Méthodes robustes de détection de parole pour la reconnaissance vocale en environnement bruité

Arnaud Martin

Résumé

Les systèmes de reconnaissance vocale se composent généralement d'un module de reconnaissance et d'un module de détection de parole. Le module de détection indique les périodes de parole au module de reconnaissance. Les non-détections de parole sont donc des erreurs graves, ainsi que les détections de bruit qui peuvent être reconnues comme des mots du vocabulaire. Dans le cadre de la parole continue une plus grande précision des frontières des phrases détectées est de plus exigée afin d'éviter les insertions ou omissions de mots qui peuvent perturber la reconnaissance de toute la phrase de part l'application d'un modèle de langage. Dans la première partie, nous proposons une méthode d'évaluation rigoureuse pour examiner en détail les influences d'un module de détection sur un système de reconnaissance. L'évaluation est effectuée à la fois au niveau des résultats de détection et du système de reconnaissance. Nous montrons que les performances du module de détection sont insuffisantes d'une part pour des communications bruitées, et d'autre part dans le cas de la détection de parole continue. Il est aussi montré que les réglages du module de détection diffèrent selon le réseau d'appel, le niveau de bruit et le vocabulaire. Afin de pallier à ces insuffisances nous évaluons dans la seconde partie une méthode de réduction de bruit qui s'avère performante quand le bruit est stationnaire et d'un niveau élevé. Nous cherchons ensuite à affiner l'estimation de la distribution de l'énergie à l'aide des statistiques d'ordre supérieur. Enfin, pour réduire les détections de bruit de courte durée et améliorer la détection de parole continue, nous étudions l'intégration dans le module de détection d'un paramètre de voisement, puis des coefficients cepstraux fusionnés par une analyse discriminante. Ces approches permettent une amélioration significative. L'approche fondée sur le paramètre de voisement donne les meilleurs résultats et réduit de plus la sensibilité de réglage du module de détection.

Abstract

Speech recognition systems generally contain a speech detection module that detects speech periods to be recognized by the speech recognition module. The non-detections of speech signals are major errors as well as noise detections that may be recognized as a vocabulary word. Moreover continuous speech recognition requires more precise boundaries of detected sentences to avoid insertions or omissions of words, that may disturb the recognition of the whole sentence due to the application of the language model. In the first part, we propose an evaluation framework to study in depth the influence of a detection module on a speech recognition system. The evaluation considers both speech detection and recognition errors on several large databases. We show the need to reduce errors in noisy environment and to improve continuous speech detection. It is also shown that the tuning of the detection module depends on the noise level, the transmission system and the vocabulary. In order to solve these problems we first evaluate in the second part a noise reduction module to improve results in noisy conditions. We propose a better estimation of the energy distribution using high order statistics. We integrate into the speech detection module a voicing parameter and also Mel Frequency Cepstrum Coefficients (MFCCs) using a linear discriminant analysis. The noise reduction system is significantly efficient with high level, stationary noise. The integration of the voicing parameter or the MFCCs is efficient to reduce non-stationary noise detections and to improve the continuous speech detection. The best approach is provided by using the voicing parameter, and it reduces the sensitivity of the adjustment of the detection module.

Table des matières

Table des figures

Liste des tableaux

Abréviations utilisées

ABP : Automate Bruit/Parole.

ACP : Analyse en Composantes Principales.

AFD : Analyse Factorielle Discriminante.

CART : Classification et arbre de décision (de l'anglais *Classification And Regression Trees*).

DAV : Détection d'Activité Vocale.

DBP : Détection Bruit/Parole.

DTW : Alignement temporel dynamique (de l'anglais *Dynamic Time Warping*).

FFT : Transformé de Fourier rapide (de l'anglais *Fast Fourier Transform*).

GSM : *Global System for Mobile.*

HMM : Modèle de Markov caché (de l'anglais *Hidden Markov Model*).

LPC : Analyse de prédiction linéaire (de l'anglais *Linear Prediction Coding*).

MFCC : Coefficients cepstraux issus de la représentation en bancs de filtres avec une échelle Mel (de l'anglais *Mel Frequency Cepstrum Coefficients*).

PARCOR : Coefficients obtenus lors de l'analyse par prédiction linéaire (de l'anglais *PARtial CORrelation*).

RSB : Rapport Signal à Bruit.

RTC : Réseau Téléphonique Commuté.

SVI : Serveur Vocal Interactif.

Notations utilisées

Bases de données

base AGORA : Base de données d'expérimentation, de parole continue pour la mise en œuvre d'une application de dialogue homme-machine, enregistrée sur le réseau RTC.

base GSM_A : Base de données de laboratoire, de mots isolés, enregistrée sur le réseau GSM, utilisée pour l'apprentissage.

base GSM_A M18 : Partie bruitée de la base GSM_A limitée aux fichiers ayant un RSB inférieur à $18dB$.

base GSM_A P18 : Partie calme de la base GSM_A limitée aux fichiers ayant un RSB supérieur à $18dB$.

base GSM_T : Base de données de laboratoire, de mots isolés, enregistrée sur le réseau GSM, utilisée pour les tests.

base GSM_T M18 : Partie bruitée de la base GSM_T limitée aux fichiers ayant un RSB inférieur à $18dB$.

base GSM_T P18 : Partie calme de la base GSM_T limitée aux fichiers ayant un RSB supérieur à $18dB$.

base RTC_A : Base de données d'exploitation d'un serveur interactif en activité, enregistrée sur le réseau RTC, utilisée pour l'apprentissage.

base RTC_A M20 : Partie bruitée de la base RTC_A limitée aux fichiers ayant un RSB inférieur à $20dB$.

base RTC_A P20 : Partie calme de la base RTC_A limitée aux fichiers ayant un RSB supérieur à $20dB$.

base RTC_T : Base de données de laboratoire, de mots isolés, enregistrée sur le réseau RTC, utilisée pour les tests.

base RTC_T_L : Partie de la base RTC_T limitée aux fichiers comportant des mots lus par le locuteur.

base RTC_T_R : Partie de la base RTC_T limitée aux fichiers comportant des mots répétés par le locuteur.

Différents critères du module de détection de parole ABP

LCT : Le critère de détection est fondé sur le rapport signal à bruit (énergie à long terme et à court terme).

SB : Le critère de détection est fondé sur les statistiques du bruit.

SBP : Le critère de détection est fondé sur les statistiques du bruit et de la parole.

SB+M3 : Le critère de détection fondé sur les statistiques d'ordre 3, et il est associé au critère SB.

SB+F_0 : Le critère de détection est fondé sur la fréquence fondamentale, et il est associé au critère SB.

SB+VP : Le critère de détection est fondé sur l'analyse factorielle discriminante sur les MFCC et nécessite le calcul des valeurs propres, et il est associé au critère SB.

Notations usuelles

λ	Facteur d'oubli.
$P(.)$	Probabilité.
$P(./.)$	Probabilité conditionnelle.
$d(.,.)$	Distance.

Sur les statistiques d'ordre supérieur

$E[.]$	Espérance mathématique.
$Var(.)$	Variance mathématique d'un estimateur.
μ_r	Moment d'ordre r.
m_r	Moment, non centré, normalisé par la variance, d'ordre r.
κ_r	Cumulent d'ordre r.
χ	Skewness.
γ	Kurtosis.
σ	Ecart-type.
$\hat{\mu}_r$	Estimation du moment d'ordre r.
$\hat{\kappa}_r$	Estimation du cumulent d'ordre r.

Sur la fusion de données

$\mathbf{x}$	Une trame, composée de p coefficients.
$\mathbf{X}$	Un tableau de données composé de n lignes et p colonnes.
x_{ij}	Une donnée du tableau de données $i = 1,\ldots,n$ et $j = 1,\ldots,p$.
$\bar{x}_j$	Moyenne de x_{ij} sur les lignes.
$\mathbf{I}$	Une classe.
$\bar{x}_{kj}$	Moyenne de x_{ij} sur les lignes de la classe I_k.

Introduction Générale

Le langage parlé est le mode de communication le plus naturel chez l'homme. En effet, la parole permet un échange rapide et efficace d'informations entre deux personnes, plus rapide que l'écriture, la gestuelle, *etc.* Ce mode de communication permet en outre une activité visuelle ou gestuelle simultanée.

Ainsi, avec l'apparition de l'automatisation de la communication homme-machine, le rêve d'une interaction vocale entre l'homme et la machine est apparu très tôt. Pour réaliser cette communication homme-machine, les recherches sur la reconnaissance automatique de la parole ont débuté dès les années 50. Et ce rêve devient de plus en plus réalité. Depuis ses débuts, les problèmes de la reconnaissance vocale n'ont cessé d'évoluer, avec notamment l'expansion de la téléphonie.

En effet, la grande complexité de la parole est modélisée de plus en plus précisément. Les personnes parlent différemment selon leur âge, leur sexe, leur accent, *etc.*, mais également selon le contexte dans lequel elles se trouvent, si elles sont anxieuses, stressées, si elles parlent dans un milieu ambiant bruité, *etc.* De la prise de son va également dépendre la qualité du signal de parole, par exemple les contraintes de prise de son pour la transmission téléphonique ne nous placent pas dans les conditions optimales de qualité sonore. Il y a ainsi un grand nombre de paramètres dont la reconnaissance doit être indépendante. Il est vrai que selon les applications recherchées, les problèmes peuvent être limités. Par exemple, les logiciels de dictée vocale utilisent une prise de son de bonne qualité, et la reconnaissance peut être adaptée pour se limiter à reconnaître la voix d'une seule personne.

Dans le cadre de la reconnaissance vocale pour des applications téléphoniques, les problèmes actuels restent la reconnaissance d'un grand nombre de mots isolés (par exemple pour un annuaire vocal), et la reconnaissance de la parole continue, c'est-à-dire la reconnaissance de phrases du langage courant. Avec l'importance croissante des téléphones portables, la qualité sonore est diminuée, d'une part par la transmission de la parole avec ces réseaux cellulaires, et d'autre part par l'environnement souvent bruité des appels. Ces téléphones sont utilisés dans n'importe quel lieu, et avec des réceptions médiocres.

Afin d'éviter de chercher à reconnaître de la parole sur des périodes de silence, il est important de détecter les périodes de parole et ainsi à la fois améliorer les performances et réduire le coût du système de reconnaissance vocale. Le système de reconnaissance est donc constitué d'un module de détection de parole et d'un module de reconnaissance de la parole. Notre étude porte sur les méthodes de détection la parole dans le bruit dans le cadre de la reconnaissance vocale, pour des applications téléphoniques. Ainsi le signal

reçu, à reconnaître, a été perturbé par l'environnement de l'appel, par la transmission, mais également par la détection d'activité vocale présente dans le système de transmission selon le réseau utilisé. Une détection robuste à tous ces paramètres perturbateurs permet une meilleure reconnaissance de la parole. Elle est d'autant plus indispensable si le milieu ambiant est bruité. Dans le cas de la reconnaissance de parole continue, il ne s'agit plus de détecter des mots isolés, mais des phrases entières. Les temps d'hésitation de la personne sont un paramètre perturbant la détection, car une phrase risque alors d'être fragmentée en deux détections. Le module de reconnaissance de la parole continue nécessite de plus une précision plus grande en début et fin de détections, que pour la reconnaissance de mots isolés. En effet le module de reconnaissance de parole continue peut reconnaître des phrases comportant un nombre indéfini de mots, une détection incluant du silence ou du bruit au début ou à la fin de la phrase peut donc entraîner des insertions de mots, principalement des mots courts comme des articles. Nous concentrons cette étude plus particulièrement sur l'amélioration des performances de la détection de parole bruitée, et de la parole continue.

Ce document se compose de deux parties principales. Dans la première partie nous présentons d'abord le problème de la détection de parole. Dans le Chapitre 1 "*Contextes de la détection de parole*", nous détaillons les approches de la détection de parole selon différents contextes afin de déterminer les différences et similitudes des méthodes de détection. Plus particulièrement, nous étudions la détection pour la reconnaissance vocale dans le Chapitre 2 "*Détection de parole pour la reconnaissance vocale*". Dans ce chapitre nous présentons le module de détection fondé sur un automate Bruit/Parole à cinq états et les trois critères qui contrôlent les transitions d'un état à l'autre. Nous étudions également les influences du module de détection sur le système de reconnaissance, ce qui nous permet de définir notre méthode d'évaluation qui est comparée à d'autres méthodes. Dans le Chapitre 3 "*Analyse des sources d'erreurs du module de détection*", nous étudions en détail les sources d'erreurs du module de détection. Ainsi, les résultats sont dégradés par rapport à une segmentation idéale dans le cas d'environnements bruités par des bruits de courte durée ou avec un niveau de bruit élevé, mais également dans le cas de la détection de parole continue, et selon le vocabulaire employé. Les principales sources d'erreurs permettent de définir les objectifs de la thèse au Chapitre 4 "*Voies envisagées pour l'amélioration du module de détection*". Dans ce chapitre, une étude comparative des trois critères du module de détection fondé sur l'automate Bruit/Parole détermine le meilleur critère qui sera le critère de base que nous cherchons à améliorer. Les objectifs définissent où doivent être apportées les améliorations; une étude des systèmes de détection existants permet de dégager les caractéristiques et méthodes envisageables pour ces améliorations. Nous définissons ainsi les axes de recherche pour obtenir le meilleur critère.

Dans la seconde partie, nous approfondissons les différents axes de recherche précédemment définis pour l'amélioration du module de détection. Ainsi dans le Chapitre 5 "*Méthode de débruitage*", les performances des trois critères du module de détection sont évaluées avec une méthode de débruitage. Le but est de réduire le niveau de bruit du signal pour permettre une meilleure détection et une meilleure reconnaissance de la parole. Cette approche s'avère performante pour les bruits stationnaires, mais ne permet pas de réduire

les bruits impulsifs ou de courte durée. Nous cherchons donc à réduire les détections de ces derniers bruits. Ceci se traduit par l'intégration d'une nouvelle décision dans l'automate. Une étude au Chapitre 6 "*Intégration d'une nouvelle condition dans l'automate*" permet de déterminer la meilleure intégration de cette nouvelle décision. Le but de cette décision est soit d'affiner l'estimation de la distribution de l'énergie, soit d'apporter plus d'informations à l'aide des caractéristiques déterminées dans la première partie de la thèse. Ainsi le Chapitre 7 "*Utilisation des statistiques d'ordre supérieur*" présente une approche à l'aide des statistiques des moments d'ordre 3 pour estimer plus précisément la distribution de l'énergie. Deux approches ont ensuite pour objectif d'apporter plus d'informations à l'aide de caractéristiques différentes de l'énergie. Le Chapitre 8 "*Utilisation d'un paramètre de voisement*" propose l'intégration d'un paramètre de voisement dans la prise décision. Afin d'intégrer un grand nombre de caractéristiques, une étude des approches de fusion de données est proposée au Chapitre 9 "*Utilisation de la fusion de données*". Nous intégrons l'analyse factorielle discriminante avec les coefficients cepstraux et les coefficients de la sortie du banc de filtres. Une combinaison des coefficients cepstraux avec le paramètre de voisement et le moment d'ordre 3 est également proposée.

Pour conclure, nous confirmons que les objectifs proposés sont atteints par l'approche présentant les meilleures performances qui est celle employant le paramètre de voisement. Puis nous présentons quelques perspectives de recherche pour la détection de parole.

Première partie

La détection de parole

Cette partie "*La détection de parole*" a pour but de poser le problème de la détection de parole pour la reconnaissance vocale et de définir les objectifs de la thèse et les moyens de les atteindre.

Dans le Chapitre 1 "*Contextes de la détection de parole*", nous décrivons les principaux contextes de la détection de parole afin de situer la détection de parole pour la reconnaissance vocale. Nous distinguons ainsi la détection d'activité vocale, la détection en segments voisés/non-voisés/silence et la détection Bruit/Parole. Ce chapitre montre les différences et similarités de la détection de parole selon les contextes.

Dans le Chapitre 2 "*Détection de parole pour la reconnaissance vocale*" nous présentons la détection de parole pour la reconnaissance vocale employée dans cette étude et le principe d'évaluation. Le module de détection de parole est fondé sur un automate, dont les transitions peuvent être contrôlées selon trois critères. Nous étudions ensuite l'influence importante du module de détection sur le système de reconnaissance. Cette étude permet de définir le principe d'évaluation du module de détection qui se compose de deux parties, d'une part l'évaluation du module de détection isolé, et d'autre part l'évaluation du module de détection dans le système de reconnaissance. Le principe d'évaluation s'avère rigoureux en comparaison des méthodes d'évaluation employées ces dernières années dans des systèmes de détection. Ce chapitre montre d'une part que le système de reconnaissance peut être rendu plus performant en diminuant les erreurs dues au module de détection, et d'autre part que l'évaluation du module de détection doit se faire par une évaluation des résultats de détection mais également par l'évaluation des résultats du système de reconnaissance.

Pour dégager les insuffisances actuelles du module de détection, nous nous proposons ensuite d'étudier le détail des erreurs de détection et de reconnaissance pour un critère particulier du module de détection. Dans le Chapitre 3 "*Analyse des sources d'erreurs du module de détection*", l'étude des erreurs débute par l'examen de l'influence du seuil de détection sur une base enregistrée en environnement RTC, puis de l'influence des environnements et du rapport signal à bruit sur des bases enregistrées en environnement RTC et GSM. Une étude de l'influence du vocabulaire est ensuite réalisée.

Nous étudions le cas particulier de la base GSM_A (présentée au paragraphe 3.2) pour un rapport signal à bruit inférieur à $18dB$. En effet, sur cette partie de la base les résultats sont particulièrement dégradés à cause d'un grand nombre de bruits impulsifs, moins important en environnement RTC. Pour préciser l'influence du bruit, nous étudions les performances du module de détection sur la partie de la base GSM_A dont les enregistrements ont un rapport signal à bruit supérieur à $18dB$, avec un ajout à différents RSB de deux types de bruits stationnaires.

Après avoir considéré le cas de la reconnaissance de mots isolés, nous détaillons le cas de la parole continue. Ce cas doit être considéré différemment. La détection recherchée n'étant plus des mots, mais des phrases, d'une part le type des erreurs du module de reconnaissance est différent, d'autre part l'influence des erreurs du module de détection sur le système de reconnaissance n'est pas le même que dans le cas des mots isolés. En effet une phrase dont le début n'est pas détecté peut entraîner la suppression d'un mot de la phrase.

Ce chapitre permet de déterminer les principales sources d'erreurs du module de détection et de présenter la méthode d'évaluation du module de détection au cours de l'étude.

Le chapitre 4 "*Voies envisagées pour l'amélioration du module de détection*" a pour but de définir les axes de recherche de cette thèse. Dans un premier temps le rappel des principales sources d'erreurs du module de détection déterminées dans le chapitre précédent permet de définir les objectifs de la thèse. Nous évaluons le critère le plus performant à l'aide d'une étude comparative des trois critères du module de détection, qui sera le critère de base que nous cherchons à améliorer. Nous présentons ensuite différents systèmes de détection pour dégager les caractéristiques et approches envisageables pour discriminer davantage le bruit et la parole. Ces caractéristiques sont plus particulièrement étudiées sur les segments de bruits et de parole des bases de données sur le réseau RTC et GSM. Nous pouvons ainsi définir les axes d'étude pour l'amélioration du module de détection.

Nous intégrons ensuite ces approches les plus adaptées à notre problème dans le module de détection à partir du critère le plus performant dans la partie II "*Amélioration du module de détection*".

Chapitre 1

Différents contextes de la détection de parole

1.1 Introduction

Ce chapitre a pour objectif de présenter les différents contextes de détection de parole afin de comprendre les différences et similarités avec la détection de parole pour la reconnaissance.

La parole est un signal complexe qui est très difficile à décrire de part sa variabilité temporelle et fréquentielle et de part sa dépendance au locuteur et à son état. La parole peut cependant être classée selon différents critères, présentés en Annexe A. Les coefficients qui parviennent aux systèmes de détection de parole sont issus de l'analyse du signal. Différents coefficients employés pour la détection de parole comme les coefficients cepstraux aussi appelés MFCC (*Mel Frequency Cepstrum Coefficients*) ou les coefficients LPC (*Linear Predictor Coding*).

Les méthodes de détection de parole dans le contexte de la reconnaissance vocale ont déjà fait l'objet de nombreux travaux. Il est cependant important de noter que le problème de détection de parole se retrouve dans d'autres domaines du traitement de parole que la reconnaissance vocale. Ces détections peuvent segmenter la parole de différentes façons selon le cadre de l'étude (phrases, mots, ou unités acoustiques), ainsi la terminologie employée diffère selon la détection recherchée.

Nous parlons de *systèmes* de détection de parole lorsque le contexte n'est pas différencié. La terminologie de *module* de détection est utilisée dans le cadre de la reconnaissance vocale.

Nous regroupons les différents systèmes de détection selon la précision attendue : la détection d'activité vocale qui doit être le plus rapide possible, la détection des segments voisés/non-voisés/silence qui recherche une détection précise au niveau des sons constituant la parole, et la détection Bruit/Parole, appliquée en reconnaissance de la parole bruitée.

Nous présentons donc dans le paragraphe 1.2 la détection d'activité vocale, dans le paragraphe 1.3 la détection en segments voisés/non-voisés/silence. Puis, le paragraphe 1.4

présente la détection Bruit/Parole.

1.2 Détection d'activité vocale

La détection d'activité vocale (DAV) a pour but de détecter avec un délai minimum les périodes de parole. Il est généralement recherché une détection sans omission, ni coupure de la parole, sans nécessairement une précision importante au niveau des frontières de la détection. La DAV est utilisée pour le codage de la parole souvent lié à la transmission mais aussi pour les modules de débruitage ou d'autres problèmes nécessitant une détection rapide. Dans les codeurs, la DAV permet de réduire le débit, les périodes de bruit n'étant pas transmises avec le même débit, tandis que pour le débruitage, la DAV permet une mise à jour des paramètres du bruit.

De nombreux systèmes de détection d'activité vocale ont été élaborés. Nous ne citons ici que quelques uns d'entre eux.

Dans [Freeman *et al.*, 1989] et [Braun *et al.*, 1990] une méthode de détection, destinée à la transmission GSM à 13 *kbit/s*, utilise deux systèmes de détection à la suite, une DAV permettant d'estimer spécifiquement les statistiques du bruit, l'autre les intégrant pour la détection finale. La première DAV est consiste en la comparaison d'un seuil avec la sortie d'un filtre inverse qui permet de réduire le bruit supposé stationnaire, et de générer parfois un bruit de confort calculé avec l'estimation du bruit. La deuxième DAV affine la détection de la première DAV par comparaison de l'énergie du signal à un seuil. Seules les caractéristiques spectrales (LPC) utilisées pour la transmission sont nécessaires.

La DAV du codeur G.729 à 8 *kbit/s* (*cf.* [ITU Recommendation, 1996]), utilise également deux DAV à la suite. Les caractéristiques acoustiques utilisées sont l'énergie, l'énergie dans les basses fréquences de 0 *Hz* à F_l *Hz*, le taux de passage par zéro et la distorsion spectrale. Les paramètres extraits sont les coefficients d'autocorrélation $R(i)$ avec $i = 0, \cdots, 12$. L'énergie est calculée par :

$$E_f = 10 \log_{10}(\frac{1}{N} R(0)), \tag{1.1}$$

où $N = 240$ est la taille de la fenêtre d'analyse LPC. L'énergie dans les basses fréquences est donnée par :

$$E_l = 10 \log_0(\frac{1}{N} \mathbf{h}^* \mathbf{R} \mathbf{h}), \tag{1.2}$$

où $\mathbf{h}$ est la réponse impulsionnelle d'un filtre avec une fréquence de coupure de F_l *Hz*, et $\mathbf{R}$ est la matrice d'autocorrélation de Toeplitz avec les coefficients d'autocorrélation sur chaque diagonale. Le taux de passage par zéro est calculé par :

$$ZC = \frac{1}{2M} \sum_{i=0}^{M-1} |sgn(x(i)) - sgn(x(i-1))|, \tag{1.3}$$

où $M = 80$, $x(i)$ est le signal d'entrée, et $sgn(x) = 1$ si x est positif et $sgn(x) = -1$ sinon. Les coefficients de prédiction linéaire sont obtenus à partir des coefficients d'autocorrélation afin de calculer la distorsion spectrale :

$$\Delta S = \sum_{i=0}^{10} (LPC_i - \overline{LPC}_i)^2, \tag{1.4}$$

où LPC_i est le vecteur des coefficients de prédiction linéaire de la trame courante et $\overline{LPC}_i$ est le vecteur de la moyenne des coefficients de prédiction linéaire calculée sur le bruit de fond.

Cette DAV est modifiée dans [Watson *et al.*, 1997], en utilisant le pitch pour l'adaptation des paramètres du bruit et du seuil sur l'énergie dans la seconde DAV. Cette adaptation est effectuée si le pitch est continu c'est-à-dire si les variations du pitch suive une certaine tolérance. Dans [Cavallaro *et al.*, 1998], les mêmes paramètres acoustiques sont employés, mais les critères de décision sont fondés sur la logique floue. Le même principe est utilisé dans [Beritelli *et al.*, 1999] pour une application avec deux microphones. Cependant dans ce cas, il est possible d'utiliser la fonction de cohérence (*cf.* [Le Bouquin-Jeannès et Faucon, 1995]) pour une réduction de bruit performante. Dans [Sohn et Sung, 1998], le seuil adaptatif est remplacé par une règle de décision fondée sur un rapport de vraisemblance, s'appuyant sur les statistiques du bruit estimées lors de la première DAV. Ce système permet d'être plus robuste face aux bruits non stationnaires. La DAV de [Doukas *et al.*, 1997], détaillée au Chapitre 7 "*Utilisation des statistiques d'ordre supérieur*", paragraphe 7.6, est fondée sur la minimisation du cumulant croisé de la sortie de deux filtres, et permet une séparation de sources du bruit et de la parole.

La détection de parole recherchée dans le cadre de la DAV est une détection trop sensible pour un module de reconnaissance. Il est important, comme nous l'avons vu au cours de l'étude du Chapitre 3 "*Analyse des sources d'erreurs du module de détection*" de diminuer les détections de bruits considérés comme de la parole. Cependant la détection de parole pour la reconnaissance vocale permet un délai plus important que pour la transmission, et donc d'employer des caractéristiques dont l'estimation peut être plus coûteuse en terme de délai.

Notons que [Hsieh, 1998] emploie une détection de parole très simple pour un système de reconnaissance similaire à une détection d'activité vocale. Cette détection est fondée sur la comparaison de l'énergie à un seuil. Les performances ne sont cependant pas démontrées dans un environnement bruité.

Une détection plus fine pour la reconnaissance de parole peut être envisageable, comme la détection en segments voisés/non-voisés/silence.

1.3 Détection de segments voisés/non-voisés/silence

Cette détection plus fine qui segmente la parole en parties voisées ou non-voisées, est utilisée pour le codage de la parole, les parties de parole voisées et non-voisées ne sont pas codées de la même façon, pour de la segmentation automatique de parole, utilisé par

exemple pour la labellisation de bases de données pour la synthèse vocale. Mais cette détection peut aussi être employée pour la reconnaissance de parole (*cf.* [Rabiner et Sambur, 1977], [Daaboul et Adoul, 1977], *etc*). Notons que d'autres systèmes de détection d'unités acoustiques plus précises sont utilisés pour la segmentation de parole. Ces systèmes sont cependant utilisés pour des traitements sans grande contrainte de temps. Par exemple, dans [Depambour *et al.*, 1997] une détection d'allophones est présentée pour la labellisation de corpus.

Les sons voisés sont produits par la vibration des cordes vocales. Les voyelles sont intrinsèquement voisées, tandis que les consonnes peuvent l'être ou non (*cf.* [Calliope, 1989]). Nous pouvons donc considérer qu'un mot est constitué d'une suite de segments voisés, de segments non-voisés et de silences brefs. Cependant toute suite de ces trois segments de base ne correspond pas à un mot, du bruit peut être constitué par des sons voisés. Un des paramètres de voisement est le pitch, qui est le terme anglais qui couvre la fréquence des cordes vocales, la fréquence laryngienne si nous voulons faire référence au processus de génération articulatoire et la fréquence fondamentale si nous nous plaçons dans le domaine acoustique (*cf.* Chapitre 8 "*Utilisation d'un paramètre de voisement*"). Les méthodes d'extraction du pitch ne sont cependant pas toujours performantes dans les environnements bruités, et d'autres caractéristiques sont employées.

Dans [Un et Lee, 1980], cette détection en segment voisés/non-voisés/silence est modélisée par un automate à trois états dont les transitions sont contrôlées par des tests sur le taux de passage par zéro du signal filtré et sur un coefficient de variation de la parole. Cette détection a été simplifiée dans [Cho et Un, 1982] pour une DAV utilisée pour le codage et la transmission de la parole.

Des méthodes fondées sur les réseaux neuronaux sont utilisées dans [Bendiksen et Steiglitz, 1990] et [Cohn, 1991] pour une classification de la parole pour la synthèse vocale. Ces méthodes sont cependant fortement dépendantes de l'apprentissage. Dans [Di Francesco, 1990], un test de convexité sur la divergence de Kullback est utilisée pour détecter les périodes de signal voisées et non-voisées, et permet d'obtenir la période du pitch. Cette segmentation est employée pour le codage de la parole. Des données statistiques, comme les moments d'ordre 3 et 4 du signal sont proposées dans [Jacovitti *et al.*, 1991], pour une application de classification de la parole en segments voisés/non-voisés/silence (*cf.* Chapitre 7 "*Utilisation des statistiques d'ordre supérieur*", paragraphe 7.6).

Dans [Daaboul et Adoul, 1977], quatre caractéristiques acoustiques de la parole : l'énergie, l'énergie normalisée, le taux de passage par zéro du signal de parole et le le taux de passage par zéro de deux trames adjacentes du signal de parole, permettent d'établir une détection de la parole en segments voisés/non-voisés/silence pour un système de reconnaissance vocale. La décision est prise à l'aide de tests conditionnels des caractéristiques par rapport à des seuils. Les évaluations effectuées sont en vue d'application monolocuteur en environnement calme.

L'approche statistique de [Cox et Timothy, 1980] consiste en un test sur la sortie d'un banc de quatre filtres répartis sur la gamme de fréquence 200-3200 *Hz*. Si les distributions des quatre échantillons de la sortie du banc de filtres sont identiques, la trame de 15 *ms* est prise comme étant du bruit ou silence, sinon elle est considérée comme de la parole. Il

est ainsi supposé que les statistiques des quatre échantillons sont identiques. Ceci limite ce système de détection à de la parole avec un bruit large bande.

Jusqu'à cinq caractéristiques acoustiques sont employées dans [Atal et Rabiner, 1976] et dans [Rabiner *et al.*, 1977], le logarithme de l'énergie du signal, le taux de passage par zéro, la corrélation des trames adjacentes de parole, le premier coefficient de prédiction linéaire et le coefficient d'autocorrélation normalisé. L'énergie est également employée, dans [Rabiner et Sambur, 1975] et dans [Rabiner et Sambur, 1977], où il est rajouté une information à partir d'une distance LPC (*Linear Predictor Coding*). Les cinq caractéristiques de [Atal et Rabiner, 1976] sont reprises dans [Sarma et Venugopal, 1978], [Mwangi et Xydeas, 1985], [Bruno *et al.*, 1987] et [Ghiselli-Cripa et El-Jaroudi, 1991] , cependant le calcul de la distance minimum fondée sur un critère de maximum de vraisemblance est modifié. Dans [Sarma et Venugopal, 1978], l'hypothèse de distribution multi-gaussienne sur trois caractéristiques est faite, et la décision est prise à l'aide de seuils. [Mwangi et Xydeas, 1985] utilise la théorie des ensembles flous, mais n'améliore pas les performances de [Atal et Rabiner, 1976]. La décision bayesienne introduite dans [Bruno *et al.*, 1987] semble apporter une amélioration en environnement calme à la détection de [Atal et Rabiner, 1976]. Dans [Ghiselli-Cripa et El-Jaroudi, 1991] un réseau de neurones est employé avec les mêmes caractéristiques acoustiques que [Atal et Rabiner, 1976], et obtient de meilleurs résultats que celui-ci avec une restriction monolocuteur.

Une telle détection en segments voisés/non-voisés/silence a été une des premières approches appliquée à la reconnaissance vocale. En effet, au début de la reconnaissance vocale, les modèles de rejet étant moins performants, une détection précise était cruciale. Cependant les exigences par rapport à l'environnement étaient moins importantes.

Pour la reconnaissance, la séparation des trames de parole voisée et non-voisée, d'une part n'est pas indispensable, et d'autre part n'est souvent applicable qu'en milieu peu bruité. Ainsi, avec de plus des modèles de rejet performants dans le module de reconnaissance, la détection en segments voisés/non-voisés/silence a été remplacée par la détection Bruit/Parole plus adaptée à des applications en reconnaissance vocale.

1.4 Détection Bruit/Parole

La détection Bruit/Parole (DBP) ou détection des début et fin de mots, est utilisée pour délimiter un mot, ou une requête à reconnaître. Le système de reconnaissance de mots isolés présuppose que nous avons isolé le mot dans une communication qui peut être bruitée. La qualité du système de reconnaissance dépend donc de la DBP. Dans le cas de la reconnaissance de parole continue, une bonne DBP est également nécessaire, pour remédier aux modèles de rejet moins performants. Depuis une trentaine d'années que le problème se pose, les techniques évoluent pour être de plus en plus robustes aux milieux bruités.

La DBP est très liée au système de reconnaissance à laquelle elle est associée. Le système de reconnaissance est composé d'un module de détection et d'un module de reconnaissance. Il existe un grand nombre de modules de reconnaissance, qu'il est possible

de regrouper en deux classes : les modules fondés sur les principes d'alignement temporel de formes acoustiques (DTW *Dynamic Time Warping*) et les modules fondés sur les chaînes de Markov cachés (HMM *Hidden Markov Model*) (*cf.* Annexe B).

Pour classer les modules de détection, dans [Lamel *et al.*, 1981] il est fait référence à trois sortes de modules de DBP. Il y a tout d'abord les systèmes *explicites*, qui sont des modules de détection en amont des modules de reconnaissance (par exemple dans [Rabiner et Sambur, 1975]). C'est le cas de la plupart des modules de DBP. Ils sont indépendants du module de reconnaissance, et lui indiquent la détection des début et fin de mots. Les systèmes *implicites* sont au contraire inclus dans le module de reconnaissance (par exemple dans [Ney, 1981]). Il y a ainsi interaction entre les deux modules, le résultat du module de reconnaissance est utilisé pour la détection. Nous trouvons ensuite les systèmes *hybrides* qui réalisent une partie de la détection avant le processus de reconnaissance, puis en interagissant avec le module de reconnaissance. Il y a donc une première estimation des début et fin de mots par un système explicite, puis correction par un système implicite (par exemple dans [Lamel *et al.*, 1981]). C'est à dire qu'un système hybride est la combinaison d'un système explicite et d'un système implicite.

Dans [Lamel *et al.*, 1981], après implémentation de trois types d'algorithmes, un implicite, un explicite et un hybride, il est conclu que l'algorithme le plus performant est l'hybride. Notons tout de même que le système hybride de [Lamel *et al.*, 1981] peut être vu comme un système explicite (*cf.* [Kuroiwa *et al.*, 1999]). Cependant dans [Junqua *et al.*, 1994], le système hybride de [Lamel *et al.*, 1981], corrigé par [Reaves, 1991] est comparé à quatre autres systèmes explicites qui s'avèrent plus performants. Un algorithme est fondé sur l'information du pitch introduit dans [Hamada *et al.*, 1990], deux algorithmes introduits dans [Junqua *et al.*, 1991] sont fondés sur l'énergie et le taux de passage par zéro employés avec différentes règles heuristiques, et un algorithme (*cf.* [Mak *et al.*, 1992]) est fondé sur le taux de passage par zéro et sur un paramètre temps-fréquence qui représente l'énergie dans la bande de fréquence 250-3500 *Hz*, et le logarithme de la racine carrée de la moyenne des carrés des observations de l'énergie, sur une fenêtre donnée. Ce dernier algorithme est le plus performant dans des environnements bruités.

Le système implicite de [Wilpon et Rabiner, 1987] est fondé sur l'alignement produit par des modèles de Markov cachés (HMM *Hidden Markov Model*) pour les mots et le bruit, la reconnaissance est faite avec un autre modèle de Markov. Le meilleur alignement ainsi obtenu en terme de maximum de vraisemblance, permet de déterminer à la fois début et fin du mot, mais aussi le résultat de la reconnaissance. Cependant, cette approche très coûteuse ne pouvait pas être utilisée ainsi. Dans [Acero *et al.*, 1993], une approche en temps réel fondée sur un rapport de vraisemblance calculé à partir des modèles de Markov cachés de la parole et du bruit, est proposée. Ce système est comparé à celui de [Lamel *et al.*, 1981], et donne de meilleurs résultats. Dans [Takeda *et al.*, 1995], le même principe que dans [Acero *et al.*, 1993] est appliqué, avec un modèle de Markov plus performant. Une adaptation de cette approche pour la reconnaissance de parole continue est proposée dans [Yamamoto *et al.*, 1997], reprise par [Kuroiwa *et al.*, 1999]. Ces travaux sont détaillés dans [Naito *et al.*, 1998]. La dernière solution apportée dans [Segawa *et al.*, 2001] est un système de reconnaissance de parole continue sans détection de parole. L'utilisation des

modèles de Markov reste coûteuse, et le modèle de rejet du module de reconnaissance permet de rejeter les détections de bruit. L'introduction des modèles de Markov dans le module de détection de France Télécom R&D (*cf.* [Mauuary, 1994]) n'a en outre pas apporté une amélioration significative.

Cependant, un système implicite peut avantageusement être appliqué pour confirmer la fin de la détection. Dans [Ney, 1981] et [Haltsonen, 1984] une approche implicite est proposée pour un algorithme fondé sur les principes d'alignement temporel de formes acoustiques (DTW *Dynamic Time Warping*), qui permet d'ajouter ou supprimer quelques trames en début ou fin de mots pour ajuster la détection. Cependant dans [Ney, 1981], l'enregistrement ne peut contenir qu'un mot, et dans [Haltsonen, 1984] le vocabulaire utilisé est composé de l'alphabet et des chiffres. Il est montré dans [Hanel et Jouvet, 2000] que l'utilisation des modèles de Markov pour la détection de la fin d'une épellation de noms, permet d'améliorer les performances de reconnaissance. Dans le cas de la reconnaissance de parole continue, la détection de fin de phrases peut être confirmée par un autre système de détection pour remédier au problème de baisse énergétique et de pauses inter-mots longues. Ainsi, le résultat intermédiaire de la reconnaissance peut être utilisé, comme dans [Hariharan *et al.*, 2001] où il est associé à l'énergie dans plusieurs sous-bandes contrôlée par des seuils, pour la détection de fin de phrases. Notons également que [Ariyoshi, 2000] propose une détection d'intensité (sur l'énergie) intégrée au système de reconnaissance pour permettre le calcul d'une mesure de similarité par mot. Cette détection intégrée est une simple DAV indispensable au calcul de cette mesure.

À la différence de la DAV, où la détection est fondée sur les deux états de parole et non-parole, les systèmes de DBP peuvent s'appuyer sur des automates à trois états (*cf.* [Li *et al.*, 2001]), à quatre états (*cf.* [Ganapathiraju *et al.*, 1996] et [Reaves, 1997]), ou à cinq états (*cf.* [Mauuary et Monné, 1993]).

1.5 Conclusion

Ce chapitre qui présente les différents contextes de détection de parole, montre les similitudes et différences de la détection de parole selon le cadre de l'étude. La principale différence vient du but recherché par chaque système de détection. La DAV est une détection avec un délai minimum, il est important de ne pas omettre de parole, mais la précision des frontières n'est pas recherchée. La détection de segments voisés/non-voisés/silence est une détection plus précise de la parole selon les sons voisés et non-voisés. Selon l'application ce type de système de détection ne requière pas un faible délai. Cette détection plus précise est cependant peu performante pour des communications bruitées. La DBP doit être un détection la plus précise possible au niveau des frontières, le délai peut être plus long que la DAV. La DBP peut omettre quelques mots, mais éviter de trop détecter des bruits. Les méthodes employées peuvent être différentes selon la détection recherchée, cependant les caractéristiques permettant la détection sont souvent les mêmes. Les approches souvent simples de la DAV et la précision recherchée de la détection en segments voisés/non-voisés/silence peuvent être employées pour la DBP.

La détection de parole qui fait l'objet de cette étude est la détection Bruit/Parole pour un système de reconnaissance. Nous avons ici présenté différentes approches de DBP. Dans le Chapitre 2 "*Détection de parole pour la reconnaissance vocale*", nous présentons les particularités du module de détection employé pour le système de reconnaissance de France Télécom R&D. Nous étudions ensuite l'influence du module de détection sur le système de reconnaissance qui nous permet d'établir un principe d'évaluation.

Chapitre 2

Détection de parole pour la reconnaissance vocale

2.1 Introduction

Le *système* de reconnaissance utilisé dans cette étude est composé d'un *module de détection* et d'un *module de reconnaissance*. Le module de détection de parole détermine les périodes du signal où la parole est présente. La structure du système décrit sur la figure 2.1 est le système explicite développé à France Télécom R&D. Le module de reconnaissance est désigné par RECO et le module de détection de Bruit/Parole par DBP.

L'analyse du signal de parole permet d'extraire des coefficients pertinents pour le module de détection de parole et pour le module de reconnaissance. Dans le système décrit ici, et que nous utilisons dans la suite de l'étude, les modules de détection et de reconnaissance fonctionnent simultanément. Le signal de parole et le module de l'analyse du signal qui permet de dégager des coefficients pertinents sont présentés en Annexe A. Le module de reconnaissance utilisé dans cette étude pour la reconnaissance de mots isolés et de parole continue est décrit en Annexe B. La DBP précise le début, puis la fin de parole au module de reconnaissance. À la fin d'une détection de parole, le système de reconnaissance indique ce qui a été reconnu.

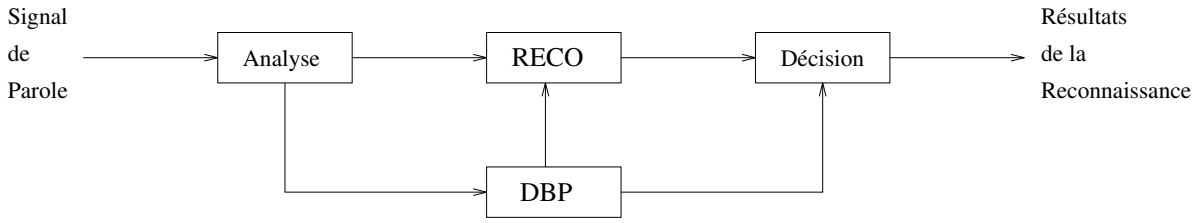


FIG. 2.1 – *Structure du système de reconnaissance étudié.*

Ce chapitre a pour but de présenter le module de détection dans le système de recon-

naissance. Le module de détection doit être considéré comme une partie intégrante du système de reconnaissance. C'est pourquoi son évaluation qui peut se faire en ne considérant que les résultats de la détection, doit être abordée en considérant le système de reconnaissance dans son ensemble. Ce chapitre montre également qu'il est important d'améliorer les performances du module de détection pour obtenir une meilleure reconnaissance vocale.

Le paragraphe 2.2 présente les différents critères du module de détection fondé sur un automate Bruit/Parole à cinq états. Ce module de détection est désigné par *module de détection ABP*. Il est déterminant de bien comprendre les conséquences du module de détection ABP sur le système de reconnaissance, au paragraphe 2.3, pour ensuite définir les principes d'évaluation utilisés pour une étude comparative des différents critères du module de détection ABP abordés dans cette étude, au paragraphe 2.4. Ce dernier paragraphe situe notre méthode d'évaluation par rapport à celles employées ces dernières années.

2.2 Le module de détection

Nous décrivons dans ce paragraphe le module de détection ABP fondé sur un automate Bruit/Parole, qui est la base de notre étude. Dans un premier temps nous décrivons l'automate Bruit/Parole au paragraphe 2.2.1. Nous présentons ensuite les trois critères développés antérieurement à nos travaux à France Télécom R&D : un critère fondé sur le rapport signal à bruit au paragraphe 2.2.2, un critère fondé sur les statistiques du bruit au paragraphe 2.2.3 et un critère fondé sur les statistiques du bruit et de la parole au paragraphe 2.2.4. Ces critères sont évalués au chapitre suivant, afin de déterminer le critère présentant les meilleurs résultats.

2.2.1 L'automate Bruit/Parole

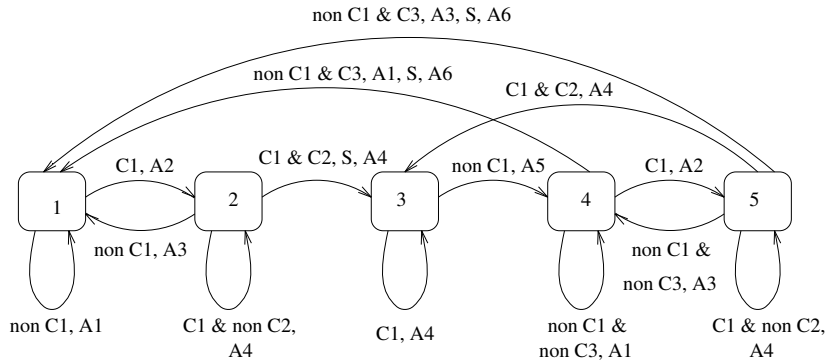
Le module de détection ABP est employé en vue de l'appliquer à un module de reconnaissance fondé sur les chaînes de Markov cachées (*cf.* Annexe B et [Jouvet, 1988] pour plus de détails sur le module de reconnaissance).

Le module de détection ABP utilise un automate à cinq états, qui sont :

bruit ou silence, *présomption de parole*, *parole*, *plosive non voisée ou silence*, *reprise possible de parole*.

Le fonctionnement de l'automate donné par la figure 2.2 est décrit dans [Mauuary, 1994]. Notre étude étant fondée sur cet automate, nous détaillons ici son fonctionnement.

Dans la version initiale de l'algorithme de DBP, les passages d'un état à un autre sont conditionnés par un seuil sur l'énergie du signal et par des contraintes structurelles de durée (durée minimum d'une voyelle et durée maximum d'une plosive). D'autres méthodes de passage d'un état à un autre ont été étudiées et sont présentées ci-dessous aux paragraphes 2.2.2, 2.2.3 et 2.2.4. Les passages dans l'état *parole* déterminent les frontières de début de la parole dans le signal. Le module de reconnaissance prend en compte ces données avec une marge de sécurité sur la frontière de début de 160 *ms*.



CONDITIONS :

C1 : Énergie > Seuil de détection

C2 : Durée Parole (DP) >= Parole Minimum

C3 : Durée Silence (DS) >= Silence Fin

ACTIONS :

A1 : DS = DS +1 A4 : DP = DP + 1

A2 : DP = 1 A5 : DS = 1

A3 : DS = DS + DP A6 : DS = DP = 0

S : sauvagarde de frontières

VALEURS INITIALES :

ÉTAT = BRUIT OU SILENCE

DS = DP = 0

Silence Fin est généralement choisi entre 240 ms (mots isolés) et 640 ms (mots connectés),
Silence Fin représente le maximum entre la durée maximum d'une tenue de plosive (fixé arbitrairement à 240 ms)
et la durée maximum d'une pause entre mots (jusqu'à 64oms).

N.B. : & est le *et* logique

FIG. 2.2 – *Automate de détection Bruit/Parole.*

L'état *bruit ou silence* est l'état initial de l'algorithme. Nous faisons ainsi l'hypothèse que la communication débute par une trame de bruit ou de silence. L'automate reste dans cet état tant qu'il n'y a pas de trame énergétique (*i.e.* une trame dont l'énergie est supérieure au seuil).

Lors de la première trame énergétique, l'automate passe dans l'état *présomption de parole*. Dans cet état, une trame non énergétique le fait retourner à l'état *bruit ou silence*. Après être resté un nombre minimum de trames (*Parole Minimum*) dans l'état *présomption de parole*, l'automate passe à l'état *parole*.

Il reste dans l'état *parole* tant que les trames sont énergétiques. Il passe à l'état *plosive non voisée ou silence*, dès que la trame courante est non énergétique.

Dans l'état *plosive non voisée ou silence*, un certain nombre (*Silence Fin*) de trames non énergétiques confirment le silence et le retour dans l'état *bruit ou silence*. L'action *A1* fournit le silence après la dernière trame de parole détectée, tandis que l'action *A6*

réinitialise la durée de silence *Durée Silence*. Si dans l'état *plosive non voisée ou silence*, la trame courante est énergétique, l'automate passe dans l'état *reprise possible de parole*.

Dans ce dernier état, une trame non énergétique le fait retourner dans l'état *plosive non voisée ou silence* ou dans l'état *bruit ou silence* si la durée de silence (*Durée Silence*) qui représente le temps passé dans l'état *plosive non voisée ou silence* et dans l'état *reprise possible de parole*, est supérieure à un certain nombre de trames (*Silence Fin*). Lors du retour à l'état *plosive non voisée ou silence*, l'action *A3* permet de préciser le nombre de trame de silence après la dernière trame de parole de l'état *parole* pour déterminer la frontière de fin de parole. Lors du retour à l'état *bruit ou silence* les actions *A3* et *A6* sont effectuées. Après être resté un nombre minimum (*Parole Minimum*) de trames énergétiques dans l'état *reprise possible de parole*, l'automate retourne dans l'état *parole*.

Les trois états *présomption de parole*, *plosive non voisée ou silence* et *reprise possible de parole* sont introduits pour modéliser les variations énergétiques du signal de parole et des bruits. L'état *présomption de parole* permet de ne pas détecter des bruits impulsifs qui sont énergétiques mais de courte durée (*i.e.* quelques trames). L'état *plosive non voisée ou silence* modélise les passages peu énergétiques dans le mot ou la phrase, tels que les silences intra-mot ou les plosives. Dans le cas de la reconnaissance de parole continue, l'état *reprise possible de parole* modélise les silences inter-mots, plus longs que les silences intra-mot.

Le passage d'un état à l'autre peut se faire selon plusieurs critères, qui correspondent à l'utilisation de différents tests possibles pour la condition C1. Nous présentons ci-dessous trois critères qui ont été développés antérieurement à nos travaux et qui sont évalués dans le chapitre suivant.

2.2.2 Critère de détection fondé sur le rapport signal à bruit

Nous considérons le logarithme de l'énergie pour diminuer la dynamique de l'énergie. Le seuil sur le logarithme de l'énergie qui permet le passage d'un état à l'autre est adaptatif. Nous cherchons à comparer des estimations à court-terme et à long-terme du logarithme de l'énergie du signal. Le logarithme de l'énergie à court-terme est le logarithme de la moyenne des carrés des échantillons, sur une fenêtre de 32 *ms*, et noté E. Le logarithme de l'énergie à long-terme (ELT) est calculé, dans les périodes de silence, de façon récursive :

$$ELT(n) = ELT(n-1) + (1-\lambda)(E(n) - ELT(n-1)), \tag{2.1}$$

où λ est un facteur d'oubli (en général fixé à 0.99 pour cet automate, ce qui correspond à une constante de temps de 1600 *ms*), et n est l'indice de la trame. L'estimation à long-terme du logarithme de l'énergie ELT est une estimation de la moyenne du logarithme de l'énergie dans les périodes de silence. Une estimation du rapport signal à bruit est donnée par la différence des logarithmes de l'énergie à long-terme et à court-terme, qui est comparée à un seuil de détection donné afin de décider de la présence ou non d'une trame de parole :

$$C1 : E(n) - ELT(n) > \textit{Seuil de détection.} \tag{2.2}$$

Ce critère fondé sur le rapport signal à bruit, qui nécessite une estimation à long-terme et à court-terme du logarithme de l'énergie, est appelé *critère LCT*.

2.2.3 Critère de détection fondé sur les statistiques du bruit

Nous faisons l'hypothèse que le logarithme de l'énergie du bruit suit une loi normale de paramètres (μ,σ^2). La figure 4.12 du paragraphe 4.6 montre que cette hypothèse est valide, et par ailleurs souvent considérée dans l'état de l'art (*cf.* paragraphe 4.5.2). Les statistiques du logarithme de l'énergie du bruit sont estimées lorsque l'automate est dans l'état *bruit ou silence*. La moyenne est estimée, de même que dans l'équation (2.1) pour le critère LCT, par :

$$\hat{\mu}(n) = \hat{\mu}(n-1) + (1-\lambda)(E(n) - \hat{\mu}(n-1)), \tag{2.3}$$

et l'écart-type par :

$$\hat{\sigma}(n) = \hat{\sigma}(n-1) + (1-\lambda)(|E(n) - \hat{\mu}(n-1)| - \hat{\sigma}(n-1)), \tag{2.4}$$

où n est l'indice de la trame. Ces estimations se font dans l'état *bruit ou silence* de l'automate. Le facteur d'oubli λ a été optimisé empiriquement dans [Karray, 1998b]. Pour l'estimation de la moyenne $\lambda = 0.99$, ce qui correspond à une constante de temps de 1600 *ms*, et pour l'estimation de l'écart-type $\lambda = 0.995$, ce qui correspond à une constante de temps de 3200 *ms*. L'estimation de l'écart-type de l'équation (2.4) est une approximation de l'écart-type de la loi gaussienne par la loi laplacienne. L'estimation de l'écart-type de la loi gaussienne sur une fenêtre exponentielle est donnée par :

$$\hat{\sigma}_G^2(n) = \hat{\mu}_2(n) - \hat{\mu}^2(n), \tag{2.5}$$

où $\hat{\mu_2}$ est l'estimation de la moyenne des carrés du logarithme de l'énergie. Ainsi un biais est introduit et l'estimation de l'écart-type de la loi gaussienne peut s'écrire : $\hat{\sigma_G} = b\hat{\sigma}$, où b est le biais. Cette hypothèse est faite pour simplifier l'estimation de l'écart-type, l'estimation de $\hat{\mu_2}$ n'est pas nécessaire. Le logarithme de l'énergie de chaque trame est considéré et nous cherchons à vérifier l'hypothèse que nous sommes dans l'état *bruit ou silence*, qui correspond à l'absence de parole. La décision sera prise en fonction de l'écart du logarithme de l'énergie de cette trame par rapport à la moyenne estimée du bruit, c'est-à-dire selon la valeur du rapport critique $r_{SB}(E(n)) = \frac{E(n)-\hat{\mu}(n)}{\hat{\sigma}(n)}$, comparé à un seuil :

$$C1 : r_{SB}(E(n)) > \textit{Seuil de détection.} \tag{2.6}$$

Ce critère fondé sur l'estimation des statistiques du bruit est appelé *critère SB*.

2.2.4 Critère de détection fondé sur les statistiques du bruit et de la parole

Cette approche découle d'une approche Bayesienne (*cf.* [Karray et Monné, 1998]). Nous testons deux hypothèses :

-H_0 : la trame courante est du bruit ou du silence,
-H_1 : la trame courante est de la parole (bruitée).

Pour chaque observation $E(n)$, le logarithme de l'énergie à la trame n, nous cherchons à comparer le maximum de vraisemblance $P(H_i/E)$ de chaque hypothèse. C'est-à-dire que le rapport de vraisemblance $r_{SBP}(E) = \frac{P(H_0/E)}{P(H_1/E)}$ est comparé à 1. Si $r_{SBP}(E) \leq 1$ la trame est alors considérée comme étant de la parole bruitée, sinon elle sera du bruit. En pratique, la condition $C1$ s'écrit :

$$C1 : E(n) > \textit{Seuil de détection}, \tag{2.7}$$

où *Seuil de détection* $= s * \alpha$, s est une solution de l'équation $r_{SBP}(x) = 1$, où $x = E(n)$, et α est un facteur d'interpolation permettant de remédier aux erreurs d'approximations dues aux hypothèses et évaluations. Ce facteur d'interpolation est associé au seuil de détection dans les autres chapitres. Une fois s déterminé, α varie, faisant ainsi varier le seuil de détection.

Pour déterminer s, nous supposons d'abord H_0 et H_1 équiprobables, ainsi nous avons $r_{SBP}(x) = \frac{P(x/H_0)}{P(x/H_1)}$, où $x = E(n)$. Dans un premier temps nous supposons que les deux distributions sont gaussiennes, la figure 4.12 du paragraphe 4.6 suggère qu'il est possible d'approcher les deux distributions par des distributions gaussiennes. Nous avons, pour $i = 0{,}1$:

$$P(x/H_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}. \tag{2.8}$$

En considérant (2.8), l'équation $r_{SBP}(x) = 1$ devient :

$$(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2})x^2 - 2\left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right)x + \frac{\mu_0^2}{\sigma_0^2} - \frac{\mu_1^2}{\sigma_1^2} + 2\ln\frac{\sigma_0}{\sigma_1} = 0. \tag{2.9}$$

Cette équation possède deux solutions, nous choisissons celle située la plus proche du milieu de $[\mu_0,\mu_1]$. S'il n'existe pas de solution dans cet intervalle, nous prenons le milieu de $[\mu_0,\mu_1]$. Notons cependant que pour certains enregistrements très bruités l'intersection des deux gaussiennes peut se trouver en dehors de cet intervalle. Ces deux solutions de l'équation (2.9) approchent assez bien, graphiquement, les points réels d'intersection des courbes représentatives de l'énergie.

Pour simplifier cette équation, nous faisons l'hypothèse que les distributions sont laplaciennes. Au lieu de l'équation (2.8), nous avons alors, pour $i = 0{,}1$:

$$P(x/H_i) = \frac{1}{\sqrt{2}\sigma_i} e^{-\frac{\sqrt{2}|x-\mu_i|}{\sigma_i}}. \tag{2.10}$$

Ainsi l'équation $r_{SBP}(x) = 1$ s'écrit :

$$\frac{\sigma_0}{\sigma_1} e^{-\sqrt{2}(\frac{|x-\mu_0|}{\sigma_0} - (\frac{|x-\mu_1|}{\sigma_1})} = 1. \tag{2.11}$$

En faisant l'hypothèse que $s \in [\mu_0,\mu_1]$ et considérant (2.10), s est l'unique solution de l'équation (2.11), et nous avons :

$$\left(\frac{1}{\sigma_0}+\frac{1}{\sigma_1}\right)x = \frac{\mu_0}{\sigma_0}+\frac{\mu_1}{\sigma_1}-\frac{1}{\sqrt{2}}\ln\frac{\sigma_0}{\sigma_1}. \tag{2.12}$$

Le seuil de décision est donc la solution de (2.12) :

$$s = \frac{\frac{\mu_0}{\sigma_0}+\frac{\mu_1}{\sigma_1}-\frac{1}{\sqrt{2}}\ln\frac{\sigma_0}{\sigma_1}}{\frac{1}{\sigma_0}+\frac{1}{\sigma_1}}. \tag{2.13}$$

Les statistiques de l'énergie de la parole sont estimées dans l'état *parole* (μ_1 et σ_1), de la même façon que les statistiques de l'énergie du bruit (μ_0 et σ_0) le sont dans l'état *bruit ou silence*. Le facteur d'oubli a été optimisé à 0.95 pour la parole et à 0.99 pour le bruit, le bruit étant supposé plus stationnaire (*cf.* [Karray, 1998a]). Ces facteurs d'oubli correspondent à des constantes de temps de 320 *ms* pour la parole et de 1600 *ms* pour le bruit.

Ce critère fondé sur l'estimation des statistiques du bruit et de la parole est appelé *critère SBP*.

Vérifions expérimentalement, que nous avons bien en général, l'hypothèse $s \in [\mu_0,\mu_1]$. Nous avons calculé la moyenne et l'écart-type du logarithme de l'énergie sur les périodes de bruit et de parole sur les bases RTC_A et GSM_A (décrites en Annexe D). Nous avons $\mu_0 < \mu_1$ et $\sigma_0 < \sigma_1$ (*cf.* figure 4.12 du paragraphe 4.6). Vérifions donc que $s \in [\mu_0,\mu_1]$, et que s est bien solution de l'équation (2.11), nous avons trois cas possibles.

