser, J. F. (avril 1990) On a simple algorithm to calculate the 'energy' of a signal. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, p. 381–384.

Keum, J.-S. et Lee, H.-S. (décembre 2006) Speech/music discrimination using spectral peak feature for speaker indexing. *International Symposium on Intelligent Signal Processing and Communications*, p. 323–326.

Klapuri, A. (mars 1999) Sound onset detection by applying psychoacoustic knowledge. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, , n° 6, p. 3089–3092.

Kleijn, W. B., Ramachandran, R. P. et Kroon, P. (mars 1992) Generalized analysis-by-synthesis coding and its application to pitch prediction. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, p. 337–340.

Lemyre, C., Jelinek, M. et Lefebvre, R. (avril 2008) New approach to voiced onset detection in speech signal and its application for frame error concealment. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, p. 4757 – 4760.

Levine, S. N. (Décembre 1998) *Audio representations for data compression and compressed domain processing.* Thèse de doctorat, Standford University.

Lu, L., Zhang, H.-J. et Jiang, H. (octobre 2002) Content analysis for audio classification and segmentation. *IEEE transactions on speech and audio processing*, volume 10, n° 7, p. 504–516.

Maragos, P., Kaiser, J. F. et Quatieri, T. F. (mars 1992) On separating amplitude from frequency modulations using energy operators. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, p. 1–4.

Maragos, P., Kaiser, J. F. et Quatieri, T. F. (octobre 1993) Energy separation in signal modulations with application to speech analysis. *IEEE transactions on signal processing*, volume 41, n° 10, p. 3024–3051.

Markel, J. D. (décembre 1972) Sift algorithm for fundamental frequency estimation. *IEEE transactions on audio and electroacoustics*, volume AU-20, n° 5, p. 367–377.

Martin, P. (mai 1982) Comparison of pitch detection by cepstrum and spectral comb analysis. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, p. 180–183.

Nahumi, D. et Kleijn, W. B. (septembre 1995) An improved 8 kb/s rcelp coder. *1995 IEEE Workshop on Speech Coding for Telecommunications*, p. 39–40.

Noll, A. M. (février 1967) Cepstrum pitch determination. *The journal of the acoustical society of america*, volume 41, n° 2, p. 293–309.

Paksoy, E., Srinivasan, K. et Gersho, A. (avril 1993) Variable rate speech coding with phonetic segmentation. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, p. 155–158.

Panagiotakis, C. et Tziritas, G. (février 2005) A speech/music discrimination based on rms and zero-crossing. *IEEE transactions on multimedia*, volume 7, n° 1, p. 155–166.

Perkins, C., Hodson, O. et Hardman, V. (septembre-octobre 1998) A survey of packet loss recovery techniques for steaming audio. *IEEE Network*, volume 12, n° 5, p. 40–48.

Rabiner, L. R. (février 1977) On the use of autocorrelation analysis for pitch detection. *IEEE transactions on acoustics, speech, and signal processing*, volume ASSP-25, n° 1, p. 24–33.

Ross, M. J., Shaffer, H. L., Cohen, A., Freudberg, R. et Manley, H. J. (octobre 1974) Average magnitude difference function pitch extractor. *IEEE transactions on acoustics, speech, and signal processing*, volume ASSP-22, n° 5, p. 353–362.

Samad, S. A., Hussain, A. et Fah, L. K. (2000) Pitch detection of speech signals using the cross-correlation technique. *Proceedings TENCON 2000*, volume 1, p. 283–286.

Saunders, J. (mai 1996) Real-time discrimination of broadcast speech/music. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, p. 993–996.

Scheirer, E. D. (janvier 1998) Tempo and beat analysis of acoustic musical signals. *J. Acoust. Soc. Amer.*, volume 103, n° 1, p. 588–601.

Schroeder, M. R. (avril 1968) Period histogram and product spectrum : New methods for fundamental-frequency measurement. *The Journal of the Acoustical Society of America*, volume 43, n° 4, p. 829–834.

Schroeder, M. R. et Atal, B. S. (avril 1985) Code-excited linear prediction (celp) : high quality speech at very low bit rates. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 10, p. 937–940.

Tammi, M. et Jelinek, M. (septembre 2002a) Signal modification for coding purely voiced sections in a wideband acelp speech coder. *Eusicop*.

Tammi, M. et Jelinek, M. (octobre 2002b) Signal modification for voiced wideband speech coding and its application for is-95 system. *IEEE Workshop Proceedings Speech Coding*, p. 35–37.

Tammi, M., Jelinek, M. et Ruoppila, V. T. (septembre 2005) Signal modification method for variable bit rate wide-band speech coding. *IEEE transactions on speech and audio processing*, volume 13, n° 5, p. 799 – 810.

Tan, H. L., Zhu, Y., Chaisorn, L. et Rahandja, S. (juin 2010) Audio onset detection using energy-based and pitch-based processing. *Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS)*, p. 3689–3692.

Tosun, L. et Kabal, P. (septembre 2005) Dynamically adding redundancy for improved error concealment in packet voice coding. *Proc. European Signal Processing Conference.*

Vafin, R., Heusdens, R., van de Par, S. et Kleijn, W. B. (octobre 2001) Improved modeling of audio signals by modifying transient locations. *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, p. 143–146.

Wah, B. W., Su, X. et Lin, D. (décembre 2000) A survey of error-concealment schemes for real-time audio and video transmissions over the internet. *Proceedings of International Symposium on Multimedia Software Engineering*, p. 17–24.

Wang, S. et Gersho, A. (mai 1989) Phonetically-based vector excitation coding of speech at 3.6 kbps. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, p. 49–52.

Zhou, R., Mattavelli, M. et Zoia, G. (novembre 2008) Music onset detection based on resonator time frequency image. *IEEE transactions on audio, speech and language processing*, volume 16, n° 8, p. 1685–1695.