- Premièrement, si $s < \mu_0$, nous avons aussi $s < \mu_1$, et en remplaçant s par sa valeur donnée par (2.13) dans ces deux inégalités, nous obtenons :

$$\begin{cases} \sqrt{2}(\mu_0-\mu_1)-\sigma_0\ln\frac{\sigma_0}{\sigma_1} < 0 \\ \sqrt{2}(\mu_1-\mu_0)-\sigma_1\ln\frac{\sigma_0}{\sigma_1} < 0 \end{cases}, \tag{2.14}$$

  et d'après l'équation (2.11), nous obtenons :

$$\sigma_1\ln\frac{\sigma_0}{\sigma_1} = \sqrt{2}(\mu_1-\mu_0), \tag{2.15}$$

  ceci étant expérimentalement incorrect, car $\mu_0 < \mu_1$ et $\sigma_0 < \sigma_1$, d'après les valeurs expérimentales.
- Deuxièmement, si $s \in [\mu_0,\mu_1]$, en remplaçant s par sa valeur donnée par (2.13) dans cette appartenance, nous obtenons :

$$\begin{cases} \sqrt{2}(\mu_0-\mu_1)-\sigma_0\ln\frac{\sigma_0}{\sigma_1} < 0 \\ \sqrt{2}(\mu_1-\mu_0)-\sigma_1\ln\frac{\sigma_0}{\sigma_1} > 0 \end{cases}, \tag{2.16}$$

dans ce cas d'après l'équation (2.11), nous avons :

$$\mu_0 - \mu_1 < \frac{\sigma_0}{\sqrt{2}} \ln \frac{\sigma_0}{\sigma_1}, \tag{2.17}$$

qui est correct d'après les valeurs expérimentales.

- Troisièmement, si $s > \mu_1$, nous avons aussi $s > \mu_0$, et en remplaçant s par sa valeur donnée par (2.13) dans ces deux inégalités, nous obtenons :

$$\begin{cases} \sqrt{2}(\mu_0 - \mu_1) - \sigma_0 \ln \frac{\sigma_0}{\sigma_1} > 0 \\ \sqrt{2}(\mu_1 - \mu_0) - \sigma_1 \ln \frac{\sigma_0}{\sigma_1} > 0 \end{cases}, \tag{2.18}$$

ce qui, d'après l'équation (2.11), conduit à :

$$\sigma_0 \ln \frac{\sigma_0}{\sigma_1} = \sqrt{2}(\mu_0 - \mu_1). \tag{2.19}$$

Cette condition est fausse en général, d'après les valeurs expérimentales de μ_0, μ_1, σ_0 et σ_1.

Nous avons ainsi vérifié que les hypothèses faites sont correctes expérimentalement, et que la solution trouvée vérifie bien l'équation de départ. Cependant l'approximation de la distribution de la parole par une laplacienne peut être discutée (*cf.* paragraphe 4.5.2). Il est noté dans [Karray, 1998a] que les résultats obtenus sont aussi bons en remplaçant la condition donnée dans (2.7) simplement par :

$$C1 : E(n) > \alpha \hat{\mu}_0(n) + (1 - \alpha)\hat{\mu}_1(n), \tag{2.20}$$

où α est un facteur d'interpolation ($0 < \alpha < 1$) qui est optimisé empiriquement. Cette approche est cependant très dépendante du facteur d'interpolation, et ne permet pas une adaptation à la variabilité de l'écart-type du signal de parole et du bruit.

2.2.5 Remarque sur l'estimation

Dans les trois critères du module de détection ABP proposés ci-dessus, l'estimation de la moyenne ou de la variance, du bruit ou de la parole se fait dans les états détectés comme étant de la parole ou du bruit par l'automate. Ainsi l'automate contrôle lui-même les estimations qui servent à la détection. Il est donc important que la détection soit suffisamment fiable pour pouvoir contrôler ses propres estimations.

2.3 Influence du module de détection sur le système de reconnaissance

Un module de détection doit fournir les débuts et fins de parole au module de reconnaissance. C'est sur cette période du signal que le module de reconnaissance cherche à

reconnaître ce qui a été prononcé (*cf.* Annexe B). Il s'agit avant tout de ne pas perdre de périodes de parole utile. Nous ne cherchons pas une grande précision au niveau des frontières de la détection des mots (ou des phrases), du moment que le mot (ou la phrase) n'est pas tronqué à droite ou à gauche, et qu'il ne soit pas élargi de trop à gauche ou à droite (c'est-à-dire que la détection de début ne se fasse pas trop tôt ou que la détection de fin ne se fasse pas trop tard, quelques trames en début en fin de détection sont gardées comme marge de sécurité). En effet, un modèle de silence de début et de fin de mots, dans le module de reconnaissance permet une détection plus large au niveau des frontières, sans perturber les performances du système de reconnaissance. Cependant les détections trop longues de silences entraînent des erreurs dans le cas de reconnaissance de parole continue (*cf.* paragraphe 3.8).

Ainsi le module de détection a une influence importante sur le système de reconnaissance. Avant de présenter les conséquences des erreurs du module de détection sur les erreurs du module de reconnaissance au paragraphe 2.3.3, il nous faut présenter les différentes erreurs d'une part du module de détection au paragraphe 2.3.1 et d'autre part du module de reconnaissance au paragraphe 2.3.2. Il est important de bien définir les différents types d'erreurs du module de détection et du module de reconnaissance, et de comprendre les conséquences des premières sur les secondes, afin d'établir une méthode d'évaluation rigoureuse au paragraphe 2.4.

2.3.1 Erreurs du module de détection

Le module de détection doit donc fournir toutes les périodes de parole prononcée, sans les tronquer à droite ou à gauche, et sans en omettre. Il doit de plus éviter de détecter des périodes de bruit. Il est également préférable de ne pas détecter les périodes de parole moins énergétique d'une tierce personne qui parle en arrière plan. En effet même si ces détections peuvent être rejetées par le module de reconnaissance, elles ne le sont pas à chaque fois, et elles augmentent les erreurs du module de reconnaissance. Ces erreurs du module de détection sont des *insertions* de bruits. De plus, il est impératif de ne pas fragmenter une période de parole utile. En effet dans le cas de la reconnaissance de mots isolés, le module de reconnaissance cherche à identifier autant de mots que de détections issus de la *fragmentation* du mot. Dans le cas de la reconnaissance de parole continue, le sens de la phrase peut être affecté par ces erreurs de *fragmentation*. Ceci entraîne un nombre d'erreurs important, car le module de reconnaissance, même s'il peut récupérer ces erreurs par un modèle de rejet, risque de reconnaître plusieurs mots, avec par exemple la première et la seconde moitié du mot fragmenté, ou des mots se rapprochant. Dans le cas de la parole continue, ces erreurs ont une influence importante sur des applications utilisant un système de dialogue homme-machine. Dans une requête, si une partie de la phrase manque, le sens peut en être considérablement atteint. De même, si le module de détection regroupe plusieurs périodes de parole utile en une seule, avec les modèles utilisés dans cette étude, le module de reconnaissance ne cherchera à reconnaître qu'une seule période de parole au lieu de plusieurs. Il s'en suit une erreur systématique de la reconnaissance.

Nous distinguons donc parmi les erreurs du module de détection :
- Les *insertions*.
- Les *omissions*.
- Les *fragmentations*.
- Les *regroupements*.
- Les *détections imprécises* : c'est-à-dire des segments tronqués ou des segments élargis.

2.3.2 Erreurs du module de reconnaissance

Nous distinguons ici les erreurs du module de reconnaissance pour les mots isolés et pour la parole continue car le module de détection n'a pas la même influence sur les erreurs dans les deux cas. En effet une erreur d'insertion du module de détection est la seule source d'erreurs qui peut entraîner le module de reconnaissance de mots isolés à reconnaître un mot alors qu'il n'y en a pas, tandis que le module de reconnaissance de parole continue peut reconnaître un mot non prononcé sur une détection correcte. C'est pourquoi nous employons une terminologie différente des erreurs de reconnaissance pour les mots isolés et la parole continue.

Erreurs du module de reconnaissance pour les mots isolés

Les erreurs du module de reconnaissance de mots isolés sont regroupées en trois types. Ces erreurs sont de plus soit des erreurs intrinsèques au module de reconnaissance, soit des erreurs dues au module de détection. Les trois types d'erreurs sont :

- Les erreurs de *rejet à tort*, qui correspondent aux segments de parole du vocabulaire rejetés par le système. Soit les segments sont rejetés par le modèle de rejet du module de reconnaissance, soit ils n'ont pas été détectés par le module de détection. De plus les erreurs de fragmentation et de regroupement peuvent provoquer des rejets à tort. Une détection imprécise peut également entraîner ce type d'erreur.
- Les erreurs de *fausse acceptation*, qui correspondent aux segments de bruit ou de parole hors vocabulaire pris pour de la parole du vocabulaire. Le locuteur prononce souvent des mots hors vocabulaire, surtout s'il ne connaît pas l'application, ceci peut produire une erreur de fausse acceptation. La détection de parole non utile produite par le locuteur, ne peut être considérée comme une erreur du module de détection. Par contre, trop de détections de bruit, ou de parole en arrière plan augmenteront ce type d'erreurs. Les erreurs de fragmentation du module de détection risquent également d'entraîner des erreurs de fausse acceptation.
- Les erreurs de *substitution*, qui correspondent aux segments comprenant un mot du vocabulaire reconnu comme un autre mot du vocabulaire. Le module de détection peut avoir une influence sur ces erreurs avec les erreurs de regroupement et de fragmentation. Si plusieurs segments sont regroupés en un seul, et que la détection regroupant ces segments est reconnue comme un mot du vocabulaire, il y aura au moins une erreur de substitution parmi les différents segments regroupés. Si un

segment est fragmenté, une partie du segment peut être reconnue comme un autre mot du vocabulaire que celui contenu dans le segment entier, surtout si cette partie ressemble phonétiquement à un autre mot du vocabulaire. Pour les mêmes raisons, les détections imprécises peuvent également provoquer une erreur de substitution, en effet un mot tronqué peut être phonétiquement proche d'un autre mot.

Erreurs du module de reconnaissance pour la parole continue

Dans le cas de la reconnaissance de parole continue, les erreurs de reconnaissance ne sont plus comptées de la même manière. Nous distinguons ici quatre types d'erreurs principales, qui sont également dues, soit au module de reconnaissance, soit au module de détection. Ces erreurs sont comptées au niveau des mots dans une requête, mais aussi au niveau de la requête, contrairement au cas de la reconnaissance de mots isolés, où les erreurs ne sont comptées qu'au niveau des segments.

- Les erreurs de *rejet à tort* correspondent aux requêtes non reconnues comme de la parole, et rejetées par le modèle de rejet. Ces requêtes peuvent être de la parole hors vocabulaire ou hors syntaxe. Les requêtes non détectées par le module de détection sont comptabilisées comme des rejets à tort. Les erreurs de fragmentation et de regroupement peuvent aussi provoquer des rejets à tort. Les rejets à tort seront exprimés en fonction des omissions de mots qu'ils engendrent.
- Les erreurs d'*omission* correspondent aux omissions de mots contenus dans une requête. Un segment tronqué peut entraîner des omissions de mots, ainsi que des erreurs de fragmentation ou de regroupement de parole. Une omission d'une requête entière, due au module de détection, entraîne inévitablement autant d'omissions de mots qu'il y a de mots dans la requête, cependant ces erreurs sont comptées au niveau des rejets à tort.
- Les erreurs d'*insertion* correspondent aux insertions de mots par rapport aux mots contenus dans une requête. Une détection trop large peut entraîner des insertions de mots, de même pour les erreurs de fragmentation ou de regroupement de parole. Une détection d'une partie du signal qui ne correspond à aucune requête, si elle n'est pas rejetée, donnera également des insertions de mots.
- Les erreurs de *substitution* correspondent aux mots reconnus comme un autre mot du vocabulaire. Le module de détection peut encore avoir une influence sur ces erreurs avec les erreurs de regroupement et de fragmentation. Plusieurs requêtes regroupées ou inversement une requête fragmentée, modifieront les probabilités des mots de début de requêtes par exemple. De même si la requête est tronquée au début, par exemple le deuxième mot d'une requête peut se retrouver en premier, avec une plus faible probabilité de commencer une phrase qu'un autre mot lui ressemblant de près ou de loin. Une détection trop large entraînant par exemple l'insertion d'un mot, le premier mot de la requête se retrouve en seconde position, avec une très faible probabilité de suivre le premier mot inséré, si le silence de début s'aligne bien sur un mot du vocabulaire.

2.3.3 Conséquences des erreurs du module de détection sur les erreurs du système de reconnaissance

Les conséquences des erreurs du module de détection ABP sur les erreurs du système de reconnaissance sont importantes aussi bien dans le cas de reconnaissance de mots isolés (*cf.* figure 2.3) que dans le cas de reconnaissance de parole continue (*cf.* figure 2.4). Nous remarquons que les erreurs de fragmentation et de regroupement peuvent avoir une forte implication sur les différents types d'erreurs du module de reconnaissance. Dans le cas de la parole continue la seule différence se trouve au niveau des détections imprécises. En effet, le modèle de rejet étant moins performant, une détection trop large peut entraîner beaucoup d'insertions de mots.

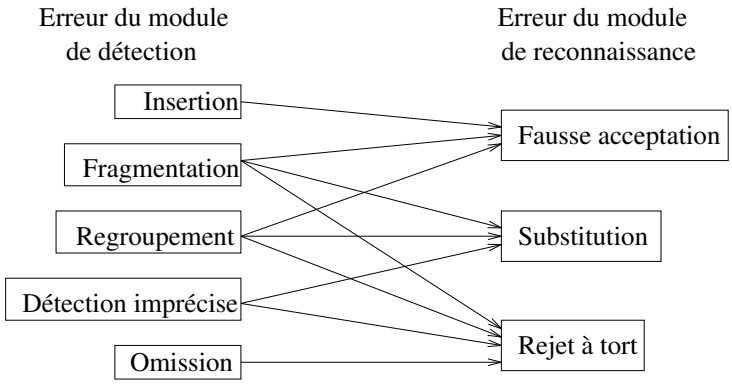


FIG. 2.3 – *Relations entre les erreurs du module de détection et les erreurs du module de reconnaissance pour la reconnaissance de mots isolés.*

Pour évaluer les performances de la détection de parole, nous comparons les résultats de reconnaissance d'une détection automatique avec ceux obtenus à partir d'une segmentation manuelle. La figure 2.5 donne les taux d'erreur de la reconnaissance de mots isolés (*cf.* figure 2.5(a)) et les taux d'erreur de la reconnaissance de parole continue (*cf.* figure 2.5(b)) pour une segmentation manuelle et une détection automatique obtenue avec le critère LCT. Les résultats obtenus pour les critères SB et SBP sont du même ordre. Les courbes sont obtenues en faisant varier le poids du rejet du module de reconnaissance.

La segmentation manuelle comprend tous les segments de *Parole*, c'est-à-dire les segments de *Parole-Voc* et de *Parole-Hors-Voc* (*cf.* Glossaire). Sur une base de mots isolés (la base RTC_A décrite en Annexe D), nous constatons (*cf.* figure 2.5(a)) que la détection automatique provoque plus d'erreurs de substitution, de rejet à tort (pour un poids de rejet fixé), et surtout un grand nombre de fausses acceptations, qui sont dues aux détections de bruits. Les taux de substitution et de fausse acceptation sont d'environ 50% plus importants avec cette détection automatique.

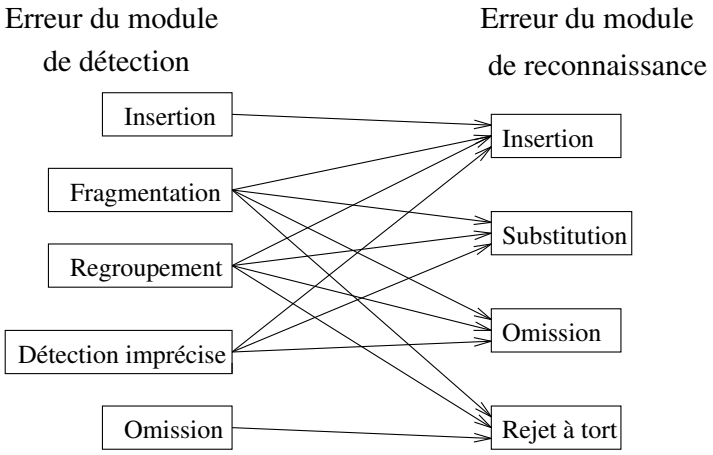


FIG. 2.4 – *Relations entre les erreurs du module de détection et les erreurs du module de reconnaissance pour la reconnaissance de parole continue.*

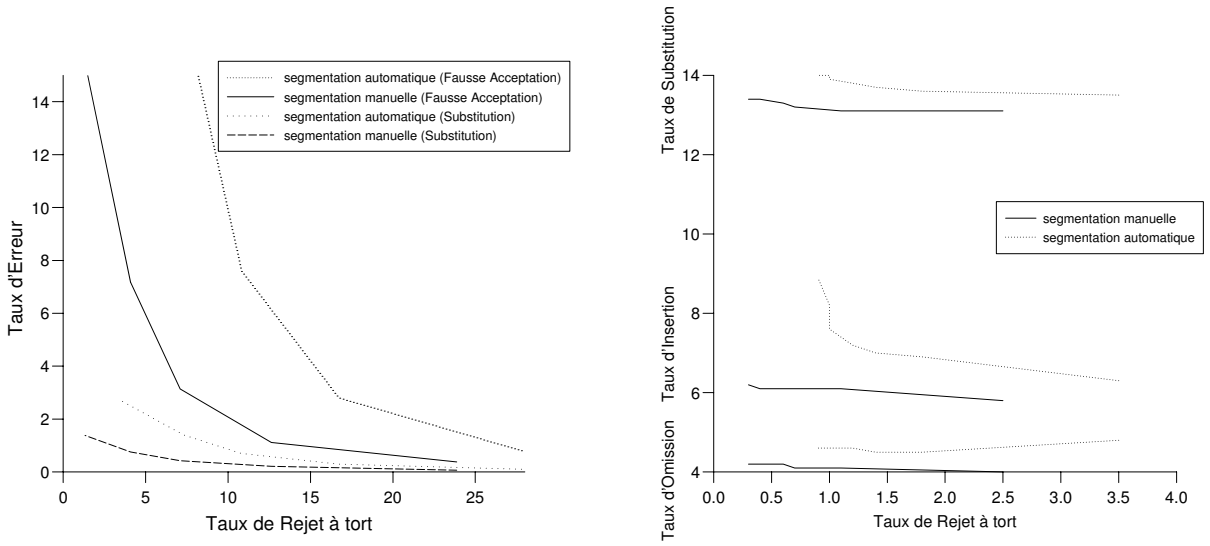
Sur la base de parole continue (la base AGORA décrite en Annexe D), nous remarquons (*cf.* figure 2.5(b)) que la détection automatique provoque plus d'erreurs d'omission, d'insertion, mais aussi de substitution de mots au niveau de la requête. En effet l'influence des premières erreurs qui peuvent être dues à la détection, peut ensuite provoquer des erreurs au milieu de la requête. Il y a également un peu plus de rejets à tort, pour un poids de rejet fixé. Notons que la différence des taux d'erreur entre la détection automatique et la segmentation manuelle est plus faible que pour la reconnaissance de mots isolés. Les taux de rejet à tort sont trois fois plus importants avec la détection automatique. Le taux d'omission est 20% plus élevé avec la détection automatique, tandis que les taux d'insertion et de substitution sont proches.

Une amélioration du module de détection ABP est donc nécessaire pour faire diminuer principalement les erreurs de fausse acceptation du module de reconnaissance de mots isolés, et tous les types d'erreurs pour la reconnaissance de parole continue. Pour mesurer l'amélioration du module de détection ABP il faut avant tout choisir un principe d'évaluation.

2.4 Principe d'évaluation de la détection de parole

L'évaluation de la détection de parole diffère énormément selon les auteurs. Elle doit permettre de rendre compte des performances de la détection, et une comparaison aisée des différents détections de parole. Notons que pour que l'évaluation ait réellement un sens, la base de test doit être suffisamment réaliste pour une application concrète.

L'évaluation d'un module de détection Bruit/Parole peut-être effectuée par comparaison à une segmentation supposée idéale, faite manuellement. Le résultat de ce module



(a) Sur base de mots isolés.

(b) Sur base de parole continue.

FIG. 2.5 – *Comparaison des résultats de reconnaissance avec la segmentation manuelle et une détection automatique.*

doit donc s'approcher de celui obtenu manuellement. Les différences de la détection automatique avec la segmentation manuelle donnent lieu aux erreurs précédemment définies, qui sont comptabilisées.

Cependant, même si ces erreurs sont définies en vue de caractériser le module de reconnaissance, le but du module de détection Bruit/Parole étant d'améliorer les performances du système de reconnaissance, un moyen de tester les limites de la détection Bruit/Parole est d'examiner les performances du système de reconnaissance obtenues avec le module de détection (*cf.* [Mauuary et Karray, 1997]). Nous évaluons ainsi les conséquences (présentées au paragraphe 2.3) des erreurs du module de détection sur le système de reconnaissance.

Dans le cadre de la détection de parole continue, la reconnaissance vocale est typiquement utilisée pour une application de dialogue homme-machine. Dans ce cas, une évaluation du module de détection pourrait se faire par une évaluation du système de dialogue qui comprend le système de reconnaissance, avec le module de détection à évaluer. Même si dans notre étude, la base de données utilisée pour la reconnaissance de parole continue est une base de données pour une application de dialogue homme-machine, nous n'évaluons pas le module de détection par l'évaluation des performances du système de dialogue. En effet, d'une part, il est possible de se placer dans un contexte plus vaste de reconnaissance de parole continue pour un système de dialogue quelconque, d'autre part le problème d'évaluation d'un système de dialogue est un problème complexe à lui seul.

Nous présentons donc l'évaluation de la détection par comparaison à la segmentation manuelle, puis l'évaluation de la détection dans le système de reconnaissance en différen-

ciant la reconnaissance de mots isolés et de parole continue. En effet, le paragraphe 2.3 montre que l'évaluation se fait différemment.Nous donnons ensuite la méthode retenue dans cette étude pour valider les résultats, et ainsi déterminer si une amélioration est significative. Enfin le dernier paragraphe permet de situer notre méthode d'évaluation parmi celles employées ces dernières années.

2.4.1 Évaluation de la détection par rapport à la segmentation manuelle

Les bases de données ont été segmentées manuellement, les périodes de parole et des bruits ont été étiquetées. Pour évaluer la détection à l'aide de la segmentation manuelle, il faut comparer les segments de test issus de la détection et les segments de référence issus de la segmentation. Un exemple de mise en relation de segments de référence issus de la segmentation manuelle (de R_1 à R_6) et des segments de test issus de la détection automatique (de T_1 à T_6) est donné par la figure 2.6.

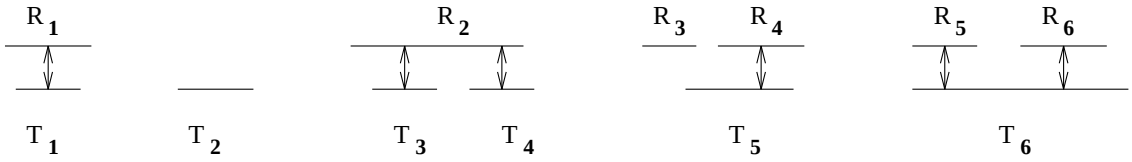


FIG. 2.6 – *Exemple de mise en relation des segments de référence et des segments de test.*

La règle de mise en relation est :

$$R_i \leftrightarrow T_j \quad si \quad ||L_{R_i \cap T_j}|| \geq \frac{\min(L_{R_i}, L_{T_j})}{2}, \tag{2.21}$$

où L_{R_i} et L_{T_j} sont respectivement les longueurs (en nombre de trames) des segments de référence et de test. Ainsi les segments référence et test sont reliés si leur recouvrement est plus grand que la moitié d'un des segments. Sur cet exemple nous avons :

$$\begin{array}{cccc} R_1 \leftrightarrow T_1 & R_2 \leftrightarrow T_3 & R_3 & R_5 \leftrightarrow T_6 \\ T_2 & R_2 \leftrightarrow T_4 & R_4 \leftrightarrow T_5 & R_6 \leftrightarrow T_6 \end{array}$$

Ainsi pour comptabiliser les erreurs, il faut considérer les segments de référence étiquetés, et les segments de test. Pour simplifier l'étude nous considérons que tous les segments de référence étiquetés sont de la parole. Les règles de comptage sont les suivantes :

- Si $R_i \leftrightarrow T_j$ alors, la détection est "correcte".
- Si $R_i \leftrightarrow rien$ alors, une omission est comptée.
- Si $rien \leftrightarrow Tj$ alors une insertion est comptée.

- Si $\begin{cases} R_i \leftrightarrow T_j \\ R_i \leftrightarrow T_{j+1} \\ \vdots \\ R_i \leftrightarrow T_{j+N} \end{cases}$ alors N erreurs de fragmentation sont comptées, et une détection est comptée "correcte".
- Si $\begin{cases} R_i \leftrightarrow T_j \\ R_{i+1} \leftrightarrow T_j \\ \vdots \\ R_{i+N} \leftrightarrow T_j \end{cases}$ alors N erreurs de regroupement sont comptées, et une détection est comptée "correcte".

Un segment de référence sera comptabilisé :
- "correct", s'il a été détecté et est correctement relié,
- "omission", s'il n'a pas été détecté,
- "fragmenté", si ce segment a été mis en correspondance avec plusieurs segments du test,
- "regroupé", si ce segment et d'autres ont été détectés comme un seul segment.

Nous avons vu au paragraphe 2.3.1 qu'il faut éviter les erreurs d'omission de parole, d'insertion de bruit ou de parole non utile, les erreurs de fragmentation et les erreurs de regroupement.

Une détection d'un segment de parole qui est correctement relié au segment de référence est considéré comme "correct", même si le segment de référence est tronqué ou élargi, à gauche ou à droite. Comme nous venons de le voir les erreurs d'insertion de bruit ou de parole peuvent être corrigées par un modèle de rejet. Nous considérerons donc les erreurs d'insertion comme des erreurs *rejetables* par le module de reconnaissance. Par contre les erreurs de regroupement, les erreurs d'omission et les erreurs de fragmentation seront considérées comme des erreurs définitives pour le module de reconnaissance. Les erreurs de fragmentation peuvent être récupérées dans le cas où un des segments est reconnu correctement et que les autres sont rejetés, ce cas de figure reste cependant idéal et improbable. C'est pourquoi ces erreurs sont considérées comme des erreurs *définitives*.

Ainsi pour comparer les performances de plusieurs algorithmes de détection Bruit/Parole sur une même base, nous représentons le taux d'erreur définitive en fonction du taux d'erreur rejetable par le module de reconnaissance. Le taux d'erreur rejetable est calculé en divisant le nombre d'erreurs rejetables par le nombre de segments *Parole* (*cf.* Glossaire), et multiplié par cent. De même, le taux d'erreur définitive est calculé en divisant le nombre d'erreurs définitives par le nombre de segments *Parole* (*cf.* Glossaire), et multiplié par cent.

2.4.2 Évaluation de la détection dans le système de reconnaissance

Nous distinguons le cas de la reconnaissance de mots isolés et de parole continue, car nous avons défini une terminologie différente pour les erreurs de la reconnaissance de mots

isolés et de parole continue. Comme nous l'avons vu au paragraphe 2.3.2 l'influence des erreurs du module de détection sur les erreurs de la reconnaissance de mots isolés et sur les erreurs de la reconnaissance de parole continue est différente.

Cas de la reconnaissance de mots isolés

L'évaluation des systèmes de reconnaissance se fait en considérant les trois types d'erreurs précédemment définis.

Ces erreurs du système de reconnaissance sont plus ou moins graves du point de vue de l'ergonomie du système. En effet les erreurs de rejet à tort, si elles ne sont pas trop nombreuses ne gêneront pas l'utilisateur du système qui devra simplement répéter sa commande. Par contre les erreurs de fausse acceptation et de substitution sont des erreurs graves. Elles vont entraîner une incompréhension du système.

Nous remarquons, au vu de la figure 2.3, que les erreurs d'omission n'entraînent que des erreurs de rejet à tort, considérées comme moins graves que les autres erreurs. Les omissions sont cependant considérées comme des erreurs définitives pour l'évaluation du module de détection par rapport à la segmentation manuelle, car elles conduisent à une erreur systématique du module de reconnaissance.

Afin d'évaluer les performances d'un module de détection intégré dans un système de reconnaissance nous représentons donc, les taux de fausse acceptation, ou de substitution, ou une combinaison de ces deux types d'erreurs en fonction du taux de rejet à tort (moins graves). En effet, il y a une relation entre les erreurs de rejet à tort et les erreurs de fausse acceptation et de substitution. Le taux de rejet à tort peut être considéré comme inversement proportionnel aux taux de fausse acceptation et de substitution. En effet lorsque le modèle de rejet entraîne beaucoup d'erreurs de rejet à tort, les erreurs de fausse acceptation et de substitution qui étaient commises sur ces segments rejetés ne le sont plus.

Les taux de rejet à tort et de substitution sont respectivement le nombre d'erreurs de rejet à tort et de substitution divisé par le nombre de segments de *Parole-Voc* (*cf.* Glossaire), puis multiplié par cent. Le taux de fausse acceptation est le nombre d'erreurs de fausse acceptation divisé par le nombre de segments *Parole-Hors-Voc* et *Non-Parole* (*cf.* Glossaire), puis multiplié par cent. La combinaison des taux de substitution et de fausse acceptation se fait par une somme pondérée, dont le coefficient de pondération pour le taux de substitution est le nombre de segments de *Parole-Voc* et *Non-Parole* divisé par le nombre total de segments de référence (*Parole-Voc*, *Parole-Hors-Voc* et *Non-Parole*), et le coefficient de pondération pour le taux de fausse acceptation est le nombre de segments de *Parole-Hors-Voc* et *Non-Parole* divisé par le nombre de total de segments référence.

Cas de la reconnaissance de parole continue

L'évaluation du système de reconnaissance de parole continue avec le module de détection de parole, se fait en considérant les quatre types d'erreurs définies au paragraphe 2.3.2.

Dans le cas de la parole continue les erreurs de rejet à tort sont également moins importantes du point de vue de l'utilisateur. Cependant le taux acceptable pour la reconnaissance de mots isolés, doit être revu à la baisse pour la parole continue. Il est en effet plus contraignant à l'utilisateur de répéter des phrases entières plutôt que des mots.

Les erreurs d'omission, d'insertion ou de substitution sont graves dans la mesure où la compréhension de la phrase peut être modifiée. Cependant s'il s'agit de mots courts, comme les articles, bien souvent le système de dialogue homme-machine associé à la reconnaissance ne sera pas perturbé.

Afin de disposer d'une représentation simple permettant d'évaluer le système de reconnaissance dans le cadre de la parole continue, nous adoptons une représentation des taux d'omission, des taux d'insertion, et des taux de substitution, ou une combinaison des trois, en fonction des taux de rejet à tort qui sont exprimés en omission de mots.

Les taux d'insertion, d'omission, de substitution et de rejet à tort sont calculés respectivement en divisant le nombre d'erreurs d'insertion, d'omission, de substitution et de rejet à tort par le nombre de mots total, puis multiplié par cent.

2.4.3 Signification des résultats

Afin de comparer deux modules de détection, avec une évaluation par comparaison à la segmentation manuelle, ou avec une évaluation du système de reconnaissance, les représentations des taux d'erreur précédemment citées fournissent une réponse. En effet la simple comparaison des deux courbes, si elles ne se croisent pas, permet de savoir lequel des deux modules présente le moins d'erreurs.

Cependant, les évaluations conduisent à estimer différents taux d'erreur, pour la détection ou la reconnaissance. Ces estimations se font sur des bases de données qui comportent un nombre important d'échantillons. Ce ne sont cependant que des estimations, et la question de leur validité se pose donc.

En Annexe C, nous présentons les principales approches permettant la validation des résultats dans le cadre de notre problème. Pour valider les résultats obtenus, nous utiliserons le calcul de l'intervalle de confiance à 95% en faisant l'hypothèse que l'espace d'observation des erreurs est un espace de Bernouilli : nous supposons que pour chaque segment observé, il y a erreur ou pas, que ce soit pour les tests de détection ou de reconnaissance.

2.4.4 Avantages de notre principe d'évaluation

La comparaison des méthodes de détection proposées par les auteurs de systèmes de détection est parfois délicate. En effet, la détection dépend beaucoup des conditions de la prise de son. Elle est d'autant plus aisée que l'environnement est peu bruité. Certains corpus utilisés ont été enregistrés dans un environnement calme, puis bruités par des corpus de bruits, voire par des bruits artificiels. Tandis que d'autres sont composés d'enregistrements dans différents environnements. De plus, il est préférable que le corpus servant pour les tests comprenne suffisamment de mots (prononcés par un grand nombre de personnes,

hommes et femmes avec différents accents, *etc.*). Lorsqu'il s'agit de détection appliquée pour la reconnaissance, les comparaisons sont soit faites uniquement sur les performances du module de détection, soit sur les performances de la reconnaissance, soit les deux.

Évaluation avec les résultats du module détection

L'évaluation des résultats de détection peut se faire par comparaison avec une segmentation manuelle ou visuellement avec le signal. La comparaison des frontières de la détection au signal du mot ou de la phrase peut donner un première idée des performances du module. Cette évaluation est cependant parfois utilisée exclusivement (*cf.* [Kobatake *et al.*, 1989], [Ying *et al.*, 1993], [Rangoussi *et al.*, 1993] et [Van Gerven et Xie, 1997]). Une comparaison visuelle sur le signal ne permet pas d'évaluer sur une grande base de données nécessaire à une évaluation rigoureuse. [Hahn et Park, 1992] et [Hariharan *et al.*, 2001] comparent le début et la fin de la détection à une segmentation manuelle, et estiment ainsi les détections trop tôt ou trop tard avec des taux d'erreur. Cependant une détection de parole pour un module de reconnaissance ne nécessite pas une détection parfaite. Ainsi dans [Puel, 1997] les erreurs sont exprimées en *ms*. Dans [Savoji, 1989] différents taux d'erreur sont classés en justes, acceptables, loin et inacceptables, selon l'éloignement des frontières avec la segmentation manuelle.

En plus de cette évaluation au niveau des frontières, les erreurs d'omission de parole sont également calculées dans [Dermatas *et al.*, 1991]. Il est important d'évaluer aussi les erreurs d'insertion qui peuvent provoquer des erreurs du module de reconnaissance, ainsi que les erreurs de regroupement et de fragmentation de la parole, qui sont un problème majeur pour le module de reconnaissance (*cf.* [De Souza, 1983]).

Cependant l'évaluation faite uniquement sur les résultats de détection ne permet pas d'évaluer le système de reconnaissance.

Évaluation avec les résultats du module reconnaissance

L'évaluation du module de détection peut se faire uniquement à l'aide des taux de reconnaissance (*cf.* [Shin *et al.*, 2000] et [Singh *et al.*, 2001]).

Dans [Junqua *et al.*, 1991] quatre modules de détection sont comparés à l'aide de deux systèmes de reconnaissance différents. En effet, un module de détection peut être plus robuste qu'un autre appliqué à un module de reconnaissance, alors que c'est le contraire s'ils sont utilisés par un autre module de reconnaissance. Un module est fondé sur les chaînes de Markov cachées (HMM), l'autre est fondé sur les principes d'alignement temporel de formes acoustiques (DTW), qui sont les principales approches des modules de reconnaissance avec les réseaux de neurones (*cf.* [Jouvet, 1988]). La comparaison est faite sur les performances de la reconnaissance uniquement, avec un corpus à RSB variable.

Cependant, cette évaluation ne permet pas de distinguer les erreurs dues seulement à la reconnaissance, ou au contraire une correction des erreurs de détection par le module de reconnaissance. Les erreurs de fausse acceptation permettent néanmoins d'évaluer l'effet des insertions du module de détection sur le module de reconnaissance (*cf.* [Ganapathiraju *et al.*, 1996]).

Évaluation avec les résultats des modules de détection et de reconnaissance :

Pour évaluer un module de détection il est donc important de l'évaluer d'une part à l'aide d'une comparaison à la segmentation manuelle, d'autre part à l'aide des résultats

du module de reconnaissance. Dans [Li *et al.*, 2001], une simple représentation de la détection et du signal permet dans un premier temps d'évaluer la position des frontières, puis un taux d'erreur du module de reconnaissance est donné. Dans [Seok et Bae, 1999], l'évaluation est faite plus précisément à l'aide de taux d'erreur de début et fin de frontières et du pourcentage de bonne reconnaissance. Pour une évaluation plus en relation avec le module de reconnaissance la déviation des frontières (exprimée en *ms*), puis les taux d'erreur du module de reconnaissance sont considérés dans [Shen *et al.*, 1998], [Huang et Yang, 2000] et [Yang et Hsieh, 2000].

Cependant ces évaluations ne permettent pas de préciser les différentes erreurs du module de détection (omissions, insertions, fragmentations et regroupements), mais aussi les erreurs du module de reconnaissance dues au module de détection (*cf.* [Mauuary et Karray, 1997] et [Karray et Martin, 2001]).

C'est pourquoi, notre protocole d'évaluation présenté précédemment est composé de deux parties. La première consiste en une évaluation de la détection par comparaison à la segmentation manuelle, la seconde est une évaluation de la détection à l'aide des résultats du module reconnaissance pour lequel la DBP est conçue. Les bases de données présentées en Annexe D, composées d'un grand nombre d'enregistrements dans différents environnements par un grand nombre de locuteurs permettent une évaluation rigoureuse. Pour obtenir des conditions plus critiques du niveau de bruits ambiants, que ces bases n'offrent pas, il a été ajouté des bruits enregistrés dans différents environnements (*cf.* paragraphe 3.7).

2.5 Conclusion

Ce chapitre présente le module de détection ABP fondé sur un automate à cinq états, ainsi que les trois critères du module détection ABP dans le contexte de la reconnaissance vocale utilisé à France Télécom R&D. Notre étude est fondée sur ce module de détection.

Les conséquences importantes du module de détection sur les performances du système de reconnaissance, montrent que nous ne pouvons étudier le module de détection en dehors de son contexte applicatif, la reconnaissance vocale. Il est de plus important d'améliorer les performances du module de détection pour une meilleure reconnaissance vocale. La figure 2.5 montre que la segmentation manuelle donne des taux d'erreur inférieur de près de 50%, en comparaison de la détection automatique obtenue avec la version LCT. Dans le cas de la reconnaissance de parole continue, la différence au niveau des taux de rejet à tort est de plus de 60%, et de près de 20% pour les taux d'omission.

Une fois les différentes erreurs définies, l'étude des conséquences des erreurs commises par le module de détection ABP sur le système de reconnaissance nous ont conduit à définir une méthode d'évaluation du module de détection. Il est en effet important d'évaluer un module de détection au sein du système de reconnaissance en plus d'une évaluation propre de la détection. Cette évaluation se divise donc en deux parties : la première partie est celle de l'évaluation des erreurs intrinsèques du module de détection et la seconde partie celle de l'évaluation des erreurs du système de reconnaissance utilisant le module

de détection. Cette évaluation diffère dans le cadre de la reconnaissance de mots isolés et dans le cadre de la reconnaissance de parole continue. Nous avons également choisi une méthode indispensable pour déterminer si les différences lors de l'évaluation sont significatives. La méthode utilisée dans cette étude est fondée sur l'intervalle de confiance à 95%. Notre méthode d'évaluation s'avère très complète en comparaison de celles employées ces dernières années.

Ce chapitre montre qu'il est important d'améliorer les performances d'un des trois critères actuels du module de détection ABP afin d'améliorer les performances du système de reconnaissance. Ainsi, à l'aide du protocole d'évaluation établi, nous étudions dans le Chapitre suivant 3 "*Analyse des sources d'erreurs du module de détection*" les principales sources d'erreurs du module de détection ABP. Ce chapitre permet ainsi de préciser où doivent être apportés les modifications afin de diminuer les erreurs commises par le module de détection ABP qui perturbent le système de reconnaissance.

Chapitre 3

Analyse des sources d'erreurs du module de détection

3.1 Introduction

Ce chapitre a pour but de déterminer les principales sources d'erreurs du module de détection ABP pour définir où doivent être apportées des améliorations. Ainsi, nous cherchons d'abord à dégager les principales erreurs commises par le module de détection, en particulier dans le cas de signal très bruité par des bruits stationnaires ou impulsifs et dans le cas de la détection de phrases. Ces deux cas particuliers de la détection s'avèrent être les plus délicats.

Notre étude est ici restreinte à l'étude du critère LCT du module de détection ABP (présenté au paragraphe 2.2.2). En effet, les sources d'erreurs du module de détection ABP sont les mêmes avec les deux autres critères SB et SBP.

Dans le paragraphe 3.2 nous présentons les différentes bases de données employées au cours de cette étude.

Les sources d'erreurs qui peuvent influencer le module de détection sont de plusieurs types. Tout d'abord, pour une application donnée, le choix du seuil de détection est important. Nous étudions donc l'influence du seuil de détection sur les erreurs de détection, puis sur les erreurs de reconnaissance (*cf.* paragraphe 3.3). L'environnement et le rapport signal à bruit (RSB) ont également une influence importante sur le fonctionnement du module de détection (*cf.* paragraphe 3.4). La détection de parole doit aussi être robuste au type de mots du vocabulaire de l'application (*cf.* paragraphe 3.5). Par exemple, certains mots courts, ou peu énergétiques sont plus souvent omis.

Nous abordons ensuite deux cas difficiles qui entraînent beaucoup d'erreurs, le cas de la détection de parole dans un milieu très bruité, et le cas de la reconnaissance continue. Dans un milieu très bruité une détection de la parole est plus délicate, d'une part à cause de bruits impulsifs (de courte durée), d'autre part à cause du faible RSB (*cf.* paragraphe 3.6). Pour réaliser l'étude de la détection de parole dans un milieu très bruité selon le RSB, nous bruitons une base de donnée avec différents niveaux de bruit, en considérant deux types de bruits stationnaires (*cf.* paragraphe 3.7).

Dans le cas de la reconnaissance de parole continue, une détection plus précise des frontières des périodes de parole est plus importante pour le module de reconnaissance (*cf.* paragraphe 3.8). En effet, le modèle de rejet pour la reconnaissance de parole continue est moins performant que dans le cas de la reconnaissance de mots isolés étudiée, à cause d'un plus grand nombre de mots de vocabulaire et notamment d'un grand nombre de petits mots. Ainsi, un segment tronqué ou élargi, peut entraîner facilement l'omission ou l'insertion d'un mot changeant le sens général de la phrase.

3.2 Présentation des bases de données

Dans ce chapitre, nous utilisons différentes bases de données. Toutes les bases de données sont décrites en détail en Annexe D. Nous présentons ici les principales caractéristiques de ces bases.

Les bases RTC_A et GSM_A ne sont utilisées que pour l'apprentissage des paramètres pour les différents critères du module de détection ABP. Les tests pour l'évaluation des résultats se font sur les bases GSM_T et RTC_T. Une dernière base utilisée dans ce chapitre est la base AGORA de parole continue enregistrée sur le réseau RTC. Nous l'utilisons principalement comme une base de test. Les bases de données sont constituées de communications enregistrées en continue, puis segmentées manuellement pour obtenir les segments de référence. Nous considérons trois étiquettes issues de la segmentation manuelle :

- *Parole-Voc* : qui sont les étiquettes correspondant aux mots du vocabulaire,
- *Parole-Hors-Voc* : qui sont les étiquettes correspondant à la parole hors vocabulaire,
- *Non-Parole* : qui sont les étiquettes correspondant à tous types de bruits.

3.2.1 La base RTC_A

La base RTC_A est une base d'exploitation enregistrée sur le réseau RTC à partir d'une application de serveur interactif vocal (SVI). Elle est composée de 999 appels pour une durée totale de 32 *h* 25 *min*. Le vocabulaire est composé de 25 mots. Le nombre de répétitions de chaque mot dépend de l'utilisateur du SVI. La segmentation manuelle est constituée de 58% de *Parole-Voc*, 13% de *Parole-Hors-Voc* et 29% de bruits divers (bruits de fond, rires, toux, bruits de combiné, *etc.*). Le nombre total de segments de référence issu de la segmentation manuelle est de 10021.

3.2.2 La base GSM_A

La base GSM_A est une base de laboratoire enregistrée sur le réseau GSM dans quatre environnements différents : intérieur, extérieur, véhicule à l'arrêt et véhicule roulant. Les locuteurs doivent répéter 53 mots de vocabulaire. Normalement chaque mot est répété une seule fois, excepté si il y a un bruit important pendant la prononciation du mot. Les nombres d'occurrences de chaque mot sont donc sensiblement identiques. Le corpus

est composé de 68% de *Parole-Voc*, 4% de *Parole-Hors-Voc* et 28% de bruits divers. Le nombre total de segments de référence issu de la segmentation manuelle est de 32042, donc trois plus important que la base RTC_A.

3.2.3 La base GSM_T

Cette base est également une base de laboratoire enregistrée sur le réseau GSM dans quatre environnements différents : intérieur, extérieur, véhicule à l'arrêt et véhicule roulant. Le mode d'enregistrement est le même que pour la base GSM_A. Le vocabulaire est constitué de 65 mots. Les 29558 segments issus de la segmentation manuelle sont répartis en 85% de *Parole-Voc*, 3% de *Parole-Hors-Voc* et 11% de bruits. Cette base employée pour les évaluations est très proche de la base GSM_A.

3.2.4 La base RTC_T

Cette base est une base de laboratoire enregistrée sur le réseau RTC. Pour la moitié de la base (notée RTC_T_L) les locuteurs lisent les mots du vocabulaire, pour l'autre moitié (notée RTC_T_R) les mots sont répétés. Le vocabulaire est constitué de 68 mots qui sont en grande partie identiques à ceux de la base GSM_T. L'écho très important sur cette base a été supprimé. Les segments issus de la segmentation manuelle sont au nombre de 13850, et répartis en 91% de *Parole-Voc*, 3% de *Parole-Hors-Voc* et 6% de bruits. Cette base est également employée pour les évaluations.

3.2.5 La base AGORA

La base AGORA est une base d'expérimentation d'une application de dialogue homme-machine, enregistrée sur le réseau RTC. C'est une base de parole continue. Nous l'utilisons principalement comme une base de test. Elle est composée de 64 enregistrements pour l'expérimentation d'une application de dialogue. Les 3115 segments de référence comprennent 12635 mots. Le vocabulaire du modèle de reconnaissance est de 1633 mots. Il n'y a pas de segments *Parole-Hors-Voc* pour cette base. Les segments *Parole* constituent 81% des segments de référence et les segments de bruits 19%.

3.3 Influence du seuil de détection

Nous présentons ici l'influence du seuil de détection du module de détection ABP, d'une part sur les erreurs de la détection, d'autre part sur les erreurs de la reconnaissance. En effet les erreurs varient énormément en fonction du seuil de détection de l'énergie. Notre étude est ici restreinte à l'étude des erreurs pour les fichiers de la base RTC_A. Nous finissons par une discussion sur ces résultats.

3.3.1 Les erreurs de détection

Nous présentons ici les principales erreurs de détection en comparaison à la segmentation manuelle du module de détection.

Dans un premier temps, nous détaillons les erreurs de détection en fonction du seuil. Dans un second temps, nous résumons une partie de ces erreurs à l'aide de la différenciation en erreurs rejetables et erreurs définitives.

Erreurs de détection détaillées

- Sur la figure 3.1 sont représentés les nombres d'erreurs en fonction du seuil de détection exprimé en dB. Nous considérons ici les erreurs de regroupement de segments de parole du vocabulaire *Parole-Voc*, les omissions de segments de parole du vocabulaire *Parole-Voc* et de parole hors vocabulaire *Parole-Hors-Voc*, les fragmentations de *Parole-Voc* et de *Parole-Hors-Voc* et les insertions de bruits (qui correspondent à des détections de segments de bruits étiquetés *Non-Parole* plus des insertions de segments non étiquetés).
- Nous constatons d'après cette figure que plus le seuil est grand, plus les nombres des insertions et des regroupements sont petits, et plus les nombres des omissions et des fragmentations sont grands. Comparativement aux autres erreurs, les erreurs d'insertion sont en très grand nombre. Nous remarquons qu'il y a un nombre d'omissions de parole hors vocabulaire (*Parole-Hors-Voc*) plus important que d'omissions de parole du vocabulaire (*Parole-Voc*), ces erreurs ne sont cependant pas des erreurs du point de vue du système de reconnaissance. En effet, une omission de *Parole-Hors-Voc* n'entraîne pas d'erreurs de reconnaissance, et peut même éviter une erreur de fausse acceptation. De même le nombre de fragmentations de *Parole-Hors-Voc* est plus important que le nombre de fragmentations de *Parole-Voc* et n'est pas une erreur pour le système de reconnaissance. La faible valeur des fragmentations de *Parole-Hors-Voc* pour les seuils élevés est due au fait qu'une des parties fragmentées n'est plus détectée.

Positionnement des frontières des détections

- La figure 3.2 est l'histogramme cumulé des segments selon le positionnement des frontières détectées des mots du vocabulaire (*Parole-Voc*) par rapport à la détection manuelle selon le seuil de détection exprimé en dB. Les courbes de gauche donnent le pourcentage cumulé de segments élargis (côté négatif) ou tronqués (côté positif), selon qu'ils sont respectivement avant ou après la trame 0 qui représente la frontière gauche de référence issue de la segmentation manuelle. De même, les courbes de droite donnent le pourcentage cumulé de segments élargis (côté positif) ou tronqués (côté négatif), selon qu'ils sont respectivement après ou avant la trame 0 qui représente la frontière droite de la segmentation manuelle. Nous rappelons cependant qu'une marge de sécurité est ajoutée en début (160 *ms*) et fin (240 *ms*) de détection.

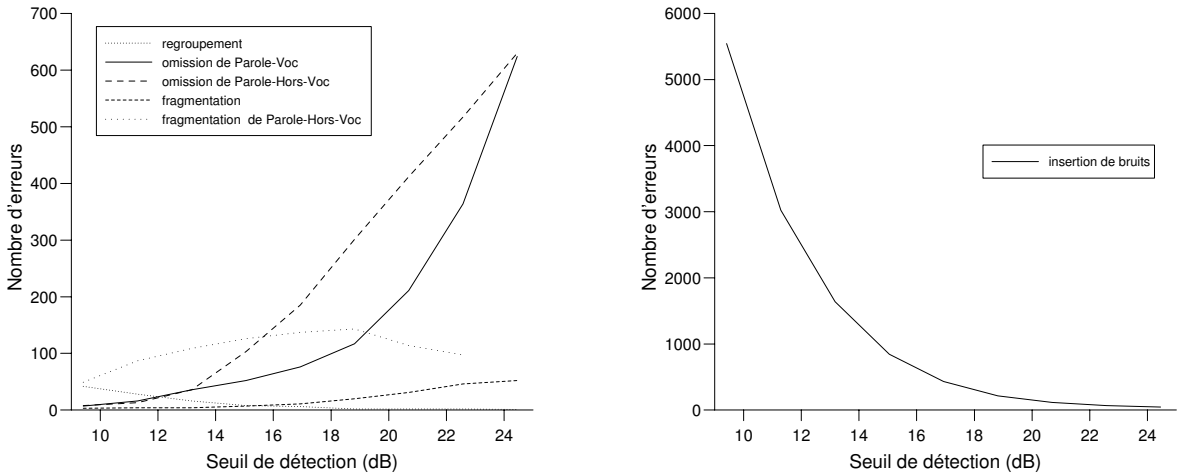


FIG. 3.1 – *Erreurs de détection détaillées, sur la base RTC_A.*

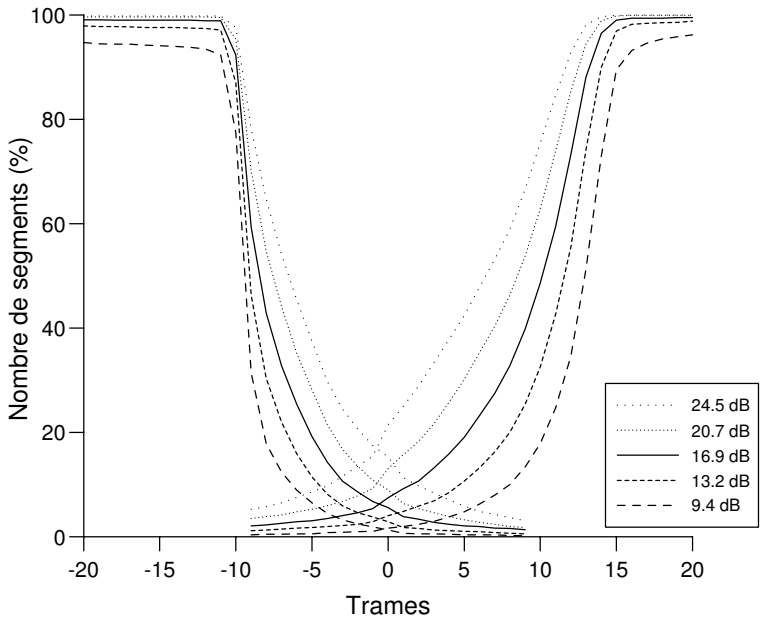


FIG. 3.2 – *Positionnement des frontières des détections sur la base RTC_A.*

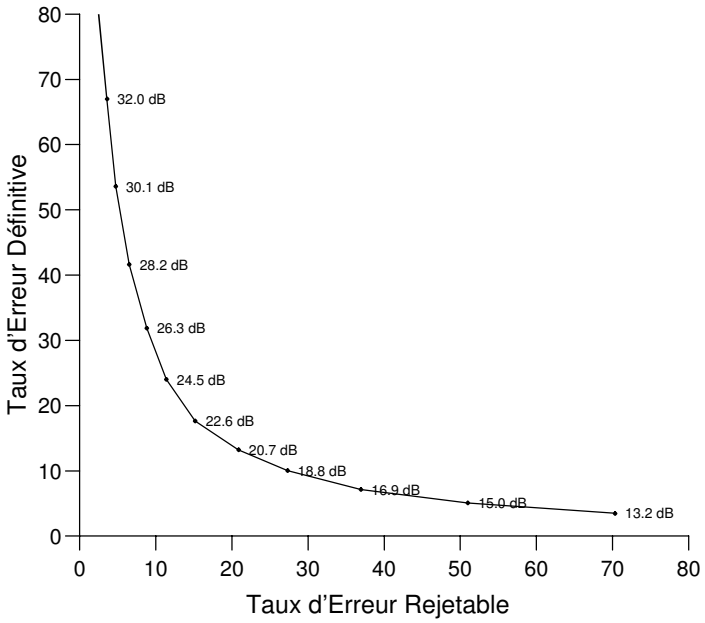


FIG. 3.3 – *Résultats de détection sur la base RTC_A.*

– Nous constatons que pour un seuil petit, le nombre de mots tronqués à gauche et à droite est petit, mais le nombre de mots élargis est grand. Au contraire pour un seuil élevé le nombre de mots tronqués est plus grand, mais le nombre de mots élargis est petit. Nous remarquons également qu'il y a plus de mots tronqués à droite qu'à gauche.

L'étude des erreurs, en fonction du seuil de détection pour cet algorithme, ne nous permet pas de donner un seuil optimal (qui minimise les erreurs) pour l'application au module de reconnaissance. Cependant la représentation simplifiée de la figure 3.3 des taux d'erreur définitive en fonction des taux d'erreur rejetable obtenue en faisant varier le seuil de détection indiqué sur la courbe et exprimé en dB, permet de situer ce seuil optimal entre 18.8 *dB* et 24.5 *dB*. Le minimum des taux d'erreur définitive et rejetable additionnée (taux d'erreur associée) est ici atteint pour un seuil de détection de 22.6 *dB* (*cf.* tableau 3.1 au paragraphe 3.4). Ce seuil correspond au seuil qui donne le minimum des taux d'erreur associée de la détection, or du point de vue de la reconnaissance les erreurs rejetables sont moins importantes que les erreurs définitives. Ce n'est donc probablement pas le seuil qui donne le minimum des taux d'erreur associée de la reconnaissance. Il est difficile de déterminer ce seuil uniquement à l'aide des résultats de la détection. Dans le paragraphe suivant nous allons voir pour quel seuil nous obtenons de meilleurs résultats de reconnaissance.

3.3.2 Les erreurs de reconnaissance

Le modèle de reconnaissance utilisé ici et par la suite est à base de modèles de phonèmes en contexte construit à partir de la description phonétique du vocabulaire (*cf.* paragraphe B.4 de l'Annexe B). Les figures 3.4(a), 3.4(b) et 3.4(c) sur la figure 3.4 représentent respectivement le taux de substitution, le taux de fausse acceptation, et une somme pondérée des deux, en fonction du taux de rejet à tort. Les courbes sont obtenues en faisant varier le poids du rejet du module de reconnaissance de -800 (M800) à 800 (indiqué sur une courbe). Sur ces figures apparaissent les résultats uniquement pour les principaux seuils de détection. Elles montrent que du point de vue de la reconnaissance, c'est le seuil 18.8 *dB* qui est optimal. Les seuils 16.9 *dB* et 20.7 *dB* sont cependant très proches. De plus, pour un poids de rejet fixé, tous les seuils représentés ont des résultats très proches.

La figure 3.4(a), montre qu'il y a très peu d'erreurs de substitution, de plus la figure 2.5 (du Chapitre 2 "*Détection de parole pour la reconnaissance vocale*") indique que le module de détection a peu d'influence sur ces erreurs dans le cas de reconnaissance de mots isolés. Nous adopterons dans le reste de cette étude, la représentation du taux de substitution combiné avec le taux de fausse acceptation. Cette représentation permet d'une part, de trouver le seuil qui donne le minimum des taux d'erreur associée de reconnaissance, et d'autre part, de comparer aisément les différents critères du module de détection ABP (*cf.* paragraphe 4.4).

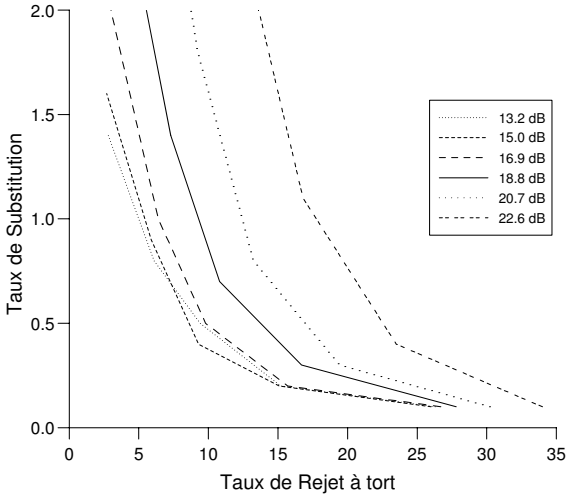
Nous remarquons que plus le poids du rejet est important, moins il y a de rejets à tort, au détriment d'une augmentation des fausses acceptations, et des erreurs de substitution.

Pour étudier plus en détail les différentes erreurs du module de reconnaissance issue du module de détection, nous nous limitons au seuil de détection 18.8 *dB*, pour un poids de rejet de 400 (qui présente le minimum des taux d'erreur associée pour un taux de rejet à tort inférieur à 10%, *cf.* figure 3.4). Pour ce seuil de 18.8 *dB*, 8302 segments ont été détectés, 119 segments de vocabulaire (2%) et 2017 (48%) des segments parole hors vocabulaire et bruit, n'ont pas été détectés.

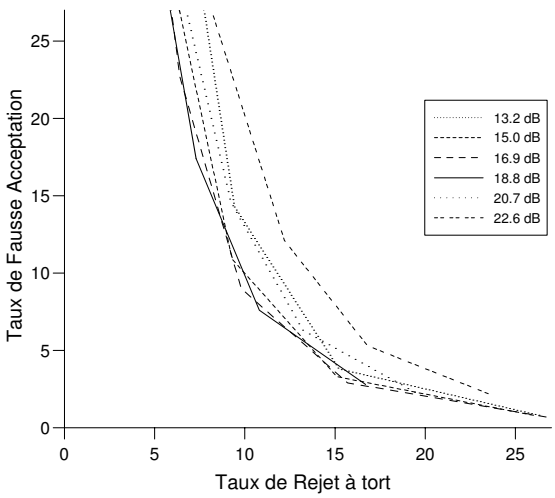
Erreurs de reconnaissance selon les résultats de détection

La figure 3.5 représente les erreurs de reconnaissance (substitutions, fausses acceptations et rejets à tort) en fonction du type de la détection (correct, regroupée (Reg.), fragmentée (Frag.), omission, insertion), sur la base RTC_A pour un seuil optimal de 18.8 *dB* et un poids de rejet de 400.

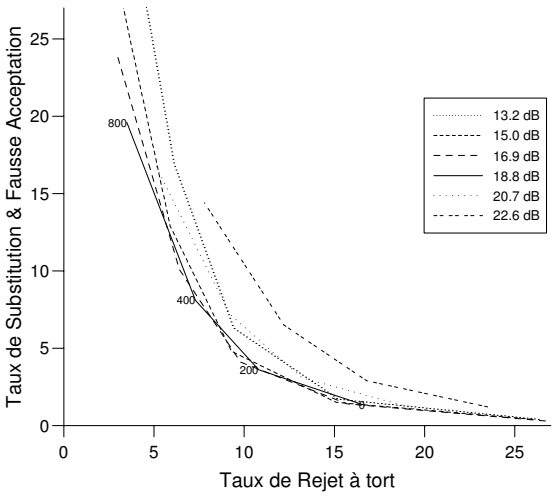
- Nous pouvons noter que les erreurs de reconnaissance associées lors d'un regroupement ou d'une fragmentation sont peu nombreuses. Cependant ces types d'erreurs de détection ne sont pas importantes pour ce seuil, en proportion il y aura donc plus de rejets à tort et de substitutions sur des détections regroupées ou fragmentées.
- Les omissions donnent des erreurs de rejet à tort, qui représentent près du tiers des rejets à tort totaux. Ces erreurs ne sont dues qu'au module de détection.
- Les erreurs de fausse acceptation qui proviennent uniquement des insertions restent en nombre important. Le modèle de rejet du module de reconnaissance n'a pu



(a) Taux de Substitution.



(b) Taux de Fausse Acceptation.



(c) Taux de Substitution et Fausse Acceptation.

FIG. 3.4 – *Résultats de reconnaissance sur la base RTC_A.*

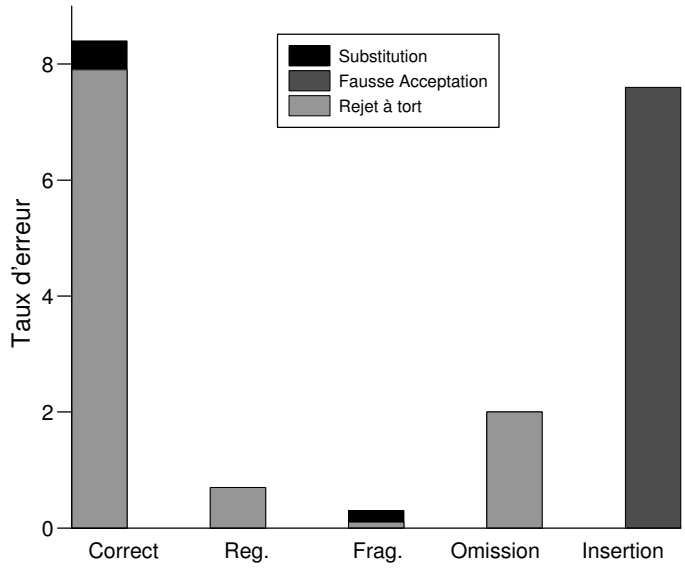


FIG. 3.5 – *Erreurs de reconnaissance selon les résultats de détection sur la base RTC_A.*

supprimer toutes les détections de bruits ou de parole hors vocabulaire.

– Notons encore une fois que les erreurs de substitution sont des erreurs marginales. Elles se produisent en majorité sur des détections correctement reliées.

Erreurs de reconnaissance selon le positionnement des frontières

Parmi les détections correctement reliées de *Parole-Voc* sont comptées les détections imprécises. Il est important de détailler les erreurs de reconnaissance en fonction du positionnement des frontières des détections de mots du vocabulaire. La figure 3.6 représente les pourcentages d'erreurs de reconnaissance (substitutions et rejets à tort) selon le positionnement des frontières : bien placée (B.P.), Tronquées à gauche (T.G.), c'est-à-dire que la détection a débuté après le début de la segmentation manuelle, tronquées à droite (T.D.), pour la fin du segment et tronquées à gauche et à droite (T.G.D.). Ces erreurs sont données pour un seuil de optimal de 18.8 *dB* et un poids de rejet de 400, sur la base RTC_A.

– Les erreurs de substitution sont proportionnellement plus présentes sur des segments tronqués à gauche et tronqués à gauche et à droite. Ceci peut s'expliquer par le vocabulaire employé ; par exemple "Suite" et "Huit".

– Les rejets à tort se produisent beaucoup plus sur les détections imprécises. Elles sont en proportion plus nombreuses sur des segments tronqués à gauche, et surtout à gauche et à droite où plus de 70% de ces segments sont rejetés à tort.

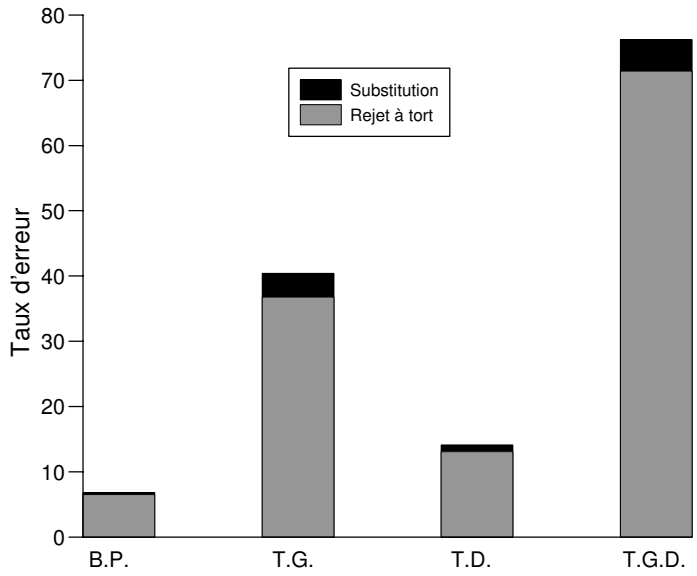


FIG. 3.6 – *Erreurs de reconnaissance selon le positionnement des frontières sur la base RTC_A.*

3.3.3 Sensibilité du seuil de détection

Du fait que le seuil de détection fasse varier les résultats de reconnaissance, il est important pour une application de pouvoir mesurer la sensibilité du seuil de détection au changement de base, au changement du réseau d'appel ou au niveau de bruit. Nous avons étudié la sensibilité du seuil de détection en Annexe H. Un critère de non-sensibilité est définit dans chaque cas. Le principe consiste dans un premier temps à calculer les taux d'erreur associée de reconnaissance sur une base A (ou partie de base), obtenus avec le seuil "optimal" de la base A et d'une base B (ou partie de base). Dans un second temps, nous déterminons l'intervalle de confiance à 95% du taux d'erreur associée de reconnaissance obtenu avec le seuil "optimal" de la base B. L'appartenance du taux d'erreur associée de reconnaissance obtenu avec le seuil "optimal" de la base A à l'intervalle de confiance définit le critère de non-sensibilité. L'intervalle de confiance est calculé à partir des segments *Parole-Voc*. Selon les bases ou parties de base A et B, nous pouvons déterminer la sensibilité du seuil de détection du critère LCT au changement de base, au changement du réseau d'appel ou au niveau de bruit.

Les résultats donnés en Annexe H montrent que le seuil de détection de ce critère est sensible au changement de base, au changement du réseau d'appel et au niveau de bruit.

3.3.4 Discussion

Pour l'évaluation du système de reconnaissance avec le module de détection, la représentation des substitutions et des fausses acceptations en fonction des rejets à tort, permet de déterminer le seuil qui donne le minimum des taux d'erreur associée de reconnaissance, et permet également une comparaison aisée. Le poids du rejet est à déterminer selon l'application. Un taux de rejet à tort de 10% est généralement accepté. Cependant cette évaluation est à compléter avec une représentation des erreurs de reconnaissance en fonction de la détection (*cf.* histogrammes 3.5 et 3.6).

Le seuil de détection fait varier les erreurs de détection et les erreurs de reconnaissance. Plus le seuil de détection est élevé, plus il y a d'omissions, et donc moins il y a d'insertions ; il y a ainsi plus d'erreurs définitives et moins d'erreurs rejetables. Le seuil fait aussi varier le positionnement des frontières de la détection. Nous avons vu que lorsque le mot détecté est tronqué, le module de reconnaissance commet plus d'erreurs surtout au niveau des rejets à tort, et en particulier pour les segments tronqués à gauche et à droite.

Le seuil donnant les meilleurs résultats de reconnaissance n'est pas le seuil optimal fournissant le minimum des taux d'erreur associée de détection, mais un seuil inférieur donnant davantage d'erreurs rejetables (insertions), qui sont en partie rejetées par le module de reconnaissance.

Nous avons également défini un critère de sensibilité du seuil de détection au changement de base, au niveau de bruit et au réseau d'appel.

Dans le paragraphe suivant, nous étudions les erreurs du module de détection, selon l'environnement et le RSB.

3.4 Influence de l'environnement et du RSB

Nous nous contentons ici d'étudier les environnements de la base GSM_A, *intérieur*, *extérieur*, *véhicule à l'arrêt* et *véhicule roulant* (*cf.* Annexe D). Le calcul du RSB défini en Annexe D, au paragraphe D.5, nous permet d'étudier les erreurs également en fonction des différents RSB.

En effet, les différents environnements traduisent plus ou moins bien des RSB différents. Par exemple l'environnement *intérieur* contiendra en général des appels ayant un RSB plus grand (c'est-à-dire moins bruités).

Pour cette étude nous allons procéder de la même manière que dans le paragraphe précédent. Nous étudions d'une part les erreurs de détection, et d'autre part les erreurs de reconnaissance ce qui amène une discussion sur ces résultats.

3.4.1 Les erreurs de détection

Pour simplifier l'étude, nous cherchons d'abord le seuil de détection qui donne le minimum des taux d'erreur associée de détection, pour étudier plus en détail les erreurs pour ce seuil.

	Seuils de détection (en *dB*)								
Bases	9.4	11.3	13.2	15.0	16.9	18.8	20.7	22.6	24.5
RTC_A	147.0	101.4	73.8	56.1	44.1	37.3	34.1	**32.8**	35.4
GSM_A *intérieur*	30.6	23.2	**20.8**	22.9	26.8	35.5	45.9	58.3	68.8
GSM_A *extérieur*	25.6	**22.4**	24.6	31.5	41.3	53.3	65.2	77.4	86.9
GSM_A *arrêt*	28.6	23.4	21.6	**20.9**	22.7	26.0	31.7	39.0	48.9
GSM_A *roulant*	24.5	**21.8**	24.1	28.8	34.2	44.8	55.1	64.2	73.1
RTC_A P20	123.7	82.0	55.4	40.0	29.7	23.7	21.4	**19.7**	21.3
RTC_A M20	177.7	127.0	98.1	77.4	63.1	55.4	50.8	**50.1**	54.0
GSM_A P18	20.7	15.4	**13.2**	14.4	16.8	21.7	28.9	38.9	49.9
GSM_A M18	37.8	**29.7**	31.6	36.3	43.6	55.2	66.6	76.7	85.3

TAB. 3.1 – *Taux d'erreur associée de détection selon le seuil sur les bases RTC_A et GSM_A.*

Taux d'erreur associée de détection

Le tableau 3.1 donne les taux d'erreur associée de détection en fonction du seuil de détection, pour les bases RTC_A et GSM_A selon l'environnement et le RSB.

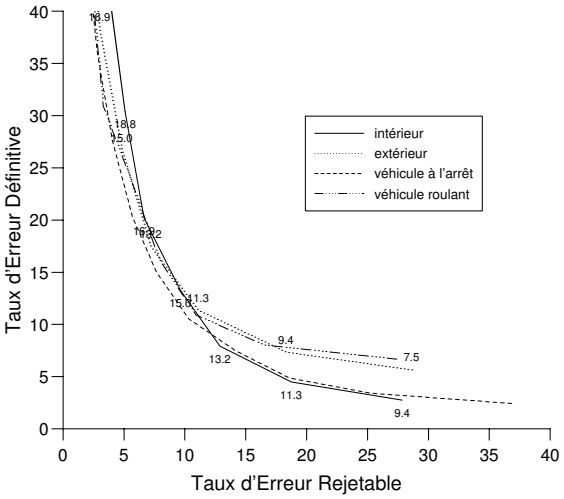
Ce tableau montre que le minimum des taux d'erreur associée n'est pas atteint pour un même seuil selon l'environnement. Plus l'environnement est bruité, plus le seuil doit être petit.

Nous constatons que pour la base RTC_A le minimum des taux d'erreur associée est plus élevé que pour la base GSM_A. Cela vient du fait que les fichiers enregistrés sur le réseau RTC_A sont enregistrés en contexte applicatif. La différence va donc se faire au niveau du nombre de détections de parole hors vocabulaire plus important pour la base RTC_A, des temps de pause plus long entre les mots (il faut attendre que le système réponde, tandis que pour la base de laboratoire, seul le mot à répéter est émis par le système), *etc.*

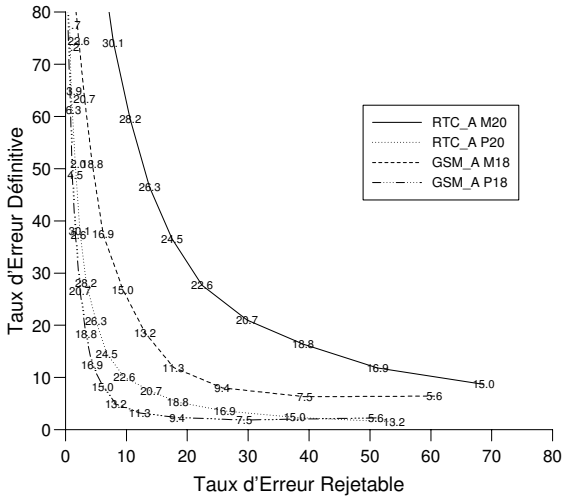
Les fichiers RTC_A avec moins de 20 *dB* de RSB et les fichiers GSM_A avec plus de 18 *dB* de RSB sont d'un niveau de bruit comparable (*cf.* figure D.2), d'après notre critère de calcul du RSB (*cf.* paragraphe D.5), cependant le minimum des taux d'erreur associée n'est pas atteint pour un même seuil, et le nombre d'erreurs est plus élevé pour la base RTC_A. Ceci s'explique par le fait que la base GSM_A est une base de laboratoire, tandis que la base RTC_A est une base d'exploitation. Pour un même RSB, il est difficile de comparer les deux bases.

Taux d'erreur définitive en fonction du taux d'erreur rejetable

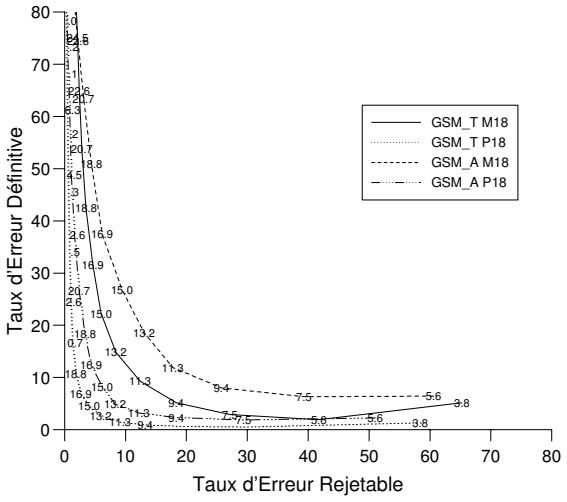
Nous représentons le taux d'erreur définitive en fonction du taux d'erreur rejetable sur la figure 3.7 pour les 4 environnements GSM_A, figure 3.7(a), pour les différents RSB de la base RTC_A et GSM_A, figure 3.7(b), et pour les différents RSB des deux bases GSM_A. Cette représentation permet une comparaison simplifiée des environnements et des RSB.



(a) Différents environnements de la base GSM_A.



(b) Différents RSB pour RTC_A et GSM_A.



(c) Différents RSB pour GSM_A et GSM_T.

FIG. 3.7 – *Résultats de détection sur les bases RTC_A, GSM_A et GSM_T.*

La figure 3.7(a) montre que seuls les environnements extérieur et véhicule roulant se démarquent des autres par des taux d'erreur définitive plus élevés, pour des taux d'erreur rejetable dépassant 10%. Ceci est bien sûr dû aux fichiers plus bruités contenus dans ces environnements.

La différence entre les 4 environnements d'appel est cependant moins marquée que sur la figure 3.7(b), où les fichiers sont classés selon le RSB. Rappelons que cette figure ne nous permet pas une comparaison entre les bases RTC_A et GSM_A. En effet, la base RTC_A est une base d'exploitation tandis que la base GSM_A est une base de laboratoire.

La figure 3.7(c) nous permet de comparer les bases GSM_A et GSM_T de laboratoire. Cette comparaison est intéressante, car nous constatons une différence entre deux classes de RSB. Même si les deux bases sont très proches, cette différence peut s'expliquer par un taux de mot de une syllabe plus important dans la base GSM_A, cette base contient aussi quelques signaux DTMF dans la communication qui peuvent être perturbateurs. Nous aborderons le problème important de l'influence des mots sur les erreurs de détection et de reconnaissance au paragraphe 3.5.

L'influence de l'environnement, et donc du niveau de bruit, a une grande importance au niveau de la détection de parole. La figure 3.7(a), représentant les erreurs pour les différents environnements d'appel, dégage d'une façon moins nette cette influence. Chaque environnement comporte des appels calmes, et d'autres plus bruités. Dans la suite de l'étude nous étudions l'influence du RSB qui présente plus d'intérêt que l'influence de la répartition de l'environnement d'appel de la base GSM_A.

Plus précisément, nous étudions les différentes erreurs du module de détection, en se limitant à l'étude des seuils qui donnent le minimum des taux d'erreur associée (notés en gras sur le tableau 3.1), pour chaque RSB de la base RTC_A et de la base GSM_A (*cf.* histogramme 3.8).

Erreurs de détection détaillées

L'histogramme 3.8 donne les erreurs de la détection sur les bases RTC_A et GSM_A pour différents RSB pour le seuil optimal de chaque cas considéré. Nous distinguons les mêmes erreurs de détection qu'au paragraphe 3.3 : les insertions, les regroupements de *Parole-Voc* (reg.), les omissions de *Parole-Voc* (omis. [PV]) et de *Parole-Hors-Voc* (omis [PHV]), et les fragmentations de *Parole-Voc* (frag. [PV]) et *Parole-Hors-Voc* (frag. [PHV]). Les pourcentages des insertions sont calculés en fonction du nombre de segments *Non-Parole* ; les pourcentages des regroupements, des omissions et des fragmentations de *Parole-Voc* en fonction des segments *Parole-Voc* ; et les pourcentages des omissions et des fragmentations de *Parole-Hors-Voc* en fonction des segments *Parole-Hors-Voc*.

- Nous remarquons d'abord que les pourcentages d'insertion et de regroupement sont très faibles quelque soit la base et le RSB.
- Nous constatons de plus que plus l'environnement est bruité plus il y a d'erreurs d'omission. Ceci n'est cependant valable que sur une même base, la base RTC_A, ou la base GSM_A.

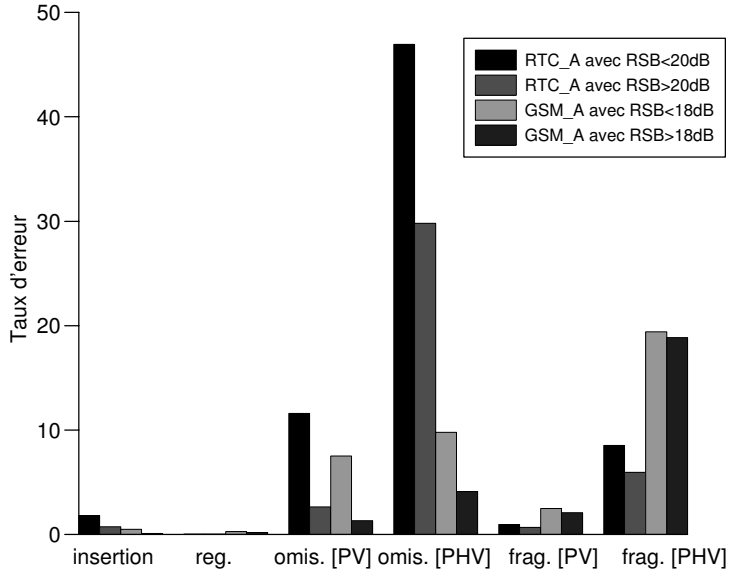


FIG. 3.8 – *Erreurs de détection détaillées sur les bases RTC_A et GSM_A.*

- Par contre les erreurs de fragmentation de *Parole-Voc* sont peu différentes sur une même base avec des RSB différents. Le RSB ne semble donc pas influencer beaucoup les erreurs de fragmentation de *Parole-Voc*.
- Sur les segments étiquetés *Parole-Hors-Voc*, plus le RSB est grand plus les erreurs de fragmentation sont faibles. Ceci peut provenir du fait que la compréhension est plus délicate lorsque le RSB est faible, l'utilisateur est amené à faire des commentaires plus longs. Notons également que la différence reste peu importante, surtout sur la base GSM_A.
- Notons encore qu'il y a plus d'omissions de *Parole-Hors-Voc* sur la base RTC_A. Cette différence s'explique car la base RTC_A étant une base enregistrée sur un système en activité, les personnes ont un comportement différent (apartés moins énergétiques, hésitations, ou plus de mots hors vocabulaire).

3.4.2 Les erreurs de reconnaissance

Nous étudions donc ici uniquement l'influence du RSB. Nous nous contentons de la représentation des substitutions combinées avec les fausses acceptations pour comparer les différents RSB. Le tableau F.3 en Annexe F donne les seuils optimaux pour la reconnaissance. La figure 3.9 représente pour ces seuils les erreurs de substitution et de fausse acceptation en fonction des erreurs de rejet à tort, selon la base et le RSB. Il est intéressant de noter que les résultats sont sensiblement les mêmes pour la base RTC_A avec moins de 20 *dB*, et pour la base GSM_A avec plus de 18 *dB*. En effet ces deux partitions des bases correspondent à un même niveau de RSB. Par contre pour les fichiers très bruités

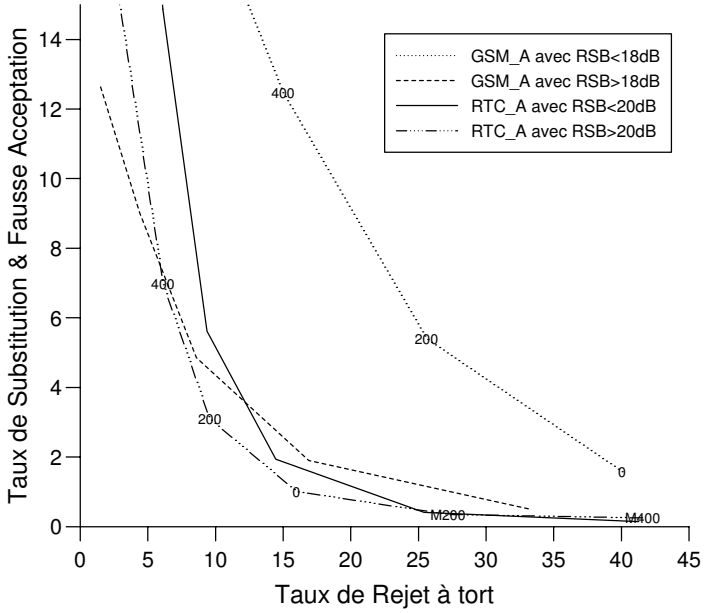


FIG. 3.9 – *Résultats de reconnaissance sur les bases RTC_A et GSM_A.*

de la base GSM_A (moins de 18 *dB*), les résultats sont dégradés. Cette dégradation est en partie due à la détection comme nous l'avons remarqué dans le paragraphe 3.4.1, mais aussi en partie due au module de reconnaissance. Une amélioration de la détection particulièrement pour ces fichiers bruités rendrait le système de reconnaissance plus robuste au bruit.

Par exemple, pour un poids de rejet fixé, les taux de rejet à tort sur la base RTC_A et sur la partie la moins bruitée de la base GSM_A sont proches et beaucoup moins importants que le taux de rejet à tort sur la partie bruitée de la base GSM_A. Ceci s'explique d'une part par le taux d'omission de segments plus important sur cette partie de la base, d'autre part par une moins bonne reconnaissance lorsque la parole est bruitée.

Erreurs de reconnaissance selon les erreurs de détection

Pour détailler ces résultats, l'histogramme 3.10, compare les résultats de reconnaissance pour les bases RTC_A et GSM_A selon RSB, et selon les résultats de détection pour un poids de rejet fixé à 400.

- Notons que la majorité des erreurs proviennent des insertions, qui provoquent des erreurs de fausse acceptation.
- Les erreurs de reconnaissance (rejets à tort et substitutions) sur les détections correctement reliées sont également importantes, en particulier pour la partie bruitée de la base GSM_A. Remarquons qu'il y a plus d'erreurs sur la partie calme de la base RTC_A que sur la partie bruitée. Ceci vient du choix du seuil optimal qui en contre partie donne moins d'erreurs de reconnaissance sur les segments insérés.

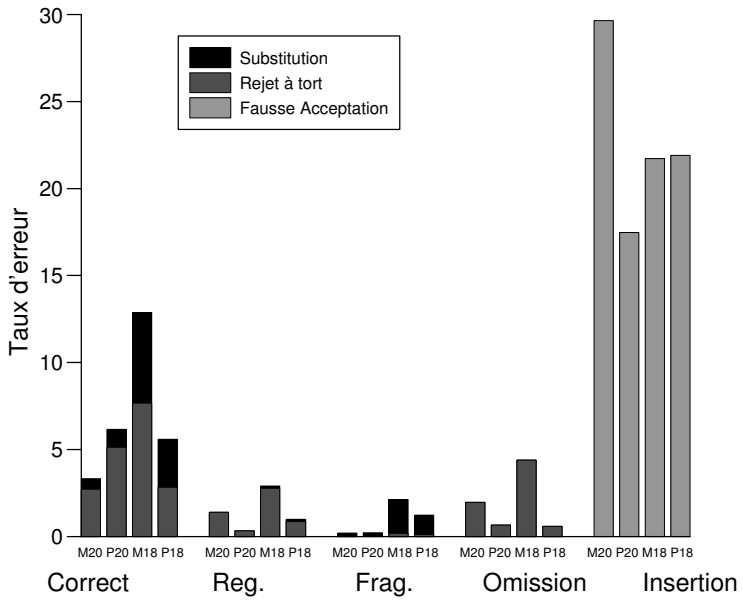


FIG. 3.10 – *Erreurs de reconnaissance selon les résultats de détection sur les bases RTC_A et GSM_A.*

- Sur la partie bruitée de la base GSM_A il y a également plus d'erreurs dues aux omissions et aux regroupements.
- Lorsque le RSB est petit, il y a plus d'erreurs de reconnaissance pour toutes les types de détections (excepté pour les détections correctement reliées de la base RTC_A).

3.4.3 Discussion

L'hétérogénéité des bases ne permet pas une comparaison entre elles. En effet, les conditions d'enregistrement (en exploitation ou en laboratoire, ce qui entraîne une différence comportementale des usagers du système), le type des mots du vocabulaire (par exemple différence du nombre de mots à une syllabe, *cf.* paragraphe 3.5), ont une influence importante sur les résultats de détection et de reconnaissance. Par contre les résultats précédents montrent que pour une même base l'influence du RSB est déterminante, et il est intéressant de comparer les résultats d'enregistrements ayant des RSB différents.

Les figures 3.7(c) et 3.9 montrent que dans le cas de la base GSM_A, pour un RSB inférieur à 18 *dB*, les résultats de la détection de parole et surtout du système de reconnaissance sont très dégradés. La différence des résultats de détection entre la partie calme de la base GSM_A de plus de 18 *dB*, et celle bruitée de moins de 18 *dB* est proportionnellement moins importante que la différence entre la partie calme et bruitée des résultats de reconnaissance. C'est-à-dire que les erreurs provoquées par la détection sur la partie bruitée n'expliquent pas toutes les erreurs provoquées par le module de reconnaissance.

Dans le paragraphe 3.6 nous étudions en détail ce cas critique.

3.5 Influence des types de mots du vocabulaire

Le vocabulaire est composé de différents types de mots, selon le nombre de syllabes ou bien selon les phonèmes qui constituent les mots. Ces types de mots ont une influence importante sur les résultats de la détection. En effet, certains mots peuvent créer plus d'omissions, comme le mot "Six" qui est souvent peu énergétique et qui est un mot d'une syllabe, ou de fragmentations comme le mot "Écouter". Des mots trop proches phonétiquement peuvent aussi être confondus par le module de reconnaissance, par exemple si la différence phonétique est perdue par une détection qui tronque un segment. Nous étudions d'abord l'influence du vocabulaire sur les erreurs de détection, puis sur les erreurs de reconnaissance, nous discutons enuiste de ces résultats.

3.5.1 Les erreurs de détection

Nous présentons dans cette partie les résultats de détection pour quelques mots de la base GSM_A avec un RSB supérieur à 18 *dB*. Ces résultats sont décrits dans le tableau E.1.

Les mots étudiés sont ceux qui présentent le plus d'erreurs et constituent des types de mots différents.

- Les mots courts et peu énergétiques selon la prononciation des locuteurs ont un taux plus fort d'omission ("Six", "Suite"). Cependant des mots courts de une à deux syllabes, mais énergétiques sont peu omis ("Trois", "Pause", "Quatre", "Sept", "Neuf", "Un").

  Nous remarquons également que "Supprimer" est souvent omis, ceci peut s'expliquer par un bruit important uniquement pendant la prononciation du mot sur quelques fichiers. Le mot devra donc être répété trois fois pendant cette période bruitée, ce qui suffit à accentuer les omissions de ce mot.
- Les mots longs avec un phonème peu énergétique au milieu, en général une plosive ("Consultation", "Consulter", "Modifier", "Écouter", "Enregistrer", "Répéter", "Répétition") donnent un taux de fragmentation plus important. Lorsque ces mots ne sont pas comptés fragmentés, ils peuvent être comptés comme tronqués ("Écouter", "Modifier" et "Répétition"). Les mots courts de une à deux syllabes tels que "Cinq", "Fin", "Un", "Guide", "Neuf", "Non", "Oui", "Pause" ou "Trois" ne sont pas fragmentés.
- Les mots commençant par des fricatives ou occlusives sont souvent tronqués à gauche ("Quitter", "Supprimer", "Stop", "Fin").
- Inversement les mots finissant par des fricatives ou occlusives sont plus souvent tronqués à droite ("Huit", "Neuf", "Sept", "Quatre", "Quarante Quatre", "Suite", "Cinq", "Six"). Les mots finissant par des "e" muet sont également tronqués à droite ("Pause", "Reprise"), ainsi que les mots finissant par "tion" ("Consultation", "Validation", "Répétition", "Information"), cette syllabe étant peu énergétique.

– Les mots commençant par des fricatives ou des occlusives et finissant par des fricatives, des occlusives ou des "e" muet, sont ainsi plus souvent tronqués à gauche et à droite ("Stop", "Sept", "Six", "Suite", "Cinq").

3.5.2 Les erreurs de reconnaissance

Les résultats de la reconnaissance par mot comprennent les bonnes reconnaissances du mot, les rejets à tort du mot (rejets à tort), et les substitutions du mot. Les résultats de reconnaissance par mots en fonction de la détection sont donnés dans les tableaux E.2 et E.3 en Annexe E. Dans le tableau E.2 sont rappelées les erreurs d'omission dues au module de détection uniquement et donnant des erreurs de rejet à tort.

– Tout d'abord notons que le pourcentage de reconnaissance est bien différent selon les mots. Ceci vient des différences de détection selon les types de mots, mais aussi des erreurs de substitution et de rejet à tort qui varient également selon les types de mots. Certains mots comme le mot "Quitter" ont des taux de reconnaissance beaucoup plus faibles que les autres mots.
– Les mots qui ont été beaucoup fragmentés, provoquent sur ces segments des erreurs de rejet à tort ("Répétition", "Répéter", "Écouter", "Enregistrer" et "Consulter") ou des erreurs de substitution ("Répétition", "Écouter"), et sont de fait mal reconnus.
– Les taux de rejet à tort sont importants pour certains mots, malgré une détection correctement reliée. Ces erreurs peuvent être dues au système de reconnaissance, mais aussi à une détection imprécise comme nous le verrons par exemple pour le mot " Quitter".
– Les erreurs de substitution sont en partie dues au module de reconnaissance pour la reconnaissance de mot isolé (par exemple "Répéter" et "Répétition"), et restent peu nombreuses. Notons tout de même quelles sont plus fréquentes pour les mots cours phonétiquement proches d'autres mots courts du vocabulaire. Par exemple les mots "Fin" et "Un" sont souvent substitués l'un à l'autre, sur des segments correctement reliés. Le mot "Fin" produit également beaucoup de substitutions avec le mot "Cinq", par contre le mot "Cinq" n'est que très peu substitué avec le mot "Fin". Le mot "Stop" est substitué avec le mot "Quatre" en majorité, le mot "Suite" avec "Sept", "Six" et "Huit", et le mot "Oui" avec "Huit" ou "Ouais", qui leur sont phonétiquement proches. Cependant ces erreurs de substitution peuvent être aggravées par le positionnement des frontières.
– En effet le tableau E.3, montre que les substitutions du mot "Quitter" viennent des segments tronqués à gauche tandis que celles des mots "Répétition" et "Consultation" viennent des segments tronqués à droite. En revanche les substitutions des mots "Oui" et "Huit" et des mots "Fin" et "Un" viennent en grande partie du module de reconnaissance uniquement. Une faible partie des substitutions de "Fin" (avec "Un") vient cependant des segments tronqués à gauche et une partie de celle de "Huit" (avec "Oui") viennent des segments tronqués à droite.

– Lorsque les mots sont tronqués, ils sont le plus souvent rejetés. Nous avons constaté précédemment que les troncatures dépendent des caractères phonétiques des débuts et des fins de mots. Ainsi les mots commençant par des fricatives ("Stop", "Supprimer") ou des occlusives ("Quitter") sont plus rejetés. Les segments tronqués à gauche du mot "Écouter" ont aussi été rejetés en grand nombre. Les mots tronqués à droite étant en nombre moins important le taux de rejet à tort reste faible, excepté pour le mot "Huit". Par contre les erreurs de rejet à tort sont en proportion plus importante sur les mots finissant par des fricatives ou des occlusives. De même pour les mots commençant et finissant par des fricatives ou des occlusives sont également plus rejetés.

3.5.3 Discussion

Ce paragraphe permet de mettre en évidence l'importance du choix des types de mots du vocabulaire pour un système en exploitation. Les types de mots influencent aussi bien les résultats de la détection des mots, que leur reconnaissance.

Les mots courts et peu énergétiques provoquent des omissions, les mots longs avec une syllabe peu énergétique en milieu de mot (occlusive) provoquent des fragmentations, les types de mots commençant par des fricatives ou des occlusives, ou finissant par des fricatives, des occlusives, des "e" muets ou "tion" sont plus souvent tronqués. Ces erreurs du module de détection entraînent des erreurs du module de reconnaissance ; si le mot tronqué ou fragmenté est proche phonétiquement d'un autre mot du vocabulaire il est davantage substitué sinon il est davantage rejeté. Le module de reconnaissance se trouvera également en difficulté face à des mots correctement reliés, mais proches phonétiquement d'autres mots du vocabulaire, produisant des erreurs de substitution.

La simple différence de quelques mots entre deux bases différentes, avec les mêmes conditions d'enregistrement, ne nous permettra pas une comparaison aisée des bases. Il est cependant difficile d'évaluer la robustesse d'un module de détection face aux types phonétiques de mots utilisés dans le vocabulaire. En effet il faut d'abord définir précisément les types phonétiques de mots provoquant des erreurs, ce que nous avons fait brièvement, et évaluer ensuite la robustesse d'un ou plusieurs paramètres de l'algorithme agissant sur ces erreurs. Nous nous contentons ici de cette brève étude pour faire apparaître le problème de l'influence des types de mots du vocabulaire.

Nous avons étudié en priorité les types de mots du vocabulaire de la partie calme de la base GSM_A afin de ne pas subir l'influence du bruit. En effet sur la partie bruitée de la base, nous observons un grand nombre d'erreurs de détection et de reconnaissance dues aux bruits. Dans le paragraphe suivant 3.6 nous étudions ce cas particulier.

3.6 La détection sur une base bruitée

Nous avons remarqué dans le paragraphe 3.4 que les résultats sur la base GSM_A pour un RSB inférieur à 18 *dB* sont très dégradés. Dans ce paragraphe nous étudions

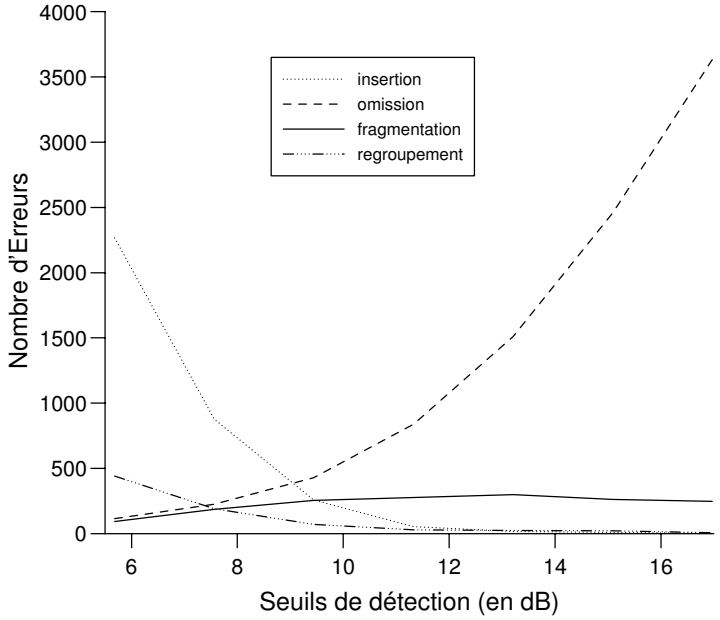


FIG. 3.11 – *Erreurs de détection détaillées sur la partie bruitée de la base GSM_A.*

la provenance des principales erreurs pour les fichiers de cette partie bruitée de la base GSM_A, décrite en Annexe D.

Comme précédemment nous étudions d'abord en détail les résultats de la détection puis de la reconnaissance, puis nous analysons ces résultats dans leur globalité.

3.6.1 Les erreurs de détection

Rappelons que la partie bruitée de la base contient 10414 segments de *Parole-Voc*, 662 de *Parole-Hors-Voc* et 4569 de *Non-Parole*.

Le tableau 3.1 montre que le taux d'erreur associée de détection est bien plus important sur cette partie de la base que pour des enregistrements moins bruités. L'histogramme 3.8 montre que ceci vient surtout des omissions pour les erreurs définitives et des insertions également plus nombreuses.

La figure 3.11 présente les erreurs de détection selon le seuil de détection, et l'histogramme 3.12, le positionnement des frontières sur les détections correctement reliées de *Parole-Voc*.

Le seuil de détection fait varier les différents types d'erreur de la même façon que sur la base RTC_A.

La figure 3.11 montre que les erreurs d'omission et d'insertion qui évoluent inversement sont les plus nombreuses. Les erreurs de fragmentation varient peu mais restent en nombre non négligeable. Les erreurs de regroupement restent peu nombreuses.

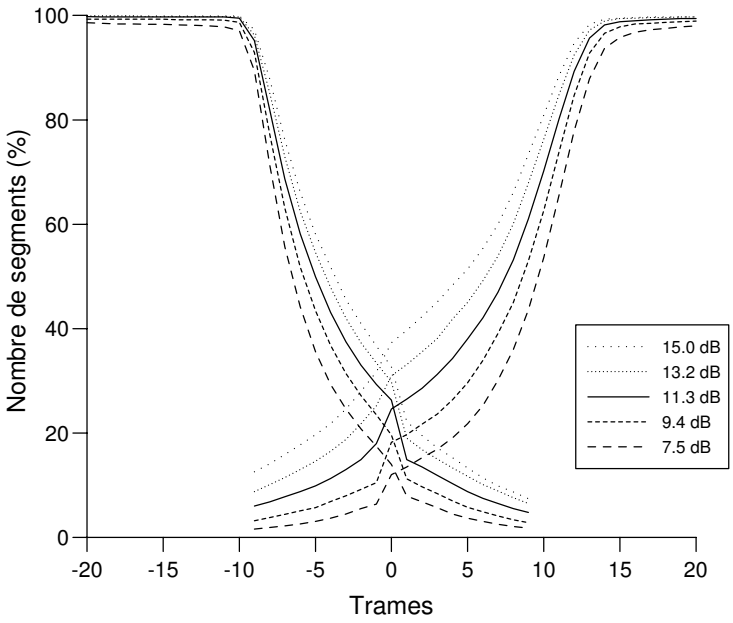


FIG. 3.12 – *Positionnement des frontières des détections sur la partie bruitée de la base GSM_A.*

L'histogramme 3.12 montre qu'il y a un grand nombre de détections imprécises quel que soit le seuil. Notons qu'il y a plus de segments tronqués à droite que de segments tronqués à gauche, et également plus de segments élargis à droite que de segments élargis à gauche. La détection à gauche est cependant peu précise.

Ces erreurs de détection vont avoir une influence sur les erreurs du système de reconnaissance. Nous étudions dans ce qui suit les erreurs de reconnaissance selon la détection.

3.6.2 Les erreurs de reconnaissance

Nous présentons ici les résultats de reconnaissance pour un poids de rejet fixé à 400. Le poids du rejet a été choisi de façon à ne pas produire trop de rejets à tort. Le taux de rejet à tort produit est cependant très important pour cette partie de la base (*cf.* figure 3.9). La variation du poids du rejet fera varier les résultats de la même façon que pour la base RTC_A, précédemment étudiée au paragraphe 3.3.2.

Les taux d'erreur de reconnaissance de l'histogramme 3.13 sont présentés en fonction du seuil de détection exprimé en *dB* (5.6, 7.5, 9.4, 11.3, 13.2, 15.0), sur la base GSM_A avec un RSB inférieur à 18 *dB* pour un poids de rejet de 400.

Les erreurs de substitution, surtout présentes lors de détections correctement reliées, augmentent avec le seuil, car les segments sont alors plus tronqués (*cf.* figure 3.12). Les erreurs de substitution sur les détections fragmentées évoluent peu avec le seuil.

Les erreurs de rejet à tort sont importantes sur cette partie de la base pour des détections correctement reliées, mais aussi à cause des omissions dues au module de détection.

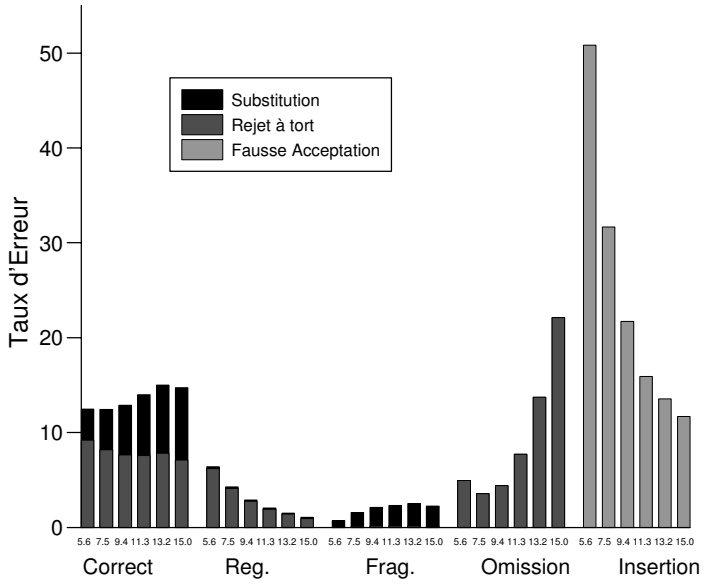


FIG. 3.13 – *Erreurs de reconnaissance selon les résultats de détection sur la partie bruitée de la base GSM_A.*

Les erreurs de regroupement et de fragmentation ne provoquent proportionnellement au reste que peu d'erreurs de rejet à tort.

Les erreurs de fausse acceptation sont nombreuses et sont la première cause d'erreurs du module de reconnaissance. Le modèle de rejet du système de reconnaissance a permis de rejeter un grand nombre des insertions dues au module de détection, il est cependant moins performant sur un signal bruité.

L'histogramme 3.14 permet d'étudier les erreurs de reconnaissance selon le positionnement des frontières des détections correctement reliées des segments de *Parole-Voc* selon le seuil de détection exprimé en *dB* (5.6, 7.5, 9.4, 11.3, 13.2, 15.0), sur la base GSM_A avec un RSB inférieur à 18 *dB* pour un poids de rejet de 400. Les taux sont des pourcentages par rapport au nombre total des détections B.P., T.G., T.D. et T.G.D.

Nous avons vu sur la figure 3.12 que le nombre de segments tronqués augmente avec le seuil mais le taux d'erreur associée de reconnaissance sur ces segments tronqués diminue. En effet, même si nous observons une augmentation du taux de substitution avec le seuil, le taux de rejet à tort diminue davantage. Effectivement, le nombre de segments tronqués est plus faible pour les seuils petits, mais les segments fortement tronqués sont en proportion plus nombreux (*cf.* figure 3.12), ainsi le taux de rejet à tort sera plus important sur les segments tronqués pour les seuils faibles.

Nous remarquons que les erreurs de rejet à tort et de substitution sur les segments tronqués sont proportionnellement plus importantes que sur les détections bien placées. En effet les erreurs de reconnaissance sur les segments tronqués représentent environ

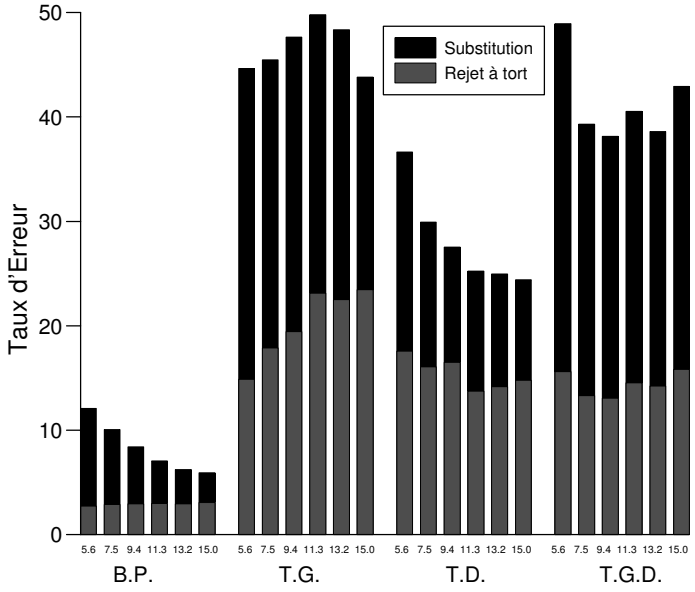


FIG. 3.14 – *Erreurs de reconnaissance selon le positionnement des frontières sur la partie bruitée de la base GSM_A.*

55% des erreurs des détections correctement reliées, alors que les segments tronqués ne représentent que 32% des détections correctement reliées.

3.6.3 Discussion

Sur cette partie de la base, où il y a un grand nombre de bruits impulsifs (de quelques trames), les insertions sont nombreuses. En partie rejetées par le modèle de rejet du module de reconnaissance, elles provoquent cependant un grand nombre de fausses acceptations. C'est la source principale des erreurs de reconnaissance. Cependant, l'histogramme 3.10 montre que les erreurs de reconnaissance dues aux insertions ne sont pas plus nombreuses sur cette partie de la base. Par contre les erreurs dues aux omissions, et les erreurs de reconnaissance sur les détections correctement reliées, qui sont en grande partie dues aux détections imprécises (*cf.* histogramme 3.14) sont bien plus importantes avec le RSB inférieur à 18 *dB*.

Nous allons maintenant répondre à la question : "Pourquoi les résultats sont si dégradés sur cette partie plus bruitée de la base?". Dire simplement que c'est parce que justement, c'est plus bruité, ne nous permet pas d'envisager une solution. En étudiant les fichiers, nous constatons que même si la moyenne de l'énergie de la parole est un peu plus importante dans des conditions bruitées, car l'utilisateur va parler plus fort (Effet Lombard), la différence de la moyenne du bruit et de la parole devient plus réduite sur cette partie de la base. Rappelons également que la prise de son entraîne une diminution du RSB sur le réseau GSM. Cependant à partir d'un niveau de bruit, ici atteint, l'écart entre l'énergie

de la parole et du bruit devient insuffisant pour une détection fondée sur le RSB utilisé avec ce critère. Le niveau de bruit va provoquer des erreurs d'omission et des détection imprécises entraînant des erreurs de rejet à tort et de substitution. De plus les insertions dues aux bruits impulsifs (de quelques trames), sont plus nombreuses sur cette partie de la base. Les insertions ne provoquent cependant pas plus d'erreurs que sur la partie calme, même si elles restent en très grand nombre.

Afin de préciser l'effet du bruit sur les erreurs du module de détection, et sur le système de reconnaissance, nous proposons dans le paragraphe suivant une étude de ces erreurs avec un ajout contrôlé de deux bruits types. Ceci nous permet de définir un critère de robustesse du module de détection face au bruit (*cf.* paragraphe H.2).

3.7 La détection selon le niveau de bruit

Pour évaluer l'influence du niveau de bruit sur les performances du module de détection, nous ajoutons deux types de bruit sur la partie calme de la base GSM_A (*cf.* paragraphe D.2), avec différents rapport signal à bruit (RSB). Un premier bruit est le bruit de conversations (*babble*), il n'est pas possible de comprendre ce qui est dit. Ce bruit est stationnaire. Le second bruit est le bruit enregistré à l'intérieur d'une voiture (*car*). Ce bruit est également stationnaire, mais avec des zones non-stationnaires.

Le RSB est ici aussi calculé sur les périodes de parole étiquetées manuellement et sur le bruit existant sur le signal d'origine et le bruit ajouté. Nous considérons toujours que la parole est bruitée par un bruit moyen calculé sur les périodes de signal non étiquetées (*cf.* Annexe D.5). Le bruit n'est ajouté qu'à la partie calme de la base GSM_A (*i.e.* la partie dont les fichiers ont un RSB supérieur à 18 *dB*).

Nous présentons les erreurs du module de détection, d'une part les erreurs de détection, d'autre part les erreurs de reconnaissance, ce qui amène à une discussion.

3.7.1 Les erreurs de détection

La figure 3.15 montre que plus le RSB est faible plus les taux d'erreur sont grands. L'étude peut se réduire aux RSB de 15 *dB* à 10 *dB*, au delà le module de détection ne permet pas de détecter la parole avec des performances suffisantes pour une application de reconnaissance vocale. En effet pour un RSB de 10 *dB* le taux d'erreur définitive ne descend pas en dessous de 35%. Les seuils indiqués sur la figure correspondent aux seuils qui donnent le minimum des taux d'erreur associée de détection (erreurs rejetables et définitives). Nous remarquons que pour les deux bruits ajoutés à différents RSB, le seuil optimal ne varie pas, ou très peu (pour le bruit *car* avec un RSB de 10 *dB*).

Nous pouvons déjà constater que le bruit *car* dégrade davantage les résultats de détection que le bruit *babble*.

Les figures 3.16 détaillent les erreurs du module de détection selon le seuil de détection pour un RSB de 10 *dB* à 15 *dB* (les insertions sur la figure 3.16(a), les omissions sur la figure 3.16(b), les fragmentations sur la figure 3.16(c) et les regroupements sur la figure

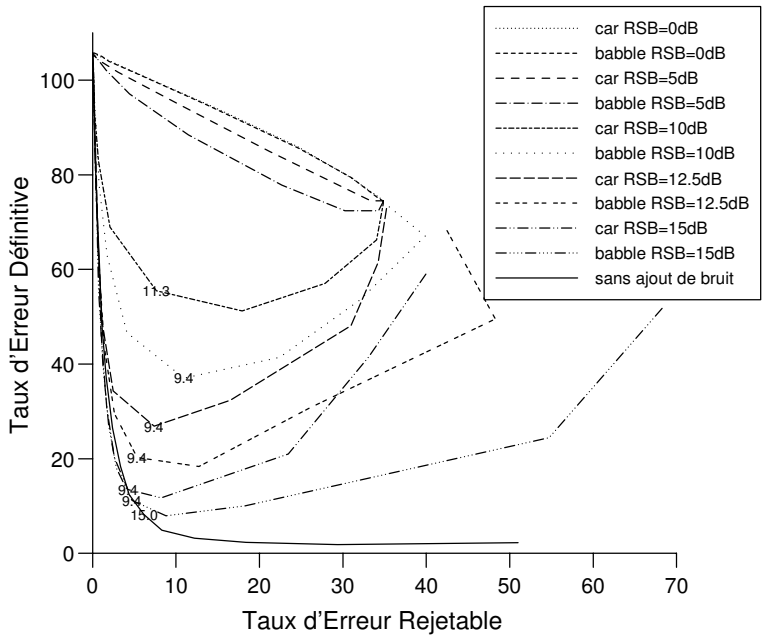
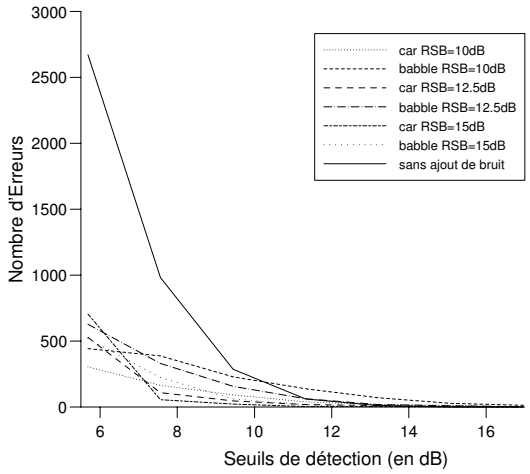


FIG. 3.15 – *Résultats de détection sur la base GSM_A bruitée.*

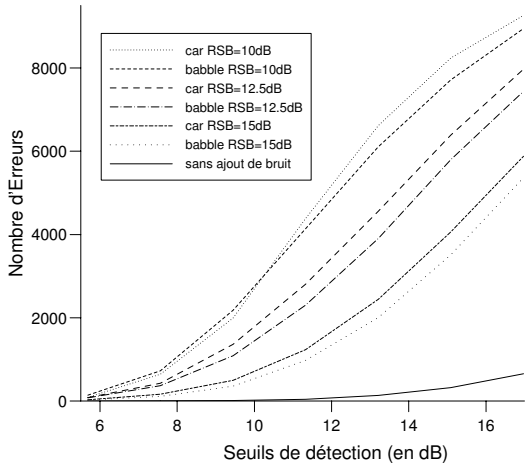
3.16(d)).

Résultats de détection détaillés

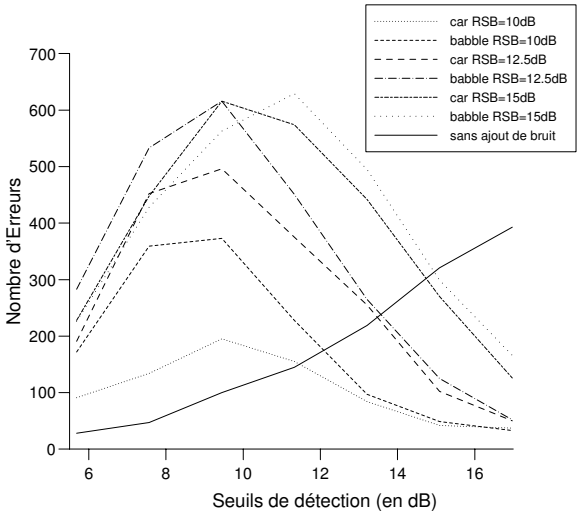
- Nous remarquons d'abord que le bruit *babble* donne moins d'erreurs que le bruit *car* pour les erreurs d'omission et de regroupement, alors que c'est l'inverse pour les erreurs d'insertion et de fragmentation qui restent peu nombreuses.
- Les erreurs d'omission et de regroupement de *Parole-Voc* sont en très grand nombre comparées aux erreurs de fragmentation et d'insertion. Ces erreurs augmentent proportionnellement avec le RSB, et l'écart entre le bruit *car* et *babble* reste le même. Les erreurs d'omission augmentent plus vite avec le seuil pour un RSB faible, alors que les erreurs de regroupement diminuent plus vite pour un RSB fort. C'est principalement sur ces deux erreurs que le niveau de bruit agit. Ces erreurs sont faibles lorsque aucun bruit n'est ajouté.
- Les erreurs de fragmentation restent peu nombreuses. Plus c'est bruité, moins il y a d'erreurs de fragmentation. Ceci peut s'expliquer par un nombre plus important d'omissions, lorsque le signal est bruité. En effet, pour un RSB donné, le maximum d'erreurs de fragmentation marque le début de la croissance plus importante des erreurs d'omission, quel que soit le bruit. Nous pouvons remarquer que ces erreurs ont un comportement différent suivant le seuil de détection lorsque aucun bruit n'est ajouté. Elles sont moins importantes et en nombre croissant.
- Les erreurs d'insertion diminuent avec le niveau de bruit. Elles diminuent avec le seuil de détection d'autant plus vite que le niveau de bruit est faible. Le bruit *babble* donne plus d'erreurs d'insertion, car ce bruit est physiquement plus proche



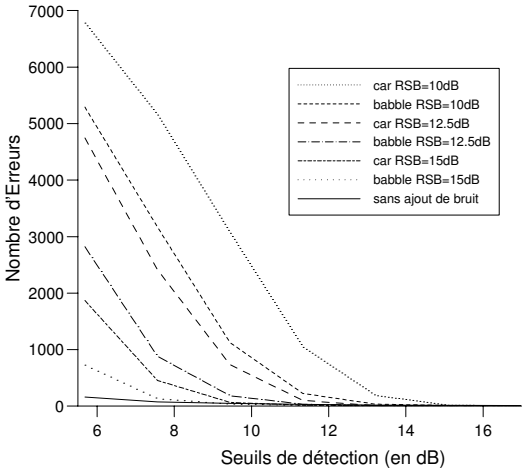
(a) Insertions.



(b) Omissions.



(c) Fragmentations.



(d) Regroupements.

FIG. 3.16 – *Erreurs de détection détaillées sur la base GSM_A bruitée.*

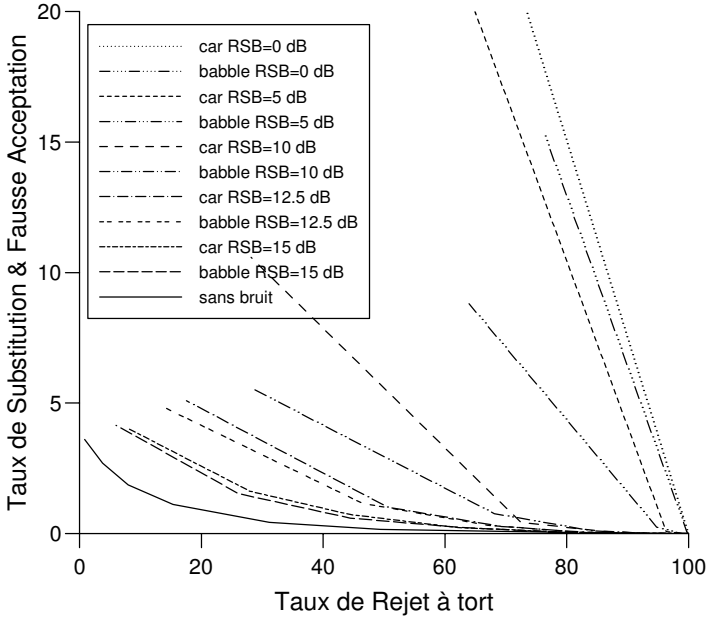


FIG. 3.17 – *Résultats de reconnaissance d'une détection idéale sur la base GSM_A bruitée.*

de la parole. Cependant ces erreurs sont en faible nombre et les différences entre les différents RSB pour les deux bruits sont faibles, excepté dans le cas où aucun bruit n'est ajouté. En effet si le niveau de bruit moyen est plus faible, les bruits impulsifs ou de courte durée sont plus marqués et provoquent ainsi plus d'insertions.

Nous allons à présent étudier l'influence du niveau de bruit sur les erreurs de reconnaissance.

3.7.2 Les erreurs de reconnaissance

Dans un premier temps nous représentons les erreurs du module de reconnaissance avec une détection idéale (issues de la segmentation manuelle) pour les deux bruits précédemment étudiés *car* et *babble* selon le RSB (*cf.* figure 3.17).

Nous remarquons que le module de reconnaissance ne permet pas de reconnaître la parole pour un RSB inférieur à 10 *dB*, même avec une détection idéale. Le modèle utilisé est cependant le même que précédemment lorsque la base n'est pas bruitée. Pour un RSB de 10 *dB*, le taux de rejet à tort reste supérieur à 30%. Nous restreindrons donc de nouveau l'étude au RSB de 10 *dB* à 20 *dB*. Notons de plus que le bruit *car* donne de moins bons résultats que le bruit *babble*, et que l'écart est d'autant plus grand que le niveau de bruit est grand.

La figure 3.18 donne les résultats de reconnaissance avec le module de détection ABP pour les seuils optimaux de reconnaissance donnés dans le tableau F.4 de l'Annexe F sur la base GSM_A bruitée par les bruits *car* et *babble*. Nous constatons conformément aux résultats de détection, que plus le RSB est petit plus les taux d'erreur sont importants.

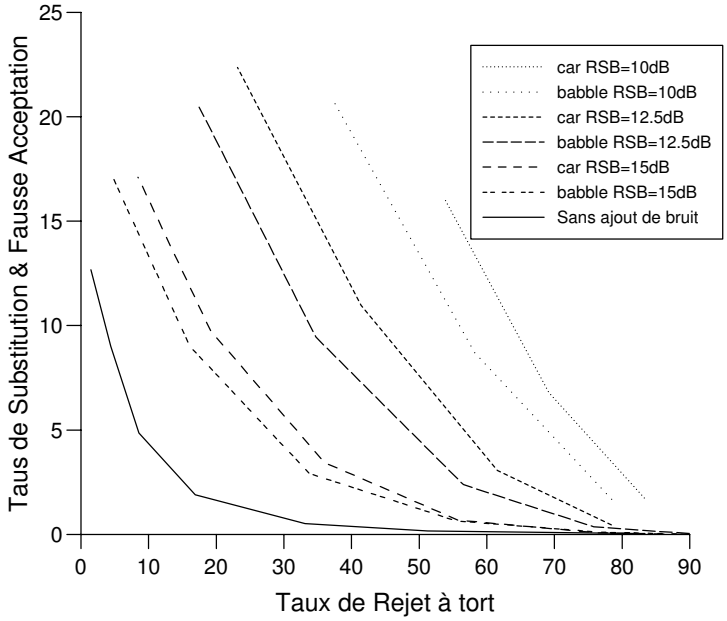


FIG. 3.18 – *Résultats de reconnaissance sur la base GSM_A bruitée.*

La différence entre le bruit *car* et *babble* est d'autant plus grande que le RSB est petit.

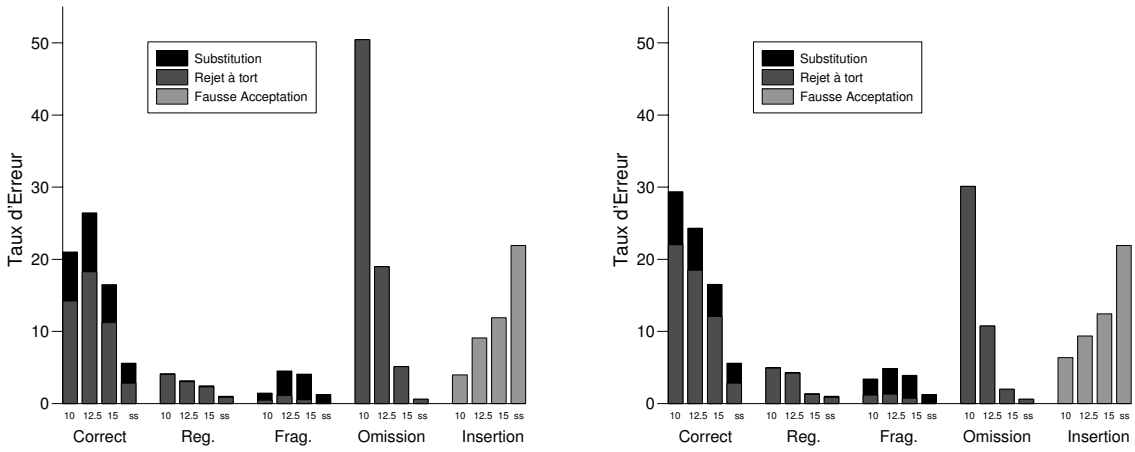
Résultats de reconnaissance selon la détection

Pour détailler ces résultats selon les erreurs de la détection, les histogrammes 3.19(a) et 3.19(b) présentent les erreurs de reconnaissance pour les deux bruits selon la détection, selon le RSB, et sans ajout de bruit (ss).

- Nous remarquons que la principale source d'erreurs lorsque le RSB est petit est due aux omissions qui provoquent des rejets à tort. En contre partie les insertions entraînent des erreurs de fausse acceptation. Les erreurs d'insertion étant moins importantes lorsque le RSB est faible, elles entraînent moins d'erreurs de fausse acceptation.
- Nous constatons également que les erreurs de reconnaissance sur les segments correctement reliés sont en nombre plus important lorsque le RSB est petit.

3.7.3 Discussion

Dans cette étude, le bruit ajouté artificiellement fait que l'effet Lombard est inexistant. Lorsque le bruit est ainsi ajouté avec un RSB inférieur à 10 *dB*, les résultats du module de détection, mais aussi du module de reconnaissance sont fortement dégradés. Le bruit *car* donne de moins bons résultats que le bruit *babble*. En effet le bruit *car* est un bruit moins stationnaire que le bruit *babble* qui est de plus un bruit de conversation donc plus proche de la parole utile. Ceci montre qu'en outre le RSB, le type de bruit (stationnarité,



(a) Bruit *car*.

(b) Bruit *babble*.

FIG. 3.19 – *Erreurs de reconnaissance selon les résultats de détection sur la base GSM_A bruitée.*

proche de la parole) dégrade différemment les performances des modules de détection et de reconnaissance.

Une méthode de débruitage peut améliorer les performances du module de détection et de reconnaissance dans le cas de bruits stationnaires, même avec un RSB petit.

Nous avons jusqu'à présent étudié les résultats de détection dans le cas de la reconnaissance de mots isolés. Pour la reconnaissance de parole continue, le problème doit être abordé différemment. Nous étudions le cas de la parole continue dans le paragraphe suivant.

3.8 La détection de la parole continue

Il faut considérer le cas de la parole continue à part. Bien sûr le seuil de détection, le niveau du RSB et le vocabulaire sont aussi influents que dans le cas de la reconnaissance de mots isolés. Cependant l'algorithme de détection de la parole précédemment utilisé doit subir quelques modifications pour son adaptation au contexte de la parole continue. Il ne s'agit plus de détecter des mots isolés, mais des phrases, le temps de parole est donc plus long, et les pauses inter-mots sont également plus importantes que les pauses inter-syllabes ou inter-phonèmes. Ainsi le module de détection doit pouvoir détecter des temps de silence plus longs entre les mots.

Nous avons donc augmenté la durée du silence de fin de parole, initialement à 240 *ms*, à 960 *ms*. Cette seule modification entraînerait un nombre de segments élargis très

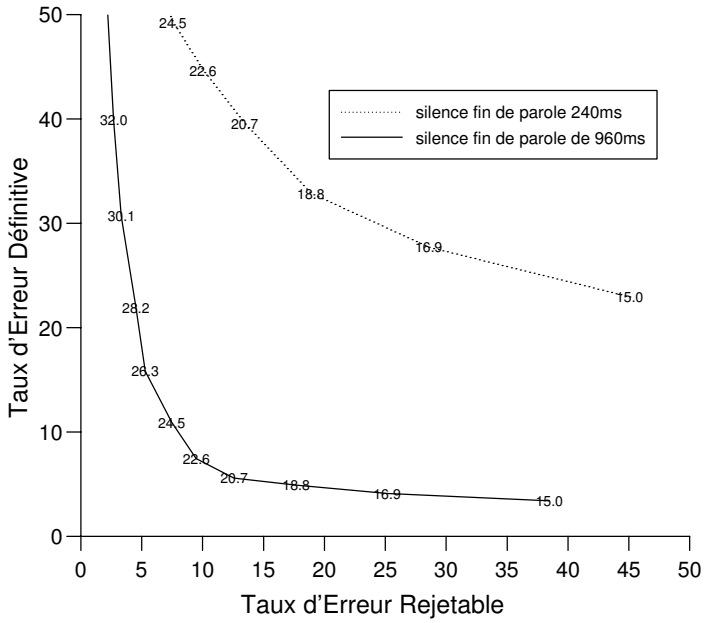


FIG. 3.20 – *Résultats de détection sur la base AGORA avec différentes valeurs du silence fin de parole.*

importants en fin de phrase, ce qui perturberait le module de reconnaissance. Après les 960 *ms* de silence suivant la fin de parole, nous nous ramenons donc à 240 *ms*, plaçant la frontière de fin de parole 720 *ms* avant la fin de parole trouvée.

La figure 3.20 représente la différence des erreurs de détection de l'algorithme sans modifications, et avec les modifications apportées, de fin de parole et de retour en arrière.

Comme précédemment, nous étudions dans un premier temps les résultats de détection, et dans un second temps les résultats de reconnaissance.

3.8.1 Les erreurs de détection

Rappelons que le nombre total de segments étiquetés *Parole-Voc* est 2520 et étiquetés *Non-Parole* 1018.

La figure 3.21 donne les erreurs de détection en comparaison de la segmentation manuelle, selon les principaux seuils.

– Le nombre de regroupements de *Parole-Voc* reste faible. Rappelons que les erreurs de regroupement sont obtenues relativement à la segmentation manuelle, qui reste objective quand aux choix des requêtes. Les erreurs de regroupement sont cependant pénalisantes pour la reconnaissance de parole continue et pour certains systèmes de dialogue associé, qui traitent les requêtes séparément. Deux requêtes regroupées peuvent entraîner des substitutions importantes dues au modèle de langage, de plus le système de dialogue s'attend à plusieurs requêtes, et non à une seule.

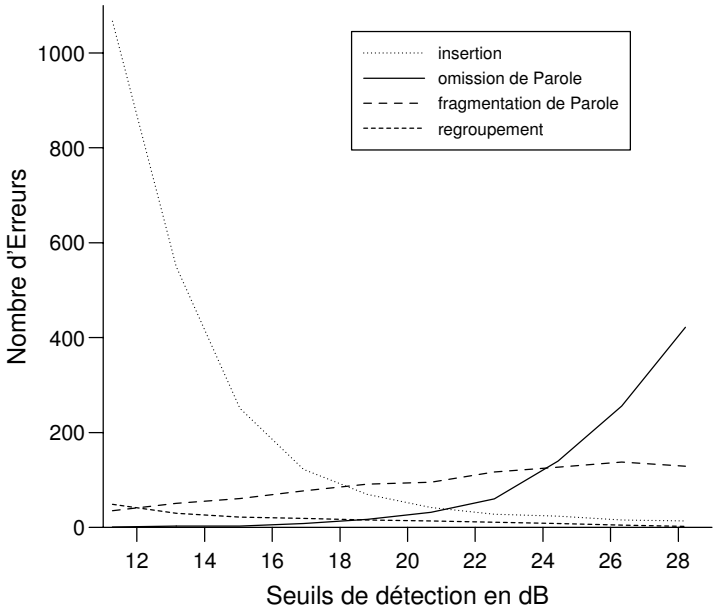


FIG. 3.21 – *Erreurs de détection détaillées sur la base AGORA.*

- Nous remarquons que le nombre de fragmentations est en proportion beaucoup plus important que pour les bases de mots isolés précédemment étudiées. Ces erreurs peuvent être très pénalisantes pour le système de dialogue même sans erreurs du module de reconnaissance de parole continue. En effet, les requêtes sont traitées les unes à la suite des autres, une fragmentation d'un segment en deux sera donc traitée comme deux requêtes différentes. Ces erreurs de détection viennent des hésitations, les pauses entre mots dans une même requête peuvent être longues selon le locuteur. Le taux de fragmentation va croissant avec le seuil.
- Le nombre d'omissions croît également avec le seuil de détection. Nous constatons que le nombre d'omissions de segments est moins important que sur les bases de mots isolés proportionnellement au nombre de segments *Parole-Voc*. Les segments de parole, étant à présent des phrases, sont plus longs, et il est normal d'en omettre moins. Cependant après le seuil 22.6 *dB* le nombre d'omissions croît rapidement.
- Le nombre de segments correctement reliés va décroissant avec le seuil, nous pourrions donc penser que le seuil le plus intéressant est un seuil bas, qui donne aussi peu d'omissions et peu de fragmentation. Le problème réside dans le nombre d'insertions qui entraînent des erreurs rejetables, mais pas systématiquement rejetées. Le nombre d'insertions est cependant moins important que sur la base RTC_A proportionnellement aux segments *Non-Parole*.

Le seuil présentant le minimum des taux d'erreur associé (erreurs rejetables et définitives) est ici 22.6 *dB*.

L'histogramme 3.22 détaille le positionnement des frontières pour la base AGORA pour les détections de *Parole-Voc* correctement reliées selon le seuil de détection. Contrai-

rement à la reconnaissance de mots isolés, la reconnaissance de la parole continue permet de reconnaître un nombre de mots non prédéfini par détection. Une détection précise pour la reconnaissance de la parole continue est donc importante. Le vocabulaire contient des mots courts tels que "je", "le", "oui", *etc.* L'insertion ou l'omission des mots courts au début ou à la fin de la détection est fréquente. De plus au début de la détection, une insertion ou une omission est pénalisante pour la suite de la reconnaissance, car les probabilités de reconnaissance des mots suivants induites par le modèle de langage, vont être modifiées.

Nous remarquons que la proportion de segments tronqués est plus importante que sur les bases de mots isolés. Les segments sont davantage tronqués à droite qu'à gauche. L'augmentation du seuil va entraîner une augmentation des segments tronqués, mais une baisse des segments élargis. Rappelons qu'il est plus aisé pour le module de reconnaissance, de traiter les erreurs de segments élargis, avec un modèle de silence ou/et de bruit, que les erreurs issues de segments tronqués.

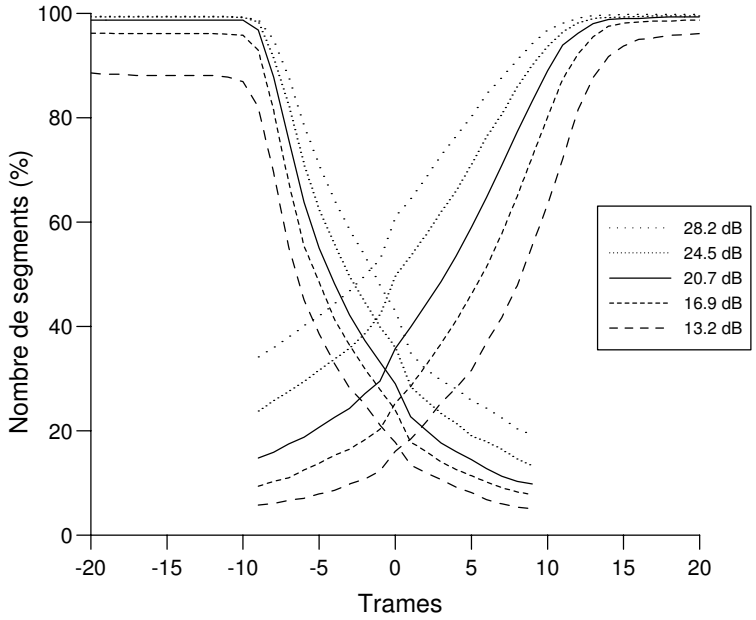


FIG. 3.22 – *Positionnement des frontières des détections sur la base AGORA.*

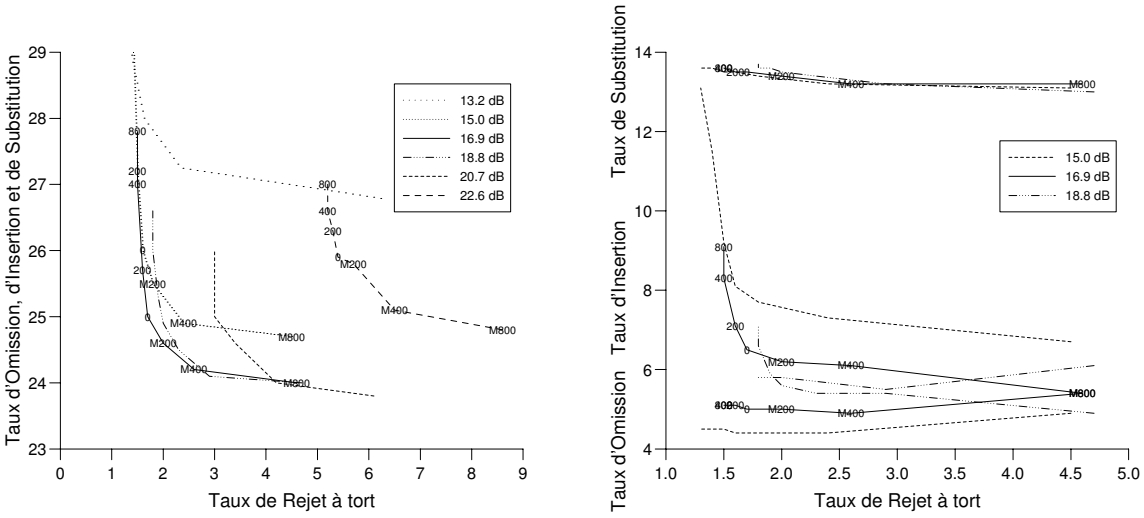
Il reste à voir l'influence des erreurs de la détection sur les résultats du système de reconnaissance.

3.8.2 Les erreurs de reconnaissance

Nous présentons ici les erreurs de reconnaissance sur la base de parole continue. Le modèle utilisé est un modèle flexible associé à un modèle de langage fondé sur un bigramme (*cf.* Annexe B). Nous ajoutons au modèle de bruit pour rejeter une requête, un modèle

de bruit en début et en fin de phrase, pour pouvoir éviter d'éventuelles insertions de mots en début et en fin de détection dues aux segments élargis.

Rappelons que les types d'erreurs en reconnaissance de parole continue diffèrent de ceux utilisés pour la reconnaissance de mots isolés (*cf.* paragraphe 2.3). Nous distinguons donc les erreurs d'omission, d'insertion et de substitution de mots, et les rejet à tort exprimés en omissions de mots. Les rejets à tort sont constitués des omissions de mots sur toute la requête, dues à l'omission de la requête par le module de détection, et dues aux rejets de la requête par le module de reconnaissance.



(a) Erreurs regroupées.

(b) Erreurs séparées.

FIG. 3.23 – *Résultats de reconnaissance sur la base AGORA.*

La figure 3.23 représente les erreurs d'omission, d'insertion et de substitution en fonction des erreurs de rejet à tort, pour différents seuils de détection. Sur la figure 3.23(a) les erreurs d'omission, d'insertion et de substitution sont additionnées, sur la figure 3.23(b) elles sont séparées pour uniquement les trois principaux seuils de détection. Cette figure fait apparaître l'importance des erreurs de substitution.

Le taux de rejet à tort devient important pour les seuils de 20.7 *dB* et 22.6 *dB*. Le seuil de 16.9 *dB* donne le minimum des taux d'erreur associée, ce seuil est cependant le seuil moyen entre les seuils 15.0 *dB* et 18.8 *dB* (*cf.* figure 3.23(b)). Il semble que plus le seuil est élevé moins il y a d'insertions, mais plus il y a d'omissions de mots. Les erreurs de substitution, les plus importantes (environ la moitié des erreurs totales sur les mots), varient peu avec le seuil de la détection. En effet, les erreurs de substitution ne se produisent pas davantage en début ou en fin de segment, là où la détection peut être critique. *A priori* un segment tronqué qui entraîne une omission de mot peut provoquer une substitution de mot à cause du modèle de langage. Ces résultats montrent que ces

erreurs sont essentiellement dues au module de reconnaissance.

Pour étudier plus en détail les erreurs de reconnaissance, nous fixons le poids de rejet, nous prenons un poids nul, et nous considérons les trois seuils 15.0 *dB*, 16.9 *dB*, 18.8 *dB*.

Nous présentons les résultats de la reconnaissance en fonction du module de détection à l'aide des histogrammes 3.24 et 3.25. Sur l'histogramme 3.24 les erreurs sont ici exprimées en pourcentage du nombre total des mots de référence, tandis que sur l'histogramme 3.25 elles correspondent aux mots omis, insérés ou substitués pour un seuil et un placement donné, exprimées en pourcentage du nombre de mots total selon le positionnement des frontières des détections correctement reliées des segments *Parole-Voc*. Sur cet histogramme, nous étudions également l'influence des segments élargis (à gauche E.G., à droite E.D., et à gauche et à droite E.G.D.). Ainsi, nous considérons qu'un segment est élargi si la détection donne une frontière déplacée de plus de 160 *ms* à gauche, ou de plus de 240 *ms* à droite. Ce choix a été fait d'après les paramètres de silence en début et fin de l'automate Bruit/Parole.

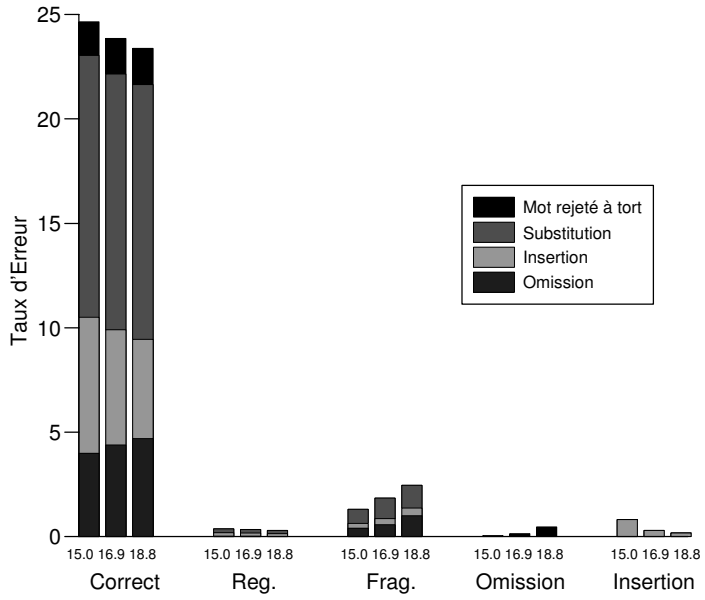


FIG. 3.24 – *Erreurs de reconnaissance selon les résultats de détection sur la base AGORA.*

Résultats de reconnaissance selon la détection

L'observation de l'histogramme 3.24 amène les remarques suivantes :

– La majorité des erreurs de reconnaissance se produisent sur des segments reliés correctement (74.84% des erreurs de reconnaissance pour le seuil de 16.9 *dB*). Nous verrons dans l'interprétation de l'histogramme 3.25 qu'une partie de ces erreurs est due aux détections imprécises.

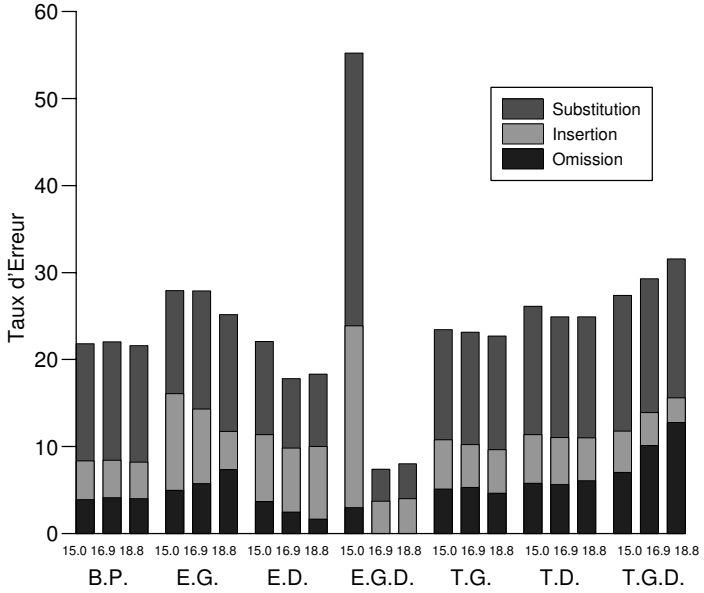


FIG. 3.25 – *Erreurs de reconnaissance selon le positionnement des frontières sur la base AGORA.*

Les principales erreurs sont les substitutions de mots lors d'une détection correctement reliée, qui ne varient pas vraiment avec le seuil de détection. Les insertions et les omissions de mots dans ce cas sont également en forte proportion.

Les erreurs de regroupement diminuent très légèrement avec le seuil. Ces erreurs restent cependant en faible nombre. Néanmoins les erreurs de reconnaissance commises sur les segments regroupés sont en proportion bien plus importantes que sur les requêtes reliées correctement (*cf.* figure 3.21).

Les erreurs de fragmentation augmentent avec le seuil. Sur les segments fragmentés, les erreurs de reconnaissance sont aussi en proportion plus importante que sur les requêtes reliées correctement.

Les omissions de segments qui augmentent avec le seuil ne représentent qu'une faible part des omissions de mots. Ceci est dû au fait que les omissions de segments de parole sont des omissions de phrases courtes voire de phrases contenant un seul mot. En moyenne le nombre de mots par segment omis passe de 1.7 mots pour le seuil 15.0 *dB* à 3.4 mots pour le seuil 18.8 *dB*.

Les insertions de segments sont bien rejetées par le module de reconnaissance, ils ne donnent ainsi que peu d'insertions de mots. De plus ces segments insérés sont courts, c'est pourquoi le nombre de mots insérés reste faible.

D'après l'histogramme 3.25, nous observons :

– En proportion il y a un peu plus d'erreurs sur les segments avec des frontières mal positionnées. Le cas des segments élargis à gauche et à droite n'est pas interprétable compte tenu du faible nombre de ce type d'erreurs.

Les segments élargis entraînent en pourcentage davantage d'insertions que les détections bien placées. Remarquons également que les segments élargis à gauche provoquent plus d'omissions que sur les détections bien placées. Ceci peu s'expliquer par le fait qu'une insertion à gauche sur un segment élargi peu perturber le modèle de langage et ainsi entraîner une omission.
Les segments tronqués à gauche ou à droite vont bien sûr provoquer des omissions, mais également un grand nombre d'insertions et de substitutions, qui sont en pourcentage légèrement plus important que lors de détections bien placées. Cependant lorsque le segment est tronqué à gauche et à droite, les omissions de mots deviennent importantes. Plus le seuil est grand et plus il y aura d'omissions provoquées sur ces segments. En effet, nous avons vu (*cf.* figure 3.25) que plus le seuil est grand plus les troncatures seront importantes en nombre, mais aussi en trames.
À l'aide du nombre des segments correctement reliés, nous constatons que 12.5% des erreurs (pour le seuil de 16.9 *dB*) se produisent lorsque les détections sont bien placées. En effet il y a un nombre plus grand de segments correctement reliés, et donc plus de mots concernés. Ce taux d'erreur directement dues au module de reconnaissance reste important. Le module de détection peut néanmoins diminuer une partie des erreurs des détections mal placées.

3.8.3 Discussion

Ce paragraphe nous a permis de dégager les principales sources d'erreurs du module de détection et du système de reconnaissance sur la base de parole continue.

Au niveau des résultats de la détection :
Retenons que les taux de fragmentation et de regroupement de parole sont plus importants que sur les bases de mots isolés. Ces erreurs peuvent cependant venir d'une segmentation manuelle qui reste subjective.

Au niveau des résultats de la reconnaissance :
- 75% des erreurs se produisent sur des segments reliés correctement.
- 50% des erreurs sont des substitutions de mots.
- Sur les segments reliés correctement, 87.5% des erreurs se produisent lors de détections imprécises qui représentent 41% de ces segments.
- Les rejets à tort et les omissions de segments, qui sont des omissions de phrases courtes, ne provoquent que peu d'erreurs (environ 1%).
- Les fragmentations et les regroupements provoquent en proportion plus d'erreurs de reconnaissance que les segments correctement reliés. Ces erreurs sont cependant en faible nombre (environ 2% pour les fragmentations et 0.4% pour les regroupements).
- Les insertions de segments ne provoquent également que peu d'erreurs d'insertion de mots (environ 0.5%).

Les principales erreurs du système de reconnaissance dues au module de détection sont donc des erreurs d'omission et d'insertion de mots, qui proviennent de détections de phrases imprécises, ce qui correspond à 15% des erreurs de reconnaissance.

3.9 Conclusion

Ce chapitre permet d'abord de mettre en œuvre la méthodologie pour l'évaluation, et la comparaison des techniques de détection de parole dans le cadre de la reconnaissance. Nous pouvons ainsi étudier les erreurs dans leur globalité par des représentations simples, ou lorsqu'il est nécessaire, détailler les erreurs pour une plus grande précision. Ce mode d'évaluation est repris dans la suite de cette étude.

Nous avons vu l'influence des différentes sources d'erreurs du module de détection, au niveau des résultats de détection et au niveau des résultats de reconnaissance. Les relations entre la détection et la reconnaissance sont très étroites. Les différents paramètres de l'algorithme doivent être choisis selon l'application (vocabulaire, environnement, *etc.*).

- Le choix du seuil de détection doit se faire à l'aide des résultats de reconnaissance, mais aussi de détection, selon l'application recherchée.
- Le vocabulaire influe sur la détection et sur la reconnaissance. Il est important de choisir un vocabulaire en fonction des caractéristiques phonétiques des mots pour obtenir une détection et une reconnaissance plus robuste.
- Le niveau du RSB est un paramètre qui dépend de l'environnement d'appel. Lorsque l'environnement est bruité, les performances sont fortement dégradées.
- Les résultats du paragraphe 3.8, montrent que le problème de la détection de parole continue doit être considéré différemment de la détection de mots isolés. La précision de la détection a une plus grande importance, et joue un rôle important sur les erreurs de reconnaissance.

Nous avons également étudié la sensibilité du seuil de détection au changement de base, au niveau de bruit et au réseau d'appel.

Le lien entre les erreurs rejetables et définitives montre qu'une augmentation des erreurs rejetables fait diminuer les erreurs définitives. Améliorer l'algorithme en diminuant les erreurs rejetables peut sembler inintéressant du point de vue de la reconnaissance. En effet le modèle de rejet du système de reconnaissance fondé sur les chaînes de Markov cachées est un outil performant pour supprimer ses erreurs. Cet outil est cependant coûteux. De plus un trop grand nombre d'erreurs rejetables entraîne un taux de fausse acceptation important. Ainsi améliorer l'algorithme pour diminuer un type d'erreur permet d'obtenir un nombre total d'erreurs moins important. Nous avons également remarqué que les meilleurs résultats de reconnaissance ne sont pas obtenus pour le seuil donnant le minimum des taux d'erreur associée de détection. Il vaut mieux privilégier plus d'erreurs rejetables dans une proportion raisonnable, et donc moins d'erreurs définitives.

Ainsi pour améliorer les performances du module de détection et donc du système de reconnaissance, il faut chercher essentiellement à diminuer les erreurs rejetables. Cette étude permet dans le Chapitre 4 " *Voies envisagées pour l'amélioration du module de détection* de définir les objectifs de cette thèse.

Chapitre 4

Voies envisagées pour l'amélioration du module de détection

4.1 Introduction

Ce chapitre a pour but de présenter les voies de recherche explorées dans cette étude pour l'amélioration du module de détection ABP.

Les chapitres précédents ont montré les insuffisances du module de détection ABP, et qu'il est important d'améliorer ces performances pour diminuer les erreurs du système de reconnaissance. Particulièrement le Chapitre 3 "*Analyse des sources d'erreurs du module de détection*" permet de dégager les principales sources d'erreurs du module de détection. Ces principales sources d'erreurs sont rappelées au paragraphe 4.2 pour permettre au paragraphe 4.3 de définir les objectifs de cette étude précisément. Nous y présentons également l'approche choisie pour évaluer si les objectifs sont atteints.

Nous avons présenté au Chapitre 2 "*Détection de parole pour la reconnaissance vocale*" les trois critères du module de détection ABP. Afin d'améliorer ce module de détection, nous déterminons au paragraphe 4.4 le meilleur de ces trois critères, qui sert par la suite de critère de base. C'est donc à partir de ce critère que nous cherchons à améliorer le module de détection. Pour satisfaire à nos objectifs, un grand nombre d'approches sont *a priori* envisageables. Le paragraphe 4.5 permet de dégager les principales approches envisageables pour l'amélioration du module de détection ABP. En particulier nous dégageons certaines caractéristiques qui sont étudiées plus précisément dans le paragraphe 4.6 pour permettre une meilleure détection. Enfin nous présentons dans le paragraphe 4.7 les axes de recherche retenus pour diminuer les erreurs du module de détection et qui seront développés dans la deuxième partie de cette étude.

4.2 Les principales sources d'erreurs du module de détection

Nous rappelons dans ce paragraphe les principales sources d'erreurs du module de détection ABP déterminées au Chapitre 3 "*Analyse des sources d'erreurs du module de détection*", qui permettent de définir les objectifs de la thèse.

Nous avons vu dans ce chapitre qu'il est important d'étudier le module de détection dans le cadre de son application qui est la reconnaissance vocale. Ainsi différentes sources d'erreurs du module de détection dégradent les performances du système de reconnaissance.

Nous avons ainsi déterminé :

- L'influence du seuil de détection.
  Nous avons montré au paragraphe 3.3 que les erreurs d'omission augmentent lorsque le seuil de détection augmente, tandis que les erreurs d'insertion diminuent. Ainsi les erreurs rejetables diminuent lorsque le seuil de détection augmente. Nous avons observé que le seuil qui donne le minimum d'erreurs du module de reconnaissance pour un poids de rejet fixé est inférieur au seuil qui donne le minimum d'erreurs du module de détection (somme des erreurs rejetables et des erreurs définitives). En effet une partie des erreurs rejetables est rejetée par le modèle de rejet du module de reconnaissance.
- L'influence du réseau d'appel.
  Le seuil de détection qui donne le minimum d'erreurs aussi bien au niveau du module de détection qu'au niveau du module de reconnaissance dépend également du réseau d'appel. Le module de détection est moins performant sur le réseau GSM (avec la base GSM_A) que sur le réseau RTC (avec la base RTC_A). En effet les résultats sur le réseau GSM sont dégradés par les bruits qui proviennent de l'environnement d'appel, mais aussi par la dégradation plus importante sur ce réseau de la qualité du signal vocal due à la transmission. Une autre raison de la dégradation peut provenir de la différence du vocabulaire des bases GSM_A et RTC_A.
- L'influence du vocabulaire.
  Le seuil de détection est donc sensible au changement de base et donc de vocabulaire. Les résultats varient également selon les types phonétiques des mots du vocabulaire. Selon la prononciation des mots, ils peuvent être plus difficiles à détecter, en particulier ceux commençant ou finissant par des fricatives ou des occlusives. De plus la ressemblance de certains mots du vocabulaire ou non avec d'autres mots du vocabulaire peut entraîner des erreurs du module de reconnaissance.
- L'influence du bruit, au niveau du type de bruit (stationnaire ou impulsif) et au niveau du RSB.
  Le paragraphe 3.4 montre que le seuil de détection optimal varie en fonction de l'environnement et du RSB. De plus d'après les paragraphes 3.6 et 3.7, les résultats sont très dégradés lorsque le RSB est faible avec un bruit stationnaire et lorsque le signal

contient beaucoup de bruits impulsifs (de courte durée). Un faible RSB entraîne beaucoup d'erreurs d'omission du module de détection, car la différence énergétique entre les périodes de parole et de bruit est trop faible. Les bruits impulsifs ou de courte durée provoquent des erreurs rejetables (*i.e.* des erreurs d'insertion) en grand nombre, qui ne sont pas systématiquement rejetées par le modèle de rejet.

- L'influence de l'application.
  Selon l'application : reconnaissance de mots isolés ou reconnaissance de parole continue, le seuil de détection qui donne le minimum d'erreurs est également différent. De plus les taux d'erreur dans le cas de la reconnaissance de parole continue sont élevés. Ces taux d'erreur peuvent être diminués par le module de reconnaissance mais aussi par le module de détection.

Ces résultats nous permettent de définir différents objectifs pour améliorer le module de détection. Ces objectifs de la thèse font l'objet du paragraphe suivant.

4.3 Objectifs de la thèse

Afin d'améliorer le système de reconnaissance, nous devons chercher d'une part à améliorer les performances du module de détection au niveau des résultats de la détection et au niveau des résultats de reconnaissance dans toutes les conditions. D'autre part nous devons chercher à diminuer la sensibilité du seuil de détection au réseau d'appel, au changement de base et au niveau de bruit, rappelée au paragraphe précédent.

Nous devons chercher avant tout à améliorer les performances du module de détection ABP d'une part en se rapprochant le plus de la segmentation manuelle, d'autre part en diminuant les erreurs du système de reconnaissance. Nous avons vu que diminuer les erreurs du module de détection n'entraîne pas obligatoirement une diminution des erreurs de reconnaissance. Cependant une diminution des erreurs du module de détection entraîne généralement une diminution des erreurs de reconnaissance. Nous devons donc diminuer soit les erreurs rejetables soit les erreurs définitives. Le Chapitre 3 "*Analyse des sources d'erreurs du module de détection*" a permis de constater que les erreurs rejetables sont particulièrement nombreuses sur des enregistrements qui contiennent des bruits de courte durée. Il faut donc diminuer les erreurs rejetables du module de détection ABP, sans augmenter les erreurs définitives.

Dans le paragraphe 4.2 nous avons vu que les performances sont différentes selon le niveau de bruit, selon le vocabulaire et selon le réseau d'appel.

En effet les résultats sont très dégradés lorsque le niveau de bruit est important et lorsqu'il y a beaucoup de bruits de courte durée. Nous devons donc chercher avant tout à diminuer les erreurs pour des communications bruitées par des bruits stationnaires ou impulsifs ou de courte durée. Notre premier objectif est donc de diminuer les erreurs pour les communications bruitées.

Nous avons également constaté que les taux d'erreur du module de détection peuvent être diminués pour améliorer le système de reconnaissance dans le cadre de la reconnaissance de parole continue. Ceci est notre deuxième objectif.

Au Chapitre 3 "*Analyse des sources d'erreurs du module de détection*", l'étude du vocabulaire sur les erreurs du système de reconnaissance a montré que ces erreurs sont en grande partie dues au module de reconnaissance et aux erreurs de précision des frontières. Pour déterminer le vocabulaire d'une application il faut donc choisir des types de mots qui doivent :

- être adaptés à l'application,
- être éloignés phonétiquement des autres mots du vocabulaire,
- éviter certains phonèmes en début et en fin de détection.

Ainsi pour améliorer les résultats du système de reconnaissance le vocabulaire doit être choisi judicieusement. Les types de phonèmes à éviter en début et fin de détection sont typiquement les fricatives et occlusives. Le vocabulaire des bases de données utilisées pour la reconnaissance de mots isolés est trop petit pour pouvoir déterminer les types de mots et donc le vocabulaire qui entraîne le moins d'erreurs pour une application donnée. Nous ne cherchons donc pas particulièrement à diminuer les erreurs à l'aide du vocabulaire car nous ne pouvons pas définir précisément les types de mots à l'aide des bases à notre disposition. De bonnes performances sur les différentes bases montrent une indépendance des résultats au changement de vocabulaire.

Les résultats sur le réseau GSM sont également moins bons que sur le réseau RTC. Ceci peut être dû au fait soit que la base GSM_A utilisée contient plus de bruits que la base RTC_A, qui proviennent de l'environnement d'appel ou au réseau lui même, soit que le vocabulaire provoque plus d'erreurs que celui de la base RTC_A. Notre premier objectif atteint permet donc de diminuer les erreurs dues aux bruits. La dégradation qui peut être due au vocabulaire est supprimée si les performances sont bonnes sur les différentes bases.

Le troisième objectif est de diminuer la sensibilité du seuil de détection au niveau du bruit, au changement de base et au réseau d'appel.

Une fois ces trois objectifs définis, nous devons préciser la méthodologie permettant d'affirmer si ces objectifs sont atteints ou non.

Le premier objectif est réalisé si les évaluations au niveau du module de détection et au niveau du module de reconnaissance montrent une diminution des erreurs sur des signaux comportant du bruit stationnaire ou impulsif (de courte durée), sans augmenter les erreurs sur du signal sans bruit. Les évaluations (présentées au paragraphe 2.4) seront donc faites sur les bases RTC_T et GSM_T selon le RSB et sur la partie calme de la base GSM_A bruitée par les bruits stationnaires *car* et *babble*.

Le deuxième objectif est réalisé si les mêmes évaluations sur la base de parole continue montrent une diminution des erreurs, sans augmenter les erreurs sur les autres bases. Les évaluations seront donc également faite sur la base AGORA.

Le troisième objectif est réalisé si la sensibilité du seuil de détection au niveau de bruit, au changement de base et au réseau d'appel diminue. Le critère de sensibilité du seuil de détection d'un critère du module de détection est celui présenté au paragraphe 3.3.3 et détaillé en Annexe H.

Ainsi les trois objectifs de la thèse sont :

- Premier objectif : diminuer les erreurs du module de détection pour les communica-

tions bruitées.
- Deuxième objectif : diminuer les erreurs du module de détection de la parole continue.
- Troisième objectif : diminuer la sensibilité du seuil de détection au niveau de bruit, au changement de base et au réseau d'appel.

Un module de détection qui atteint ces trois objectifs sera dit *robuste* pour la reconnaissance vocale en environnement bruité en particulier.

Afin de réduire les erreurs et la sensibilité du seuil de détection du module de détection, nous déterminons dans un premier temps le critère du module de détection ABP le plus performant en fonction des objectifs précités. Ainsi le paragraphe 4.4 suivant est une étude comparative des trois critères existants du module de détection.

4.4 Étude comparative des trois critères existants du module de détection

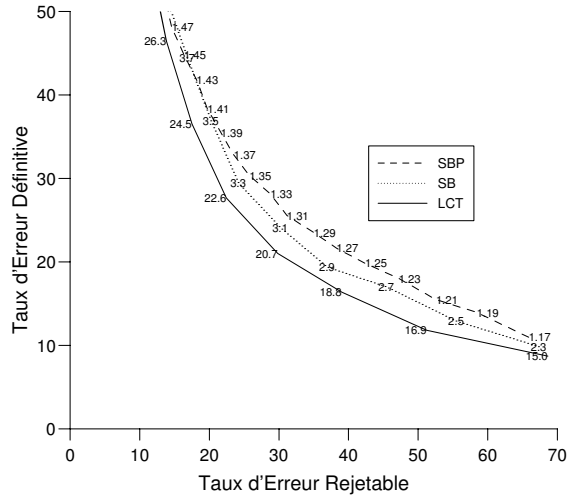
Nous comparons les trois critères présentés au paragraphe 2.2, le critère LCT fondé sur le rapport signal à bruit, le critère SB fondé sur les statistiques du bruit, et le critère SBP fondé sur les statistiques du bruit et de la parole. Cette étude est faite sur les bases RTC_A, GSM_A et AGORA. Notons que les seuils pour chacun des trois critères ne sont pas du même ordre de grandeur. La comparaison de ces différents critères utilisés dans le module de détection ABP a fait l'objet d'une étude dans [Karray et Martin, 2001].

Nous gardons la même méthodologie que dans les paragraphes précédents. Nous étudions d'abord les erreurs de détection, puis les erreurs de reconnaissance.

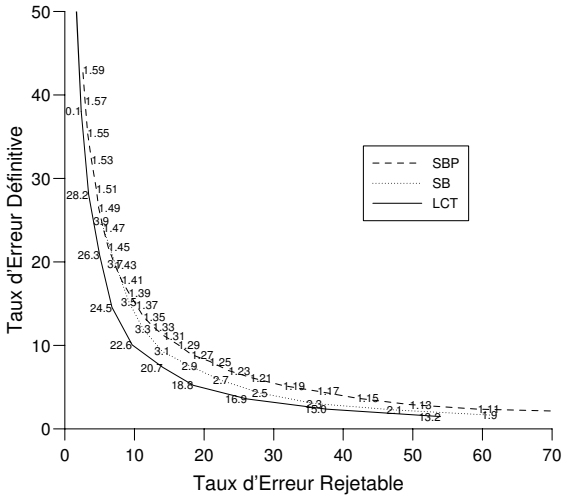
4.4.1 Résultats de détection

D'après les figures 4.1 et 4.2, nous constatons que les trois algorithmes ont des résultats assez proches. Cependant, le critère LCT donne significativement des taux d'erreur moins importants sur la base RTC_A (*cf.* tableau G.2 en Annexe G), alors que sur la base GSM_A, ce critère donne les moins bons résultats, même si la différence n'est pas flagrante sur la partie la plus calme de la base (*cf.* figure 4.1(d)). Le critère SBP donne les meilleurs résultats sur la base GSM_A, la différence n'est significative que sur la partie la plus bruitée (*cf.* tableau G.2 en Annexe G). Cependant le taux d'erreur rejetable sur la base RTC_A reste très élevé. De plus cette base comporte des fichiers de courte durée, ce qui explique que les critères SB et SBP soient moins performants que le critère LCT. En effet les estimations de statistiques de l'énergie du bruit et de la parole de ces deux critères nécessitent plus de trames car les facteurs d'oubli sont plus grands. Enfin le critère SB se situe entre les critères LCT et SBP.

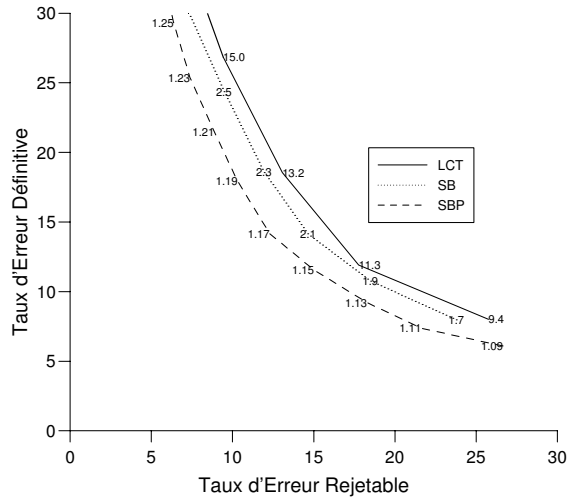
Pour la base AGORA de parole continue, les critères SB et SBP ont été modifiés de la même façon que le critère LCT au paragraphe 3.8. Nous avons donc prolongé le temps de silence de fin de phrase à 960 *ms*, tout en se ramenant à 240 *ms*, en plaçant la frontière de fin de parole 720 *ms* avant, pour éviter un nombre trop important de segments



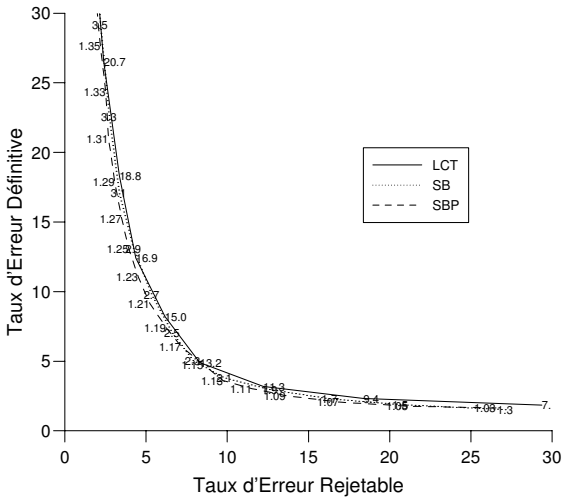
(a) Base RTC_A - RSB inférieur à 20 *dB*.



(b) Base RTC_A - RSB supérieur à 20 *dB*.



(c) Base GSM_A - RSB inférieur à 18 *dB*.



(d) Base GSM_A - RSB supérieur à 18 *dB*.

FIG. 4.1 – *Résultats de détection des trois critères sur les bases RTC_A et GSM_A.*

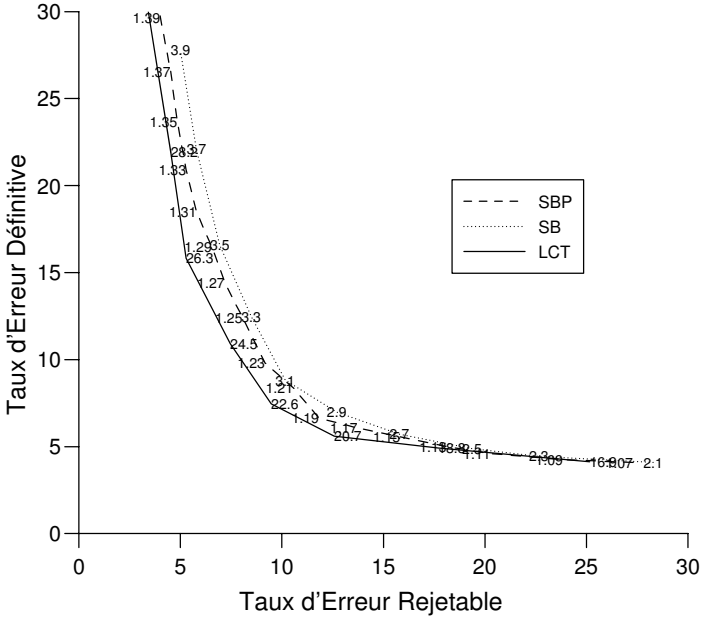
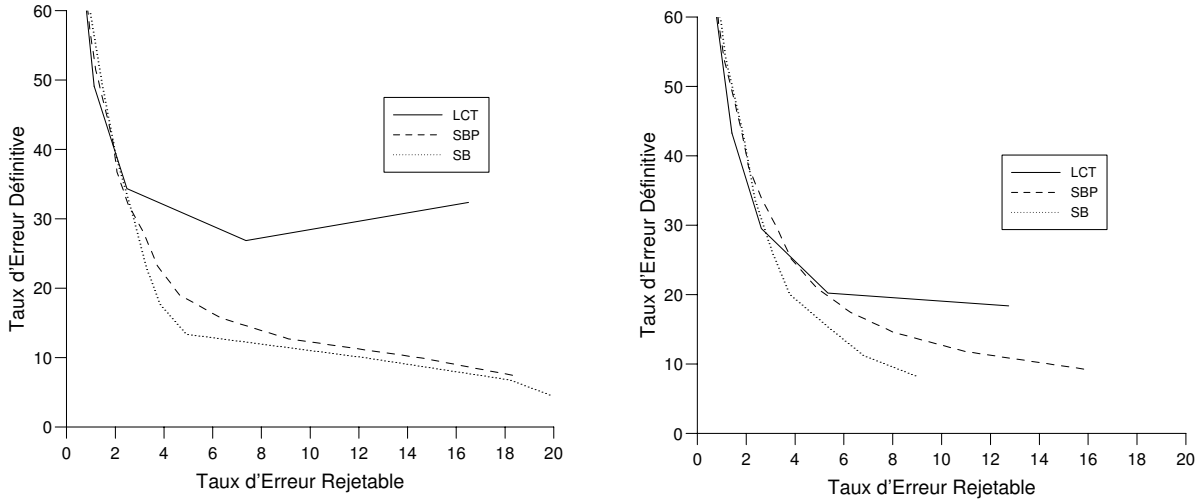


FIG. 4.2 – *Résultats de détection des trois critères sur la base AGORA.*

élargis en fin de phrase. De plus pour éviter de rester trop longtemps dans l'état *parole* de l'automate à cause d'une mauvaise estimation des paramètres du bruit, l'estimation de ces paramètres est reprise dans l'état *parole* lorsque l'automate est resté dans cet état plus de 960 *ms*. Cette durée est à modifier pour la parole continue, nous prenons 19200 *ms* qui constitue une borne supérieure de la longueur des phrases. Le critère LCT donne le minimum des taux d'erreur sur cette base, la différence n'est cependant pas significative (*cf.* tableau G.2 en Annexe G).

La figure 4.3 présente les résultats des trois critères sur la partie calme de la base GSM_A bruitée par les deux types de bruit *car* et *babble* présentés au paragraphe 3.7 à un RSB de 12.5 *dB*. Nous constatons que le critère SB donne moins d'erreurs que les deux autres critères. Les résultats sont identiques à d'autres niveaux de bruit, mais plus le niveau de bruit est élevé, plus la différence est grande. Cette différence est significative, quel que soit le niveau de bruit, et pour les deux bruits étudiés (*cf.* tableau G.4 en Annexe G). La différence entre le critère SB et le critère SBP, ici plus importante que dans les cas précédemment étudiés, s'explique par le fait que les statistiques de la parole calculées dans le critère SBP sont trop influencées par le niveau de bruit important. Ainsi les insertions seront plus importantes.

Pour comprendre davantage ces trois critères, nous détaillons les erreurs pour le seuil présentant le minimum des taux d'erreur associée de détection (erreurs rejetables et définitives) pour les différents RSB sur les deux bases RTC_A et GSM_A et sur la base de parole continue AGORA (*cf.* tableau F.1 en Annexe F). Nous présentons ces résultats sur les histogrammes 4.4(a), 4.4(b) et 4.4(c).



(a) Bruit *car*.

(b) Bruit *babble*.

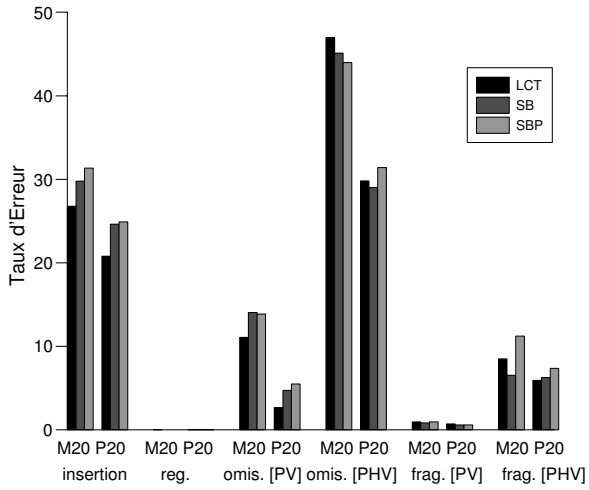
FIG. 4.3 – *Résultats de détection des trois critères sur la base GSM_A bruitée.*

Erreurs de détection détaillées

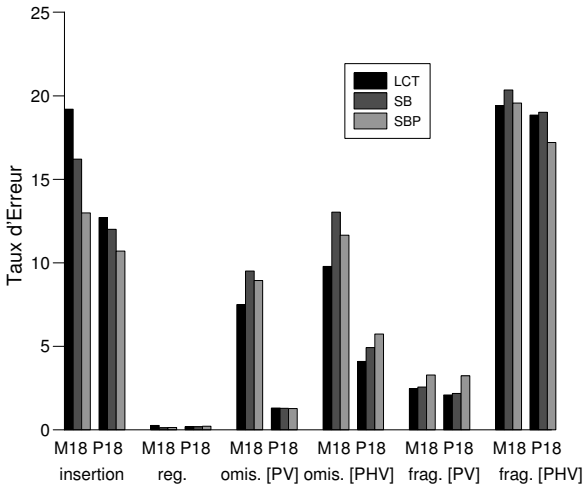
Notons tout d'abord que les erreurs de regroupement restent peu nombreuses par rapport aux autres erreurs surtout sur les bases RTC_A et GSM_A (*cf.* histogrammes 4.4(a) et 4.4(b)).
Le critère SB donne plus d'omissions de *Parole-Voc* que les deux autres critères en environnement bruité, et sur la base AGORA.
Le critère LCT donne moins d'erreurs que le critère SBP sur la base RTC_A quel que soit le RSB, excepté pour les omissions de *Parole-Hors-Voc*, qui sont des erreurs ne pouvant qu'améliorer les résultats de reconnaissance. Sur la base GSM_A le critère LCT donne plus d'erreurs d'insertion (erreurs rejetables) que le critère SBP, cependant les erreurs d'omission et de fragmentation sont moins importantes. Par contre, sur la base AGORA, il y a plus d'erreurs d'insertion et moins d'erreurs d'omission pour le critère LCT que pour le critère SBP. Cependant les erreurs d'omission de segments sur la base AGORA restent faibles, les segments *Parole-Voc* sont en effet plus long et sont donc plus difficilement omis.
Les erreurs de fragmentation, qui jouent un rôle important dans la reconnaissance de parole continue, sont moins importantes pour le critère LCT que pour le critère SBP.

Nous ne pouvons cependant pas dégager d'écart important entre ces trois critères sur ces taux d'erreur, pour les bases étudiées.

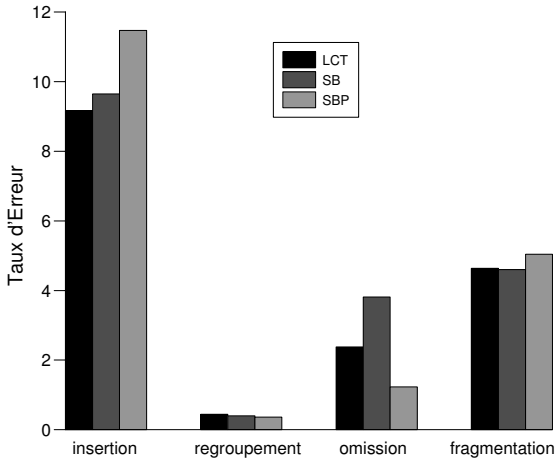
Les histogrammes 4.5 donnent le positionnement des frontières des détections correctement reliées. Il en ressort que le critère LCT se démarque par légèrement moins de segments tronqués que les deux autres critères, sur toutes les bases et quel que soit le



(a) Base RTC_A.

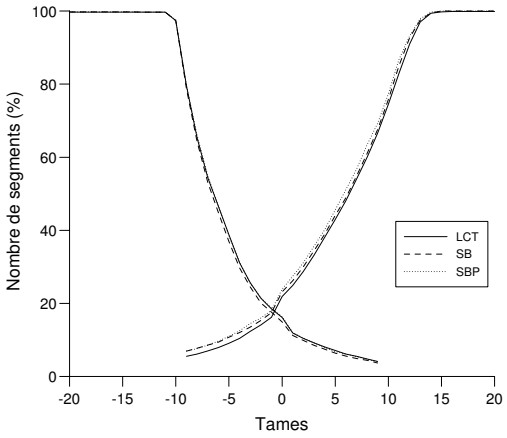


(b) Base GSM_A.

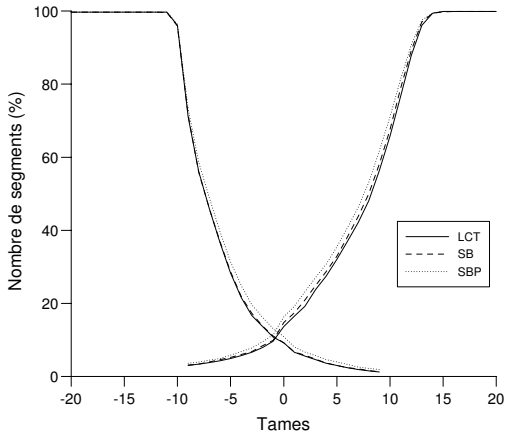


(c) Base AGORA.

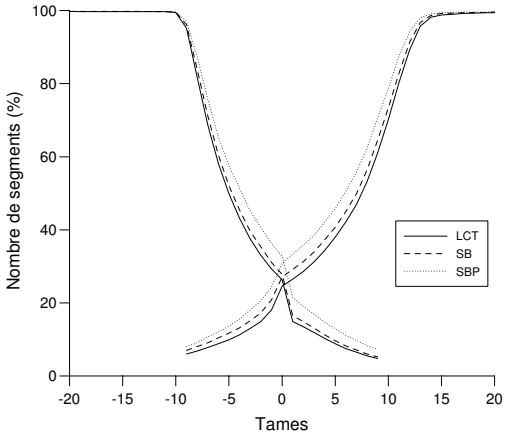
FIG. 4.4 – *Erreurs de détection détaillées des trois critères sur les bases RTC_A, GSM_A et AGORA.*



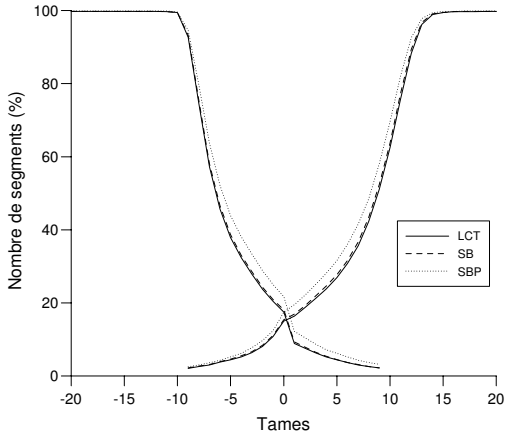
(a) Base RTC_A, RSB inférieur à 20 *dB*.



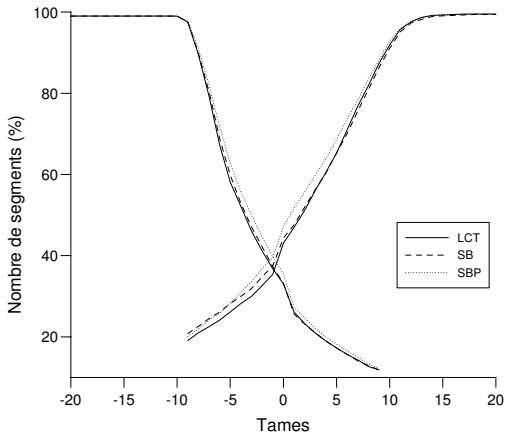
(b) Base RTC _A, RSB supérieur à 20 *dB*.



(c) Base GSM_A, RSB inférieur à 18 *dB*.



(d) Base GSM_A, RSB supérieur à 18 *dB*.



(e) Base AGORA.

FIG. 4.5 – *Positionnement des frontières des détections des trois critères sur les bases RTC_A, GSM_A et AGORA.*

RSB. Il n'y a que très peu de segments élargis sur ces bases de données.

Même si ces erreurs ne sont pas prises en considération lors de la représentation graphique des erreurs de détection, nous avons étudié leur influence sur les résultats de reconnaissance. Nous rappelons que le nombre de segments tronqués peut être diminué selon l'application, en augmentant le nombre de trames *Parole Minimum* et *Silence Fin* (*cf.* figure 2.2 au Chapitre 2 "*Détection de parole pour la reconnaissance vocale*"). Cependant il s'ensuit plus de segments élargis.

La comparaison de trois critères au niveau de la détection fait apparaître que le critère LCT donne des résultats meilleurs sur la base RTC_A, alors que sur la partie bruitée de la base GSM_A, le critère SBP donne moins d'erreurs. Rappelons que le critère LCT a été optimisé sur la base RTC_A, tandis que le critère SBP a été optimisé sur la base GSM_A. Sur la base AGORA, il n'y a pas de différences significatives entre les trois critères. Nous comparons maintenant ces trois critères au niveau de la reconnaissance.

4.4.2 Résultats de reconnaissance

Le tableau F.3 en Annexe F donne les seuils de détection optimaux pour la reconnaissance. Pour ces seuils, nous comparons les trois critères sur la base RTC_A et GSM_A (*cf.* figure 4.6) et sur la base AGORA (*cf.* figure 4.7).

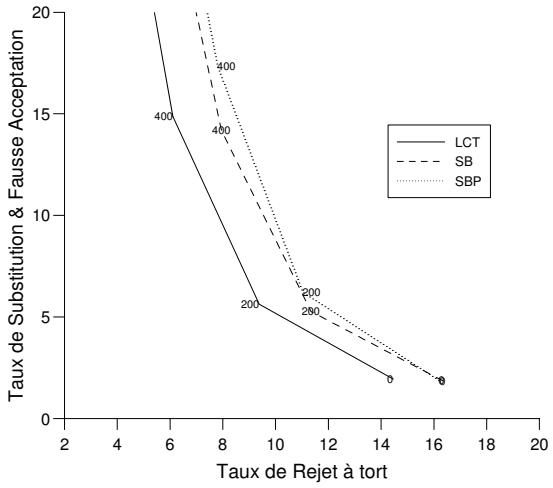
Sur les figures 4.6(a) et 4.6(b), il apparaît clairement que le critère LCT est plus performant sur la base RTC_A pour les deux RSB, et que le critère SB est meilleur que le critère SBP. Cette différence est significative d'après le tableau G.6 en Annexe G. Rappelons que cette différence s'explique par le fait que le critère LCT a été optimisé sur cette base.

Pour la base GSM_A les figures 4.6(c) et 4.6(d), montrent que la différence entre les trois critères est très faible. Sur les deux parties, c'est le critère SBP qui donne moins d'erreurs de reconnaissance. Le critère SBP est cependant significativement meilleur que le critère LCT (*cf.* tableau G.6 en Annexe G).

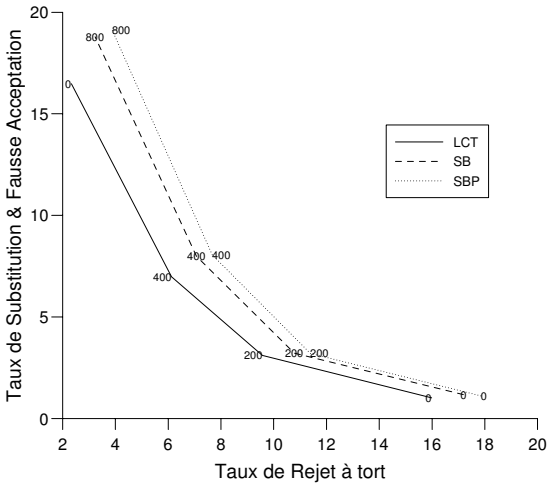
Sur la base AGORA (*cf.* figure 4.7) la figure 4.7(a) montre que sur l'ensemble des erreurs c'est le critère LCT qui donne de meilleurs résultats, pour un rejet à tort inférieur à 2.5. La figure 4.7(b) révèle cependant que la différence la plus marquée se trouve au niveau des taux d'insertion et de rejet à tort. La différence sur le taux d'erreur associée n'est cependant pas significative (*cf.* tableau G.6 en Annexe G). Cette base étant peu bruité, les critères SB et SBP ne sont pas ici plus performants.

Sur la partie calme de la base GSM_A bruitée par les deux bruits *car* et *babble* avec un RSB de 12.5 *dB*, il apparaît clairement sur la figure 4.8 que le critère LCT donne de moins bons résultats. Le critère SB est le plus performant, et la différence des taux d'erreur pour le poids de rejet fixé à 800 est significative quelque soit le niveau de bruit (*cf.* tableau G.8 en Annexe G). Cette différence est expliquée par le fait que pour le critère SB, les statistiques de la parole bruitée ne sont pas considérées. En effet le critère SBP, performant pour discriminer des bruits impulsifs de la parole, le devient moins pour un bruit stationnaire, car les statistiques de la parole sont estimées sur la parole bruitée.

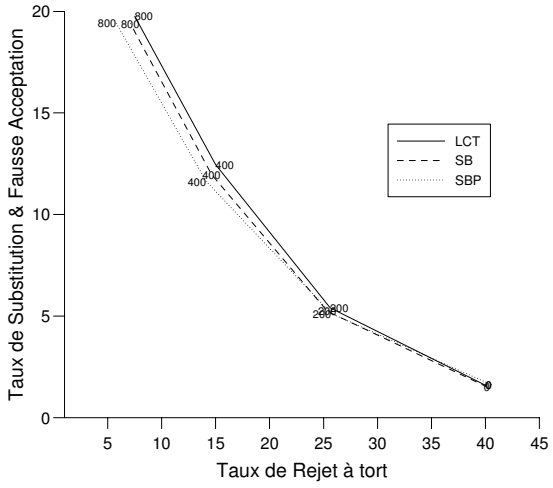
Pour bien définir les différences des caractéristiques de chaque critère, nous allons



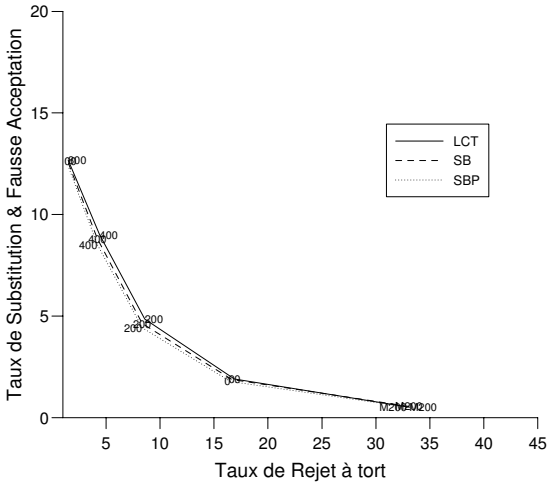
(a) Base RTC_A - RSB inférieur à 20 *dB*.



(b) Base RTC_A - RSB supérieur à 20 *dB*.

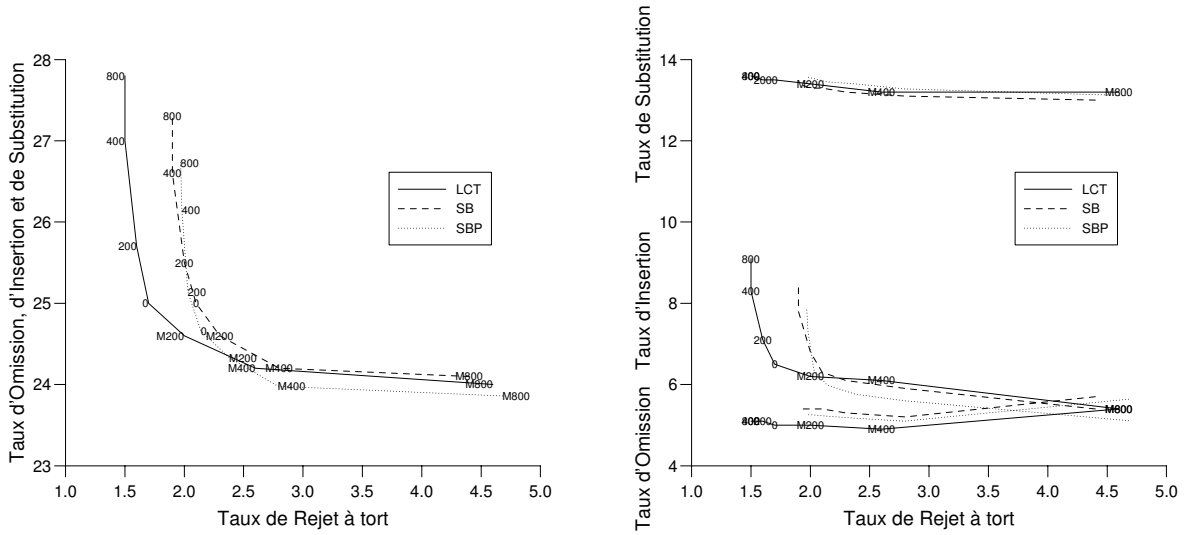


(c) Base GSM_A - RSB inférieur à 18 *dB*.



(d) Base GSM_A - RSB supérieur à 18 *dB*.

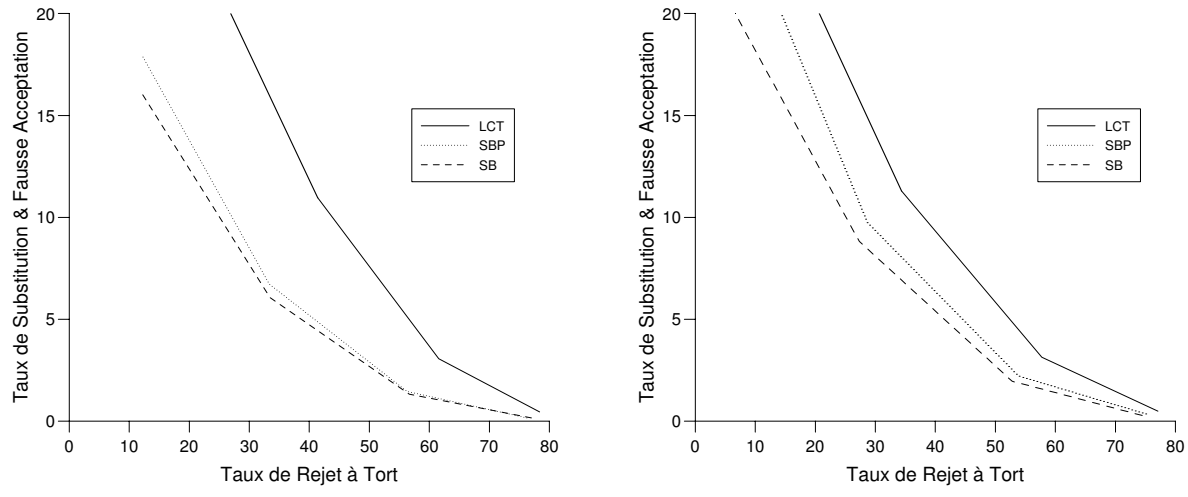
FIG. 4.6 – *Résultats de reconnaissance des trois critères sur les bases RTC_A et GSM_A.*



(a) Les taux sont regroupés.

(b) Les taux sont séparés.

FIG. 4.7 – *Résultats de reconnaissance des trois critères sur la base AGORA.*



(a) Bruit *car*.

(b) Bruit *babble*.

FIG. 4.8 – *Résultats de reconnaissance des trois critères sur la base GSM_A bruitée.*

détailler les résultats de la reconnaissance pour le seuil optimum précédemment défini, et pour un poids de rejet de 400 sur les bases RTC_A et GSM_A. Ce rejet nous permet d'avoir un taux de rejet à tort inférieur à 10%, excepté pour la partie la plus bruitée de la base GSM_A, où le taux de rejet à tort reste supérieur à 20%. Le taux de rejet à tort étant faible sur la base AGORA, le poids de rejet est de -200, qui donne le minimum des taux d'erreur associée. Nous présentons ces résultats à l'aide des histogrammes 4.9(a) et 4.9(b) pour la base RTC_A, 4.9(c) et 4.9(d) pour la base GSM_A, et 4.9(e) pour la base AGORA .

Erreurs de reconnaissance selon les résultats de détection

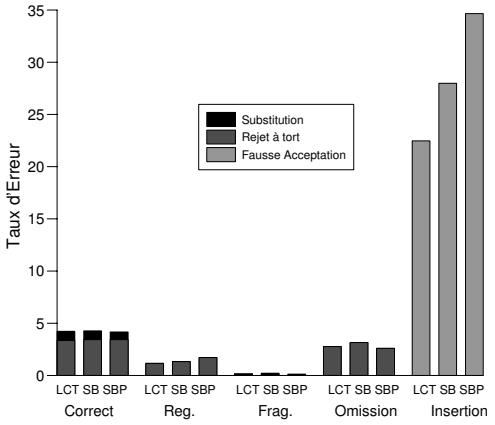
- Sur la base RTC_A (*cf.* histogrammes 4.9(a) et 4.9(b)), les trois critères sont très proches. Notons tout de même que le taux de fausse acceptation est plus élevé, en particulier sur la partie bruitée avec le critère SBP. Le taux de rejet à tort provoqué par les omissions est plus élevé sur la partie calme pour le critère SBP, et légèrement moins important sur la partie bruitée. Les différences ne sont cependant pas significatives. Les taux de substitution sont sensiblement les mêmes.
- Sur la partie calme de la base GSM_A (*cf.* histogramme 4.9(c)), il n'apparaît très peu de différence. Sur la partie bruitée, nous notons que le taux de rejet à tort pour le critère SBP, est plus élevé sur les détections correctement reliées aux segments de référence et moins élevé avec les omissions, que les deux autres critères. Les erreurs de rejet à tort et de substitution sont sensiblement les mêmes sur les erreurs de regroupement et de fragmentation du module de détection. La plus grande différence se trouve sur les erreurs de fausse acceptation, qui sont plus importantes pour le critère LCT.

  L'étude précédente sur la partie la plus bruitée de la base GSM_A montrait que c'était la plus critique. Aucun des critères ne semble apporter de réponses sur cette partie de la base.

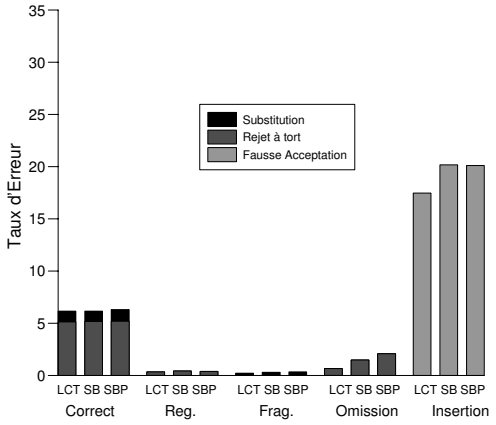
  Les différences des taux d'erreur des trois critères sur cette base ne sont toujours pas significatives.
- Sur la base AGORA (*cf.* histogramme 4.9(e)), encore une fois il n'y a pas de différences significatives au sens de l'intervalle de confiance des résultats entre les trois critères. Notons légèrement moins d'erreurs sur les détections correctement reliées et sur les fragmentations pour le critère LCT.

Pour étudier plus en détail les erreurs sur les détections correctement reliées, les histogrammes 4.10(a), 4.10(b) et 4.10(c) présentent les erreurs de reconnaissance selon le positionnement des frontières issu du module de détection, respectivement pour la base RTC_A, GSM_A et AGORA.

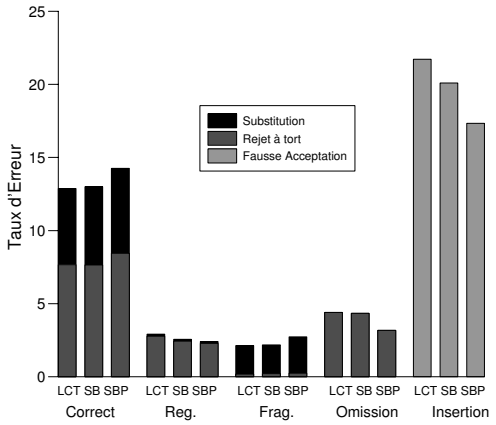
La faible différence des erreurs de reconnaissance selon le positionnement des frontières ne donne pas de différence significative sur les trois bases. Sur la base RTC_A, le critère LCT donne de meilleurs résultats excepté sur les segments tronqués à gauche et à droite, qui restent en faible nombre. Sur la base GSM_A, le critère SBP donne des résultats



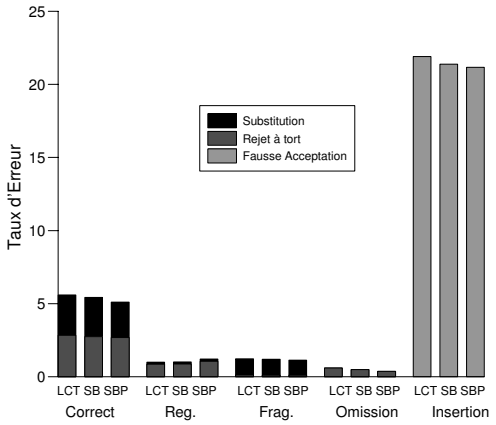
(a) Base RTC_A, RSB inférieur à 20 *dB*.



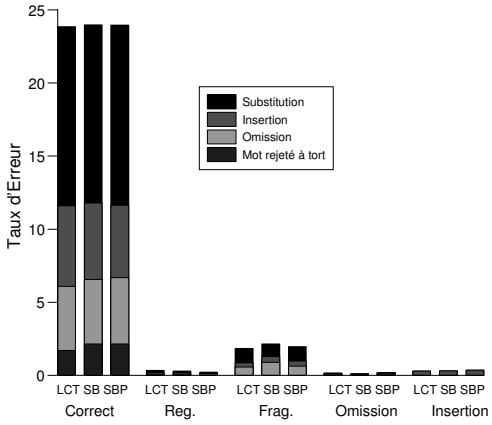
(b) Base RTC_A, RSB supérieur à 20 *dB*.



(c) Base GSM_A, RSB inférieur à 18 *dB*.

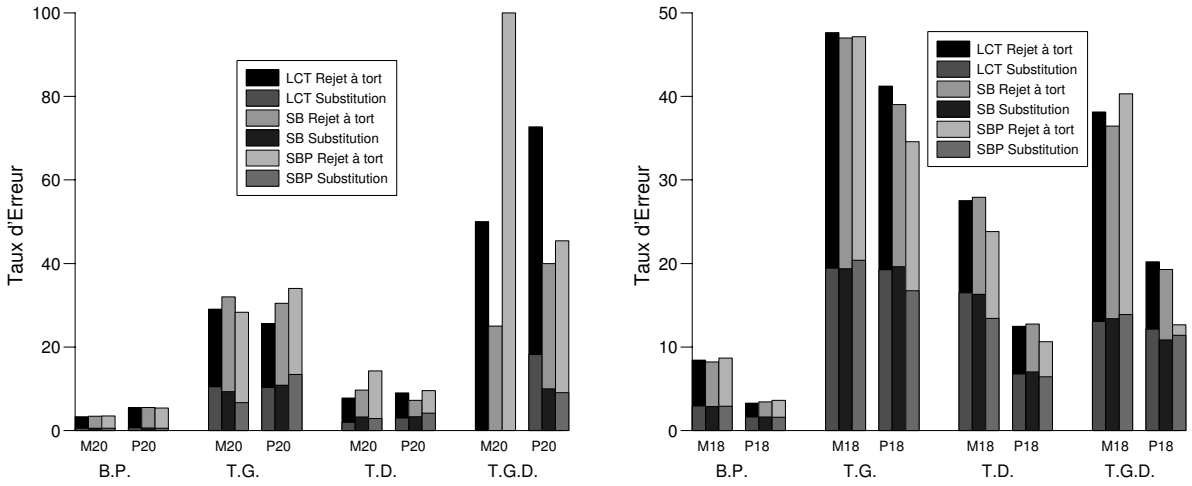


(d) Base GSM_A, RSB supérieur à 18 *dB*.



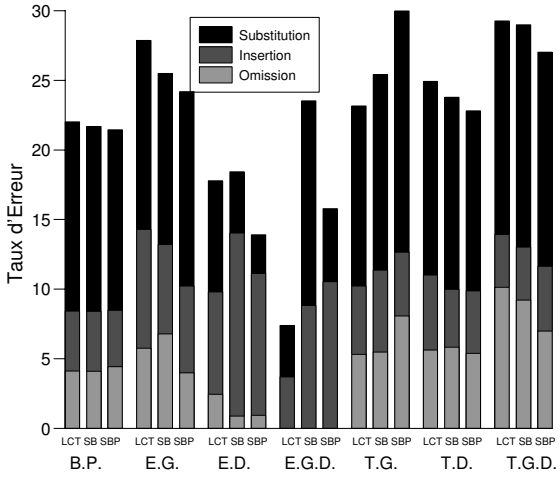
(e) Base AGORA.

FIG. 4.9 – *Erreurs de reconnaissance selon les résultats de détection des trois critères sur les bases RTC_A, GSM_A et AGORA.*



(a) Base RTC_A.

(b) Base GSM_A.



(c) Base AGORA.

FIG. 4.10 – *Erreurs de reconnaissance selon le positionnement des frontières des détections des trois critères sur les bases RTC_A, GSM_A et AGORA.*

légèrement meilleurs. Pour la base AGORA, le critère SBP se démarque de nouveau par moins d'erreurs de reconnaissance, excepté sur les segments tronqués à gauche qui sont en nombre important et les segments élargis à gauche et à droite.

4.4.3 Sensibilité du seuil de détection

Nous venons de voir que le seuil de détection a une influence importante sur les résultats de détection, mais aussi de reconnaissance, selon la base et le niveau de bruit. D'un point de vue applicatif, il est important d'obtenir un module de détection qui ne soit pas trop sensible au seuil de détection. En effet il est souvent difficile de devoir régler des paramètres une fois le système mis en application. Un module de détection qui n'est pas sensible au seuil de détection permet d'éviter ces réglages.

En Annexe H sont présentés différents résultats sur la sensibilité des seuils des trois critères au changement de base au paragraphe H.1, au niveau de bruit au paragraphe H.2, et au changement de réseau d'appel (RTC et GSM) au paragraphe H.3.

La sensibilité du seuil est mesurée à partir de l'intervalle de confiance obtenu avec le seuil "optimisé" sur une base ou partie de base d'apprentissage, et évalué sur une autre base ou autre partie de base. Si le taux d'erreur de reconnaissance avec le seuil "optimisé" et évalué sur cette autre base ou partie de base n'appartient pas à l'intervalle de confiance, alors le critère est dit sensible.

- Le tableau H.1 en Annexe H montre que les trois critères ne sont pas sensibles au changement de base, excepté le critère LCT avec un RSB supérieur à 18 *dB*.
- Pour la sensibilité au niveau de bruit, le tableau H.2 en Annexe H montre des résultats différents sur la base RTC_A et GSM_A. En effet les trois critères ne sont pas sensibles au niveau de bruit sur la base GSM_A, excepté le critère LCT pour un RSB supérieur à 18 *dB*, tandis que sur la base RTC_A les trois critères sont sensibles au seuil optimisé sur la partie calme et testé sur la partie bruitée et inversement. Cette différence peut s'expliquer par un grand nombre de bruits impulsifs sur la partie ayant un RSB inférieur à 20 *dB*, alors que sur l'autre partie il y en a beaucoup moins. La base GSM_A ne possède pas une telle différence de bruits impulsifs entre la partie bruitée et la partie calme. Rappelons que la base RTC_A est une base d'exploitation, ce qui explique la différence entre les deux bases.
- Le tableau H.3 en Annexe H montre que les trois critères sont sensibles au changement de réseau d'appel (RTC et GSM), excepté le critère SBP sur la base RTC_T_R avec le seuil optimisé sur la base GSM_T.

Il n'y a donc pas de différence importante au niveau de la sensibilité au seuil pour les trois critères. Les trois critères ne sont pas sensibles au changement de base, ni au niveau de bruit sur la base GSM_A, mais sont sensibles au niveau de bruit sur la base RTC_A, et au changement de réseau d'appel (RTC et GSM).

4.4.4 Discussion

Tout d'abord notons que les critères SB et SBP sont très proches, et légèrement meilleurs que le critère LCT, sur la base GSM_A. Par contre sur la base RTC_A, où les communications sont de durée moyenne, le critère LCT donne de meilleurs résultats. Sur la base de parole continue AGORA, qui est très peu bruitée, aucun des trois critères ne se distingue significativement. La différence la plus marquante se trouve sur la partie calme de la base GSM_A bruitée par les deux bruits *car* et *babble*. Les deux critères SB et SBP sont nettement plus performants que le critère LCT. De plus le critère SB est significativement meilleur que le critère LCT (*cf.* tableau G.4 et G.8 en Annexe G).

Nous avons également étudié dans ce paragraphe la sensibilité du seuil des trois critères. Il n'y a pas de différence importante au niveau de la sensibilité du seuil.

Ainsi les performances des critères SB et SBP étant très proches l'une de l'autre, et donnant moins d'erreurs que le critère LCT, c'est l'un des ces critères qui va être choisi pour la suite de cette étude. Ces deux critères ont été conçus pour la détection dans le cas de communications bruitées. Le critère SBP estime les statistiques de la parole sur les temps de parole détectée qui doivent être correctement détectée et d'une durée suffisante pour une bonne estimation des statistiques. Il donne des résultats légèrement meilleurs pour des bruits impulsifs, tandis que le critère SB provoque moins d'erreurs pour des bruits stationnaires (ajout de bruit). La différence étant significative pour les bruits stationnaire, notre choix se porte sur le critère SB.

Afin de situer ces trois critères du module de détection, et pour permettre de dégager les principales caractéristiques et approches de détection de parole envisageables pour améliorer le module de détection et ainsi le système de reconnaissance, nous présentons dans le paragraphe 4.5 différents systèmes de détection décrits dans la littérature. Nous présentons principalement des systèmes de détection de parole pour la reconnaissance vocale.

4.5 Systèmes existants de détection de parole

Les algorithmes des systèmes de détection doivent vérifier quelques critères importants, qui diffèrent selon les auteurs. Ils doivent être selon [Rabiner et Sambur, 1975], simples, fiables, et applicables à diverses conditions d'utilisation, selon [De Souza, 1983], indépendants du locuteur et du vocabulaire, selon [Savoji, 1989], robustes, en temps réel, et n'utilisant pas les connaissances *a priori* sur le bruit.

Toutes les approches font appel à différentes caractéristiques de la parole qui sont soit d'ordre structurel, soit d'ordre statistique, ou bien le plus souvent correspondant à une combinaison des deux.

* L'approche structurelle est définie par un apprentissage, le plus souvent déduit de connaissances *a priori*. Voici quelques exemples de règles sur la structure de la parole :

- Les maxima d'énergie de la parole ne durent pas plus de 2 *s* (*cf.* [Lynch *et al.*, 1987]).

- Un silence intra-mot (tenue de plosive) ne dure pas plus de 150 *ms* (*cf.* [Lynch *et al.*, 1987]), et pas plus de 80 *ms* (*cf.* [Seok et Bae, 1999]).

- Un mot est constitué d'un ou plusieurs maxima d'énergie (*cf.* [Lamel *et al.*, 1981]).

- Plus la distance entre deux maxima est grande, moins il est probable qu'ils appartiennent au même mot (*cf.* [Lamel *et al.*, 1981]).

- La durée minimale d'une voyelle est de 20 *ms* (*cf.* [Dermatas *et al.*, 1991]).

- La durée maximale d'un phonème est de 40 *ms* (*cf.* [Seok et Bae, 1999]).

- La durée minimale d'un mot est de 160 *ms* (*cf.* [Seok et Bae, 1999])

- Il y a une baisse énergétique en fin de phrase (*cf.* [Hariharan *et al.*, 2001]).

Notons aussi que d'après [Un et Lee, 1980], la parole ne représente environ que 40% de la durée des communications téléphoniques.

* L'approche statistique consiste à discriminer le signal de parole, avec des outils statistiques appliqués à des caractéristiques du signal, soit pour aider à la décision, soit pour fusionner ces caractéristiques (*cf.* [Cox et Timothy, 1980]).

Nous présentons dans le paragraphe suivant, différentes caractéristiques acoustiques parmi les plus employées dans des systèmes de détection, puis quelques méthodes statistiques pour la décision et la fusion de ces mêmes caractéristiques, en vue de la détection de la parole.

4.5.1 Caractéristiques acoustiques

De nombreuses caractéristiques de la parole sont employées pour détecter la parole. Bien souvent, elles sont couplées à des règles heuristiques fondées sur les connaissances structurelles de la parole, précédemment présentées.

- **Le spectre**
  Le spectre du signal permet de fournir de nombreuses caractéristiques du signal, il n'est toutefois pas utilisé directement. En effet il est cherché des coefficients en nombre moins important et le plus discriminant possible. Cependant, l'utilisation des coefficients de la sortie du banc de filtres Mel, décrits en Annexe A, est envisageable, comme par exemple dans [Yoma *et al.*, 1996].
- **L'énergie**
  L'énergie est calculée à partir des coefficients issus de l'analyse (par exemple par l'équation (1.1)). Elle est la caractéristique la plus utilisée, mais dans les environnements très bruités, le rapport signal à bruit peut devenir très faible, voire négatif, et l'énergie s'avère alors peu performante. De plus, une énergie forte ne représente bien que les voyelles et quelques consonnes. C'est pourquoi, même si elle reste discriminante, il paraît nécessaire de prendre en compte d'autres caractéristiques.

Cependant l'énergie reste le plus souvent la principale caractéristique des systèmes de détection. Pour diminuer la dynamique de cette caractéristique, c'est le logarithme de l'énergie qui est utilisé en général, en comparant l'énergie à court-terme et l'énergie à long terme (*cf.* [De Souza, 1983], [Martin *et al.*, 1989], [Mauuary et Monné, 1993], [Van Gerven et Xie, 1997]).
Dans [Rabiner et Sambur, 1977] une distance euclidienne normalisée sur le logarithme de l'énergie est introduite :

$$D_k^E = \frac{E(n) - \mu_k}{\sigma_k}, \tag{4.1}$$

où $E(n)$ est le logarithme de l'énergie à la trame n, μ_k et σ_k sont respectivement la moyenne et l'écart-type à long-terme du logarithme de l'énergie dans la classe k (k = 1,2,3 pour une détection en segments voisés/non-voisés/silence).
La moyenne et la variance du logarithme du signal sont reprises dans [Van Gerven et Xie, 1997], où trois méthodes différentes fondées sur le logarithme de l'énergie sont présentées. Les différences sont au niveau des seuils et du calcul de l'énergie à court-terme et à long-terme. Ces approches sont plus ou moins bien adaptées selon l'intensité des bruits de fond des environnements d'appel. La première version n'est pas réalisable en temps réel, elle nécessite le calcul du logarithme de l'énergie sur tout le signal afin de déterminer un seuil fixe pour la détection. La deuxième approche est une comparaison du logarithme de l'énergie à deux seuils obtenus par addition de deux constantes à la moyenne du logarithme de l'énergie. La troisième version, qui apporte de meilleurs résultats en environnement bruité, est fondée sur le même principe. Les deux seuils sont cependant obtenus à l'aide de l'estimation de la moyenne et de la variance du bruit. Un tel calcul de seuil est très proche de celui du critère SB présenté au Chapitre 2 "*Détection de parole pour la reconnaissance vocale*".
Différentes approches ont cependant été étudiées pour utiliser au mieux cette caractéristique. L'énergie peut être utilisée dans différentes bandes de fréquence (*cf.* [Bendiksen et Steiglitz, 1990], [Cohn, 1991], [Shin *et al.*, 2000], [Hariharan *et al.*, 2001]), ou calculée de façon à tenir compte de l'amplitude et de la fréquence (*cf.* [Ying *et al.*, 1993]), ou bien encore filtrée (*cf.* [Li *et al.*, 2001]).
Dans [Reaves, 1997] un automate à quatre états est dirigé par la comparaison de la variance de l'énergie limitée en fréquence à deux seuils. Mais les statistiques d'ordre supérieur de l'énergie peuvent aussi apporter une précision supplémentaire à la décision. Les cumulants d'ordre 3 et 4 donnent des informations sur la symétrie et l'aplatissement de la distribution statistique du signal. Le cumulant d'ordre 3 normalisé (ou *skewness*) est nul pour une densité symétrique, c'est le cas des gaussiennes, et pour une gaussienne le cumulant d'ordre 4 normalisé (ou *kurtosis*) vaut 3. Les cumulants pouvant s'exprimer en fonction des moments, [Jacovitti *et al.*, 1991] propose d'intégrer les moments normalisés calculés sur le signal, dans un algorithme en vue d'une détection en segments voisés/non-voisés/silence (*cf.* paragraphe 7.6). Au Chapitre 3 "*Analyse des sources d'erreurs du module de détection*", nous

avons vu qu'une discrimination énergétique plus précise du bruit et de la parole peut permettre une diminution des insertions. Au Chapitre 7 "*Utilisation des statistiques d'ordre supérieur*", les statistiques d'ordre supérieurs sont étudiées dans cette optique.

- **Le taux de passage par zéro**

  Souvent utilisé avec l'énergie, le taux de passage par zéro du signal est un paramètre calculé facilement (par exemple par l'équation (1.3)) et qui donne une information importante au niveau de la distribution spectrale du signal.

  Le taux de passage par zéro permet de détecter les fricatives de faible énergie aux frontières des mots.

  Nous avons vu que ce paramètre a été très utilisé dans les systèmes de détection en segments voisés/non-voisés/silence.

  Citons [Junqua *et al.*, 1994] et [Savoji, 1989] où sont développés des algorithmes faisant appel à l'énergie et au taux de passage par zéro. Dans [Savoji, 1989], ces deux caractéristiques sont combinées par des mesures utilisant des connaissances statistiques du signal. Dans [Ganapathiraju *et al.*, 1996], ces caractéristiques contrôlent les transitions d'un automate à quatre états. Un taux de passage par zéro modifié a été introduit dans [Hahn et Park, 1992], qui est associé à l'énergie et au taux de passage par zéro classique par un jeu de tests avec des seuils fixes.

  Cependant ces détections sont mises en œuvre pour des systèmes de reconnaissance en milieu calme. D'après [Shin *et al.*, 2000] et [Huang et Yang, 2000], le taux de passage par zéro est instable sur les parties du signal bruitées. C'est pourquoi nous n'avons pas étudié cette caractéristique dans ce travail.

- **Le pitch ou paramètre de voisement**

  Le pitch souvent confondu par abus de langage à la fréquence fondamentale, représente la périodicité du signal et sa structure harmonique. C'est en ce sens que nous pensons qu'il est un paramètre discriminant de la parole et du bruit. Le problème principal est l'extraction de la valeur du pitch dans le signal, déterminant l'existence de segments périodiques. Des études comparatives de l'estimation du pitch sont données dans [Hess, 1983] et [Bagshaw, 1994].

  Pour une détection en segments voisés/non-voisés/silence, une méthode d'extraction du pitch à l'aide de la divergence de Kullback est présentée dans [Di Francesco, 1990]. Dans [Hamada *et al.*, 1990] une estimation simple est utilisée, mais elle est dite peu précise pour extraire le pitch. Cette méthode est fondée sur les pics énergétiques du signal. L'intervalle de temps entre ces pics est évalué et permet d'estimer la périodicité du signal. Cette valeur est ensuite comparée à un seuil, les faibles valeurs déterminent les états de non-parole, la parole étant supposée périodique.

  Le problème réside dans le fait que les périodes de bruit peuvent contenir du voisement, notamment pour les bruits de fonds. Cependant nous verrons que cette caractéristique de la parole peut permettre une suppression d'une partie importante des insertions et une meilleure détection de la parole continue (*cf.* Chapitre 8 "*Utilisation d'un paramètre de voisement*", où d'autres systèmes de détection sont

présentés, [Ramana Rao et Srichand, 1996], [Strom, 1995] et [Sakurai et Hirose, 1996]).

- **L'abscisse curviligne**

L'abscisse curviligne du signal temporel est une caractéristique qui représente à la fois l'amplitude du signal et sa fréquence. Ceci regroupe ainsi l'information énergétique et le taux de passage par zéro. L'abscisse curviligne du signal temporel s_n est calculée par :

$$abs(N) = \sum_{n=1}^{N} |s_n - s_{n-1}|, \tag{4.2}$$

où N est l'indice des échantillons. C'est en fait une dérivée lissée du signal δabs calculée sur une fenêtre de 32 ou 64 échantillons, qui est utilisée par [André-Obrecht *et al.*, 1993]. Cette caractéristique permet d'indiquer si le signal est de la parole (δabs fort) ou du bruit (δabs faible) par comparaison à deux seuils, qui introduisent une zone floue traitée avec des connaissances heuristiques du signal (*cf.* [Puel, 1997]).

- **Les coefficients d'autocorrélation**

Différents coefficients d'autocorrélation ont été introduits dans des systèmes de détection, pour permettre de prendre en compte l'évolution du signal au cours du temps. Ces caractéristiques sont reliées au voisement. Ainsi, ces coefficients ont été utilisés pour la détection en segments voisés/non-voisés/silence. Ils sont définis par :

$$\phi_i = \sum_{n=i}^{N} s_n s_{n-i}, \quad \text{où} \quad i = 1, \cdots, p, \tag{4.3}$$

où N est l'indice des échantillons et s_n le signal temporel. Il est montré dans [Atal et Rabiner, 1976] que ϕ_1, normalisé par :

$$\sqrt{\sum_{n=1}^{N} s_n^2 \sum_{n=0}^{N-1} s_n^2}, \tag{4.4}$$

est proche de 1 lorsque le son est voisé, car dans ce cas l'énergie est concentrée dans les basses fréquences et la corrélation de deux trames adjacentes est forte, sinon ϕ_1 est proche de zéro. Ce coefficient est repris dans [De Souza, 1983]. Les douze premiers coefficients de corrélation avec quatre autres caractéristiques dans plusieurs systèmes de détection sont testés dans [Rabiner *et al.*, 1977]. Seuls les quatre premiers sont retenus dans les meilleurs jeux de caractéristiques.

Cependant il est possible d'utiliser un grand nombre de coefficients de corrélation, en considérant la matrice entière d'autocorrélation. [Rangoussi *et al.*, 1993] emploie la matrice d'autocorrélation, et calcule les valeurs propres de cette matrice. Sous une hypothèse très forte que le bruit est un bruit blanc, ces valeurs propres sont toutes égales à la variance du bruit si la trame courante est du bruit, et sont toutes

supérieures à la variance du bruit, si la trame courante est de la parole bruitée. Les tests faits uniquement avec l'ajout d'un bruit blanc ne permettent pas de montrer l'efficacité de la méthode pour des bruits stationnaires d'une autre nature. De plus cette approche ne permet pas d'apporter une solution à la diminution des insertions qui sont des détections de bruits de courte duré.

Dans [Zhu et Chen, 1999] la matrice d'autocorrélation est également employée, normalisée par $\frac{1}{N-i}$, en calculant une distance de cette matrice d'autocorrélation à court-terme du signal à une estimation de la matrice d'autocorrélation du bruit. Cette distance comparée à un seuil permet la décision. Elle est définie par :

$$d_\phi = \min_a \frac{\sum_{i=1}^{p} (\phi_i - a\phi_i^B)^2}{\sum_{i=1}^{p} (\phi_i)^2}, \tag{4.5}$$

où ϕ_i et ϕ_i^B sont respectivement les matrices d'autocorrélation à court-terme du signal et du bruit.

- **L'analyse LPC (Linear Predictive Coding)**

L'analyse LPC décrite en Annexe A, permet d'obtenir différents coefficients employés pour la détection de parole.

Le premier coefficient LPC, et l'erreur de prédiction normalisée sont utilisés par [Atal et Rabiner, 1976], parmi cinq paramètres. Une distance sur le logarithme de l'énergie est associée à une distance de covariance sur les coefficients LPC, par [Rabiner et Sambur, 1977] pour une détection en segments voisés/non-voisés/silence. La distance de covariance est de la forme :

$$D_k^{LPC} = \frac{(\mathbf{a} - \mu_\mathbf{k})^* \Phi (\mathbf{a} - \mu_\mathbf{k})}{\mathbf{a}^* \Phi \mathbf{a}}, \tag{4.6}$$

où $\mathbf{a}$ est le vecteur des coefficients LPC, $\mu_\mathbf{k}$ la moyenne du vecteur pour la classe k du signal (k = 1,2,3 représente le silence, la parole non-voisée, et la parole voisée) et Φ est la matrice de covariance pour la trame courante. D_k^{LPC} est essentiellement une covariance pondérée de coefficients LPC. Cette distance est combinée avec la distance sur l'énergie par un produit ou une somme. Le but de cette distance est d'utiliser l'information spectrale du signal.

Dans [Kobatake *et al.*, 1989] une DBP sensible à la détection des consonnes faibles, également fondée sur l'analyse LPC est proposée. Cette détection est composée d'une première détection des parties non stationnaires du signal, puis d'une discrimination de ces parties en parole ou bruit. La première détection est fondée sur le principe que l'erreur de prédiction finale décroît en fonction du temps lorsque le signal est stationnaire et croît en fonction du temps lorsqu'il est non stationnaire. La méthode de classification ensuite utilisée, considère cinq paramètres : la périodicité, la fréquence fondamentale, l'ordre optimal du modèle LPC, et une distance LPC

minimale. Ces caractéristiques sont comparées à des seuils pour chaque trame sur les parties détectées non stationnaires. La détection est de la parole, si le pourcentage de trames qui vérifient les cinq conditions est supérieur à un seuil prédéterminé. Cette approche nécessite un grand nombre de seuils, de plus l'évaluation est faite uniquement par comparaison des frontières. Cette évaluation ne permet donc pas de mettre en évidence les pourcentages des erreurs d'insertion et d'omission de la DBP. De plus, le bruit ajouté est un bruit blanc.

Dans de nombreuses autres études des coefficients issus de l'analyse LPC sont utilisés en vue de la détection pour un système de reconnaissance utilisant également les coefficients LPC. Cependant dans notre étude, le module de reconnaissance n'utilisant pas les coefficients LPC, le calcul de ces coefficients pour la DBP entraînerait un coût important pour le système de reconnaissance. Cependant nous disposons des coefficients cepstraux qui représentent également l'enveloppe spectrale, mais dans un autre espace.

- **Les coefficients cepstraux**

Les coefficients cepstraux sont les composants du cepstre. Le calcul de ces coefficients est rappelé en Annexe A.

Les coefficients cepstraux ou les MFCC, qui sont les coefficients cepstraux calculés après la sortie du banc de filtres de l'échelle Mel, sont très souvent utilisés comme paramètres dans les modules de reconnaissance (*cf.* [Jouvet, 1988], [Hamada *et al.*, 1990], [Héon *et al.*, 1998], [Huang et Yang, 2000], *etc.*). Il est donc intéressant d'utiliser ces coefficients qui sont déjà calculés pour le module de reconnaissance.

La DAV proposée dans [Haigh et Mason, 1993], est fondée sur un apprentissage des coefficients cepstraux du bruit et de la parole. La décision est prise à l'aide d'une mesure de similarité, qui est une distance euclidienne des coefficients de la trame courante avec les modèles de bruit et de parole. Le problème de cette méthode souligné par les auteurs, est la sensibilité au changement de statistiques des périodes de non-parole.

Pour la réduction de bruit avant un système de reconnaissance vocale, les informations cepstrales peuvent également être employées (par exemple à l'aide de réseau de neurones dans [Héon *et al.*, 1998], *cf.* paragraphe 5.2).

Au Chapitre 9 "*Utilisation de la fusion de données*", nous proposons une combinaison des coefficients cepstraux pour permettre la réduction des insertions du module de DBP de France Télécom R&D.

- **L'entropie**

Nous pouvons définir de différentes manières l'entropie, qui est une mesure de stabilité du signal.

- Un algorithme fondé sur l'entropie de Shannon est implémenté dans [Abdallah *et al.*, 1997a] (*cf.* également [Abdallah *et al.*, 1997b]). Il est défini une séquence $\{ x_{[0,N-1]}(n)\}$ pour $0 \leq n \leq N-1$ de N échantillons du signal $x(k)$, par rapport à la base $\{ X_{[0,N-1]}(k)\}_{0 \leq k \leq N-1}$ des coefficients de la transformée de Fourier discrète,

c'est-à-dire, pour $0 \leq k \leq N-1$:

$$X_{[0,N-1]}(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x_{[0,N-1]}(n) e^{-\frac{2\Pi i}{N} nk}. \tag{4.7}$$

L'entropie de Shannon est définie comme une entropie spectrale par :

$$E_{[0,N-1]} = -\sum_{k=0}^{N-1} |X_{[0,N-1]}(k)|^2 \ln |X_{[0,N-1]}(k)|^2. \tag{4.8}$$

Un critère entropique local (CEL) est ensuite défini. Le CEL est une fonction sensible aux variations du spectre du signal, et qui est plus précisément une mesure sur les variations de la concentration d'énergie du spectre à court-terme du signal :

$$CEL(m) = \frac{E_{[0,N-1]} - (E_{[0,\frac{N}{2}]} + E_{[\frac{N}{2},N-1]})}{|E_{[0,\frac{N}{2}]} + E_{[\frac{N}{2},N-1]}|}, \tag{4.9}$$

où m est le milieu de la fenêtre spectrale [0,$N-1$]. La dépendance temporelle est obtenue en faisant glisser la fenêtre d'analyse point par point sur le signal. La séquence $\{ x_{[0,N-1]}(k)\}$ pour $0 \leq k \leq N-1$ est considérée comme *entropiquement instable* et son milieu m constituera un point d'instabilité si :

$$CEL(m) > 0 \quad (\Leftrightarrow E_{[0,N-1]} > E_{[0,\frac{N}{2}]} + E_{[\frac{N}{2},N-1]}), \tag{4.10}$$

dans le cas contraire la séquence est considérée comme *entropiquement stable*. Plusieurs changements dans le spectre à court-terme du signal se traduisent par la génération de plusieurs points d'instabilité contigus constituant une *zone d'instabilité*. Il est fait l'hypothèse qu'une zone d'instabilité traduit un changement ou une rupture dans l'évolution du signal. Cette mesure permet de localiser les parties entropiquement stables du signal dans le bruit, qui sont supposées être les périodes de parole. Cette hypothèse n'est cependant pas vérifiée dans le cas de bruits stationnaires qui sont entropiquement stables. Cette approche est comparée dans [Renevey, 2000] avec une approche fondée sur le logarithme de l'énergie comparé à deux seuils adaptatifs. L'entropie permet une meilleure détection dans les cas de communications très bruitées.

- À la différence de l'entropie de Shannon (équation (4.8)) dans [Shen *et al.*, 1998] une entropie spectrale est calculée sur le spectre normalisé considéré comme une densité de probabilité. La fonction densité de probabilité est définie par :

$$p_i = \frac{s(f_i)}{\sum_{k=0}^{N-1} s(f_k)}, \quad i = 0,...,N-1, \tag{4.11}$$

où $s(f_i)$ est l'énergie spectrale pour la composante fréquentielle f_i et N est le nombre total de composantes fréquentielles dans la transformée de Fourier. Les auteurs restreignent les composantes fréquentielles entre 250 et 6000 *Hz*. Ainsi :

$$s(f_i) = 0, \text{ si } \; f_i < 250\,Hz \; \text{ ou } \; f_i > 6000\,Hz. \tag{4.12}$$

De plus, ils restreignent la densité de probabilité :

$$p_i = 0, \text{ si } \; p_i < \delta_2 \; \text{ ou } \; p_i > \delta_1, \tag{4.13}$$

où les bornes δ_1 et δ_2 sont calculées empiriquement. δ_1 est utilisé pour éliminer le bruit contenu dans certaines bandes de fréquences spécifiques, tandis que δ_2 est utilisé pour oublier les bruits de valeur de densité spectrale presque constante, comme le bruit blanc. L'entropie spectrale est alors définie par :

$$H = -\sum_{k=0}^{N-1} p_k \ln p_k. \tag{4.14}$$

Un ensemble de facteurs de pondération w_k est ensuite appliqué pour ajuster la composante fréquentielle à l'entropie spectrale. Nous avons alors :

$$H = -\sum_{k=0}^{N-1} w_k p_k \ln p_k. \tag{4.15}$$

L'ensemble des facteurs de pondération w_k est estimé statistiquement par apprentissage sur un grand nombre de signaux de parole. L'entropie spectrale définit les périodes de parole, lorsque H est inférieur à un seuil. Un autre jeu de seuils dérivé de l'analyse du signal, est utilisé pour affiner la détection. Cette dernière méthode est d'après les auteurs très performante dans les environnements bruités, pour la détection des débuts et fins de mots. Il paraît cependant difficile d'obtenir les facteurs de pondération et les seuils pour une détection robuste pour des applications réelles.

- Cette entropie est reprise dans [Huang et Yang, 2000] et dans [Yang et Hsieh, 2000], sans considérer l'ensemble des facteurs de pondération w_k (équation (4.14)). L'entropie H est combinée avec l'énergie par simple multiplication. Il obtient ainsi :

$$M = (E - \mu_E).(H - \mu_H), \tag{4.16}$$

où μ_E et μ_H sont les moyennes respectives de l'énergie et de l'entropie. Un nouveau coefficient est défini par :

$$EE = \sqrt{1 + |M|}. \tag{4.17}$$

EE permet de détecter les périodes de parole par comparaison à un seuil.

Comme le montrent ces approches, l'entropie est une caractéristique qui peut être difficile à mettre en œuvre dans le cas d'environnement bruité, par exemple dans le cas d'un bruit de forte intensité et stationnaire. Elle n'est pas utilisée dans cette étude.

- **Autres caractéristiques**

- Dans [De Souza, 1983], en plus du logarithme de l'énergie, du taux de passage par zéro et de l'autocorrélation, le nombre de passages par zéro de la dérivée du signal et une mesure sur le signal s_n sont employés. s_0, s_1, $\cdots$, s_{N-1} représentent les valeurs du signal digitalisé sur la trame observée. Cette mesure est définie par :

$$D = \log_{10}\left(\frac{\sum_{n=1}^{N-1}|s_n - s_{n-1}|}{\sqrt{\sum_{n=0}^{N-1}s_n^2}}\right), \tag{4.18}$$

qui représente le caractère "ombré" (*shade*) du signal, que nous percevons à l'oeil. Cette mesure est très proche de l'abscisse curviligne introduite dans [André-Obrecht *et al.*, 1993], équation (4.2). Le nombre de passages par zéro de la dérivée du signal donne une information sur le "cisaillement" (*jaggeness*) du signal. Notons que les cinq caractéristiques précédentes sont utilisées par un test statistique fondé sur la moyenne et la covariance, pour un algorithme silence/non-silence dans un environnement calme.

- Une mesure de non-stationnarité du signal est introduite dans [Yoma *et al.*, 1996] :

$$NST = 20\ln\frac{\sqrt{\sum_{k=1}^{14}(E_{i,k} - E_{i-1,k})^2}}{\sqrt{\sum_{k=1}^{14}(E_k^b)^2}}, \tag{4.19}$$

où $S_i = (E_{i,1},E_{i,2},\cdots,E_{i,14})$ et $S_i^b = (E_1^b,E_2^b,\cdots,E_{14}^b)$ représentent respectivement les estimations spectrales du signal à la trame i et du bruit à la sortie d'un banc de 14 filtres de l'échelle Mel. Cette caractéristique est associée à une distance spectrale de l'estimation spectrale à la sortie de chaque filtre de la trame courante à l'estimation à long-terme de l'énergie spectrale du bruit. Mais cette approche ne semble pas d'après les auteurs apporter une amélioration dans un environnement bruité.

- Dans [Agaiby et Moir, 1997] la méthode de Tucker élaborée à partir d'une mesure sur la périodicité du signal voisé est présentée. Pour des raisons de délai, et de détection uniquement des périodes voisées, cette approche n'est pas utilisée. Notons également que pour éviter les interférences périodiques, un préprocesseur est employé. C'est la cohérence du signal présenté dans [Le Bouquin-Jeannès et Faucon, 1995], qui est employée dans [Agaiby et Moir, 1997]. Cette fonction n'est cependant calculable qu'avec deux microphones.

- Une nouvelle caractéristique est définie dans [Seok et Bae, 1999] à partir de la transformé discrète en ondelettes du signal. Les coefficients d'ondelette sur le troisième

niveau de décomposition représentent les sons voisés, tandis que les coefficients sur le premier niveau indique l'existence de fricatives ou plosives. La somme pondérée de l'écart-type du coefficient sur le troisième niveau avec le coefficient du premier niveau fournit une caractéristique, qui est comparée à un seuil adaptatif. Ce test, ainsi que des connaissances heuristiques permettent d'obtenir des résultats de détection performants avec un RSB faible, si la tolérance au niveau de la précision des frontières, exprimée en *ms*, n'est pas trop petite.

- Notons que nous trouvons dans [Rabiner *et al.*, 1977], l'algorithme de [Atal et Rabiner, 1976] testé avec 70 paramètres différents, pour l'optimisation du jeu de cinq paramètres à utiliser. 12 coefficients LPC, 12 coefficients de corrélation, 12 coefficients PARCOR, 12 termes d'erreur partielle LPC, *etc.*, ont été utilisés.

Une distance de probabilité pour combiner ces caractéristiques et obtenir une décision en segments voisés/non-voisés/silence est utilisée dans [Atal et Rabiner, 1976], mais d'autres méthodes statistiques de fusion de données et de décision que nous présentons ci-après sont employées par d'autres auteurs.

4.5.2 Quelques méthodes statistiques

Rappelons tout d'abord que toutes ces caractéristiques sont employées le plus souvent par comparaison à des seuils adaptatifs, qui sont optimisés par apprentissage et à partir de connaissances heuristiques du signal. Par exemple [Watanabe et Kimura, 1991] et [Hsieh, 1998] emploient deux seuils pour déterminer les variations de l'énergie. Cependant des méthodes plus systématiques de décision et de fusion de ces caractéristiques sont également utilisées. Nous présentons ci-dessous différentes méthodes statistiques, qui sont regroupées selon qu'elles sont utilisées pour la décision ou pour la fusion de données.

La décision

Il est important d'avoir le moins possible de seuils à régler dans un système de détection, afin d'obtenir des systèmes plus robustes au changement de conditions des tests. Les seules connaissances heuristiques des variations du signal de parole ne suffisent pas à obtenir une telle robustesse aux conditions d'appel. Plusieurs approches statistiques sont employées.

- **La décision bayesienne**
  Le critère statistique du rapport de vraisemblance est utilisé par [Karray et Monné, 1998], pour discriminer la parole du bruit à partir des courbes représentatives des distributions de l'énergie du bruit et de la parole (*cf.* paragraphe 2.2.4). En notant, respectivement, H_0 et H_1, les hypothèses d'état dans le bruit et dans la parole, et en les supposant équiprobables, le seuil de passage d'un état à l'autre est obtenu par la résolution de l'équation :

$$P(x/H_0) = P(x/H_1), \tag{4.20}$$

où $P(x/H_i)$, $i = 0{,}1$, est la probabilité conditionnelle de l'observation x sous l'hypothèse H_i. Il faut alors faire une hypothèse sur la distribution du bruit et de la parole. Il est supposé dans un premier temps que les distributions sont gaussiennes, l'équation (4.20) est alors du second degré, puis pour simplifier la résolution de cette équation, les distributions sont supposées laplaciennes.

Cette méthode a également été employée dans [Arslan et Hansen, 1998] et [Singh *et al.*, 2001]. Elle permet d'utiliser les statistiques d'ordre 1 et 2 du bruit et de la parole. L'hypothèse de distributions gaussiennes, également faite dans [Atal et Rabiner, 1976], [De Souza, 1983], [Bruno *et al.*, 1987], [Di Francesco, 1990] et [Van Gerven et Xie, 1997], est acceptable si l'apprentissage des statistiques est fait sur un grand nombre de trames (théorème de la limite centrale). Cependant, les statistiques de la parole sont estimées sur les détections trouvées comme étant de la parole, ce qui représente peu de trames dans le cas de mots isolés. Dans [Cox et Timothy, 1980] la distribution du bruit est supposée gaussienne tandis que la distribution de la parole est supposée laplacienne. En effet d'après [Cox et Timothy, 1980], l'estimation des statistiques d'une distribution gaussienne converge plus lentement si elle se superpose avec une autre distribution gaussienne, comme celles du bruit et de la parole. Le problème de l'estimation des paramètres de la distribution laplacienne reste cependant toujours délicat dans le cas de la parole, où peu de trames sont disponibles.

Deux procédures de décision sont proposées dans [Cox et Timothy, 1980]. La première est un test d'hypothèses classique sur la sortie d'un banc de quatre filtres. La deuxième est la procédure de décision multiple de Krusal-Wallis. La statistique de Krusal-Wallis qui suit une loi du χ^2 permet un test du χ^2. Ces deux tests d'hypothèses reviennent à la comparaison de quatre seuils pour les quatre filtres. L'hypothèse de même distribution pour les quatre filtres limite cependant l'étude aux bruits en bande large.

- **Opérateurs logiques**

  Dans [Rabiner et Sambur, 1977] une distance sur le logarithme de l'énergie est combinée avec une distance sur les coefficients LPC. Cette combinaison est faite de deux façons différentes soit en sommant les deux distances, soit en les multipliant afin d'obtenir une nouvelle distance. Cette nouvelle distance permet donc de fusionner la décision des deux distances. Sommer ces deux distances suppose que les caractéristiques de l'énergie et des coefficients LPC sont indépendantes, ce qui n'est pas le cas. Multiplier les deux distances revient, une fois normalisées, à la mesure de probabilité présentée plus loin et introduite par [Atal et Rabiner, 1976]. Cette seconde approche est plus satisfaisante d'après les auteurs, mais n'a pas été utilisée pour des raisons d'implémentation.

- **La logique floue**

  Dans [Mwangi et Xydeas, 1985] les cinq caractéristiques de [Atal et Rabiner, 1976] sont employées, pour une décision à l'aide des ensembles flous. Pour chaque caractéristique, trois classes sont définies selon la valeur de la caractéristique, faible, moyenne ou forte. Une base d'apprentissage permet d'obtenir les connaissances qui

donnent la classe de chaque caractéristique selon le segment voisé/non-voisé/silence, et les seuils qui permettent de classer chaque caractéristique dans une des trois classes. Ainsi, pour chaque segment, une règle est établie, c'est le maximum de vraisemblance (prenant une valeur de 1 à 5) qui permet la classification. Les résultats obtenus, sont néanmoins peu performants face à la méthode de [Atal et Rabiner, 1976] qui demande moins d'apprentissage.

Dans [Cavallaro *et al.*, 1998] la logique floue avec six règles et quatre caractéristiques est utilisée pour une DAV. Les quatre caractéristiques sont les caractéristiques du codeur G.729 données dans [ITU Recommendation, 1996] : l'énergie, l'énergie dans les basses fréquences, le taux de passage par zéro et la distorsion spectrale (*cf.* respectivement les équations (1.1), (1.2), (1.3) et (1.4)). Comme précédemment, ces caractéristiques sont divisées en trois classes, fort, moyen ou faible valeur.

L'utilisation de règles pour la prise de décision, revient à ordonner les connaissances heuristiques sur les caractéristiques. Cette approche méthodique permet de manipuler plusieurs caractéristiques, cependant une bonne connaissance *a priori* des variations, et de la représentation physique de ces caractéristiques est indispensable.

- **Arbres de décision - Méthodes de segmentation**

  Les arbres de décision permettent de systématiser la logique floue, en ne considérant plus qu'une seule règle. Dans [Shin *et al.*, 2000] la méthode CART (*Classification And Regression Trees*) décrite au paragraphe 9.2.2, est employée pour une détection de parole/non-parole pour la reconnaissance vocale. Six paramètres énergétiques permettent six décisions. Les modules de détection pour chaque paramètre sont les mêmes (il s'agit de comparer une caractéristique à un seuil adapté dans les périodes de non-parole), et fonctionnent en parallèle. Les décisions sont ensuite combinées conditionnellement. Ces paramètres doivent être choisis de manière à pouvoir donner une décision individuellement. Les paramètres utilisés sont : l'énergie dans toutes les bandes de fréquence, l'énergie dans la bande de fréquence audible (300-3700 *Hz*) et dans les hautes fréquences (2-4 *kHz*), une information sur les pics énergétiques, l'énergie des résidus des LPC et l'énergie du bruit filtré. Ces six paramètres sont très redondants, mais l'énergie dans les hautes fréquences permet de détecter les consonnes, les pics énergétiques permettent de détecter les parties voisées du signal, l'énergie des résidus des LPC est robuste aux bruits de basse fréquence, et l'énergie du bruit filtré permet de remédier à l'influence du bruit ambiant.

  La construction d'un arbre de décision nécessite une base d'apprentissage. Le formalisme d'arbre de décision binaire semble bien adapté dans notre cas où le nombre de classe est restreint à deux. Cette approche est discutée au paragraphe 9.2.2.

La fusion en entrée

Les approches présentées précédemment, utilisent une caractéristique à la fois. C'est-à-dire qu'une caractéristique du signal fournit une information, et les informations obtenues sont combinées de manière à donner une décision. Ces approches sont intéressantes lorsque peu de caractéristiques sont utilisées pour la détection, avec une bonne connaissance de

leurs variations.

Il est cependant intéressant d'employer un grand nombre de caractéristiques, comme les coefficients cepstraux, qui sont difficiles à interpréter physiquement, il est donc délicat de les employer séparément. Les systèmes présentés ci-dessous permettent la fusion d'un grand nombre de données.

Notons néanmoins que les approches de la logique floue et des arbres de décision présentées peuvent être considérées comme de la fusion en entrée, et adaptées pour un grand nombre de caractéristiques.

- **Mesures de probabilité**

  Les mesures de probabilité, ou densités de probabilité, sont utilisées pour la décision de différentes façons.

  [Atal et Rabiner, 1976] utilise la distance de covariance de la trame courante à une classe (voisée/non-voisée/silence) :

$$d_k = (\mathbf{x} - \mu_k)^* \mathbf{D}_k^{-1} (\mathbf{x} - \mu_k), \tag{4.21}$$

  où $\mathbf{x}$ est le vecteur de coefficients, μ_k est la moyenne de la classe k et $\mathbf{D}_k$ la matrice de covariance de la classe $k = 1,2,3$ pour voisée, non-voisé et silence. Cette distance est également appelée distance de Mahalanobis locale. Cette distance a été reprise dans [Rabiner et Sambur, 1977] (*cf.* équation (4.6)) pour intégrer les coefficients LPC. A partir de cette distance, une mesure de probabilité permettant la décision est donnée par :

$$P_1 = \frac{d_2 d_3}{d_1 d_2 + d_2 d_3 + d_1 d_3}, \tag{4.22}$$

$$P_2 = \frac{d_1 d_3}{d_1 d_2 + d_2 d_3 + d_1 d_3}, \quad \text{et} \tag{4.23}$$

$$P_3 = \frac{d_1 d_2}{d_1 d_2 + d_2 d_3 + d_1 d_3}. \tag{4.24}$$

  La plus grande probabilité détermine la classe. [Bruno *et al.*, 1987] a repris cette distance avec les mêmes caractéristiques. La mesure de probabilité est cependant affinée à l'aide du rapport de vraisemblance précédemment décrit.

  La distance de Mahalanobis locale a été employée dans [Smith *et al.*, 1999] afin d'effectuer une analyse quadratique discriminante dans le but d'une détection d'activité vocale. Ainsi la probabilité que l'observation du vecteur de coefficients $\mathbf{x}$ soit dans la classe k est donnée par :

$$P(\mathbf{x}/k) = \frac{p_k \sqrt{|D_k|}^{-1} e^{-d_k}}{\sum_l p_l \sqrt{|D_l|}^{-1} e^{-d_l}}, \tag{4.25}$$

  où p_k est la probabilité *a priori* d'être dans la classe k. Cette probabilité permet d'assigner l'observation $\mathbf{x}$ à la classe k pour la plus grande valeur de $P(\mathbf{x}/k)$.

Dans [Hörmann et Rozinaj, 1998], la matrice de covariance est utilisée pour déterminer la distance de Mahalanobis globale sur deux caractéristiques. Cette distance décrite au paragraphe 9.2.1 peut être donnée par :

$$d_g = (\mathbf{x} - \mathbf{m})^* \mathbf{D}^{-1} (\mathbf{x} - \mathbf{m}), \tag{4.26}$$

où $\mathbf{x}$ est le vecteur de coefficients, $\mathbf{m}$ est la moyenne de coefficients et $\mathbf{D}$ la matrice de covariance totale.
La première caractéristique est définie par:

$$ZRC(i) = \sum_{n=0}^{K-1} |MFCC(i,n) - \overline{MFCC}(n)|^2, \tag{4.27}$$

où $MFCC(i,n)$ représente le $n^{\text{ième}}$ coefficient cepstral du bloc i, composé de 16 trames, $\overline{MFCC}(n)$ est la moyenne du $n^{\text{ième}}$ coefficient cepstral sur le bloc, et $K = 10$. Cette caractéristique est en fait une estimation de la variance locale des coefficients cepstraux. La seconde caractéristique est donnée par la moyenne des 160 valeurs des coefficients cepstraux sur un bloc. La distance de Mahalanobis sur ces 2 caractéristiques à la trame courante permet de déterminer la présence de parole par comparaison à un seuil. La détection Bruit/Parole de [Hörmann et Rozinaj, 1998] est développée pour un système de reconnaissance de mots isolés.
L'utilisation de la matrice de corrélation permet donc de fusionner un ensemble important de données. Dans le paragraphe 9.3 nous implémentons la méthode d'analyse factorielle discriminante qui est également une manière de prendre en compte la matrice de corrélation.

- **Les réseaux neuronaux**
  Dans [Bendiksen et Steiglitz, 1990] et [Cohn, 1991] des réseaux neuronaux sont utilisés pour une détection en segments voisés/non-voisés de la parole avec en entrée des coefficients sur l'énergie. Ces méthodes sont utilisées pour des applications demandant une grande précision, mais sans contraintes de temps et avec une connaissance du bruit. Cependant cette approche peut être utilisée pour la fusion d'un grand nombre de données comme les coefficients cepstraux.
  Dans [Ghiselli-Cripa et El-Jaroudi, 1991] un réseau de neurones à deux couches cachées, est également employé pour une détection en segments voisés/non-voisés/silence avec en entrée les mêmes caractéristiques que [Atal et Rabiner, 1976]. Le fait d'utiliser deux couches cachées augmente la complexité de l'apprentissage. Les résultats comparés à ceux de [Atal et Rabiner, 1976] sont meilleurs, mais l'application testée est monolocuteur.
  Dans le paragraphe 9.2.4, nous discutons de ces avantages.

Un grand nombre de caractéristiques acoustiques, comme les coefficients cepstraux, peut être employé en vue d'améliorer les performances du module de détection. Les méthodes présentées ci-dessus, sont des approches possibles pour l'intégration de ces caractéristiques. D'autres méthodes sont envisageables. Dans le paragraphe 9.2, nous discutons sur la meilleure approche pour notre problème, et en implémentons l'une d'elles.

4.6 Études de caractéristiques du signal discriminant le bruit et la parole

Le paragraphe 4.5 permet de dégager des caractéristiques du signal qui peuvent aider à la décision particulièrement dans des environnements bruités. Nous étudions ici ces caractéristiques du signal permettant de discriminer les périodes de bruit et de parole du signal.

Nous avons vu dans le paragraphe 4.5.1 que l'énergie est la principale composante du signal utilisée pour la détection de parole. Dans un premier paragraphe 4.6.1 nous étudions l'énergie sur les parties de bruit et de parole du signal. Nous étudions ensuite la fréquence fondamentale au paragraphe 4.6.2 qui est une caractéristique employée surtout pour la détection en segments voisés/non-voisés/silence. Cependant elle peut venir en complément de l'énergie pour discriminer la parole des bruits impulsifs qui ne sont en général pas voisés. Les MFCC étudiés au paragraphe 4.6.3, sont utilisés dans le module de reconnaissance et fournissent une information détaillée du signal. Les coefficients du vocodeur sont également étudiés au paragraphe 4.6.4 car ces coefficients sont calculés pour obtenir les MFCC.

4.6.1 L'énergie du signal

L'énergie du signal est la caractéristique la plus utilisée pour détecter les périodes de parole dans le signal. Nous détaillons le comportement du logarithme de l'énergie dans les périodes de bruit et de parole pour comprendre les différences énergétiques du bruit et de la parole bruitée. Nous étudions ici le logarithme de l'énergie sur les périodes de bruit et de parole, pour les bases RTC_A et GSM_A.

La figure 4.11 représente le logarithme de l'énergie au cours du temps sur des périodes de parole (marquées ici par les segments manuels) et de bruits d'une communication de la base GSM_A. Nous constatons que le logarithme de l'énergie permet de bien discriminer la parole du silence ou bruit de fond, sur cet exemple, le bruit de fond n'étant pas très fort. Cependant certains bruits impulsifs sont de même niveau énergétique que la parole faible. L'histogramme 4.12 cumulé donne la moyenne du logarithme de l'énergie par fichiers sur les bases RTC_A et GSM_A selon la segmentation manuelle (*Parole*, *Non-Parole* et ce qui n'est pas étiqueté). Le logarithme moyen de l'énergie sur les fichiers est bien distinct pour le silence ou bruit de fond et les segments *Parole* issus de la segmentation manuelle. En revanche, pour une moyenne énergétique comprise entre 3.8 *dB* et 5.2 *dB*, il est difficile de discriminer les segments *Parole* et *Non-Parole* qui sont des segments de bruits impulsifs ou de courte durée. Sur ces segments de *Parole* et *Non-Parole*, la détection donnera des omissions et des insertions. Le rapport de ces erreurs est à déterminer pour obtenir les meilleures performances du système de reconnaissance.

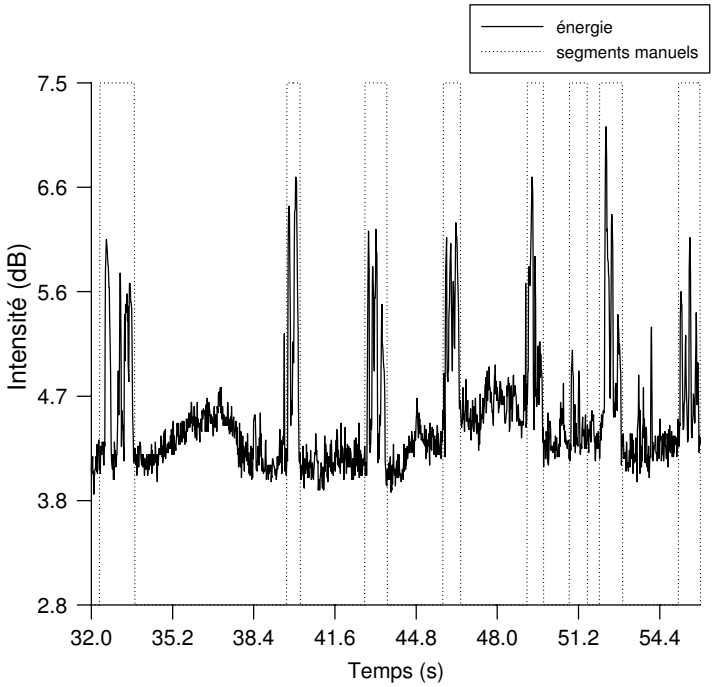


FIG. 4.11 – *Représentation du logarithme de l'énergie au cours du temps.*

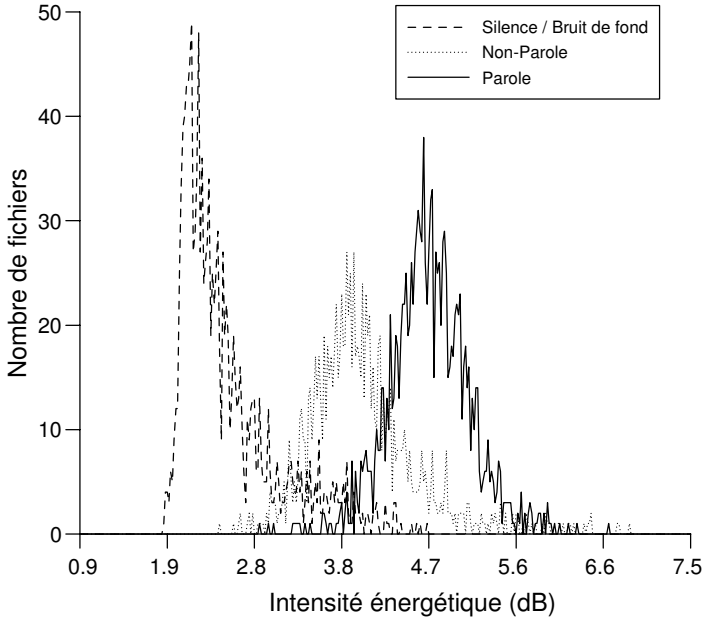


FIG. 4.12 – *Histogramme des moyennes du logarithme de l'énergie selon l'étiquetage manuel.*

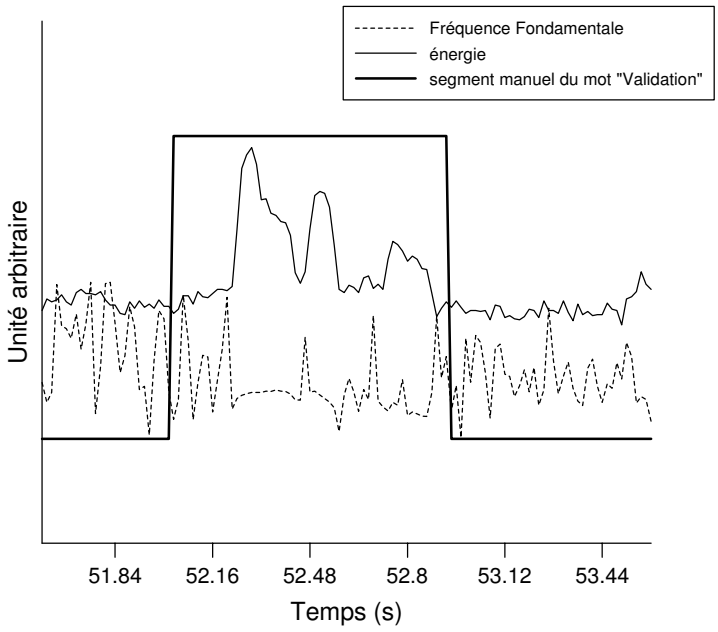


FIG. 4.13 – *Représentation du logarithme de l'énergie et de la fréquence fondamentale au cours du temps.*

4.6.2 La fréquence fondamentale

La fréquence fondamentale est utilisée dans différents systèmes de détection (*cf.* paragraphe 1.3 et paragraphe 4.5.1). Elle est surtout employée pour la détection en segments voisés/non-voisés/silence. Elle permet de déterminer les périodes de voisement, qui composent la parole. Le bruit n'est en général pas voisé, sauf pour les bruits de parole, aboiements, *etc.* Ainsi un bruit impulsif, et donc fortement énergétique peut être différencié de la parole par le fait qu'il n'est pas voisé.

La figure 4.13, représente le logarithme de l'énergie et la fréquence fondamentale calculée sur tout le signal (voisé et non-voisé) à partir d'une méthode spectrale (*cf.* Chapitre 8 "*Utilisation d'un paramètre de voisement*") sur le signal du mot "Validation". Nous utilisons ici le terme de fréquence fondamentale par abus de langage, ce n'est vraiment la fréquence fondamental que sur les périodes voisées du signal. Nous remarquons que cette caractéristique fluctue moins sur les périodes voisées du signal que sur les périodes non-voisées ou de bruit. Ainsi, il est possible d'obtenir un paramètre de voisement à l'aide de la variation de cette fréquence fondamentale. Il n'est cependant pas nécessaire d'étudier ce paramètre dans les périodes faiblement énergétiques.

4.6.3 Les coefficients cepstraux

Dans [Mauuary, 1994] l'utilisation des coefficients cepstraux (MFCC) est étudiée. Les MFCC sont combinés à l'énergie par une somme pondérée, dans le module de détection

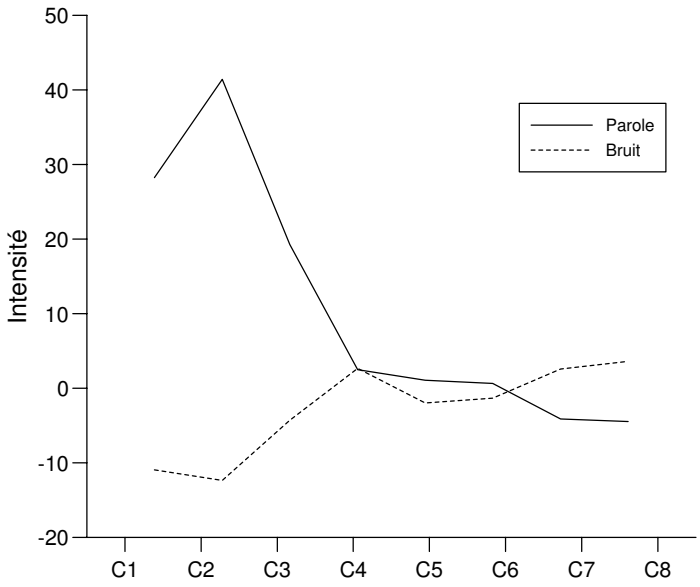


FIG. 4.14 – *Moyenne des MFCC de la parole et du bruit.*

ABP sur la base RTC_A. Il a été montré que les premiers MFCC sont discriminants et intéressants à intégrer.

La figure 4.14 montre la différence entre les moyennes des MFCC sur les étiquettes de parole et sur les étiquettes de bruit des bases RTC_A et GSM_A. Les trois premiers coefficients présentent une différence plus importante. Cependant les deux tableaux 4.1 et 4.2 des matrices de covariance des MFCC sur la parole et sur le bruit montrent que ces trois coefficients possèdent une variance plus importante que les autres coefficients. Ces tableaux permettent également de montrer que les coefficients ne sont pas indépendants, les covariances peuvent être importantes.

Il peut être intéressant d'employer ces coefficients de façon à discriminer davantage la parole et le bruit. Le problème consiste à fusionner ces caractéristiques. Cette étude est réalisée au Chapitre 9 "*Utilisation de la fusion de données*".

4.6.4 Les coefficients du vocodeur

Les coefficients du vocodeur ont également été intégrés dans l'algorithme de détection Bruit/Parole sur la base RTC_A par [Mauuary, 1994], en prenant soit le logarithme de la somme des coefficients, soit la moyenne des logarithmes des coefficients. La deuxième méthode permet de diminuer légèrement les détections de non-parole. L'auteur pense que l'amélioration de la deuxième méthode vient du fait que sur cette base de données le bruit est plus localisé en fréquence que la parole.

La figure 4.15 présente la différence entre la moyenne des énergies de la parole et du bruit dans les 24 filtres du banc, calculée sur les bases RTC_A et GSM_A. Nous

	C1	C2	C3	C4	C5	C6	C7	C8
C1	87747	1226	-13398	-2127	-1696	-3487	-3650	-2344
C2	1226	45276	12583	402	6330	1757	-595	-2303
C3	-13398	12583	34168	2792	331	1833	832	-1646
C4	-2127	402	2792	15351	2413	-823	1153	557
C5	-1696	6330	331	2413	12623	486	-210	682
C6	-3487	1757	1833	-823	486	9978	763	141
C7	-3650	-595	832	1153	-210	763	7275	-124
C8	-2344	-2303	-1646	557	682	141	-124	6294

TAB. 4.1 – *Matrice de covariance des MFCC de la parole sur les bases RTC_A et GSM_A.*

	C1	C2	C3	C4	C5	C6	C7	C8
C1	58230	3752	-4494	-265	-2038	-2308	-2397	-1234
C2	3752	27840	6277	502	2384	-31	-820	-1221
C3	-4494	6277	18654	1135	-189	696	-166	-1142
C4	-265	502	1135	10528	1739	-568	725	230
C5	-2038	2384	-189	1739	8594	292	416	525
C6	-2308	-31	696	-568	292	7191	689	-21
C7	-2397	-820	-166	725	416	689	5644	-173
C8	-1234	-1221	-1142	230	525	-21	-173	4655

TAB. 4.2 – *Matrice de covariance des MFCC du bruit sur les bases RTC_A et GSM_A.*

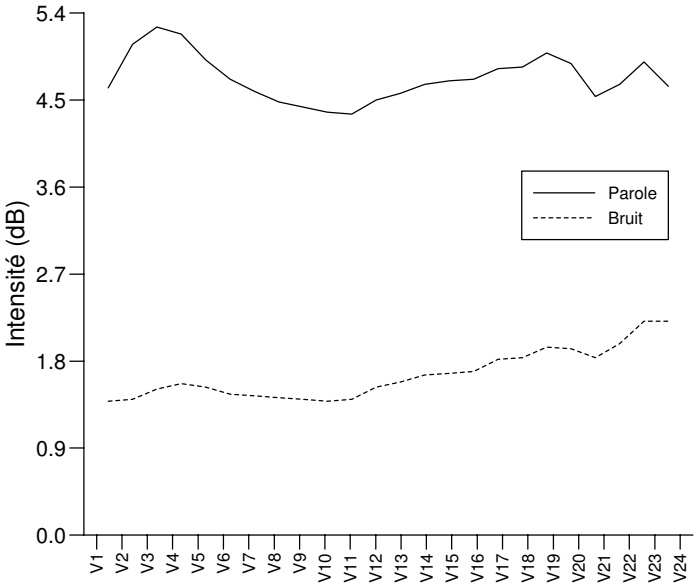


FIG. 4.15 – *Moyenne de la parole et du bruit des 24 coefficients du vocodeur.*

remarquons que les cinq premiers ont une différence plus importante. Ces énergies sont bien sûr corrélées entre elles.

Le calcul des MFCC est une transformée en cosinus inverse sur le logarithme des coefficients du vocodeur (*cf.* Annexe A). Ces coefficients sont directement employés dans le module de reconnaissance. L'utilisation des MFCC peut donc sembler plus intéressant. La fusion des coefficients du vocodeur est étudiée au Chapitre 9 "*Utilisation de la fusion de données*".

Les différentes caractéristiques étudiées dans ce paragraphe permettent de définir les axes de l'étude afin d'améliorer le module de détection de parole.

4.7 Axes d'étude

Le premier objectif étant de diminuer les erreurs du module de détection sur des signaux bruités, nous avons vu qu'une solution est l'utilisation d'un module de débruitage. Un module de débruitage permet de réduire le niveau de bruit qui perturbe beaucoup la détection (*cf.* paragraphe 3.7). Ainsi il peut permettre l'amélioration des performances aussi bien du module de détection que du module de reconnaissance. Nous proposons donc l'utilisation d'une méthode de débruitage élaborée pour la reconnaissance vocale dans le Chapitre 5 "*Méthode de débruitage*". Ce chapitre propose une étude comparative des trois critères du module de détection ABP, qui vient confirmer les résultats du paragraphe 4.4 : le critère SB reste le plus performant. Les méthodes de débruitage sont particulièrement bien adaptées pour la réduction des bruits stationnaires.

Il est donc nécessaire de réduire les détections de bruits impulsifs ou de courte durée qui correspondent aux erreurs rejetables du module de détection. Afin de diminuer ces erreurs et ainsi atteindre notre premier objectif, nous cherchons à préciser la détection en intégrant une nouvelle condition dans l'automate Bruit/Parole. Le Chapitre 6 "*Intégration d'une nouvelle condition dans l'automate*" propose différentes possibilités pour intégrer cette nouvelle condition et satisfaire à nos deux premiers objectifs : diminuer les erreurs du module de détection sur des signaux bruités et dans le cas de la détection de parole continue. La meilleure intégration de la nouvelle condition est conservée pour le reste de l'étude, en faisant l'hypothèse quelle est indépendante du critère.

À l'aide de cette condition deux possibilités sont envisageables pour améliorer le module de détection :

- soit préciser l'estimation de la distribution de l'énergie pour affiner la détection.
- soit apporter une information complémentaire du signal à l'énergie déjà employée.

Dans le Chapitre 7 "*Utilisation des statistiques d'ordre supérieur*", nous proposons de préciser l'estimation de la distribution en employant les statistiques d'ordre supérieur. Dans le paragraphe 4.5 nous avons présenté différentes possibilités d'intégration des statistiques d'ordre supérieur dans un système de détection. Le moment d'ordre 3 non centré et normalisé ici employé permet une estimation simple et de variance faible d'une statistique d'ordre 3.

Pour étudier la deuxième possibilité, les paragraphes 4.5 et 4.6 ont montré que la fréquence fondamentale peut être employée en complément de l'énergie. Le Chapitre 8 "*Utilisation d'un paramètre de voisement*" propose un paramètre de voisement obtenu à partir de la fréquence fondamentale calculée sur les périodes de signal voisé et non-voisé. Ce paramètre de voisement permet une discrimination plus fine de la parole et du bruit, et entraîne une diminution des erreurs de la détection de parole continue.

Les paragraphes 4.5 et 4.6 ont également dégagé un grand nombre de coefficients tels que les coefficients cepstraux et les coefficients de la sortie du banc de filtre qui peuvent être employés pour améliorer la détection. Ces coefficients déjà calculés pour le module de reconnaissance ne nécessitent donc pas de calculs supplémentaires. Différentes approches sont possibles afin de résoudre la difficulté de l'intégration d'un grand nombre de coefficients. Nous proposons une étude des méthodes de fusion de données, et plus particulièrement des méthode de fusion en entrée dans la Chapitre 9 "*Utilisation de la fusion de données*". La méthode d'analyse factorielle discriminante s'avère la plus adaptée à notre problème. Nous intégrons ainsi à l'aide de cette analyse les coefficients cepstraux, les coefficients cepstraux et leurs dérivées, les coefficients du vocodeur, ainsi qu'une combinaison des coefficients cepstraux, du paramètre de voisement et du moment d'ordre 3.

Ces axes d'étude définis pour l'amélioration du module de détection suivant les trois objectifs que nous avons fixés font l'objet de la deuxième partie II "*Amélioration du module de détection*".

Deuxième partie

Amélioration du module de détection

Cette partie a pour objectif d'apporter des améliorations au système de reconnaissance de France Télécom R&D en rendant plus performant le module de détection ABP.

Dans le cas d'environnement très bruité une solution pour l'amélioration du système de reconnaissance est le recours à une méthode de débruitage. Le Chapitre 5 "*Méthode de débruitage*" est une étude comparative des trois critères LCT, SB et SBP présentés au Chapitre 2 "*Détection de parole pour la reconnaissance vocale*", paragraphe 2.2. Ce premier chapitre permet de confirmer les résultats du paragraphe 4.4 au Chapitre 4 " *Voies envisagées pour l'amélioration du module de détection*" qui montrent que le critère SB est le meilleur. Cependant les méthodes de débruitage n'apportent pas d'amélioration lorsque le bruit est impulsif ou de courte durée.

Il faut donc chercher à améliorer le module de détection dans le cas d'environnements très bruités, mais aussi pour des applications de reconnaissance de parole continue. Nous avons vu dans la partie précédente le nombre important de caractéristiques du signal de parole employées dans les modules de détection. Le module de détection actuel utilise uniquement l'énergie du signal. Pour améliorer les performances du meilleur critère (critère SB), nous savons qu'il faut obtenir une détection plus précise tant au niveau des frontières qu'au niveau des détections des bruits impulsifs qui sont trop nombreuses (*cf.* Chapitre 3 "*Analyse des sources d'erreurs du module de détection*"). Pour ce faire deux possibilités sont envisageables, d'une part améliorer l'utilisation de l'énergie dans la détection, d'autre part utiliser d'autres caractéristiques du signal.

Pour améliorer le module de détection ces deux possibilités apportent une nouvelle condition à intégrer dans l'automate Bruit/Parole. Le Chapitre 6 "*Intégration d'une nouvelle décision dans l'automate*" présente différentes intégrations possibles d'une nouvelle condition dans l'automate et celle qui est retenue pour le reste de l'étude.

La première possibilité envisagée est celle d'améliorer l'utilisation de l'énergie. Dans le Chapitre 7 "*Utilisation des statistiques d'ordre supérieur*", nous cherchons à préciser la détection en tenant compte des statistiques d'ordre supérieur de l'énergie. Nous décrivons quelques éléments théoriques sur les moments et les cumulants, et donnons quelques propriétés de ces statistiques. Nous présentons ensuite plusieurs approches de l'estimation de ces statistiques. Après la présentation de quelques systèmes de détection utilisant ces statistiques, nous décrivons comment nous avons utilisé le moment d'ordre 3 normalisé. Nous donnons ensuite les résultats ainsi obtenus. Ces résultats n'étant pas satisfaisants, nous cherchons ensuite à utiliser d'autres caractéristiques.

Le Chapitre 8 "*Utilisation d'un paramètre de voisement*" propose une approche utilisant une autre caractéristique, la fréquence fondamentale, qui est une caractéristique de la prosodie de la parole. De nombreuses méthodes permettent de calculer des caractéristiques de la prosodie. Après une brève description de ces méthodes, nous présentons quelques méthodes de détection utilisant une caractéristique de la prosodie. La fréquence fondamentale calculée sur le signal voisé et non-voisé est intégrée dans le module de détection comme un paramètre de voisement. Les résultats présentés montrent une amélioration significative des performances.

Nous avons de plus cherché à employer un grand nombre de caractéristiques du signal de parole calculées pour le système de reconnaissance. Pour intégrer ces caractéristiques

dans le module de détection, il importe de les combiner au mieux sans perdre d'information. Plusieurs types de méthodes pour réaliser cette combinaison sont envisageables. Nous intégrons dans le Chapitre 9 "*Intégration de techniques de fusion de données*" une de ces méthodes, l'analyse factorielle discriminante, pour les MFCC, les MFCC et leurs dérivées, les coefficients à la sortie du banc de filtres, ainsi que la combinaison des MFCC, du moment d'ordre 3 et du paramètre de voisement présentés dans les chapitres précédents. Nous présentons les résultats obtenus avec cette méthode qui apporte également une amélioration significative des erreurs du système de reconnaissance tant au niveau de la détection de la parole continue qu'au niveau de la détection de parole dans des communications bruitées.

Chapitre 5

Méthode de débruitage

5.1 Introduction

Dans un milieu bruité, nous avons vu au Chapitre 4 "*Voies envisagées pour l'amélioration du module de détection*", que des méthodes de débruitage du signal avant le module de détection, peuvent apporter une solution pour l'amélioration des résultats du système de reconnaissance. Il est à noter que de nombreuses méthodes de débruitage pour la reconnaissance vocale ont déjà été élaborées, nous en présentons quelques unes dans ce chapitre.

L'objectif de ce chapitre est d'évaluer les trois critères LCT, SB et SBP du module de détection ABP présentés dans le paragraphe 2.2 avec une méthode de débruitage déjà existante. Nous ne proposons pas de une nouvelle méthode de débruitage ici car ce type d'amélioration ne porte pas directement sur le module de détection. La comparaison de ces trois critères se fait à l'aide du protocole de tests défini dans le paragraphe 2.4, sur la base GSM_A débruitée, ainsi que sur cette même base avec l'ajout des deux bruits *car* et *babble* du paragraphe 3.7.

Dans un premier temps, au paragraphe 5.2 nous décrivons quelques méthodes de débruitage, et plus particulièrement la méthode retenue dans cette étude. L'étude comparative des trois critères du module de détection ABP se fait ensuite dans le paragraphe 5.3 à l'aide d'une part des tests de détection dans le paragraphe 5.3.1, et d'autre part à l'aide des tests de reconnaissance dans le paragraphe 5.3.2.

5.2 Méthodes de débruitage dans le cadre de la reconnaissance vocale

Un grand nombre de méthodes de débruitage ont été développées ces dernières années. Nous ne présentons ici que les principales approches et celles qui sont employées plus particulièrement pour la reconnaissance vocale.

Les méthodes de soustraction spectrale et cepstrale ont été largement utilisées dans le domaine de la reconnaissance vocale pour obtenir une plus grande robustesse face aux

bruits additifs.

La soustraction spectrale

Certains auteurs (*cf.* [Mokbel, 1992]) ont montré que la soustraction spectrale est un bon prétraitement. Nous considérons le signal observé $x(t)$ comme un signal $s(t)$ dégradé dans le domaine temporel par un bruit additif $b(t)$, nous avons :

$$x(t) = s(t) + b(t). \tag{5.1}$$

Les signaux $s(t)$ et $b(t)$ sont supposés stationnaires à l'ordre 2 et non-corrélés. Nous obtenons la relation suivante sur les densités spectrales de puissance :

$$\gamma_x(f) = \gamma_s(f) + \gamma_b(f). \tag{5.2}$$

La densité spectrale du signal propre associée peut donc être estimée par :

$$\hat{\gamma}_s(f) = \gamma_x(f) - \hat{\gamma}_b(f), \tag{5.3}$$

et ainsi

$$|\hat{s}(f)| = \sqrt{\gamma_x(f) - \hat{\gamma}_b(f)}. \tag{5.4}$$

En pratique, l'estimation $\hat{\gamma}_b(f)$ est calculée à partir des périodes de non parole dans le signal observé, et de telle sorte que $\gamma_x(f) - \hat{\gamma}_b(f)$ reste positif. Remarquons que l'hypothèse faite sur la stationnarité du bruit n'est jamais vérifiée, notamment pour les bruits dûs à l'utilisation du réseau cellulaire. Dans [Rabiner et Juang, 1993] davantage de précisions sont présentées sur la soustraction spectrale.

La soustraction cepstrale

La soustraction cepstrale permet l'égalisation du canal contre le bruit convolutif, c'est-à-dire la correction des modifications du signal original par un filtrage linéaire. Le signal peut être modélisé par le produit de convolution de la parole $s(t)$ et par un filtre $h(t)$:

$$x(t) = s(t) \otimes h(t), \tag{5.5}$$

dans le domaine cepstral, ceci devient :

$$C_t^x = C_t^s + C_t^h. \tag{5.6}$$

L'hypothèse que l'effet du canal C_t^h ne dépend pas du temps est souvent faite.

Dans [Karray, 1998b] la soustraction spectrale et cepstrale ont donné de bonnes améliorations de la détection Bruit/Parole et de la reconnaissance de parole. Une étude comparative des trois critères du module de détection ABP avec la soustraction spectrale est donnée dans [Karray et Martin, 2001].

Un des problèmes principaux du débruitage est la DAV nécessaire pour celui-ci. Les différentes méthodes présentées au Chapitre 1 "*Différents contextes de la détection de parole*" peuvent être employées pour le débruitage.

L'emploi de plusieurs microphones (*cf.* par exemple [Martinez *et al.*, 1997], [Agaiby et Moir, 1997], [Le Bouquin-Jeannès et Faucon, 1995]) permet d'exploiter la cohérence spatiale pour la DAV du système de débruitage. Le débruitage est alors plus fiable, car nous pouvons ainsi discriminer les différentes sources émettrices de sons. Cette technique n'est cependant pas applicable à tous les contextes, notamment dans le cas de la téléphonie en général où se situe notre étude.

Dans [Shozakai *et al.*, 1998] une DAV combine, une soustraction spectrale et une soustraction cepstrale pour obtenir de bonnes performances de reconnaissance. La DAV utilisée est celle du codeur GSM.

Un réseau de neurones est utilisé dans [Héon *et al.*, 1998] pour une réduction du bruit dans le domaine cepstral. Le réseau est composé de 84 neurones avec douze sorties pour les douze coefficients cepstraux calculés. Le système de reconnaissance utilisé pour les tests est un système fondé sur les modèles de Markov cachés pour une reconnaissance de parole continue.

Dans [Wu *et al.*, 1999], un débruitage est appliqué pour améliorer les performances de la DBP donnée dans [Junqua *et al.*, 1994]. L'énergie à court-terme utilisée dans [Junqua *et al.*, 1994] est calculée à partir de l'énergie obtenue après débruitage. La méthode de débruitage est fondée sur la matrice de corrélation du bruit. L'inverse des valeurs propres de cette matrice permet d'obtenir un jeu de poids pondérant la sortie de chaque filtre. Une mesure de confiance sur le pitch est appliquée en post-traitement. L'évaluation effectuée ne permet cependant pas de mettre en évidence une amélioration au niveau des résultats de reconnaissance.

D'autres méthodes de débruitage pendant le décodage de Viterbi pour améliorer la robustesse de la reconnaissance vocale sont présentées dans [Delphin-Poulat, 1999].

Méthode de débruitage employée

La méthode de débruitage que nous utilisons dans cette étude a été élaborée récemment dans le cadre du groupe de travail ETSI/STQ/AURORA (groupe de travail dédié à la reconnaissance de la parole distribuée) pour l'amélioration des performances d'un système de reconnaissance de parole distribuée. Deux approches sont décrites dans [Noé *et al.*, 2001], dans le domaine temporel et dans le domaine fréquentiel. Les deux méthodes sont équivalentes du point de vue des résultats de reconnaissance. La méthode fréquentielle étant imbriquée avec le calcul des MFCC n'est pas directement utilisable dans notre cas. La méthode retenue est donc la méthode temporelle décrite sur la figure 5.1.

Après la soustraction de la composante continue éventuelle du signal, différents traitements sont effectués. L'analyse permet de calculer le spectre. La DAV utilisée est une simple comparaison du logarithme de l'énergie à court-terme et à long-terme. L'estimation du logarithme de l'énergie à long-terme est faite sur les périodes détectées comme n'étant pas de la parole. La détection est cependant élargie en fin de détection de 50 *ms*

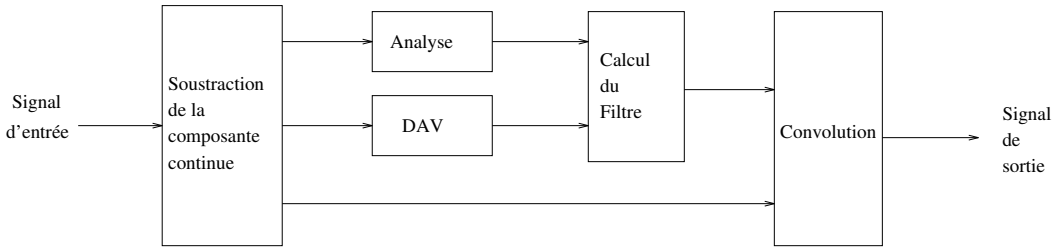


FIG. 5.1 – *Diagramme du module de débruitage.*

(*hangover*), et pour éviter la détection de bruits impulsifs, seules les détections de plus de 50 *ms* sont retenues. Cette DAV indique les trames qui serviront à l'estimation de la densité spectrale du bruit, notée $\hat{\gamma}_b$. L'estimation de la densité spectrale du signal utile, notée $\hat{S}$, est faite en deux étapes. La densité spectrale du signal utile est d'abord estimée à l'aide d'une décision dirigée (introduit dans [Ephraim et Malah, 1984]) :

$$\hat{\gamma}_{s1}(k,f) = \beta|\hat{S}(k-1,f)|^2 + (1-\beta)\ max(|X(k,f)|^2 - \hat{\gamma}_b(k,f),0), \tag{5.7}$$

où k et f sont respectivement la trame et la fréquence considérée, X la densité spectrale du signal bruité, β un coefficient optimisé à 0.98, et $\hat{\gamma}_{s1}$ est la première estimation de la densité spectrale du signal utile. La valeur optimisée du coefficient β est la même que celle trouvée dans [Ephraim et Malah, 1984]. Cette première estimation permet de calculer un premier RSB, noté RSB_1 :

$$RSB_1(k,f) = \frac{\hat{\gamma}_{s1}(k,f)}{\hat{\gamma}_b(k,f)}, \tag{5.8}$$

qui est utilisé pour le calcul de l'estimation du premier filtre, noté $\hat{H}_1$:

$$\hat{H}_1(k,f) = \frac{\sqrt{RSB_1(k,f)}}{1+\sqrt{RSB_1(k,f)}}. \tag{5.9}$$

Ce filtre est appliqué au signal pour obtenir une nouvelle estimation de la densité spectrale du signal utile, noté $\hat{\gamma}_{s2}$:

$$\hat{\gamma}_{s2}(k,f) = |\hat{H}_1(k,f)X(k,f)|^2, \tag{5.10}$$

qui permet une seconde estimation du RSB, noté RSB_2 :

$$RSB_2(k,f) = \frac{\hat{\gamma}_{s2}(k,f)}{\hat{\gamma}_b(k,f)}, \tag{5.11}$$

et ainsi l'estimation du second filtre, noté $\hat{H}_2$:

$$\hat{H}_2(k,f) = \frac{\sqrt{RSB_2(k,f)}}{1+\sqrt{RSB_2(k,f)}}. \tag{5.12}$$

L'estimation de la densité spectrale du signal utile est alors :

$$\hat{S}(k,f) = |\hat{H}_2(k,f)X(k,f)|^2. \tag{5.13}$$

La réponse impulsionnelle du filtre est ensuite calculée par la transformée de Fourier inverse. La réponse impulsionnelle, d'abord tronquée, est ensuite fenêtrée en utilisant une fenêtre de Hamming.

La réduction de bruit dans le domaine temporel est alors obtenue par convolution du signal bruité d'origine avec la réponse impulsionnelle du filtre.

Ce débruitage permet d'obtenir une nette amélioration des résultats de reconnaissance. C'est donc cette méthode que nous utilisons pour l'étude comparative des trois critères du module de détection ABP sur le signal débruité dans le paragraphe suivant.

5.3 Étude comparative des trois critères du module de détection avec une méthode de débruitage

Pour étudier l'effet du débruitage sur le module de détection, nous comparons les trois critères LCT, SB et SBP du module de détection d'une part à l'aide des tests de détection d'autre part à l'aide de tests de reconnaissance sur la base GSM_A débruitée, et sur la partie calme de cette même base ayant subi au préalable un ajout des bruits *car* ou *babble* à différents RSB (*cf.* paragraphe 3.7). Ceci nous permet d'étudier l'effet du débruitage selon la nature du bruit. En effet, la nature des bruits sur la partie de la base GSM_A ayant un RSB inférieur à 18 *dB*, est beaucoup moins stationnaire que les bruits ajoutés *car* et *babble*. De plus, il est important de considérer les résultats du module de détection après débruitage sur des fichiers moins bruités, tels que ceux contenus dans la partie de la base GSM_A avec un RSB supérieur à 18 *dB* afin d'étudier l'influence du module de débruitage sur un signal peu bruité.

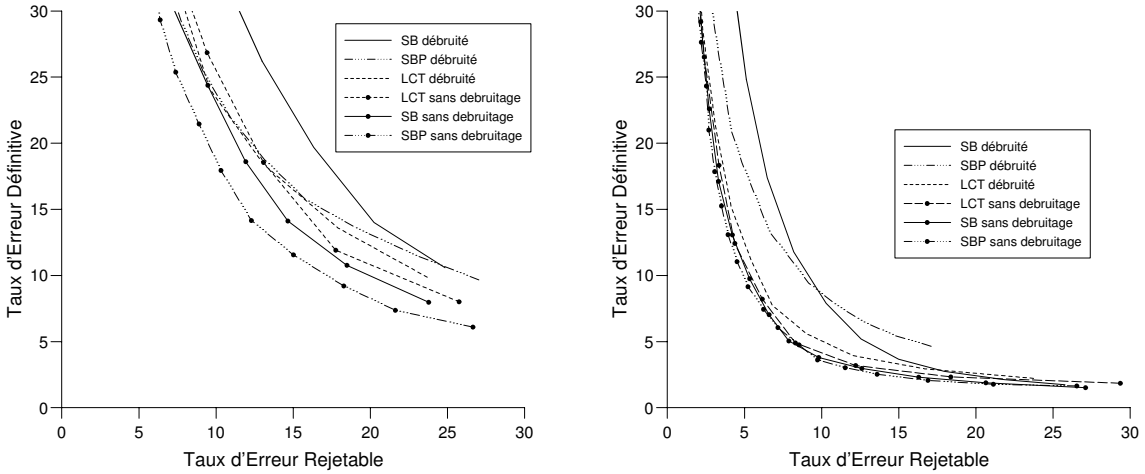
5.3.1 Résultats de détection

Nous présentons ici les résultats de détection dans un premier temps après débruitage de la base GSM_A, et dans un second temps après débruitage de la partie calme de la base GSM_A ayant subi un ajout de bruits au préalable. L'ajout de bruits est ici effectué pour obtenir un RSB de 12.5 *dB*. L'ajout du bruit est décrit au paragraphe 3.7. Les résultats avec d'autres RSB donnant des résultats comparables ne sont pas présentés. Les seuils donnant le minimum de la sommes des erreurs sont donnés pour chaque critère en Annexe F.

Débruitage de la base GSM_A

Les figures 5.2(a) et 5.2(b) donnent les résultats de détection des trois critères LCT, SB et SBP, avec et sans débruitage sur la base GSM_A, respectivement pour un RSB inférieur à 18 *dB* et pour un RSB supérieur à 18 *dB*. Nous constatons que sur la partie

calme et bruitée les résultats de détection sont en général moins bons après débruitage. La différence pour le critère LCT reste faible, mais pour les critères SB et SBP les résultats sont bien dégradés. Ainsi dans le cas débruité, le critère LCT donne de meilleurs résultats que les critères SB et SBP, alors que c'est l'inverse sans débruitage. Le critère SBP est lui-même sensiblement meilleur que le critère SB, surtout sur la partie bruitée.



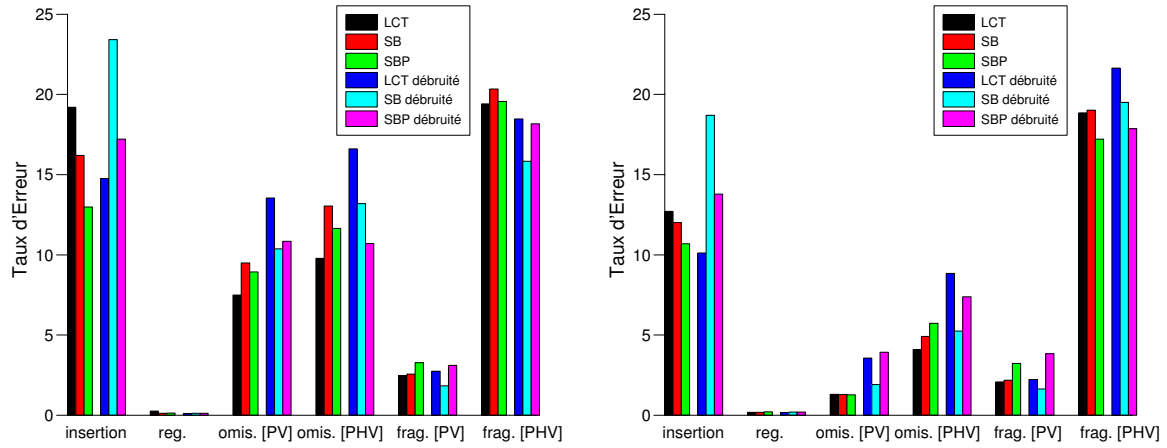
(a) RSB inférieur à 18 *dB*

(b) RSB supérieur à 18 *dB*

FIG. 5.2 – *Résultats de détection des trois critères sur la base GSM_A avec et sans débruitage.*

Pour comprendre d'où viennent les erreurs après débruitage, nous détaillons les erreurs de détection avec et sans débruitage, selon le RSB sur la figure 5.3. Nous remarquons que les erreurs d'insertion sont beaucoup plus importantes (surtout pour le critère SB). Les erreurs d'omission sont également plus importantes. Ainsi pour un choix de seuil donnant un nombre équivalent d'erreurs d'omission, les erreurs d'insertion seraient encore plus importantes. Le reste des erreurs (fragmentations et regroupements) varie peu.

Ces erreurs d'insertion plus importantes pour les critères SB et SBP que pour le critère LCT, s'expliquent par le fait que certains bruits qui n'étaient pas détectés par les critères SB et SBP, l'étaient déjà par le critère LCT. Ainsi, comme nous l'avons vu au paragraphe 4.4, le critère LCT provoque plus d'insertions sur la base GSM_A. Le module de débruitage va accentuer ces bruits détectés par rapport au bruit de fond qui reste après débruitage. Le critère LCT qui détecte déjà ces bruits sur la base non débruitée (*cf.* paragraphe 3.6), les détecte aussi sur la base débruitée, mais les critères SB et SBP qui ne les détectent pas sur la base non débruitée, vont à présent les détecter. La figure 5.4 présente l'énergie du mot "Validation" suivi de quelques bruits de courte durée. Le bruit de fond détecté par la DAV du module de débruitage est accentué sur le signal débruité.



(a) RSB inférieur à 18 *dB*

(b) RSB supérieur à 18 *dB*

FIG. 5.3 – *Erreurs de détection détaillées des trois critères sur la base GSM_A avec et sans débruitage.*

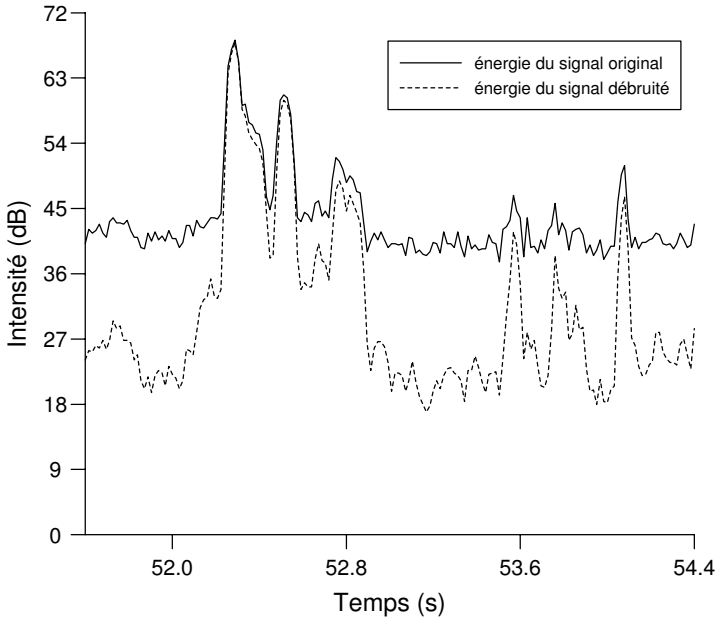


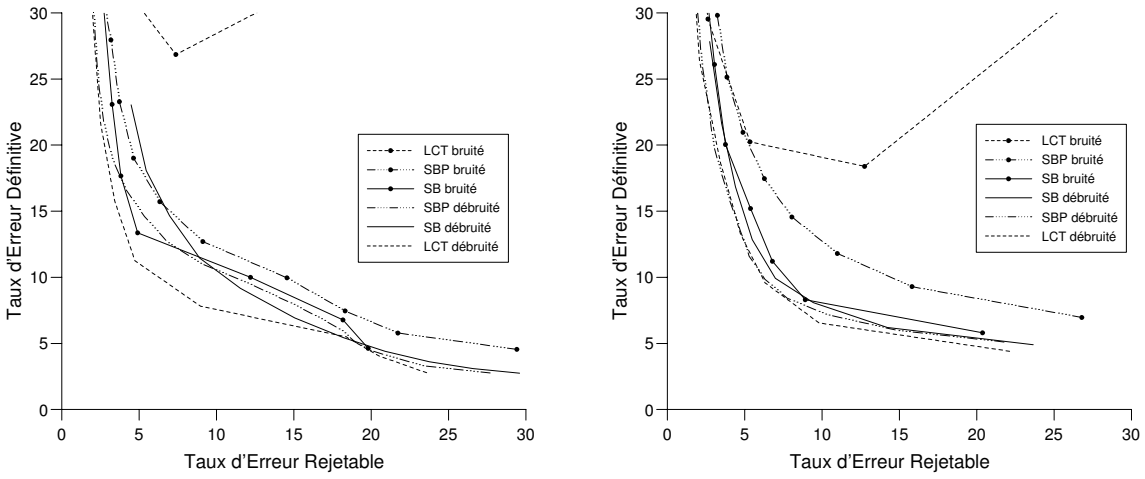
FIG. 5.4 – *Comparaison de l'énergie du signal original avec l'énergie du signal débruité.*

C'est de tels bruits que le critère LCT détecte sur le signal original et débruité, alors que les critères SB et SBP ne les détectent pas sur le signal original.

Débruitage de la base GSM_A après ajout de bruits

La figure 5.5 présente les résultats de détection pour les trois critères avec ajout des deux bruits *car* et *babble* à 12.5 *dB* sur la partie calme de la base GSM_A, puis débruitée. À la différence des bruits contenus dans la partie calme de la base GSM_A, ces deux bruits sont stationnaires. Nous observons une nette amélioration de la détection avec le critère SBP, mais surtout LCT, alors que les résultats du critère SB sont équivalents avec ou sans débruitage. Le bruit *babble* étant légèrement plus stationnaire, les résultats des trois critères sont équivalents après débruitage. Dans le cas du bruit *car*, le critère LCT donne de meilleurs résultats que les critères SB et SBP, équivalents après débruitage.

Ici encore la forte amélioration du critère LCT par rapport aux critères SB et SBP, s'explique par le fait que certains bruits sont détectés par les critères SB et SBP après débruitage, alors qu'ils ne l'étaient pas avant. Ces résultats montrent également que le module de débruitage est plus adapté à des bruits stationnaires, même avec un RSB important, qu'à des bruits impulsifs. Les résultats avec d'autres RSB sont identiques quant à la comparaison des trois critères.



(a) Bruit *car* (b) Bruit *babble*

FIG. 5.5 – *Résultats de détection des trois critères sur la base GSM_A bruitée avec et sans débruitage.*

5.3.2 Résultats de reconnaissance

Nous présentons ici les résultats de reconnaissance dans un premier temps après débruitage de la base GSM_A, et dans un second temps après débruitage de la base GSM_A ayant subi un ajout de bruits au préalable. Les seuils optimaux de reconnaissance sont donnés pour chaque critère en Annexe F.

Débruitage de la base GSM_A

Dans un premier temps pour étudier uniquement l'effet du débruitage sur les résultats de reconnaissance, nous comparons sur la figure 5.6 les résultats de reconnaissance avec et sans débruitage sur la base GSM_A pour une détection idéale, qui est la segmentation manuelle, où ne sont conservés que les segments *Parole*, c'est-à-dire les segments *Parole-Voc* et *Parole-Hors-Voc*. Nous constatons une amélioration des résultats de reconnaissance sur la partie calme et sur la partie bruitée. La différence est légèrement plus importante sur la partie bruitée.

Ainsi, le module de débruitage qui permet un débruitage des segments de parole améliore les résultats de reconnaissance.

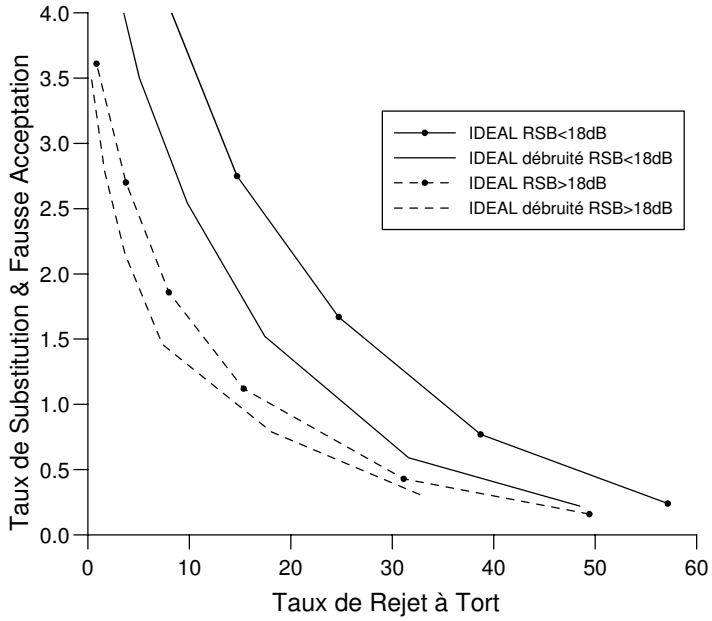
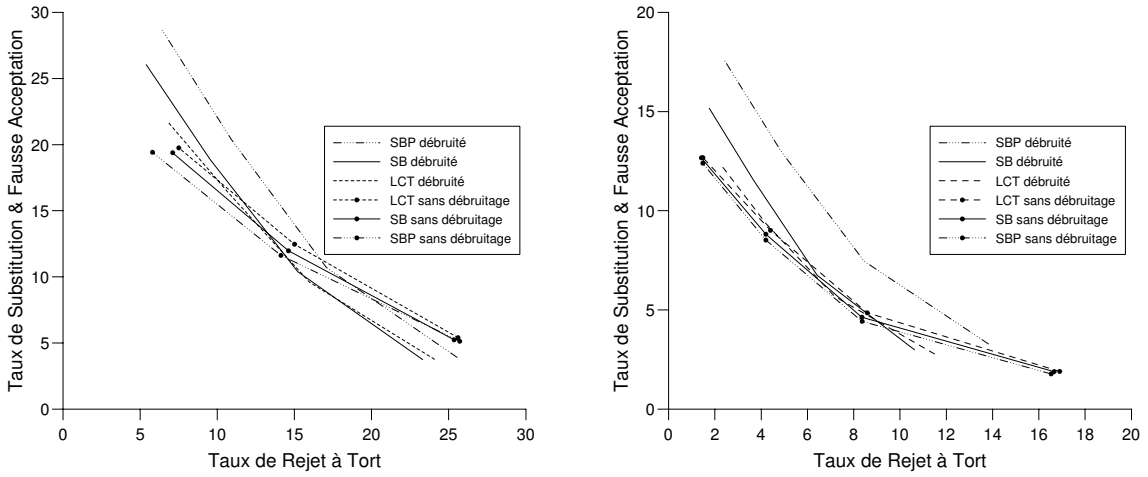


FIG. 5.6 – *Résultats de reconnaissance d'une détection idéale sur la base GSM_A avec et sans débruitage.*

Nous avons constaté que le débruitage de la base GSM_A dégrade la détection quel que soit le critère. Les figures 5.7(a) et 5.7(b) montrent que la dégradation de la détection est beaucoup moins importante au niveau des résultats de reconnaissance. En effet les insertions vont être en grande partie rejetées par le modèle de rejet du module de

reconnaissance. Cependant la dégradation des performances avec le critère SBP reste importante en particulier pour la partie calme de la base (*cf.* figure 5.7(b)). Cette dégradation s'explique par le fait que sur cette partie de la base en particulier les détections de bruits sont très importantes par le critère SBP (*cf.* figure 5.2(b)), ces détections ne sont toutes rejetées par le module de reconnaissance ce qui entraîne des erreurs de fausse acceptation. Dans le cas débruité, le critère LCT donne de meilleurs résultats sur les parties calme et bruitée que les deux autres critères.



(a) RSB inférieur à 18 *dB*

(b) RSB supérieur à 18 *dB*

FIG. 5.7 – *Résultats de reconnaissance des trois critères sur la base GSM_A avec et sans débruitage.*

Débruitage de la base GSM_A après ajout de bruits

De même que précédemment, nous étudions les résultats de reconnaissance pour une détection idéale sur la base GSM_A avec un ajout des bruits *car* et *babble*, puis avec débruitage. La figure 5.8 permet de mettre en évidence une amélioration conséquente de 50% environ pour les deux types de bruits par rapport aux résultats obtenus sur la partie calme de la base GSM_A. Nous pouvons remarquer de plus, que les résultats sont meilleurs avec l'ajout du bruit *babble* qu'avec l'ajout du bruit *car*, tandis que ce résultat est inversé après débruitage. En effet même si le bruit *car* contient des périodes non-stationnaires, le bruit *babble* reste de la parole et le module de reconnaissance entraîne donc plus d'erreurs de fausse acceptation et de substitution. Ainsi le module de débruitage est plus performant sur les bruits stationnaires et de natures différentes de la parole.

La figure 5.9 présente les résultats de reconnaissance des trois critères LCT, SB et SBP, sur la base GSM_A avec ajout des bruits *car* et *babble*, puis débruitée. Nous constatons

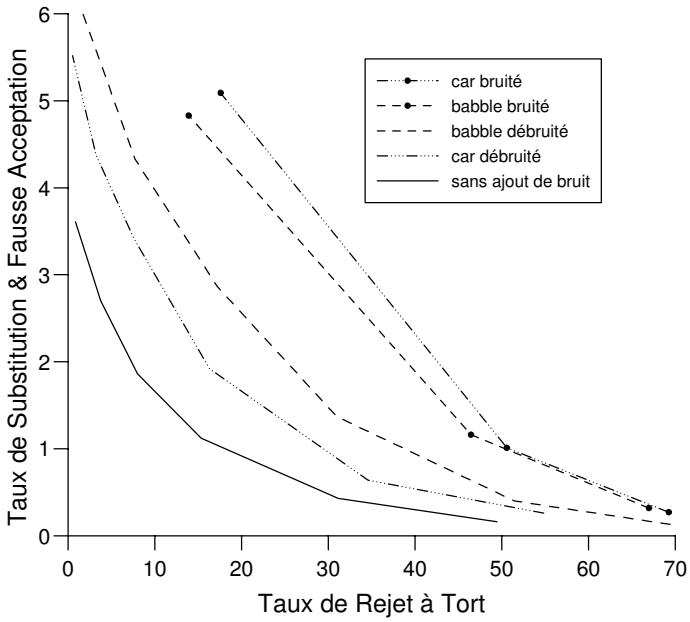
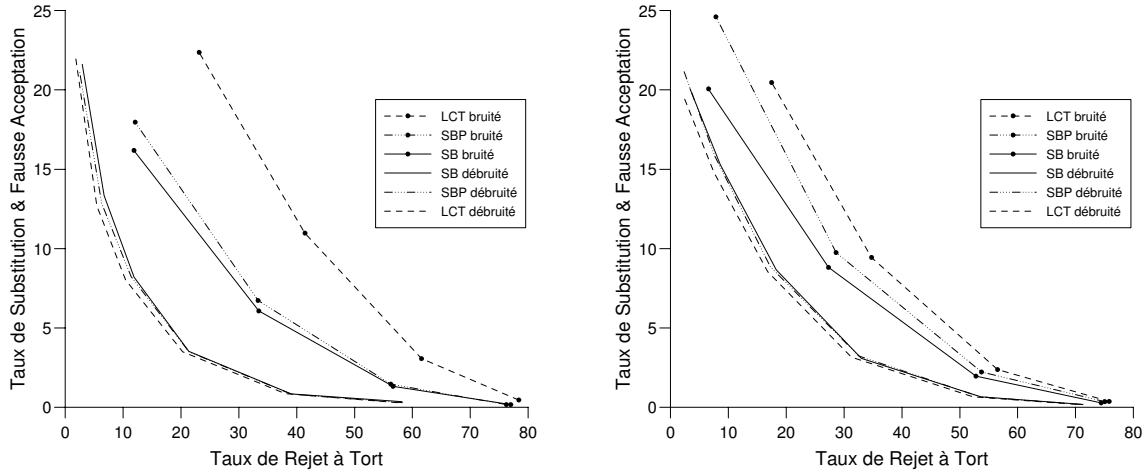


FIG. 5.8 – *Résultats de reconnaissance d'une détection idéale sur la base GSM_A bruitée avec et sans débruitage.*



(a) Bruit *car* (b) Bruit *babble*

FIG. 5.9 – *Résultats de reconnaissance des trois critères sur la base GSM_A bruitée avec et sans débruitage.*

une amélioration importante des résultats pour les deux bruits *car* et *babble* et pour les trois critères. L'amélioration est d'autant plus importante pour le critère LCT, que les résultats avec ajout de bruits sont très mauvais. Après débruitage les résultats des trois critères ne sont pas significativement différents, cependant le critère LCT est légèrement meilleur que les critères SB et SBP.

5.4 Conclusion

Les résultats présentés dans ce chapitre sont bien sûr très dépendants du choix du module de débruitage de cette étude. Ce module de débruitage s'avère plus efficace pour des bruits stationnaires, même pour des RSB faibles, et en particulier pour des bruits de natures différentes de la parole.

Le critère LCT donne de meilleurs résultats que les critères SB et SBP qui détectent les bruits impulsifs accentués par le débruitage. Dans le cas de bruits stationnaires les résultats des trois critères sont cependant très proches, et meilleurs que sans le débruitage. Par contre, les résultats des trois critères avec le module de débruitage sont dégradés dans le cas de bruits plus impulsifs. Le choix initial du critère SB peut donc être maintenu pour la suite de l'étude.

Ainsi le premier objectif est en partie atteint, puisque les erreurs du module de détection ont été diminuées pour des communications bruitées par un bruit stationnaire.

Le débruitage présenté ici cherche à ne pas dégrader le signal de parole pour que les résultats de la reconnaissance ne soient pas perturbés. Cependant un débruitage plus "agressif" pour la détection de parole pourrait apporter de meilleures perforamces tout en gardant le signal de sortie obtenue cette approche pour effectuer la reconnaissance.

Une autre possibilité d'amélioration du système de reconnaissance est, comme nous l'avons déjà vu au Chapitre 4 "*Voies envisagées pour l'amélioration du module de détection*" de réduire les détections de bruits impulsifs qui sont également le problème du module de débruitage qui les accentue.

Les méthodes de débruitage apportant des améliorations sur le signal bruité par des bruits stationnaires, nous cherchons uniquement dans la suite de cette étude à réduire les détections de bruits impulsifs ou de courte durée.

Chapitre 6

Intégration d'une nouvelle condition dans l'automate

6.1 Introduction

Dans les chapitres qui suivent nous ajoutons une condition dans l'automate Bruit/Parole présenté au paragraphe 2.2, afin d'améliorer les performances du module de détection ABP. Nous ajoutons cette nouvelle condition plus particulièrement au critère SB de l'automate caractérisé par la condition C1 :

$$C1 : r_{SB}(E(n)) > \textit{Seuil de détection.} \tag{6.1}$$

Différentes possibilités sont envisageables pour cette condition que nous nommons C4. Dans le Chapitre 7 "*Utilisation des statistiques d'ordre supérieur*" cette condition est un test sur le rapport des moments d'ordre 3 à court-terme et à long-terme. Dans le Chapitre 8 "*Utilisation d'un paramètre de voisement*" C4 est un test sur un paramètre de voisement, et dans le Chapitre 9 "*Utilisation de la fusion de données*", il s'agit d'un test sur la combinaison linéaire des MFCC obtenue par l'analyse factorielle discriminante.

Il est possible de considérer cette nouvelle condition comme une nouvelle décision de détection de la parole. Dans ce cas, le problème peut-être considéré comme un problème de fusion de décision à ajouter à la condition C1. Un grand nombre d'approches de fusion de décision sont envisageables, quelques unes sont évoquées au Chapitre 9 "*Utilisation de la fusion de données*". Cependant nous présentons ici l'ajout de cette condition C4, uniquement par les deux opérateurs logiques *et* et *ou*.

La condition C4, et la décision qui en découle peut apporter une information supplémentaire à différents niveaux de la détection selon l'endroit de l'automate où elle est considérée.

Ainsi, la condition C4 peut être ajoutée dans toutes les transitions de l'automate en complément de l'information énergétique donnée par la condition C1. Cet ajout systématique peut apporter une précision à toutes les transitions (*cf.* paragraphe 6.2). Cependant, il est possible de chercher à améliorer les débuts de détections (*cf.* paragraphe 6.3) pour

diminuer les détections de bruits. Le paragraphe 6.4 présente l'apport de la condition C4 en fin de détection.

6.2 Nouvelle condition dans toutes les transitions

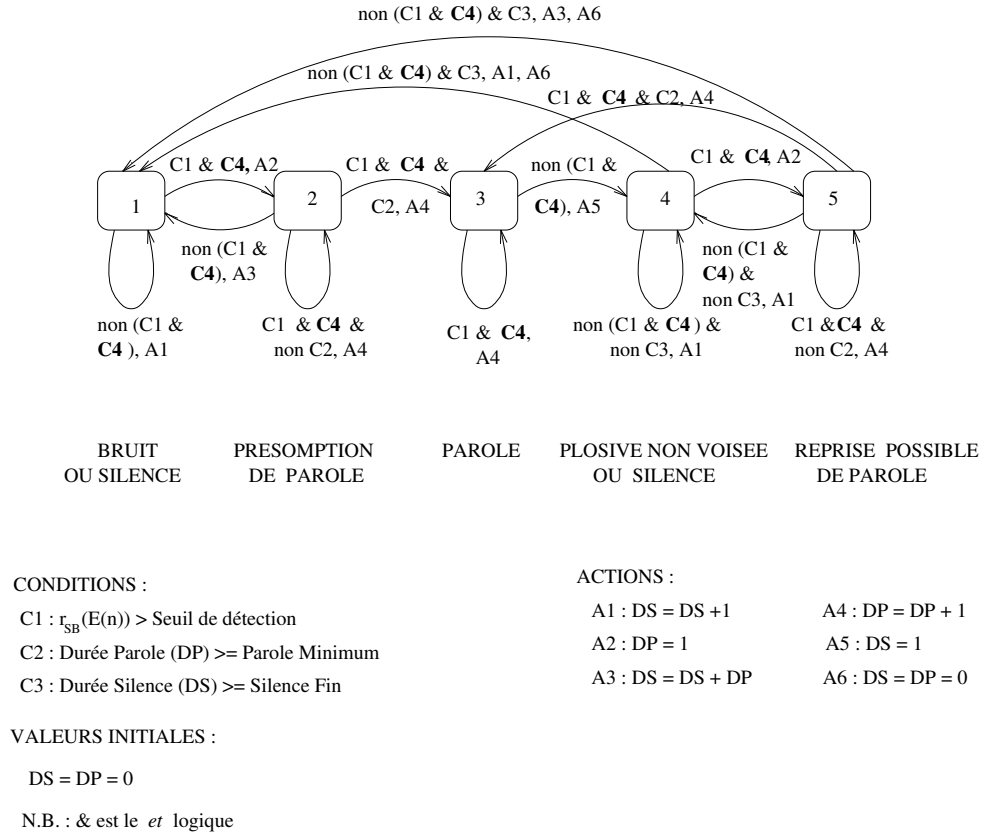


FIG. 6.1 – *Condition dans toutes les transitions.*

L'ajout d'une nouvelle condition dans toutes les transitions permet, soit de confirmer l'information énergétique déjà utilisée, si le passage d'un état à l'autre se fait avec la condition C1 *et* la condition C4, soit de considérer la nouvelle condition C4 pouvant entraîner le passage d'un état à l'autre sans avoir la condition C1 réalisée, si le passage d'un état à l'autre se fait avec la condition C1 *ou* la condition C4 (*cf.* figure 6.1, représenté avec *et*, où nous rappelons les différentes conditions et actions de l'automate).

L'ajout de C4 pour le passage à tous les états ainsi fait, place les deux conditions sur le même plan, alors que l'énergie est une caractéristique très discriminante du bruit et de la parole, considérée la plupart du temps comme la caractéristique principale (*cf.* paragraphe 4.5). Les expérimentations données en Annexe I montrent que considérer la condition C1 *ou* la condition C4 dégrade énormément les résultats. L'énergie doit rester

le paramètre principal de la détection, et l'opérateur logique *ou* ne peut donc pas être utilisé de cette façon.

6.3 Nouvelle condition pour diminuer les détections de bruits

Au Chapitre 3 "*Analyse des sources d'erreurs du module de détection*", il apparaît que les erreurs d'insertion sont les principales sources d'erreurs du système de reconnaissance. Pour permettre une diminution des insertions, la détection doit être moins sensible aux bruits impulsifs, et donc l'apport de la nouvelle condition doit se faire pour la détection du début de mot ou de requête.

Il faut donc ajouter la condition C4 au niveau de l'état *présomption de parole*. Pour ce faire plusieurs possibilités sont envisageables : la première (*cf.* figure 6.2) consiste à passer de l'état *présomption de parole* à l'état *parole* si les trois conditions C1, C2 et C4 sont réalisées, la seconde (*cf.* figure 6.3) agit en plus sur le passage de l'état *présomption de parole* à l'état *bruit ou silence*. Ce passage peut se faire si les conditions C1 ou C4 ne sont pas réalisées. À ce nouvel automate, la condition C4 peut également être ajoutée pour le passage de l'état *bruit ou silence* à l'état *présomption de parole*, ce qui peut réduire encore les insertions en évitant d'aller inutilement dans l'état *présomption de parole* (*cf.* figure 6.4).

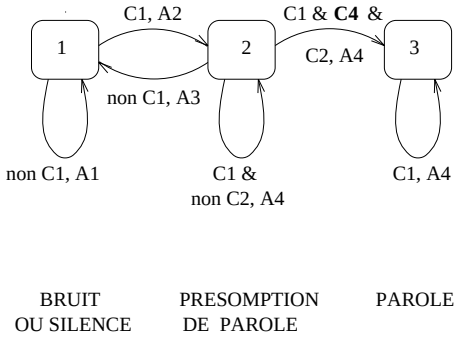


FIG. 6.2 – *Condition pour le passage de l'état "présomption de parole" à l'état "parole".*

Cependant rajouter la condition C4 de cette façon peut également provoquer davantage d'omissions. Les expérimentations réalisées en Annexe I montrent que la première solution (*cf.* figure 6.2) réduit moins le nombre d'insertions que la seconde solution (*cf.* figure 6.3). L'ajout de la condition C4 pour le passage de l'état *bruit ou silence* à l'état *présomption de parole*, n'apporte rien, et semble même dégrader légèrement les résultats, ce qui ne justifie pas son introduction.

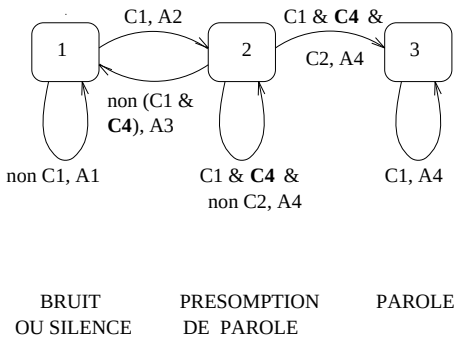


FIG. 6.3 – *Condition pour le passage de l'état "présomption de parole" à l'état "parole" et à l'état "bruit ou silence".*

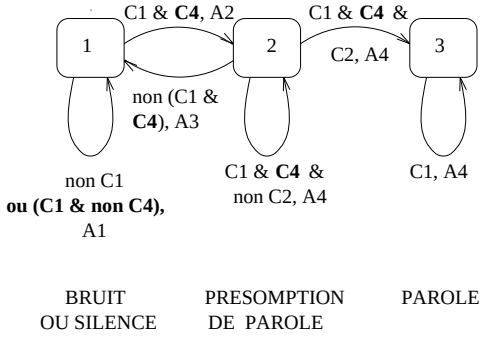


FIG. 6.4 – *Condition pour le passage de l'état "présomption de parole" à l'état "parole" et à l'état "bruit ou silence", avec le passage de l'état "bruit ou silence" à l'état "présomption de parole".*

6.4 Nouvelle condition pour améliorer la fin de la détection

Le Chapitre 3 "*Analyse des sources d'erreurs du module de détection*" a permis de montrer que le module de détection ABP provoque peu de fragmentations. Cependant la détection de fin de mot ou de requête peut être élargie ou tronquée. Les détections élargies en fin de mot ne provoquent que très peu d'erreurs de reconnaissance, en fin de requête, pour la reconnaissance de parole continue, les erreurs ainsi provoquées sont plus nombreuses, mais moins importantes qu'en début de requête. Ceci est dû au modèle de rejet plus délicat pour cette application.

Nous proposons plusieurs possibilités pour améliorer la fin de détection. La première consiste à ajouter la condition C4 à toutes les transitions de l'état *reprise possible de parole* avec les opérateurs *et* et *ou* pour C4 *et* C1 et non C4 *ou* non C1 (*cf.* figure 6.5). En effet cet état contrôle la fin de la détection. Cependant, les résultats donnés en Annexe I

montrent que l'ajout de la condition C4 ainsi faite, dégrade les résultats de détection aussi bien au niveau des erreurs rejetables et définitives, qu'au niveau des détections tronquées à droite.

Une autre approche permettant d'éviter d'augmenter les détections tronquées, est de considérer la condition C4 pour le passage de l'état *reprise possible de parole* à l'état *parole*, avec l'opérateur *ou* (*cf.* figure 6.6), et la règle C1 *ou* C4. Ainsi, de l'état *reprise possible de parole*, il devient possible de passer dans l'état *parole* si C1 *ou* C4 sont vérifiées. Les résultats de l'Annexe I montrent alors une légère amélioration tant au niveau des erreurs rejetables et définitives, que pour les détections tronquées à droite.

Il est possible de modifier cette approche en ajoutant la condition C4 pour le passage de l'état *plosive non voisée ou silence* à l'état *reprise possible de parole* par la règle C4 *ou* C1 (*cf.* figure 6.7). Cependant, même si cette approche diminue énormément les détections tronquées à droite, les détections élargies à droite sont augmentées. De plus les résultats de détection au niveau des erreurs rejetables et définitives sont dégradées (*cf.* Annexe I).

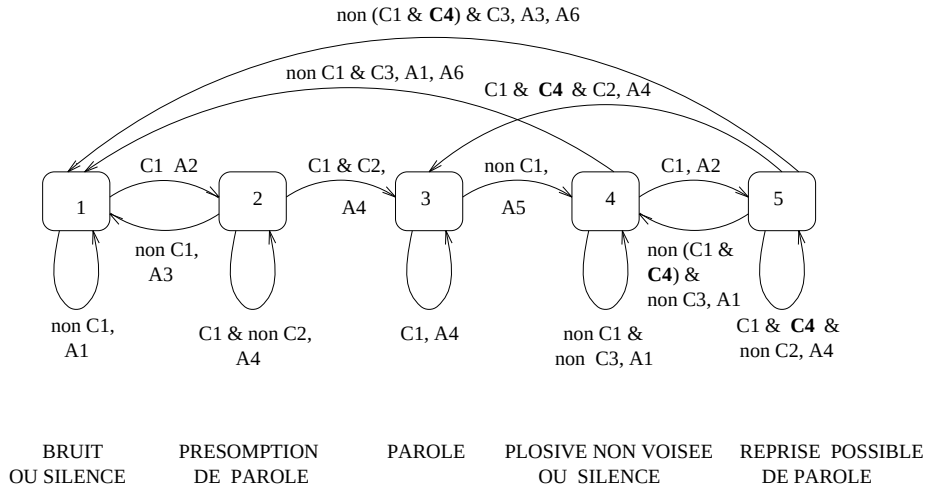


FIG. 6.5 – *Condition au niveau de l'état "reprise possible de parole".*

Ainsi l'intégration de la condition C4 à la manière de la figure 6.6 permet le mieux d'améliorer la détection de fin de mot ou de requête. Cependant pour éviter de tronquer la fin de la détection il est possible d'augmenter le *Silence Fin*, comme nous l'avons mentionné au Chapitre 2 "*Détection de parole pour la reconnaissance vocale*".

6.5 Conclusion

Parmi toutes ces intégrations possibles, l'intégration de la condition C4 au niveau de l'état *présomption de parole* (*cf.* figure 6.3) donne de meilleurs résultats de détection (*cf.* Annexe I figure I.4). Cette approche permet de réduire fortement les détections de bruits. Le Chapitre 3 "*Analyse des sources d'erreurs du module de détection*" met en évidence

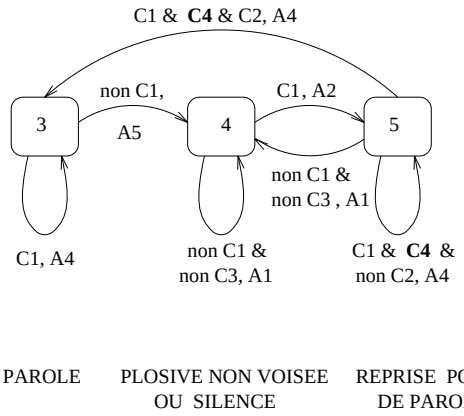


FIG. 6.6 – *Condition au niveau de l'état "reprise possible de parole" pour le passage à l'état "parole".*

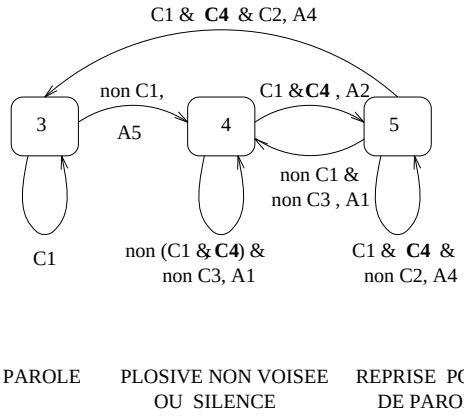


FIG. 6.7 – *Condition au niveau de l'état "reprise possible de parole" pour le passage à l'état "parole" avec l'ajout de la condition C4 au niveau de l'état "plosive non voisée ou silence" pour le passage à l'état "reprise possible de parole".*

que la diminution de ces erreurs sans augmentation des erreurs définitives, entraîne une amélioration des performances du système de reconnaissance. Dans la suite de cette étude nous intégrons donc la nouvelle condition C4 de cette manière (*cf.* Chapitre 7 "*Utilisation des statistiques d'ordre supérieur*", Chapitre 8 "*Utilisation d'un paramètre de voisement*" et Chapitre 9 "*Utilisation de la fusion de données*").

Nous constatons cependant que l'intégration cumulée de la condition C4 au niveau de l'état *présomption de parole* (*cf.* figure 6.3) et au niveau de l'état *reprise possible de parole* (*cf.* figure 6.6), permet une légère amélioration. Cette amélioration est cependant peu significative et reste au niveau des fins de détection, qui entraînent peu d'erreurs et qui peuvent être évitées en augmentant le silence de fin de parole. Nous n'étudions pas cette approche pour nous focaliser uniquement sur la diminution des détections de bruits.

Chapitre 7

Utilisation des statistiques d'ordre supérieur

7.1 Introduction

Le module de détection utilise l'énergie du signal pour discriminer le bruit et la parole. Seule la moyenne et l'écart-type de l'énergie sont considérés pour cette discrimination. Les statistiques d'ordre supérieur, souvent négligées pour des raisons de difficultés de calculs, apportent des informations importantes sur les distributions.

Nous nous proposons d'étudier ici, après une brève présentation d'éléments théoriques sur les moments et les cumulants dans le paragraphe 7.2 et leurs propriétés dans le paragraphe 7.3, quelques estimateurs de ces statistiques dans le paragraphe 7.4. Nous présentons ensuite une estimation numérique de la moyenne et de la variance de quelques estimateurs dans le paragraphe 7.5, puis l'approche de certains auteurs pour intégrer les statistiques d'ordre supérieur dans des systèmes de détection de parole dans le paragraphe 7.6. Nous tentons ensuite d'introduire ces informations dans le module de détection Bruit/Parole dans le paragraphe 7.7, et nous présentons les résultats de l'expérimentation dans le paragraphe 7.8.

7.2 Quelques éléments théoriques sur les moments et cumulants

Soit X une variable aléatoire réelle, de densité de probabilité $p_X(t)$. Sa fonction caractéristique $\phi_X(t)$, définie comme la transformée de Fourier de la densité de probabilité est donnée par :

$$\phi_X(t) = \int_{-\infty}^{+\infty} e^{itx} p_X(x) dx. \tag{7.1}$$

La fonction ϕ_X est continue, de module inférieur ou égal à 1, et $\phi_X(0) = 1$. Ainsi ϕ_X est non nulle dans un voisinage de l'origine, nous pouvons donc définir son logarithme

népérien pour t proche de l'origine :

$$\psi_X(t) = \ln(\phi_X(t)). \tag{7.2}$$

Cette fonction ψ_X est appelée *seconde fonction caractéristique. Le moment d'ordre* q est défini par la dérivée d'ordre q à l'origine, de la première fonction caractéristique :

$$\mu_q = (-i)^q . \frac{d^q \phi_X(t)}{dt^q} \mid_{t=0} = E[X^q]. \tag{7.3}$$

Le cumulant d'ordre q est défini par la dérivée d'ordre q à l'origine, de la seconde fonction caractéristique :

$$\kappa_q = (-i)^q . \frac{d^q \psi_X(t)}{dt^q} \mid_{t=0} = Cum[X,X,...,X]. \tag{7.4}$$

Cette deuxième notation ne sera employée que lorsqu'il y a confusion possible. Par l'expression $\psi_X(t) = \ln(\phi_X(t))$, nous pouvons exprimer les cumulants d'ordre q en fonction des moments d'ordre inférieur ou égal à q. Nous pouvons ainsi établir :

$$\kappa_1 = \mu_1, \tag{7.5}$$

$$\kappa_2 = \mu_2 - (\mu_1)^2, \tag{7.6}$$

$$\kappa_3 = \mu_3 - 3\mu_1\mu_2 + 2(\mu_1)^3, \tag{7.7}$$

$$\kappa_4 = \mu_4 - 4\mu_3\mu_1 - 3(\mu_2)^2 + 12\mu_2(\mu_1)^2 - 6(\mu_1)^4. \tag{7.8}$$

Remarquons que les cumulants d'ordre 1 et 2 sont respectivement la moyenne et la variance de la variable aléatoire réelle X. Les coefficients d'asymétrie (*skewness*) et d'aplatissement (*kurtosis*) sont les cumulants normalisés d'ordre 3 et 4 respectivement :

$$\chi_X = \frac{\kappa_3}{(\kappa_2)^{\frac{3}{2}}}, \tag{7.9}$$

$$\gamma_X = \frac{\kappa_4}{(\kappa_2)^2}. \tag{7.10}$$

Ces coefficients permettent de caractériser l'asymétrie et l'aplatissement de la distribution de la variable aléatoire X.

Dans le cas gaussien, nous remarquons que les cumulants d'ordre supérieur à deux sont tous nuls, et la seconde fonction caractéristique s'écrit uniquement, en fonction des moments d'ordre 1 et 2 :

$$\psi_X(t) = i\mu_1 t - \frac{1}{2}\mu_2 t^2. \tag{7.11}$$

Les variables aléatoires gaussiennes sont donc entièrement décrites par les moments du premier et second ordre. L'approximation gaussienne des distributions du bruit et de la parole, faite en considérant le théorème de la limite centrale, a restreint la recherche au second ordre.

Dans le cas des variables aléatoires multidimensionnelles ($X^T = [X_1,X_2,...,X_{n-1},X_n]$), les moments et les cumulants se déduisent des deux premières fonctions caractéristiques :

$$\phi_X(\mathbf{t}) = \int_{-\infty}^{+\infty} e^{i\mathbf{t}^T\mathbf{x}} p_X(\mathbf{x}) d\mathbf{x} = E[e^{i\mathbf{t}^T X}], \tag{7.12}$$

$$\psi_X(\mathbf{t}) = \ln(\phi_X(\mathbf{t})), \tag{7.13}$$

où $\mathbf{t}$ et $\mathbf{x}$ sont des vecteurs de dimension n.

Nous avons ainsi :

$$\mu_{q_1,...,q_n} = (-i)^q \left(\frac{\delta}{\delta t_1}\right)^{q_1} \left(\frac{\delta}{\delta t_2}\right)^{q_2} ... \left(\frac{\delta}{\delta t_n}\right)^{q_n} \phi(\mathbf{t}) \mid_{\mathbf{t}=\mathbf{0}} , \tag{7.14}$$

et

$$\begin{aligned} \kappa_{q_1,...,q_n} &= (-i)^q \left(\frac{\delta}{\delta t_1}\right)^{q_1} \left(\frac{\delta}{\delta t_2}\right)^{q_2} ... \left(\frac{\delta}{\delta t_n}\right)^{q_n} \psi(\mathbf{t}) \mid_{\mathbf{t}=\mathbf{0}} \\ &= Cum(X_1^{q_1},X_2^{q_2},...,X_n^{q_n}), \end{aligned}$$

où $q = q_1 + q_2 + ... + q_n$, et $\mathbf{t}$ est un vecteur à n dimensions.

Si les variables aléatoires sont centrées, nous avons :

$$\kappa_{i,j} = \mu_{i,j}, \tag{7.15}$$

$$\kappa_{i,j,k} = \mu_{i,j,k}, \tag{7.16}$$

$$\kappa_{i,j,k,l} = \mu_{i,j,k,l} - \mu_{i,j}\mu_{k,l} - \mu_{i,k}\mu_{j,l} - \mu_{i,l}\mu_{j,k}. \tag{7.17}$$

À partir de la définition de la seconde fonction caractéristique, il est possible d'écrire, par la formule de Leonov et Shiryayev, les relations générales liant moments et cumulants :

$$\kappa_{1,...,r} = \sum_{p=1}^{r} (-1)^p (p-1)!\, E\left[\prod_{i\in v_1} x_i\right] .E\left[\prod_{j\in v_2} x_j\right] ...E\left[\prod_{k\in v_p} x_k\right], \tag{7.18}$$

où tous les ensembles $\{v_1,v_2,...,v_p : 1 \leq p \leq r\}$ forment une partition de $\{1,2,...,r\}$, p est le nombre d'éléments composant la partition.

7.3 Quelques propriétés

a- $Cum(\alpha_1 X_1,\alpha_2 X_2,...,\alpha_n X_n) = \alpha_1\alpha_2...\alpha_n Cum(X_1,X_2,...X_n)$.

b- $Cum(X+Y,Z_1,...Z_n) = Cum(X,Z_1,...,Z_n) + Cum(Y,Z_1,...,Z_n)$.

c- Si les variables aléatoires X_i sont indépendantes, deux à deux, nous avons :

$$Cum(X_1^{q_1},X_2^{q_2},...,X_n^{q_n}) = 0. \tag{7.19}$$

Pour plus de détails, nous nous référons à [McCullagh, 1987], à [Pincibono, 1993] et à [Lacoume *et al.*, 1997].

7.4 Estimation de statistiques d'ordre supérieur

Il y a un grand nombre d'estimateurs possibles pour une statistique donnée, le choix qui est fait ci-dessous est lié à l'utilisation que nous voulons en faire. En effet le signal étudié est en général peu stationnaire, et l'estimation récursive permet de ne pas introduire de retard dans les algorithmes de détection. Les estimateurs sur des fenêtres exponentielles répondent à ces exigences. Dans tout ce paragraphe nous ferons l'hypothèse que les variables aléatoires étudiées sont indépendantes et identiquement distribuées (i.i.d.). Dans le cas du signal de parole et de bruit cette hypothèse est bien sûr abusive, mais donne un ordre d'idée quant aux moyennes et variances des estimateurs de variables corrélées.

Nous présentons ici les estimations des statistiques d'ordre supérieur sur des fenêtres exponentielles. Nous établissons les formules générales des variances des moments d'ordre 1 et 2 de ces estimateurs.

7.4.1 Moments d'ordre 1

L'estimation arithmétique de la moyenne, est la moyenne arithmétique :

$$\hat{\mu}_1(n) = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{7.20}$$

où x_i est une observation de la variable aléatoire X. L'estimateur $\hat{\mu}_1(n)$ est sans biais. Nous avons $E[\hat{\mu}_1(n)] = m$ et $Var(\hat{\mu}_1(n)) = \frac{\sigma^2}{n}$, quel que soit n, où m et σ^2 sont respectivement la moyenne et la variance théoriques de la variable aléatoire X.

Pour prendre en compte le caractère non stationnaire des signaux, l'estimation de la moyenne peut se faire sur une fenêtre exponentielle, qui joue le rôle d'un filtre passe-bas avec un facteur d'oubli. Cette estimation a de plus l'avantage de se faire de façon récursive. Nous avons ainsi :

$$\hat{\mu}_1(n) = \hat{\mu}_1(n-1) + (1-\lambda)(x_n - \hat{\mu}_1(n-1)) = (1-\lambda)\sum_{i=0}^{n} \lambda^{n-i} x_i, \tag{7.21}$$

où λ est le facteur d'oubli. Nous avons supposé $\hat{\mu}_1(0) = 0$. La moyenne de cet estimateur est :

$$E[\hat{\mu}_1(n)] = (1 - \lambda^{n+1})\, m, \tag{7.22}$$

qui est asymptotiquement sans biais. Sa variance est donnée par :

$$Var(\hat{\mu}_1(n)) \;=\; \frac{1-\lambda}{1+\lambda}(1 - \lambda^{2(n+1)})\sigma^2,$$

qui a pour limite, lorsque $n \to +\infty$: $\frac{1-\lambda}{1+\lambda}\,\sigma^2$. Cette valeur sera d'autant plus petite que le facteur d'oubli λ sera proche de 1. Le problème est que plus le facteur d'oubli λ est proche de 1, c'est-à-dire plus la fenêtre exponentielle est grande, plus λ s'applique à un signal supposé stationnaire. L'hypothèse sur le degré de stationnarité du signal détermine donc le choix du facteur d'oubli.

7.4.2 Statistiques d'ordre 2

De la même façon que pour l'estimation de la moyenne nous pouvons estimer le moment d'ordre deux par :

$$\hat{\mu}_2(n) = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \tag{7.23}$$

qui est un estimateur sans biais de moyenne :

$$E[\hat{\mu}_2(n)] = m^2 + \sigma^2, \tag{7.24}$$

et de variance :

$$Var(\hat{\mu}_2(n)) = \frac{1}{n}\,(\mu_4 - (m^2 + \sigma^2)^2), \tag{7.25}$$

où μ_4 est le moment d'ordre 4 théorique de X.

L'estimation peut aussi se faire sur une fenêtre exponentielle :

$$\hat{\mu}_2(n) = \hat{\mu}_2(n-1) + (1-\lambda)(x_n^2 - \hat{\mu}_2(n-1)) = (1-\lambda)\sum_{i=0}^{n} \lambda^{n-i} x_i^2. \tag{7.26}$$

Cet estimateur a pour moyenne :

$$E[\hat{\mu}_2(n)] = (1 - \lambda^{n+1})(m^2 + \sigma^2), \tag{7.27}$$

qui est asymptotiquement sans biais. Sa variance est :

$$Var(\hat{\mu}_2(n)) \;=\; \frac{1-\lambda}{1+\lambda}(1 - \lambda^{2(n+1)})(\mu_4 - (m^2 + \sigma^2)^2).$$

Nous avons donc :

$$\lim_{n\to+\infty} Var(\hat{\mu}_2(n)) = \frac{1-\lambda}{1+\lambda}(\mu_4 - (m^2+\sigma^2)^2). \tag{7.28}$$

Cette limite converge vers 0 lorsque le facteur d'oubli λ tend vers 1.

Remarque : Nous constatons que asymptotiquement en n, la moyenne et la variance de l'estimateur du moment d'ordre 2 avec un facteur d'oubli λ sont équivalentes à la moyenne et la variance de l'estimateur arithmétique pour $n = \frac{1+\lambda}{1-\lambda}$. C'est également le cas pour le moment d'ordre 1.

L'estimation de la variance peut se faire à partir des estimateurs précédents de différentes façons. Citons :

$$\hat{\sigma}^2(n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_1(n))^2, \tag{7.29}$$

où $\hat{\mu}_1(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$. Nous pouvons montrer que :

$$E[\hat{\sigma}^2(n)] = \frac{n-1}{n}\sigma^2, \tag{7.30}$$

et,

$$Var(\hat{\sigma}^2(n)) = \frac{n-1}{n^3}[(n-1)\mu_4 - (n-3)\sigma^4] - \frac{n-1}{n^3}[(n-1)m^4 - 2(n-3)m^2\sigma^2]. \tag{7.31}$$

Cet estimateur n'est pas sans biais, mais $\hat{\sigma^*}^2(n) = \frac{n}{n-1}\hat{\sigma}^2(n)$ est non biaisé. Si la valeur théorique de la moyenne est connue nous pouvons récrire cet estimateur comme :

$$\hat{\sigma}^2(n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 - (\hat{\mu}_1(n) - m)^2. \tag{7.32}$$

$\hat{\sigma}^2(n)$ s'écrit aussi :

$$\hat{\sigma}^2(n) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\hat{\mu}_1(n))^2. \tag{7.33}$$

Nous pouvons de même réutiliser les statistiques sur les fenêtres exponentielles. Nous avons ainsi :

$$\hat{\sigma}^2(n) = (1-\lambda)\sum_{i=0}^{n}\lambda^{n-i}x_i^2 - (\hat{\mu}_1(n))^2, \tag{7.34}$$

où à présent, $\hat{\mu}_1(n) = (1-\lambda)\sum_{i=0}^{n}\lambda^{n-i}x_i$. Nous obtenons la moyenne de cet estimateur :

$$\begin{aligned} E[\hat{\sigma}^2(n)] &= \left[1 - \frac{1-\lambda}{1+\lambda}(1-\lambda^{n+1})\right](1-\lambda^{n+1})\sigma^2 \\ &\quad +\lambda^{n+1}(1-\lambda^{n+1})m^2, \end{aligned}$$

qui est asymptotiquement biaisé, sa limite en n est $\frac{2\lambda}{1+\lambda}\sigma^2$. Nous remarquons que si λ tend vers 1, cet estimateur devient non biaisé. Nous calculons de plus, la variance de cet estimateur :

$$\begin{aligned}
Var(\hat{\sigma}^2(n)) = & \left[\frac{1-\lambda}{1+\lambda}(1-\lambda^{2(n+1)}) - 2\,\frac{(1-\lambda)^2}{1+\lambda+\lambda^2}(1-\lambda^{3(n+1)})\right. \\
& \left. + \frac{(1-\lambda)^3}{(1+\lambda)(1+\lambda^2)}(1-\lambda^{4(n+1)})\right]\mu_4 \\
& + \left\{4\,\frac{(1-\lambda)^2}{(1+\lambda)^2}\left[\lambda\,\frac{1-\lambda^{4(n+1)}}{1+\lambda^2} - \lambda^{2(n+1)}\,(1-\lambda^{2(n+1)})\right]\right. \\
& + 2\,\frac{(1-\lambda)^2}{1+\lambda+\lambda^2}(1-\lambda^{3(n+1)}) - \frac{1-\lambda}{1+\lambda}(1-\lambda^{2(n+1)}) \\
& \left. - \frac{(1-\lambda)^3}{(1+\lambda)(1+\lambda^2)}(1-\lambda^{4(n+1)})\right\}(\sigma^2+m^2)^2 \\
& - 4\,\frac{(1-\lambda)^2}{(1+\lambda)^2}\left[\lambda\,\frac{1-\lambda^{4(n+1)}}{1+\lambda^2} - \lambda^{2(n+1)}\,(1-\lambda^{2(n+1)})\right]m^4.
\end{aligned}$$

Nous avons cependant :

$$\begin{aligned}
\lim_{n\to+\infty} \quad Var(\hat{\sigma}^2(n)) = & \left(\frac{1-\lambda}{1+\lambda} - 2\,\frac{(1-\lambda)^2}{1+\lambda+\lambda^2} + \frac{(1-\lambda)^3}{(1+\lambda)(1+\lambda^2)}\right)\mu_4 \\
& + \left[4\lambda\,\frac{(1-\lambda)^2}{(1+\lambda)^2(1+\lambda^2)} + 2\,\frac{(1-\lambda)^2}{1+\lambda+\lambda^2} - \frac{1-\lambda}{1+\lambda}\right. \\
& \left. - \frac{(1-\lambda)^3}{(1+\lambda)(1+\lambda^2)}\right](\sigma^2+m^2)^2 \\
& - 4\lambda\,\frac{(1-\lambda)^2}{(1+\lambda)^2(1+\lambda^2)}m^4.
\end{aligned}$$

Cette limite tend vers zéro lorsque λ tend vers 1. Cet estimateur est donc consistant.

En supposant que la variable aléatoire suit une loi Laplacienne, l'estimation de la variance peut se faire par l'estimation de l'écart-type :

$$\begin{aligned}
\hat{\sigma}_L(n) &= \hat{\sigma}_L(n-1) + (1-\lambda)(|x_n - \hat{\mu_1}(n)| - \hat{\sigma}_L(n-1)) \\
&= (1-\lambda)\sum_{i=0}^{n}\lambda^{n-i}|x_i - \hat{\mu_1}(i)|.
\end{aligned}$$

L'estimation de la variance est alors :

$$\hat{\sigma}_L^2(n) = (\hat{\sigma}_L(n))^2. \tag{7.35}$$

Cet estimateur de la variance a l'avantage de ne faire appel qu'au moment d'ordre 1, il peut donc être moins coûteux en temps de calculs. C'est pourquoi l'hypothèse d'une distribution Laplacienne est parfois faite pour le calcul de la variance indépendamment des autres hypothèses.

7.4.3 Statistiques d'ordre supérieur à 2

De même que précédemment, nous pouvons estimer les moments d'ordre k, de façon arithmétique :

$$\hat{\mu}_k(n) = \frac{1}{n}\sum_{i=1}^{n} x_i^k. \tag{7.36}$$

Ce sont des estimateurs sans biais. Et de la même façon, sur des fenêtres exponentielles :

$$\hat{\mu}_k(n) = (1-\lambda)\sum_{i=1}^{n} \lambda^{n-i} x_i^k. \tag{7.37}$$

Ces estimateurs ont pour espérance mathématique :

$$E[\hat{\mu}_k(n)] = (1-\lambda^{n+1})\mu_k. \tag{7.38}$$

Ils sont ainsi asymptotiquement sans biais. Le calcul des variances des moments d'ordre k avec $k > 2$ pose des difficultés dans l'établissement des formules générales.

Nous pouvons estimer les cumulants d'ordre 3 et 4, d'après les formules précédemment citées :

$$\hat{\kappa}_3(n) = \hat{\mu}_3(n) - 3\hat{\mu}_1(n)\hat{\mu}_2(n) + 2\hat{\mu}_1^3(n), \tag{7.39}$$

et

$$\hat{\kappa}_4(n) = \hat{\mu}_4(n) - 4\hat{\mu}_3(n)\hat{\mu}_1(n) - 3\hat{\mu}_2^2(n) + 12\hat{\mu}_2(n)\hat{\mu}_1^2(n) - 6\hat{\mu}_1^4(n). \tag{7.40}$$

Dans le cas d'une variable aléatoire centrée, pour les estimateurs arithmétiques, nous avons :

$$E[\hat{\kappa}_4(n)] = \kappa_4 - \frac{3}{n}(\kappa_4 + 2\mu_2^2). \tag{7.41}$$

Cet estimateur est donc asymptotiquement sans biais, et a pour variance :

$$\begin{aligned} Var(\hat{\kappa}_4(n)) &= \frac{1}{n}(\kappa_8 + 16\kappa_6\kappa_2 + 48\kappa_5\kappa_3 + 34\kappa_4^2 \\ &\quad +72\kappa_4\kappa_2^2 + 144\kappa_3^2\kappa_2 + 24\kappa_2^4). \end{aligned}$$

La variance de cet estimateur converge vers zéro lorsque n tend vers l'infini, il est donc consistant.

De même le *skewness* et le *kurtosis* peuvent être estimés par :

$$\hat{\chi}(n) = \frac{\hat{\kappa}_3(n)}{(\hat{\sigma}^2(n))^{\frac{3}{2}}}, \tag{7.42}$$

et

$$\hat{\gamma}(n) = \frac{\hat{\kappa}_4(n)}{(\hat{\sigma}^2(n))^2}. \tag{7.43}$$

Ces estimateurs ont été étudiés, seulement de façon approchée, pour les grandes valeurs de n, par exemple dans [McCullagh, 1987]. Il est montré qu'ils sont biaisés au premier ordre (*i.e.* leur biais dépend des cumulants d'ordre plus élevé), et qu'ils sont corrélés. Il existe cependant des résultats exacts, dans le cas où la variable aléatoire est centrée et suit une loi gaussienne, nous avons :

$$\begin{aligned} E[\hat{\chi}(n)] &= 0, \\ E[\hat{\gamma}(n)] &= 0, \\ Var(\hat{\chi}(n)) &= \frac{6n(n-1)}{(n-2)(n+1)(n+3)} \simeq \frac{6}{n}, \\ Var(\hat{\gamma}(n)) &= \frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)} \simeq \frac{24}{n}. \end{aligned}$$

Il apparaît que plus le moment (et donc le cumulant) est d'ordre important, plus la variance de son estimateur est grande.

Une autre quantité qui peut être intéressante, est le moment, non centré, mais normalisé par la variance, *i.e.* :

$$\hat{m}_3(n) = \frac{\hat{\mu}_3(n)}{(\hat{\sigma}^2(n))^{\frac{3}{2}}}, \tag{7.44}$$

et

$$\hat{m}_4(n) = \frac{\hat{\mu}_4(n)}{(\hat{\sigma}^2(n))^2}. \tag{7.45}$$

Ces estimateurs ont une variance plus faible que le *skewness* et le *kurtosis*. Bien sûr, si la variable aléatoire considérée est centrée, nous avons :

$$\hat{\chi}(n) = \hat{m}_3(n), \tag{7.46}$$

et

$$\hat{\gamma}(n) = \hat{m}_4(n) - 3. \tag{7.47}$$

Les résultats théoriques sont souvent simplifiés, si les variables aléatoires sont centrées, or ceci n'est pas toujours le cas. Pour effectuer un centrage de manière parfaite, il faut considérer un grand nombre de données.

Sous l'hypothèse de variables i.i.d., les moments et cumulants d'ordre supérieur estimés ci-dessus, sont donc biaisés ou asymptotiquement biaisés, mais lorsque λ tend vers 1, les estimateurs ne sont plus biaisés. De plus leur variance converge vers zéro lorsque le nombre d'échantillons augmente, ils sont donc consistants. Ces estimateurs peuvent donc nous permettre de mieux caractériser les distributions des variables étudiées.

Les moyennes et les variances de ces estimateurs étant difficiles à établir théoriquement, nous allons, à présent, étudier expérimentalement les moyennes et les variances de quelques-uns d'entre eux.

7.5 Estimation des moyenne et variance de quelques estimateurs

Le calcul théorique de la moyenne et de la variance des estimateurs des statistiques d'ordre supérieur devient vite complexe. C'est pourquoi, nous présentons dans les tableaux 7.1, 7.3 et 7.4, quelques résultats expérimentaux, sur des estimateurs calculés à partir de fenêtres exponentielles, vus précédemment. Ces estimateurs ont été initialisés par leur valeur estimée arithmétiquement. Les calculs ont été faits à partir d'une variable aléatoire de loi uniforme [0,1], générée avec le générateur de Matlab. Dans un premier temps les calculs ont été réalisés pour 998 échantillons, avec une valeur de $n = 1000$. Dans la colonne "théorique" ont été reportées les valeurs des estimations, par un estimateur arithmétique.

Moyenne : $n = 1000$, test sur 998 échantillons				
	$\lambda = 0.9$	$\lambda = 0.99$	$\lambda = 0.995$	théorique
$\hat{\mu}_1(n)$	0.5006	0.5004	0.5006	0.5004
$\hat{\mu}_2(n)$	0.3330	0.3334	0.3337	0.3337
$\hat{\sigma}^2(n)$	0.0781	0.0826	0.0830	0.0833
$\hat{\sigma}^2_L(n)$	0.0590	0.0613	0.0641	0.0833
$\hat{\chi}(n)$	-0.0061	-0.0012	-0.0015	0.0000
$\hat{m}_3(n)$	12.1429	10.5749	10.4952	10.4152

TAB. 7.1 – *Moyennes expérimentales pour* $n = 1000$.

Nous remarquons que plus l'ordre de la statistique estimée est important, plus il est nécessaire de prendre un facteur d'oubli proche de 1, pour une même précision de la moyenne. De plus, l'estimateur de la variance par l'estimation de l'écart-type est moins fiable que l'estimation de la variance, pour des distributions qui ne sont pas laplaciennes.

Dans le but de leur comparaison, les valeurs numériques de la variance, sont calculées, comme une proportion de la moyenne. C'est-à-dire, la variance observée a été divisée par la moyenne observée, puis multipliée par 100. Dans le tableau 7.2 figurent les valeurs théoriques de cette proportion calculée à partir des formules théoriques présentées dans le paragraphe précédent. Les valeurs "théoriques" de la moyenne, variance, et autres moments, nécessaires pour l'obtention de ces résultats ont été calculées à partir des estimateurs arithmétiques.

Variance théorique			
	$\lambda = 0.9$	$\lambda = 0.99$	$\lambda = 0.995$
$\hat{\mu}_1(n)$	0.88	0.08	0.04
$\hat{\mu}_2(n)$	1.41	0.13	0.07

TAB. 7.2 – *Variance des estimateurs* $\hat{\mu}_1$ *et* $\hat{\mu}_2$ *à partir des formules théoriques.*

Variance : $n = 1000$, test sur 998 échantillons			
	$\lambda = 0.9$	$\lambda = 0.99$	$\lambda = 0.995$
$\hat{\mu}_1(n)$	0.86	0.08	0.04
$\hat{\mu}_2(n)$	1.38	0.13	0.06
$\hat{\sigma}^2(n)$	0.39	0.03	0.02
$\hat{\sigma}_L^2(n)$	0.37	0.04	0.02
$\hat{\chi}(n)$	1962.30	858.33	346.67
$\hat{m}_3(n)$	179.80	9.72	4.82

TAB. 7.3 – *Variance expérimentale pour* $n = 1000$.

La valeur pour $\lambda = 0.999$ peut être erratique, ceci est dû au fait que la taille de la fenêtre correspond à n. En effet, pour $\lambda = 0.999$, l'estimation se fait sur $N = \frac{1}{1-\lambda} = 1000$ valeurs. Pour remédier à ce problème, et pour calculer la variance avec des facteurs d'oubli plus importants, nous présentons les tableaux 7.3 et 7.4, pour ce dernier tableau le calcul fait avec $n = 10000$ et un nombre moins important d'échantillons.

Variance : $n = 10000$, test sur 98 échantillons			
	$\lambda = 0.999$	$\lambda = 0.9995$	$\lambda = 0.9999$
$\hat{\chi}(n)$	42.86	46.46	23.17
$\hat{m}_3(n)$	1.07	0.57	0.09

TAB. 7.4 – *Variance expérimentale pour* $n = 10000$.

Nous remarquons que plus l'ordre de la statistique estimée est faible, moins il est nécessaire d'avoir un facteur d'oubli proche de 1 pour obtenir un taux de variance faible. De plus la variance de $\hat{m}_3$ diminue plus quand le facteur d'oubli augmente que la variance de $\hat{\chi}$, alors qu'ils sont du même ordre.

Nous allons à présent voir comment ces statistiques peuvent être utilisées en vue de la détection de la parole.

7.6 Utilisation des statistiques d'ordre supérieur pour la détection de parole

[Jacovitti *et al.*, 1991] propose l'utilisation du *skewness* et du *kurtosis*, calculés sur le signal dans une perspective de détection des sons voisés et non voisés du signal. Il suppose le signal stationnaire et estime $\chi(n)$ et $\gamma(n)$ du signal sur des fenêtres rectangulaires. Les moments d'ordre supérieur étant instables, il considère des échantillons de l'ordre de 50 *ms*, pour l'estimation. Ceci entraîne un retard dans la détection.

Dans [Doukas *et al.*, 1997] le fait que le cumulant croisé de deux variables aléatoires

indépendantes est nul, est utilisé pour discriminer le signal de parole et celui du bruit à la source. Pour obtenir ce cumulant croisé, il est nécessaire d'avoir deux sources. Il filtre donc la source unique à l'aide d'un filtre passe-bas H_1 et d'un filtre passe-haut H_2, et introduit ainsi une deuxième source fictive à l'aide d'un "réseau de neurones" (*cf.* figure 7.1). Les filtres utilisés sont :

$$\begin{aligned} H_1 &= \frac{1}{2} + z^{-1} + \frac{1}{2}z^{-2}, \\ H_2 &= -\frac{1}{2} + z^{-1} - \frac{1}{2}z^{-2}. \end{aligned}$$

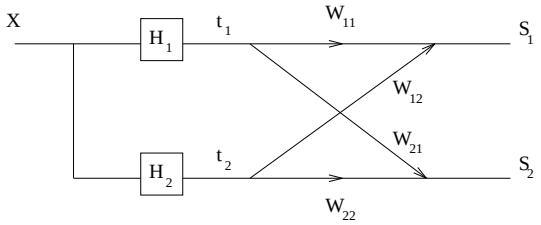


FIG. 7.1 – *Séparation de source.*

Il faut ensuite minimiser la fonction :

$$J = \sum_{i,j=0}^{+\infty} \left(E[s_1^{2i+1} s_2^{2j+1}] \right)^2 , \tag{7.48}$$

où s_1 et s_2 sont les sorties du "réseau de neurones" (*cf.* figure 7.1). En pratique la somme s'effectue pour un grand nombre de i et j. Lorsque le signal est stationnaire J reste constante, et lorsque les statistiques du signal x changent J devient très grande. J est estimée récursivement sur une fenêtre exponentielle :

$$\hat{J}(n) = \hat{J}(n-1) + (1-\lambda) \left(\sum_{i,j=0}^{+\infty} \left(E[s_1^{2i+1} s_2^{2j+1}] \right)^2 - \hat{J}(n-1) \right). \tag{7.49}$$

Cet estimateur est comparé à un seuil T_n qui est adapté par :

$$T_n = T_{n-1} + (1-\lambda_i)(\hat{J}(n-1) - T_{n-1}), \tag{7.50}$$

où λ_i pour $i = 0{,}1$, est un facteur d'oubli différent selon l'état de silence ($i = 0$) ou de parole ($i = 1$) dans lequel nous sommes. λ_0 et λ_1 sont choisis de manière à avoir $\frac{\lambda_0}{\lambda_1} = 100$. Le changement d'état se fera lorsque $\hat{J}(n) > 1 * T_n$, lorsque nous sommes dans l'état de parole, et lorsque $\hat{J}(n) < 1.4 * T_n$, lorsque nous sommes dans l'état de silence.

Cette méthode de détection d'activité vocale a été prise dans [Bouteille, 1999] en simplifiant les calculs, tout en conservant une bonne détection. Les valeurs de t_1 et de t_2

sont reprises directement sans faire appel au réseau de neurones. Pour éviter les calculs d'espérance mathématique de l'équation (7.49), t_1 et $t_2$sont intégrés dans $\hat{J}(n)$ de la façon suivante :

$$\hat{J}(n) = \hat{J}(n-1) + (1-\lambda)\left(\sum_{i,j} t_1^i t_2^j - \hat{J}(n-1)\right), \tag{7.51}$$

où $(i,j) \in \{(u,v) \in \mathbb{R}^2 : 0 \leq u \leq r,\ 0 \leq v \leq r \text{ et } u+v=r\}$, r étant l'ordre maximum des moments considérés (supérieur à 2).

Dans [Nemer *et al.*, 1999] il est montré que les cumulants d'ordre 3 sur le résidu de prédiction des coefficients LPC du signal voisé peuvent être vus comme une estimation de la fréquence fondamentale. L'étude est étendue au cumulant d'ordre 4. Il est montré que le *kurtosis* du signal voisé n'est pas nul, et peut être employé pour la détection de parole. En combinaison avec l'énergie il intègre ce cumulant dans une DAV, qu'il compare à la DAV ITU-T G.729B (*cf.* [ITU Recommendation, 1996]).

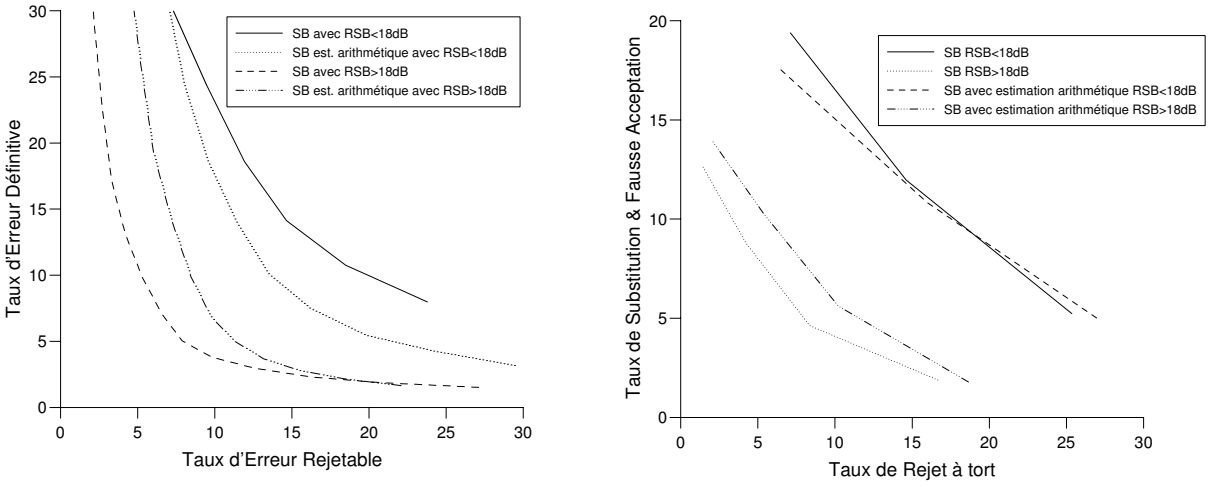
Dans les critères SB et SBP du module de détection ABP, présentés au paragraphe 2.2.3 et 2.2.4, les statistiques d'ordre 1 et 2 sont utilisées pour le calcul de la moyenne et de la variance de l'énergie du bruit et de la parole. L'estimation de ces statistiques est une estimation sur une fenêtre exponentielle, donnée par les équations (7.21) et (7.34). Pour le critère SB, la variance de l'énergie du bruit est estimée avec un facteur d'oubli $\lambda = 0.995$. Pour ce critère, nous comparons ici cette estimation avec l'estimation arithmétique de la variance de l'énergie sur une fenêtre glissante de taille équivalente $n = 200$. La figure 7.2(a) montre qu'au niveau des tests de détection sur la base GSM_A avec un RSB inférieur à 18 *dB*, il y a une amélioration des performances avec l'estimation arithmétique, alors qu'avec un RSB supérieur à 18 *dB*, il y a une dégradation. Cependant la figure 7.2(b) montre qu'au niveau des résultats de reconnaissance l'amélioration avec le RSB inférieur à 18 *dB* n'est pas conservée. De plus l'estimation arithmétique entraîne une dégradation sur la base de parole continue (*cf.* figure 7.3).

L'estimation arithmétique a l'inconvénient de donner une importance égale sur la longueur de la fenêtre glissante. Cette fenêtre étant réduite à quelques trames (ici 200), n'a pas de grande conséquence sur les résultats. Cependant les résultats présentés ici n'encouragent pas l'utilisation de cet estimateur.

Nous allons à présent voir comment intégrer les statistiques d'ordre trois au critère SB du module de détection ABP, qui est le critère le plus performant.

7.7 Intégration du moment d'ordre 3

Nous avons choisi d'intégrer en complément du critère SB du module de détection, $\hat{m}_3$, le moment d'ordre 3 non centré, mais normalisé par la variance. En effet nous avons vu que la variance de cet estimateur est plus faible que celle du *skewness*, de plus $\hat{m}_4$, le moment normalisé d'ordre 4, non centré, a une variance encore plus importante et ne



(a) Résultats de détection

(b) Résultats de reconnaissance

FIG. 7.2 – *Comparaison de l'estimation arithmétique et sur une fenêtre exponentielle sur la base GSM_A.*

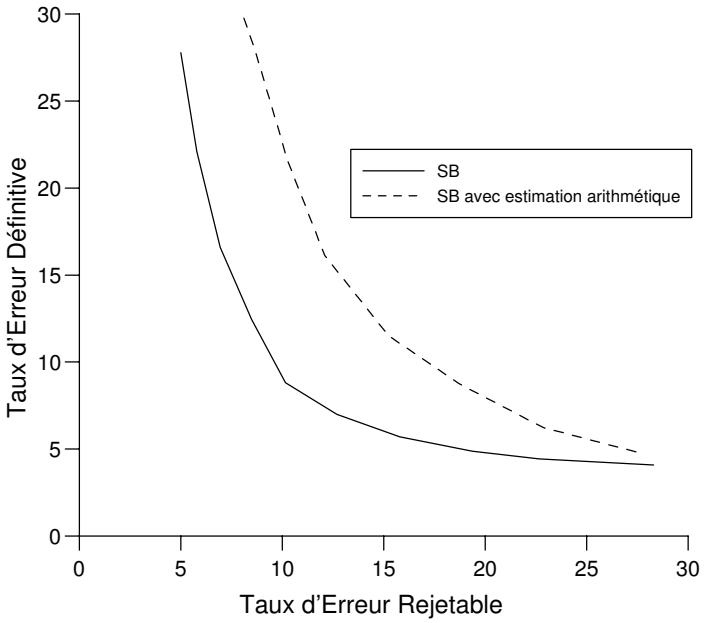


FIG. 7.3 – *Comparaison de l'estimation arithmétique et sur une fenêtre exponentielle sur la base AGORA.*

semble pas apporter beaucoup plus. La figure 7.4 montre les "fonctions de répartition" du rapport des moments d'ordre 3 et 4 calculés sur l'énergie dans les périodes de parole sur ceux calculés dans les périodes de bruit. Les rapports des moments dans les périodes de parole et dans les périodes de bruit comparés à un seuil permettent de discriminer les périodes de parole et de bruit. Les moments, non centrés, normalisés ont été calculés sur une fenêtre exponentielle. La figure 7.4 représente le nombre de fichiers dont le rapport des moments d'ordre 3 et 4 est inférieur à une borne supérieure donnée (variant de 0 à 3). Cette figure indique que pour un même facteur d'oubli $\lambda = 0.995$, le rapport des moments d'ordre 4 n'apporte pas une nette amélioration de la capacité de discrimination du bruit et de la parole par rapport au rapport des moments d'ordre 3.

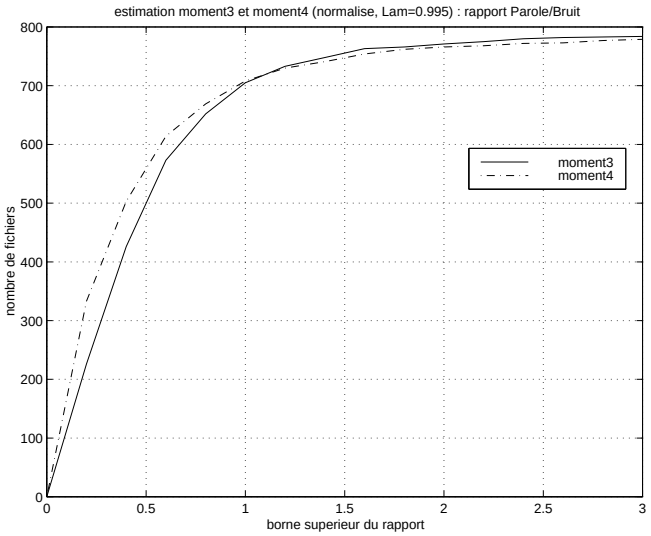


FIG. 7.4 – *Statistiques des rapports Parole/Bruit des moments d'ordre 3 et 4.*

Le moment, non centré, normalisé, d'ordre 3 a été calculé pour les premiers coefficients cepstraux, dans les périodes de silence et de parole, selon tous les environnements sur la base GSM_A. Seul pour l'énergie, il apparaît discriminant. La figure 7.5 représente le moment, non centré, normalisé, d'ordre 3 de l'énergie, calculé sur une fenêtre exponentielle de 10 trames, sur un fichier comportant du bruit de fond. La segmentation manuelle des mots est représentée ainsi que l'énergie. Nous remarquons que le moment d'ordre 3 décroît fortement pendant les périodes de parole.

Nous avons donc tenté d'affiner la discrimination de l'énergie du bruit et de la parole faite par le critère SB à l'aide du moment d'ordre 3. À ce critère, nous ajoutons le rapport de $\hat{m}_3$ à court-terme (calculé à partir des estimateurs arithmétiques sur une fenêtre glissante de 10 trames), noté $\hat{m}_{3ct}$, et de $\hat{m}_3$ à long-terme (calculé sur une fenêtre exponentielle avec un facteur d'oubli important $\lambda = 0.995$), noté $\hat{m}_{3lt}$. Le moment à court-terme peut également être calculé sur une fenêtre exponentielle (*cf.* [Martin, 2000] et [Martin *et al.*, 2000]) ce qui donne des résultats équivalents. Cependant, plus le facteur d'oubli

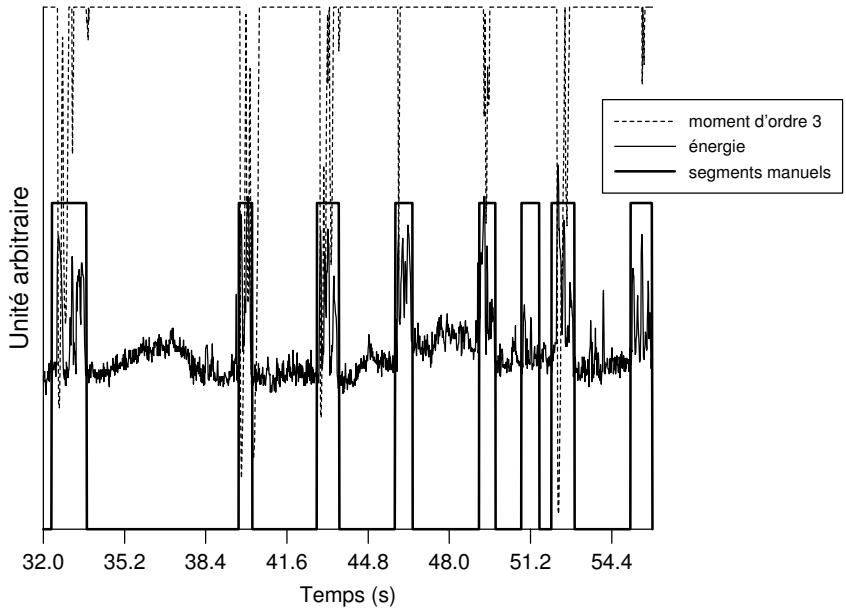


FIG. 7.5 – *Représentation du logarithme de l'énergie et du moment d'ordre 3 au cours du temps.*

du moment à court-terme est important, plus nous observons un retard sur la détection des segments de parole. Le problème réside dans le fait que la variance de cet estimateur sur une fenêtre exponentielle reste grande, contrairement à l'estimateur arithmétique. Il s'avère en fait que seul le moment à court-terme est discriminant. Le moment à long-terme est estimé dans les états *bruit ou silence* de l'automate. La décision sera prise en comparant le rapport $rap(n) = \frac{\hat{m}_{3ct}(n)}{\hat{m}_{3lt}(n)}$, des estimations à court-terme et à long-terme des moments d'ordre 3, à un seuil adaptatif. Cette décision est prise conditionnellement au test sur l'énergie de l'algorithme initial, et aux contraintes de temps (*cf.* figure 6.3). C'est-à-dire que pour le passage d'un état à l'autre, nous testons dans un premier temps $r_{SB}(x)$ au seuil de détection, qui est la condition C1 avec le critère SB (*cf.* paragraphe 2.2.3). Dans un second temps, si l'énergie est suffisamment importante, la décision est confirmée, ou non, par le test sur le rapport des moments d'ordre 3.

Le seuil adaptatif est calculé à partir de la moyenne du rapport, multiplié par un coefficient, dans les périodes de parole, prise sur une fenêtre exponentielle. Le fait de multiplier le rapport par un coefficient supérieur à 1 permet d'obtenir une borne supérieure, pour le rapport des moments d'ordre 3, $rap(n)$. Le seuil adaptatif est ainsi calculé par la formule de récurrence suivante :

$$\hat{T}(n+1) = \hat{T}(n) + (1 - \lambda_T)(coef.rap(n) - \hat{T}(n)), \tag{7.52}$$

où $rap(n)$ est le rapport des moments d'ordre 3 à court-terme et à long-terme, à l'instant n dans une période de parole, $\lambda_T = 0.99$ est le facteur d'oubli, et $coef = 3$ est le coefficient permettant d'obtenir la borne supérieure du rapport des moments d'ordre 3 dans les périodes de parole. Ces paramètres ont été optimisés expérimentalement à partir des tests de détection sur la base GSM_A. Un seuil fixe ne nous permet pas d'obtenir une adaptation indispensable du seuil au niveau de bruit ambiant qui peut changer. Ainsi la condition C4 s'écrit: $rap(n) < T(n)$.

Dans ce qui suit, nous comparons le module de détection ABP employant le critère SB avec le module de détection ABP employant le rapport des moments d'ordre 3 en complément du critère SB, critère que nous nommons SB+M3.

7.8 Expérimentations

Nous présentons ici les résultats des tests de la détection et de la reconnaissance sur la base GSM_T en fonction du RSB. L'étude présentant des résultats similaires sur les autres bases n'est pas donnée. Nous comparons le nouveau critère SB+M3 avec le critère SB, duquel il est issu.

7.8.1 Résultats de détection

La figure 7.6 donne les résultats de détection sur la base GSM_T. Nous remarquons qu'il n'apparaît pas de différences entre le critère SB+M3 et le critère SB.

Pour étudier plus en détail les erreurs produites, pour les seuils qui donnent le minimum des taux d'erreur associée (somme des erreurs rejetables et définitives) pour le critère SB et le critère SB+M3, selon les bases et les RSB. Ces seuils sont donnés dans le tableau F.1 en Annexe F.

L'histogramme 7.7 montre que les deux critères sont très similaires sur la base GSM_T. Il ne se dégage aucune réelle différence des erreurs entre les deux critères.

Ce nouveau critère ne semble donc pas pertinent tel qu'il est employé dans le module de détection. C'est pourquoi nous avons étudié d'autres possibilités d'utilisation de ce critère.

Avec l'estimation arithmétique du moment d'ordre 3 à long-terme, les résultats obtenus (*cf.* figure 7.8) montrent une dégradation importante des performances. La taille de la fenêtre glissante est de 200 trames, elle a été optimisée sur la base GSM_A.

Une autre façon de calculer le moment d'ordre 3, est de le calculer directement sur l'information temporelle, afin d'obtenir un coefficient par trame de 16 *ms*. Le moment d'ordre 3 est calculé pour chaque échantillon, la valeur gardée pour une trame est celle du dernier échantillon de la trame. Le moment ainsi calculé est plus précis et entraîne moins de retard dans l'algorithme de détection. Le moment d'ordre 3 est calculé sur une fenêtre exponentielle avec un facteur d'oubli $\lambda = 0.999$. Ce facteur a été optimisé expérimentalement sur la base GSM_A. Le facteur d'oubli pour l'estimation du seuil adaptatif ainsi que le coefficient de surévaluation restent inchangés. La figure 7.9 présente

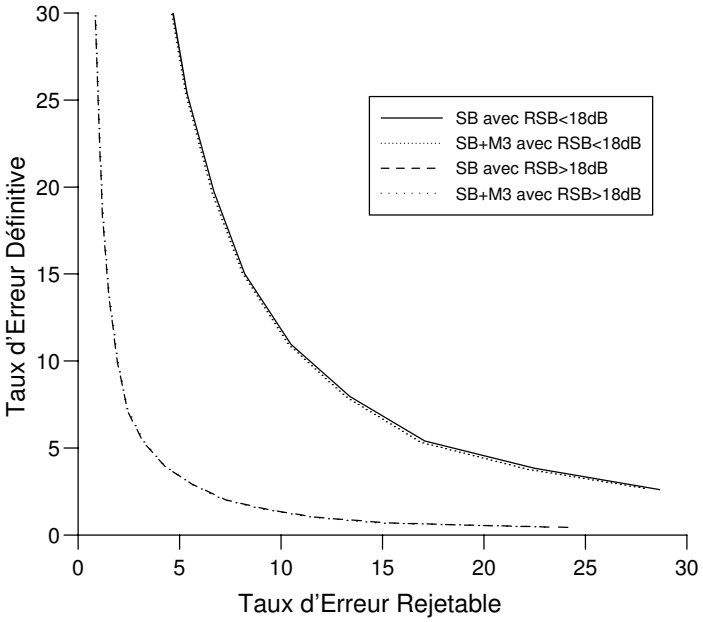


FIG. 7.6 – *Résultats de détection des critères SB+M3 et SB sur la base GSM_T.*

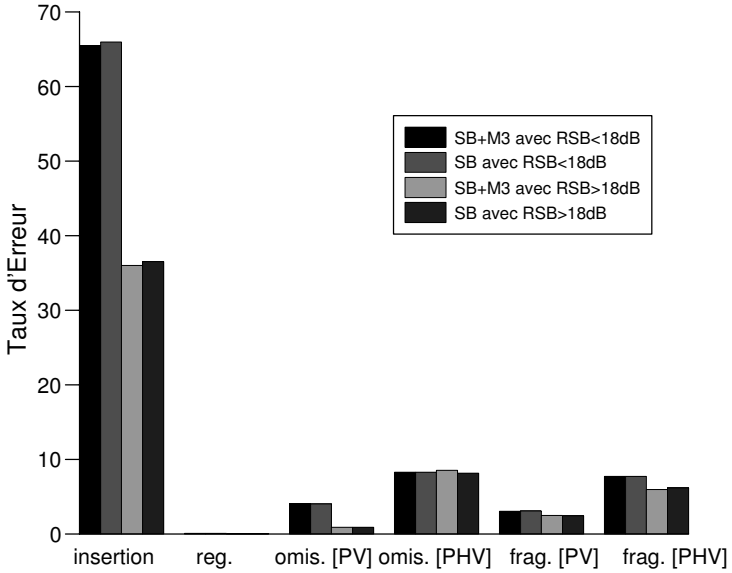


FIG. 7.7 – *Erreurs de détection détaillées des critères SB+M3 et SB sur la base GSM_T.*

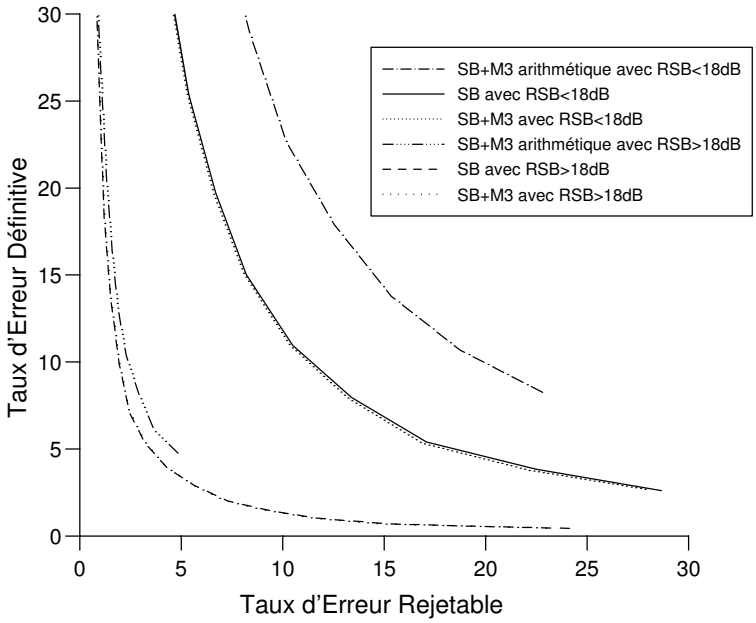


FIG. 7.8 – *Résultats de détection du critère SB+M3 avec un estimateur arithmétique et sur fenêtre exponentielle et du critère SB sur la base GSM_T.*

les résultats sur la partie la plus bruitée de la base GSM_T. Nous constatons que le calcul du moment d'ordre 3 dans le domaine temporel n'apporte pas d'amélioration significative.

7.8.2 Résultats de reconnaissance

Le paragraphe précédent montre la grande similitude des résultats du module de détection pour les deux critères SB et SB+M3. Nous présentons tout de même ici les résultats de reconnaissance. Le tableau F.3 en Annexe F donne les seuils optimaux de reconnaissance. Pour ces seuils la figure 7.10 montre qu'il n'y a pas de différences entre le critère SB et le critère SB+M3 sur la base GSM_T. Nous obtenons les mêmes résultats sur les bases RTC_T et AGORA.

7.9 Conclusion

Nous avons étudié la possibilité d'améliorer les performances du module de détection en utilisant l'énergie plus finement à l'aide des statistiques d'ordre supérieur. Les résultats présentés dans ce chapitre comparent une méthode employant les statistiques d'ordre 1 et 2 (la moyenne et la variance) avec une méthode employant en plus les moments d'ordre 3. Ce résultats montrent qu'il est difficile d'intégrer le moment d'ordre 3 de l'énergie,

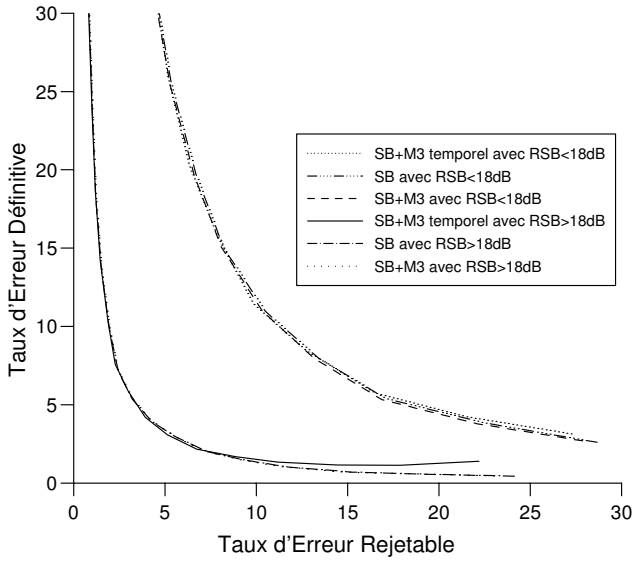


FIG. 7.9 – *Résultats de détection du critère SB+M3 dans le domaine temporel et du critère SB sur la base GSM_T.*

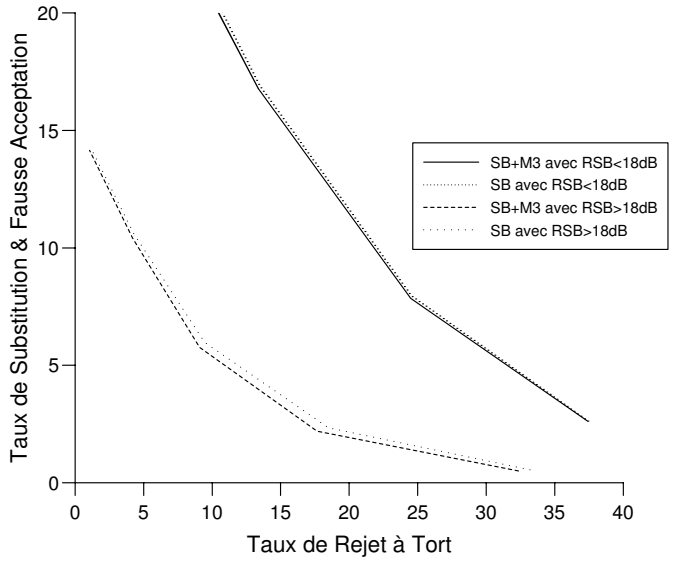


FIG. 7.10 – *Résultats de reconnaissance des critères SB+M3 et SB, sur la base GSM_T.*

pour obtenir de meilleurs performances du module de détection. Plusieurs approches de calculs et d'intégration du moment d'ordre 3 ont été étudiés. La précision donnée par la statistique sur l'énergie, telle que nous l'avons intégrée, n'apporte pas d'améliorations significatives.

Les statistiques d'ordre supérieur nécessitent une estimation qui demande un nombre important de trames, à cause de la variance des estimateurs. Le moment d'ordre 3 permet cependant une estimation sur un nombre de trames raisonnables. Il apporte une précision sur la distribution du signal de parole et de bruit par rapport à la moyenne et la variance seule. Cette information supplémentaire n'apporte cependant pas d'amélioration des performances.

Les résultats obtenus dans ce chapitre ne sont pas encourageants pour poursuivre l'étude des statistiques d'ordre supérieur pour discriminer davantage la parole et le bruit. Nous pouvons donc penser que le critère énergétique est employé en tenant compte de toute l'information utile à la discrimination de la parole et du bruit.

Pour améliorer les performances, nous avons proposé deux possibilités (*cf.* Chapitre 4 " *Voies envisagées pour l'amélioration du module de détection*") :
- soit utiliser l'énergie plus finement,
- soit utiliser d'autres caractéristiques.

Le premier point étudié dans ce chapitre n'apportant pas d'améliorations significatives, il faut donc chercher à compléter la décision faite par l'énergie à l'aide d'autres caractéristiques et critères. Une autre caractéristique importante de la parole est le voisement. Nous proposons dans le chapitre qui suit un nouveau critère utilisant cette caractéristique.

Chapitre 8

Utilisation d'un paramètre de voisement

8.1 Introduction

La prosodie de la parole est difficile à modéliser, et se limite souvent à la modélisation de la mélodie. La mélodie provient de la vibration des cordes vocales. C'est pourquoi, l'énergie est parfois associée à un paramètre de prosodie pour affiner la détection de la parole dans le bruit. Le bruit est généralement non-voisé, tandis que les voyelles et une partie des consonnes sont voisées. Nous étudions dans un premier temps les paramètres de prosodie et leur estimation dans le paragraphe 8.2, puis quelques utilisations de ces paramètres dans des systèmes de détection de parole, dans le paragraphe 8.3. Nous intégrons ensuite un paramètre de voisement dans le module de détection, dans le paragraphe 8.4, puis son évaluation au paragraphe 8.5.

8.2 Paramètres de prosodie et leur estimation

La fréquence des vibrations des cordes vocales peut se définir de différents points de vue. La fréquence de vibration laryngienne fait référence au processus de génération articulatoire. La fréquence fondamentale ou F_0 si nous nous plaçons dans le domaine acoustique, est la hauteur de la voix si nous nous plaçons dans le domaine perceptif. Le *pitch* est le terme anglais pour désigner ces trois appellations, et est souvent confondu avec la fréquence fondamentale. La prosodie est déterminée par trois paramètres, la fréquence, l'intensité et la durée des segments phonétiques qui, au cours de la prononciation du locuteur peuvent évoluer indépendamment. Il existe cependant une interaction de la fréquence, de l'intensité et de la durée qui est très complexe à définir, c'est pourquoi nous nous limitons souvent à l'étude de la fréquence F_0. Plus de détails sur ces paramètres phonétiques sont donnés dans [Calliope, 1989].

La fréquence fondamentale varie avec le sexe, l'âge, l'accent, l'état émotif du locuteur, *etc.* Ses valeurs peuvent être comprises entre 50 et 200 *Hz*.

Différentes méthodes existent pour mesurer la fréquence fondamentale, et ainsi déterminer l'existence de voisement. Il existe un grand nombre de méthodes (*cf.* [Hess, 1983] et [Bagshaw, 1994]), il ne s'agit pas ici de les énumérer. Nous pouvons cependant définir deux classes de méthodes, les méthodes dans le domaine temporel, qui entraînent généralement le calcul de la fonction d'autocorrélation, et celles dans le domaine fréquentiel qui impliquent le calcul de la transformée de Fourier, ou un calcul similaire. Notons également qu'il existe d'autres méthodes à partir du maximum de vraisemblance (*cf.* [Hess, 1983]), des méthodes à partir de l'analyse temps-échelle (*cf.* [Montrésor et Baudry, 1990]), ou encore des méthodes à partir de réseaux neuro-flou qui permettent de déterminer un trait de voisement (*cf.* [Sokol, 1996]).

8.3 Utilisation du voisement dans des systèmes de détection de parole

Nous présentons ci-dessous quelques méthodes utilisant la fréquence fondamentale dans des systèmes de détection de parole, appliqués à différentes problématiques.

La fréquence fondamentale n'est pas en général utilisée comme seule caractéristique, dans [Kobatake *et al.*, 1989] elle est employée parmi quatre caractéristiques : la périodicité, l'ordre optimal du modèle LPC et une distance LPC minimale (*cf.* paragraphe 4.5.1).

La divergence de Kullback est employée dans [Di Francesco, 1990] pour déterminer F_0 dans le domaine temporel. Il utilise le paramètre ainsi calculé pour une segmentation de parole en segment voisé/non-voisé/silence. Il définit un test à l'aide d'une mesure de convexité du rapport *a posteriori* de la divergence de Kullback. Dans un premier temps une détection de parole est effectuée à l'aide des coefficients PARCOR, la mesure de convexité permet, dans un second temps, d'affiner cette détection.

L'algorithme de [Hamada *et al.*, 1990] propose une méthode simple fondée sur l'intervalle des pics énergétiques qui est utilisée pour estimer la fréquence fondamentale. Les coefficients LPC sont combinés avec ce paramètre. Cette combinaison s'effectue à l'aide de plusieurs distances pondérées, qui sont comparées dans des environnements de différents RSB. Cette approche a été testée dans [Junqua *et al.*, 1991], il en ressort que les performances de cet algorithme sont moins concluantes que la méthode présentée par [Junqua *et al.*, 1991] (*cf.* Chapitre 4 "*Voies envisagées pour l'amélioration du module de détection*"), mais restent bonnes.

Dans [Ramana Rao et Srichand, 1996] les variations de F_0 sont intégrées pour la détection de frontières de mots. Il discute la possibilité d'utiliser cette caractéristique pour une application en reconnaissance de parole continue. Il fait l'hypothèse qu'en fin de phrase la fréquence fondamentale diminue.

L'énergie du signal et une interpolation de F_0 sont combinées dans [Strom, 1995] pour la détection d'accent et de frontières de phrase la langue allemande. Les performances de reconnaissance en sont améliorées. Ces travaux sont adaptés dans [Sakurai et Hirose, 1996] pour la reconnaissance de parole continue en japonais, en ajoutant le taux de passage par zéro. Ce système a ensuite conduit à l'utilisation de F_0 de la *more*, qui est un élément

vocalique spécifique des langues comme le japonais. Ainsi, dans [Hirose et Iwano, 1997], [Iwano et Hirose, 1998], puis dans [Iwano et Hirose, 1999] les variations de F_0 de la *more* ont été utilisées pour de la reconnaissance de parole continue de la langue japonaise.

8.4 Intégration d'un paramètre de voisement

Dans le paragraphe précédent, nous avons présenté quelques méthodes utilisant la fréquence fondamentale dans un système de détection. Nous avons choisi d'intégrer un paramètre de voisement à l'aide de la fréquence fondamentale calculée à partir d'une méthode spectrale. La méthode utilisée cherche l'harmonicité du signal par intercorrélation avec un fonction peigne. Différentes distances entre les dents de cette fonction, d'amplitude décroissante, sont employées. Cette méthode est similaire à celle décrite dans [Martin, 1982].

Cette méthode permet de calculer une valeur toutes les quatre millisecondes sur tout le signal, même dans les périodes de non-parole. Dans les périodes voisées du signal cette valeur est la fréquence fondamentale. Nous utilisons ainsi le terme de fréquence fondamentale par abus de langage pour désigner cette valeur. Toutes les quatre millisecondes, nous calculons la médiane entre la valeur courante et les deux précédentes. Nous prenons la médiane pour éviter les artefacts. Nous obtenons donc pour chaque trame n quatre valeurs $med(n_i)$ avec $i = 1,2,3,4$. Nous calculons ensuite la moyenne de la valeur absolue de la différence de la médiane courante et de la précédente :

$$\overline{\delta med}(n_i) = \frac{1}{N} \sum_{m_j=n_i-N}^{n_i} |med(m_j) - med(m_{j-1})|, \tag{8.1}$$

où N est la taille de la fenêtre arithmétique, $med(n_i)$ est la $i^{\text{ième}}$ médiane de la trame n. Cette moyenne, calculée sur les deux dernières valeurs, est un critère de la variation locale de la fréquence fondamentale. Si la fréquence fondamentale varie peu, la trame courante est supposée être une trame de parole. Nous obtenons ainsi une estimation d'un degré de voisement, $\delta med(n_i) = |med(n_i) - med(n_{i-1})|$. La figure 8.1 représente la moyenne de ce degré de voisement selon la segmentation manuelle sur les bases RTC_A et GSM_A. Nous constatons ainsi que ce degré de voisement permet de discriminer la parole des bruits impulsifs.

Ce critère de la variation de la fréquence fondamentale est intégré dans le module de détection en complément de l'énergie. Nous conservons la valeur de la moyenne de l'estimation du degré de voisement toutes les 16 *ms* $\overline{\delta med}(n_4)$, pour se ramener à la longueur de la trame utilisée. Nous comparons cette moyenne à un seuil fixe optimisé sur la base GSM_A par les tests de détection, pour le passage de l'état *présomption de parole* à l'état *parole* ou à l'état *bruit ou silence* de l'automate. Cette intégration cherche à diminuer les détections de bruit, et à obtenir une détection plus précise (*cf.* figure 6.3). Ainsi la condition C4 est : $\overline{\delta med}(n_4) < seuil_{\overline{\delta med}}$. Ce seuil a été optimisé par des tests de détection sur la partie bruitée de la base GSM_A, sa valeur est fixé à 10.

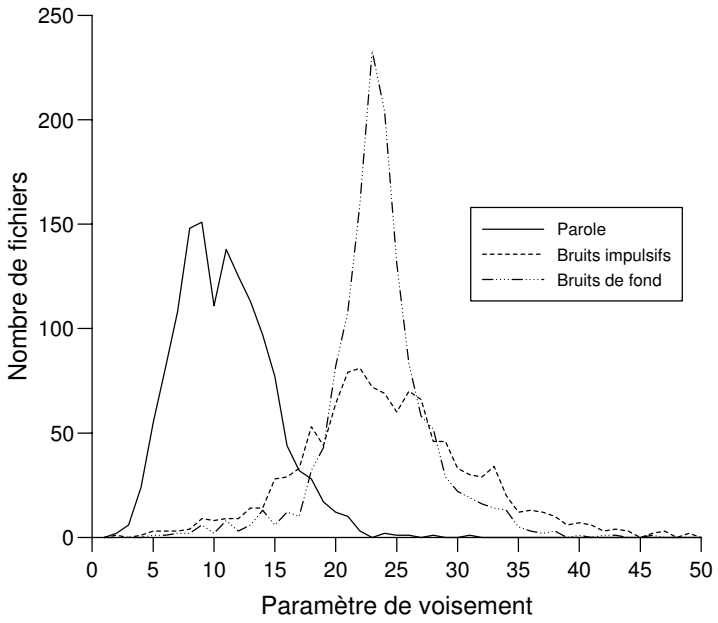


FIG. 8.1 – *Histogramme du degré de voisement sur les bases RTC_A et GSM_A.*

Il n'est donc possible de passer de l'état *présomption de parole* à l'état *parole*, que si les conditions C1, C2 et C4 sont réalisées simultanément.

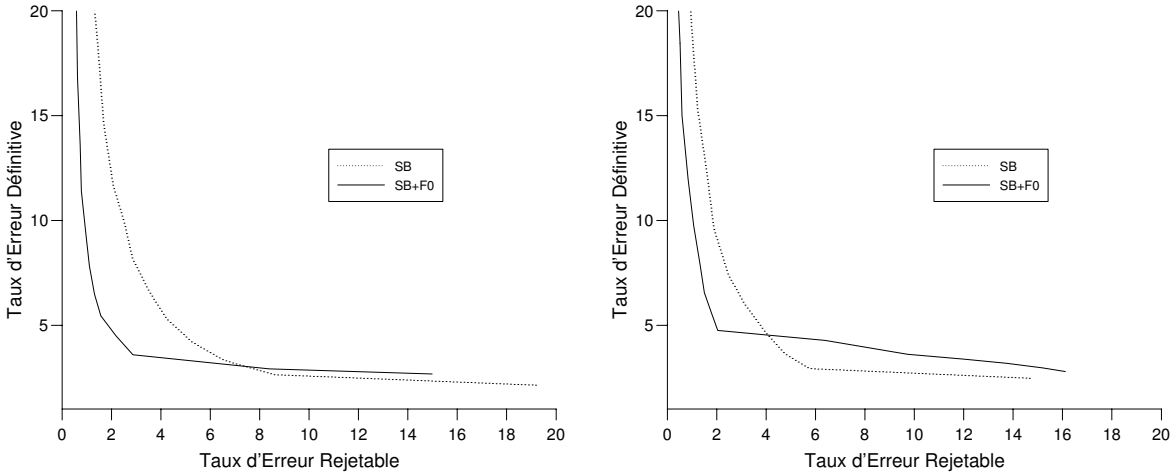
L'estimation de la moyenne du degré de voisement donnée par l'équation (8.1) est une estimation arithmétique sur une fenêtre glissante dont la taille N a été optimisée expérimentalement sur la base GSM_A. Nous partons toujours du module de détection ABP employant le critère SB pour implémenter cette nouvelle caractéristique.

8.5 Expérimentations

Nous présentons ici l'évaluation du module de détection employant ce paramètre de voisement en complément du critère SB, nous appelons cette association le critère SB+F_0. L'évaluation se fait à l'aide d'une part des résultats du module de détection, d'autre part des résultats du système de reconnaissance avec le module de détection, sur les bases de tests RTC_T, GSM_T et AGORA.

8.5.1 Résultats de détection

Nous commençons par présenter les résultats de détection avec le critère SB+F_0 en comparaison du critère SB. Les figures 8.2 et 8.3 présentent respectivement les résultats pour la base RTC_T et GSM_T pour les différents RSB qui séparent cette base en



(a) partie lue (b) partie répétée

FIG. 8.2 – *Résultats de détection des critères SB+F_0 et SB sur la base RTC_T.*

partie bruitée et non bruitée. Sur la figure 8.4 sont représentés les résultats sur la base de parole continue AGORA. Il est clair que le nouveau critère donne globalement de meilleurs résultats sur toutes les bases étudiées. Le tableau 8.1 montre que toutes les améliorations sont significatives. Pour les seuils qui donnent le minimum des taux d'erreur associée pour le critère SB et le critère SB+F_0, nous avons calculé l'intervalle de confiance à 95%. Notons que l'écart est plus important sur la partie bruitée de la base GSM_T que sur la partie calme. Nous observons également une amélioration importante sur la base AGORA.

	seuil "optimal" de SB+F_0	taux d'erreur de SB+F_0	seuil "optimal" de SB	intervalle de confiance de SB
RTC_T_L	1.7	6.46%	2.1	08.74;10.17
RTC_T_R	1.7	6.79%	1.9	07.72;09.09
GSM_T M18	1.3	15.54%	1.9	20.60;22.14
GSM_T P18	1.7	5.11%	2.5	07.79;08.69
AGORA	2.1	11.51%	3.1	17.49;20.55

TAB. 8.1 – *Taux d'erreur associée de détection du critère SB+F_0 par rapport à l'intervalle de confiance du critère SB.*

Sur les histogrammes 8.5 (*cf.* figure 8.5(a) pour la base RTC_T, figure 8.5(b) pour la base GSM_T, et figure 8.5(c) pour la base AGORA), nous détaillons les erreurs de détection, pour les deux critères SB+F_0 et SB, pour les seuils présentant le minimum des taux d'erreur (erreurs rejetables et définitives) (*cf.* tableau F.1 en Annexe F).

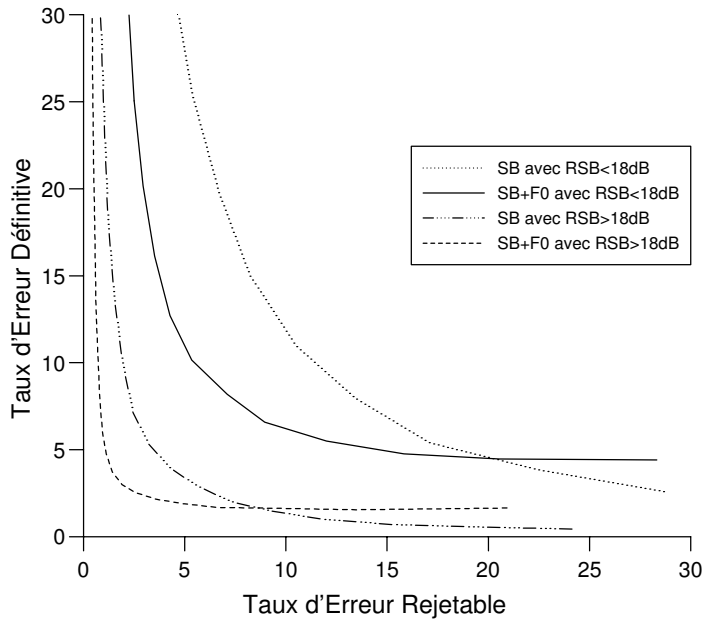


FIG. 8.3 – *Résultats de détection des critères SB+F_0 et SB sur la base GSM_ T.*

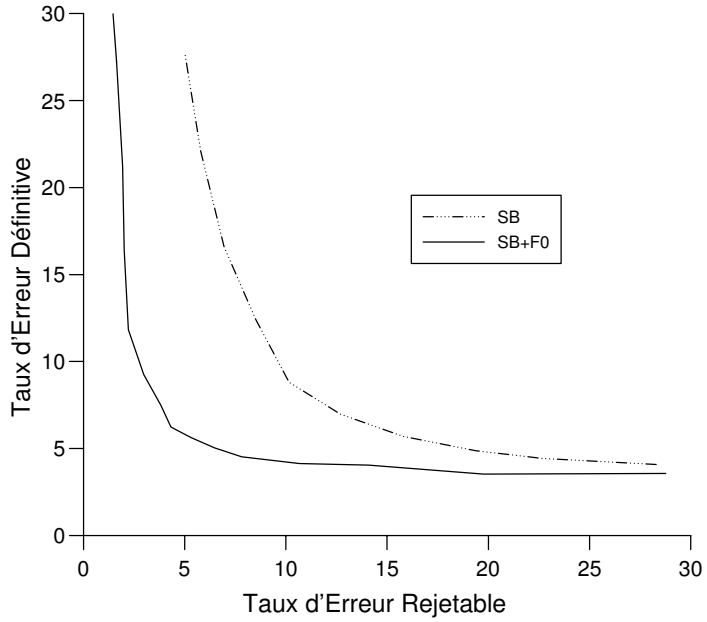
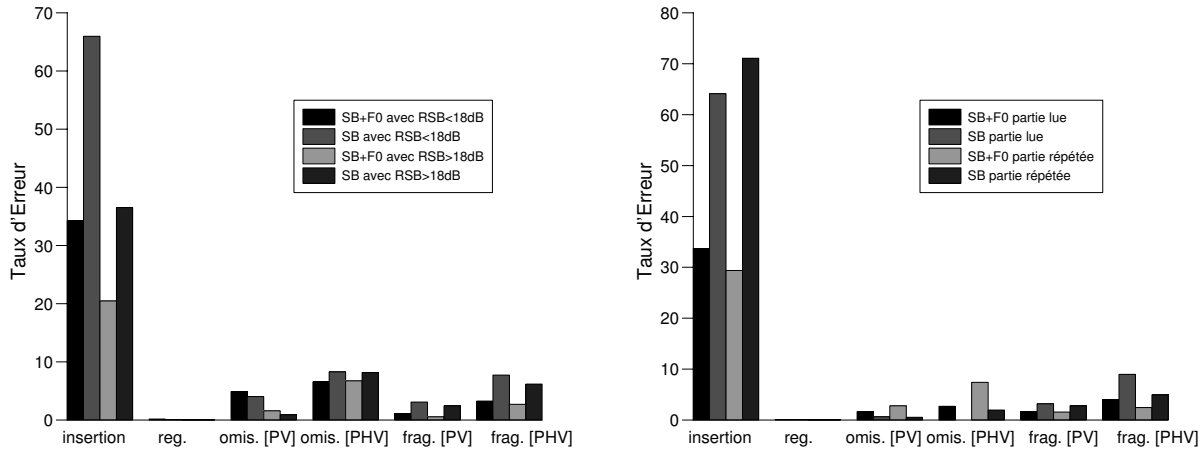
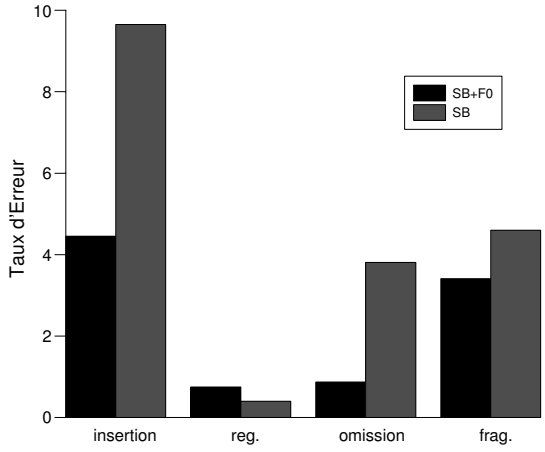


FIG. 8.4 – *Résultats de détection des critères SB+F_0 et SB sur la base AGORA.*



(a) base RTC_T.

(b) base GSM_T.



(c) base AGORA.

FIG. 8.5 – *Erreurs de détection détaillées des critères SB+F_0 et SB sur les bases RTC_T et GSM_T et AGORA.*

La différence principale qui se dégage sur toutes les bases est moins d'insertions (c'est-à-dire moins de détections de bruits), ce qui est recherché par l'apport de la fréquence fondamentale, et son implémentation. La réduction des erreurs d'insertion est de moitié, malgré une légère augmentation des omissions sur la base RTC_T_L et GSM_T.

Nous observons également une baisse des erreurs de fragmentation qui n'est pas significative et qui est due au choix du seuil (plus faible que celui du critère SB).

Les erreurs de regroupement restent très faibles, sur la base AGORA ces erreurs sont proportionnellement plus importantes. Sur cette base nous observons une augmentation de ces erreurs avec le critère SB+F_0, qui est également due au choix du seuil, plus faible que celui du critère SB.

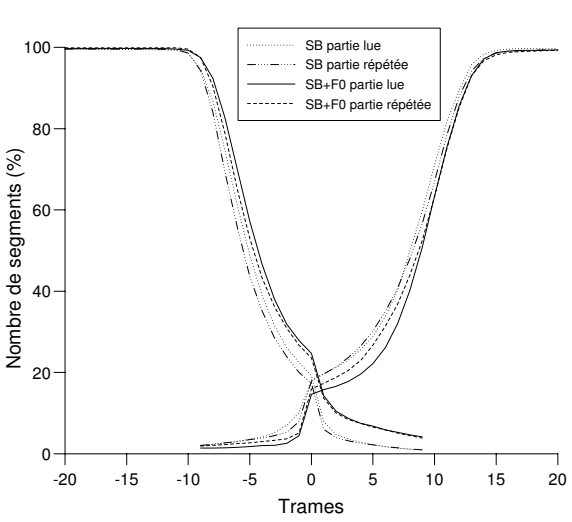
Les histogrammes 8.6 permettent d'étudier le positionnement des frontières sur les détections correctement reliées. Nous remarquons que pour les seuils étudiés (présentant le minimum des taux d'erreur associée), les segments droits élargis et tronqués sont moins importants pour le critère SB+F_0 que pour le critère SB, tandis que les segments gauches restent à peu près identique au critère SB, excepté pour la base RTC_T, où les segments élargis et tronqués sont un peu plus important pour le critère SB+F_0.

8.5.2 Résultats de reconnaissance

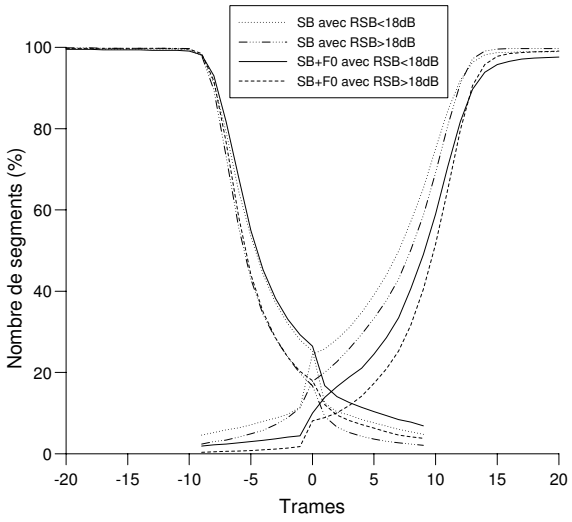
Les seuils optimaux pour la reconnaissance sur les bases RTC_T, GSM_T et AGORA sont donnés dans le tableau F.3 en Annexe F. Nous comparons le critère SB+F_0 au critère SB sur la base RTC_T (*cf.* figure 8.7), sur la base GSM_T (*cf.* figure 8.8), et sur la base AGORA (*cf.* figure 8.9).

Nous remarquons que l'amélioration reste faible sur la base GSM_T en comparaison des résultats sur les erreurs de la détection, et une dégradation sur la base RTC_T, alors que l'amélioration des résultats sur les erreurs de la détection est significative. En effet, le seuil de reconnaissance "optimal" est un seuil qui provoque plus d'erreurs rejetables que le seuil de détection "optimal" choisi pour la comparaison des résultats. Toutefois les erreurs rejetables sur cette base restent faibles. L'amélioration des résultats de détection sur la base GSM_T pouvait laisser à penser que l'amélioration de résultats de reconnaissance serait significative. Or cette faible amélioration n'est pas significative. Ce résultat n'est cependant pas inintéressant. En effet, si les résultats ne sont que faiblement améliorés, c'est parce que le modèle de rejet du module de reconnaissance permet d'éliminer un grand nombre de détections de bruit, ce qui n'est pas toujours le cas dans la pratique. Or du point de vue du coût, qui n'est pas évalué ici, il est beaucoup plus avantageux de réduire les détections de bruit à l'aide du module de détection, plutôt qu'avec le système de reconnaissance. Nous avons vu de plus que le critère SB+F_0 donne une détection plus précise que le critère SB.

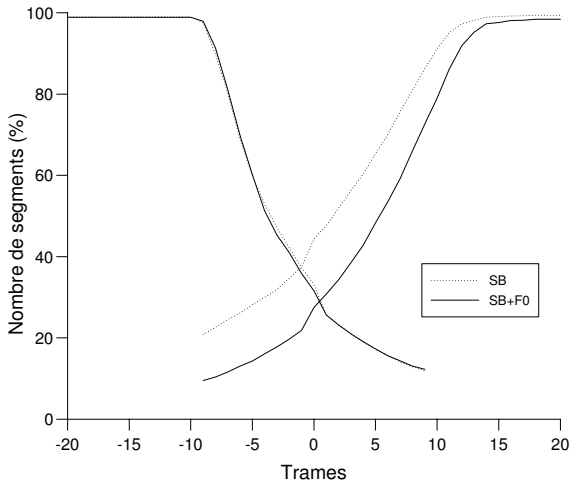
Sur la base AGORA (*cf.* figure 8.9), l'amélioration est plus marquée. En effet, pour la reconnaissance de parole continue, d'une part une détection plus précise est importante (*cf.* paragraphe 3.8), d'autre part le modèle de rejet du système de reconnaissance est moins performant. Ainsi les améliorations du critère SB+F_0 sur les résultats de détection permettent également des améliorations des résultats de reconnaissance. Le taux d'erreur



(a) base RTC_T.

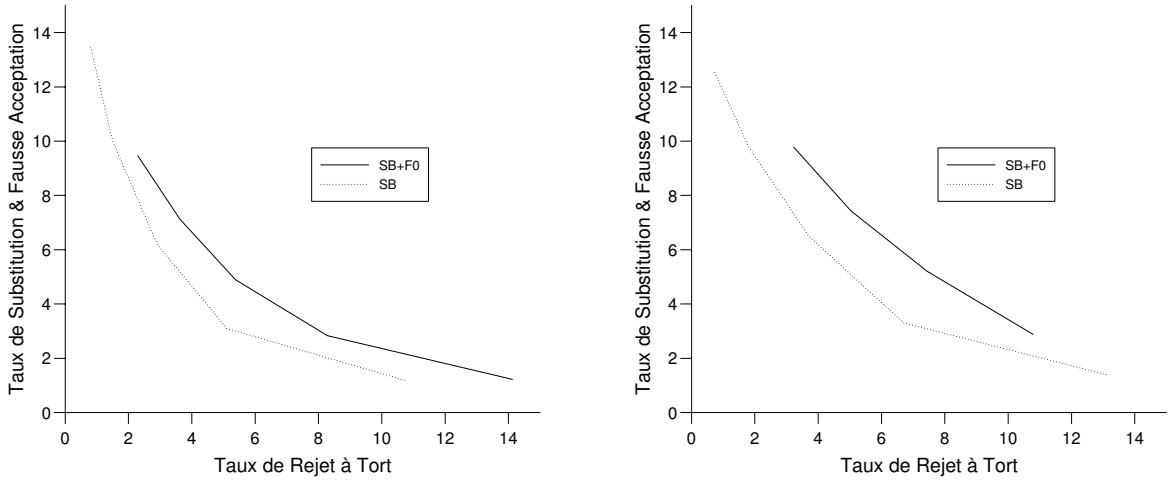


(b) base GSM_T.



(c) base AGORA.

FIG. 8.6 – *Positionnement des frontières des détections des critères $SB+F_0$ et SB sur les bases RTC_T, GSM_T et AGORA.*



(a) partie lue (b) partie répétée

FIG. 8.7 – *Résultats de reconnaissance des critères $SB+F_0$ et SB sur la base RTC_ T.*

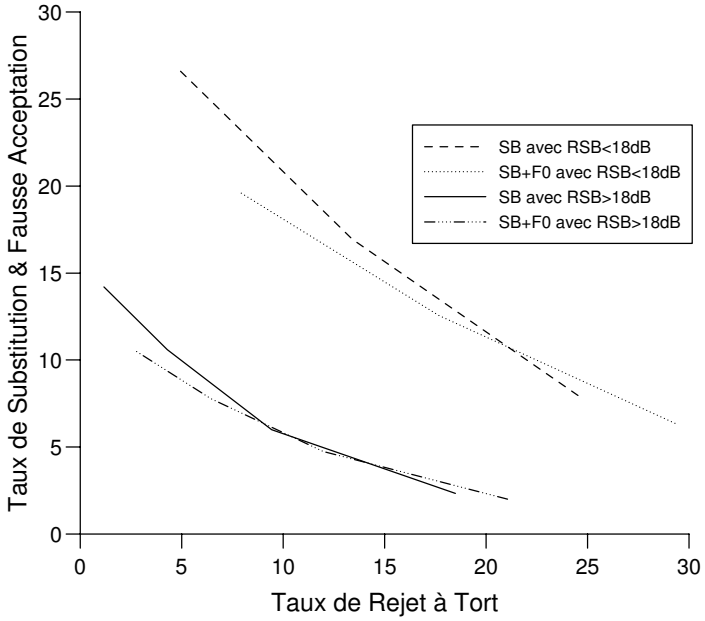
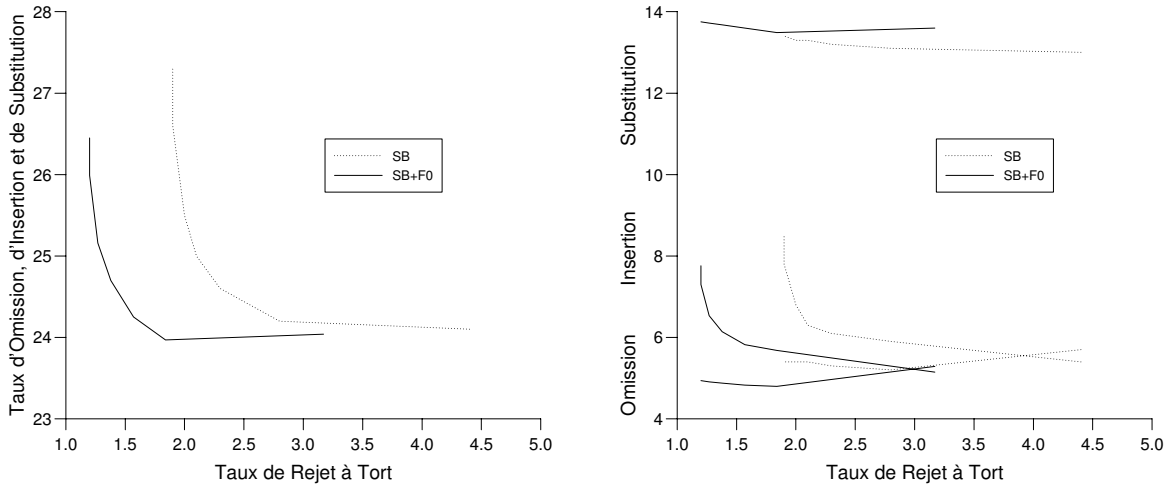


FIG. 8.8 – *Résultats de reconnaissance des critères $SB+F_0$ et SB sur la base GSM_ T.*

obtenu avec un seuil de détection optimal, et un poids de rejet nul pour la base GSM_T SB est de 27.08%, l'intervalle de confiance à 95% est : [26.31; 27.86] (*cf.* tableaux G.5 et G.6). Pour le critère SB+F_0 le taux d'erreur pour le seuil optimal avec un poids de rejet nul est 26.08%. Ainsi l'amélioration est significative au sens de l'intervalle de confiance. De plus à la baisse du taux d'erreur de reconnaissance s'ajoute une baisse des détections de bruit, comme sur les bases RTC_T et GSM_T.



(a) Les taux sont regroupés

(b) Les taux sont séparés

FIG. 8.9 – *Résultats de reconnaissance des critères SB+F_0 et SB sur la base AGORA.*

8.6 Conclusion

Ce chapitre a montré que l'utilisation de la fréquence fondamentale en complément de l'énergie améliore de façon significative les résultats de détection. Le critère de variation de la fréquence fondamentale pour le passage de l'état *présomption de parole* à l'état *parole* permet d'une part de détecter moins de bruit, et d'autre part d'obtenir une détection plus précise. Ce critère permet donc d'améliorer la discrimination entre le signal de parole et le signal de bruit.

Cette amélioration significative au niveau des résultats de détection n'entraîne pas une amélioration du même ordre au niveau des résultats de reconnaissance sur la base GSM_T (le modèle de rejet rejetant correctement les détections de bruits), et même une dégradation sur la base RTC_T. Cependant, la baisse significative des détections de bruit, nous permet d'obtenir une baisse importante de la charge, donc du coût général du système de reconnaissance.

L'amélioration des résultats de reconnaissance sur la base AGORA est significative au sens de l'intervalle de confiance à 95%. Ce module de détection, qui améliore la reconnaissance de parole continue, peut ainsi permettre de meilleurs résultats du système de dialogue associé.

L'information sur le voisement du signal permet donc une meilleure détection de la parole, et ainsi améliore en général les performances du système de reconnaissance. Ainsi les deux premiers objectifs sont atteints : les erreurs du module de détection pour les communications bruitées et pour la détection de parole continue ont diminués. De plus le module de débruitage apporte des améliorations dans le cas des bruits stationnaires *car* et *babble*. Ces résultats sont présentés en Annexe J.

Le troisième objectif est également atteint (*cf.* Annexe H) : la sensibilité du seuil de détection a diminué, en particulier la sensibilité au niveau de bruit et au réseau d'appel.

L'inconvénient du paramètre de voisement vient de l'estimation de la fréquence fondamentale qui est coûteuse en temps de calcul et augmente ainsi le coût du système de reconnaissance.

De nombreuses autres caractéristiques comme les coefficients cepstraux qui sont déjà calculés pour le module de reconnaissance peuvent être associées à l'énergie. Le problème qui se pose dès que le nombre de caractéristiques devient important, est la combinaison de celles-ci. Pour cette combinaison, nous proposons d'utiliser des technique de fusion de données. Plusieurs méthodes sont envisageables, dans le chapitre suivant, nous présentons quelques unes de ces méthodes.

Chapitre 9

Utilisation de la fusion de données

9.1 Introduction

Le signal de parole est décrit par de nombreuses caractéristiques, qui discriminent plus ou moins le bruit de la parole. Actuellement le module de détection ABP utilisé à France Télécom R&D, n'utilise qu'une de ces caractéristiques, l'énergie du signal (*cf.* Chapitre 2 "*Détection de parole pour la reconnaissance vocale*"). Le paragraphe 4.5.1 montre que d'autres caractéristiques peuvent être employées pour discriminer le bruit et la parole. Le problème réside alors dans l'intégration de ces données dans le module de détection ABP. Deux approches se présentent pour fusionner ces données :

- Soit nous cherchons à combiner directement toutes les caractéristiques dans un module de détection. Nous parlons alors de fusion en entrée. Il apparaît nécessaire d'utiliser des méthodes de discrimination, pour trouver l'espace qui discrimine le plus le bruit de la parole. Il faut ensuite affecter les nouvelles trames dans l'une ou dans l'autre des classes. Un troisième problème, la sensibilité des seuils du module de détection se pose si nous cherchons à ce que le module de détection soit performant lors des changements de conditions d'appel. Il est alors nécessaire de mettre à jour l'espace qui discrimine au mieux le bruit de la parole, ainsi que les conditions d'affectation durant l'appel. Différentes méthodes qui permettent de résoudre ces trois problèmes sont présentées dans le paragraphe 9.2.
- Soit nous considérons qu'il y a autant de modules de détection que de caractéristiques. Différents formalismes pour fusionner ces décisions sont alors envisageables. Dans ce cas si nous cherchons à adapter le module final de détection, avec les conditions d'environnement d'appel, il faut soit que chaque module de détection s'adapte à l'environnement, soit que la méthode de fusion puisse adapter les réponses de chaque module. Les méthodes de fusion de décision sont généralement employées lorsque plusieurs modules de détection de nature différente sont présents. Fusionner des modules de détection de fonctionnement identique employant chacun des caractéristiques différentes revient à une fusion en entrée de ces caractéristiques. La fusion de décision n'est donc pas abordée dans cette étude.

Ce chapitre comprend une étude détaillée des différentes possibilités de fusion en entrée adaptées à une détection de parole (*cf.* paragraphe 9.2). De toutes les méthodes présentées l'analyse factorielle discriminante s'est avérée la plus adaptée à notre problème. Nous intégrons donc cette analyse au paragraphe 9.3, avec différentes caractéristiques.

9.2 Fusion en entrée

Nous présentons ici différentes méthodes en vue de la fusion en entrée des caractéristiques du signal. Ces méthodes essaient de répondre aux trois problèmes précédemment cités : trouver l'espace qui discrimine au mieux le bruit et la parole, ou une combinaison optimale des caractéristiques pour cette discrimination ; affecter chaque nouvelle trame comme étant une trame de bruit ou de parole ; et adapter le module de détection à l'environnement.

Les méthodes factorielles sont des méthodes qui permettent de trouver des combinaisons des caractéristiques qui discriminent au mieux le bruit et la parole dans notre cas. Nous pouvons ainsi réduire l'espace des caractéristiques. Certaines méthodes permettant de déterminer si les nouvelles trames sont du bruit ou de la parole, sont appliquées aux caractéristiques. Les méthodes de segmentation non-paramétrique sont des méthodes simples permettant l'affectation de nouvelles trames en ayant réduit le nombre des caractéristiques de manière conditionnelle. Ces méthodes, dans leur forme actuelle, ne permettent cependant pas de résoudre l'adaptation à l'environnement. Les méthodes de classification et de réseaux de neurones peuvent être une alternative à ce problème.

9.2.1 Méthodes factorielles

A. Analyses factorielles

Il y a principalement trois méthodes d'analyse factorielle, l'analyse en composantes principales, l'analyse des correspondances et l'analyse des correspondances multiples. Ces méthodes réduisent les données pour des représentations simplifiées des valeurs numériques, que sont les caractéristiques du signal. Cependant l'analyse des correspondances et l'analyse des correspondances multiples, qui en est une extension, ne sont pas adaptées à notre problème, puisqu'elles s'appliquent à des tableaux de contingences. Nous ne présentons donc que l'analyse en composantes principales.

A.1. **Analyse en composantes principales**

L'analyse en composantes principales (ACP) est très utilisée dans le domaine de la parole pour réduire le grand nombre des caractéristiques du signal. Dans [Batlle *et al.*, 1998] l'ACP est utilisée dans le cadre de la reconnaissance de la parole en vue de décorréler les caractéristiques du signal. Dans [Wark et Sridharan, 1998] l'ACP est abordée pour l'extraction des caractéristiques de l'image des lèvres qu'il utilise pour la reconnaissance du locuteur. Pour la détection Bruit/Parole, l'ACP permet de regrouper les individus (ici les trames) par une distance dans leur représentation

selon les premiers axes principaux. Il existe plusieurs variantes selon les transformations effectuées sur les données initiales. Parmi ces variantes l'ACP normée est la plus utilisée.

Principe de la méthode :

Nous disposons d'un tableau de données $\mathbf{X}$ de n lignes qui correspondent aux trames et p colonnes qui représentent les p caractéristiques d'une trame. Nous définissons une distance entre deux trames i et i' :

$$d^2(i,i') = \sum_{j=1}^{p} \left(\frac{x_{ij} - x_{i'j}}{s_j \sqrt{n}} \right)^2 , \tag{9.1}$$

où s_j est l'écart-type de la caractéristique j :

$$s_j^2 = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 , \tag{9.2}$$

avec

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}. \tag{9.3}$$

Nous obtenons ainsi un tableau de données $\mathbf{Y}$ normées, des variables centrées réduites, où

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j \sqrt{n}}. \tag{9.4}$$

Nous cherchons à présent à ajuster le nuage des trames dans l'espace des caractéristiques. C'est-à-dire qu'il faut trouver le sous-espace vectoriel pour lequel la projection des trames sur cet espace soit de distance minimale (au sens des moindres carrés). Nous trouvons ainsi le sous espace vectoriel le plus représentatif des caractéristiques. Ceci revient à diagonaliser le produit $\mathbf{C} = \mathbf{Y}^*\mathbf{Y}$. En effet nous avons :

$$c_{jj'} = \sum_{i=1}^{n} y_{ij} y_{ij'} = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})}{s_j s_{j'}}. \tag{9.5}$$

$\mathbf{C}$ n'est autre que la matrice des corrélations empiriques entre les caractéristiques (grâce à l'introduction du $\sqrt{n}$ dans l'équation (9.4)). Ainsi le sous-espace de dimension p' qui représente au mieux (au sens des moindres carrés) le nuage des trames dans $\mathbb{R}^p$ est engendré par les p' premiers vecteurs propres de la matrice $\mathbf{C}$ correspondant aux p' plus grandes valeurs propres. Un moyen d'obtenir ces valeurs propres est de diagonaliser la matrice de corrélations.

Notons que dans [Charlet, 1997] il est montré qu'il existe une certaine similarité entre les coefficients cepstraux (MFCC) et les coefficients obtenus par l'ACP sur les coefficients de la sortie du banc de filtres.

A.2. Analyse factorielle discriminante

L'analyse factorielle discriminante, ou analyse linéaire discriminante, est une méthode qui est à la fois descriptive et prédictive. Elle peut être présentée comme une extension de la régression multiple (*cf.* [Lebart *et al.*, 1995]).

principe de la méthode :

Nous considérons toujours n trames décrites par p caractéristiques. Soit les données $\{x_{i,j}\}_{i=1,..,n\,;\,j=1,..p}$, à répartir en q classes $\{I_k\}_{k=1,..,q}$ de $\{n_k\}_{k=1,..,q}$ trames. Cette analyse consiste, dans un premier temps, à chercher les combinaisons linéaires des caractéristiques dont la variance entre les classes (inter-classes) est maximale, et la variance dans chaque classe (intra-classes) est minimale. Nous cherchons ainsi les combinaisons qui séparent au mieux les q classes. Notons $\mathbf{T}$ la matrice des covariances totales :

$$t_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}) = \frac{1}{n}\sum_{k=1}^{q}\left[\sum_{i\in I_k}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})\right], \tag{9.6}$$

où $\bar{x}_j = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_{ij}$. Décomposons la matrice $t_{jj'} = cov(x_j,x_{j'})$ en somme de covariances intra-classes et covariances inter-classes. Nous avons :

$$(x_{ij} - \bar{x}_j) = (x_{ij} - \bar{x}_{kj}) + (\bar{x}_{kj} - \bar{x}_j), \tag{9.7}$$

où $\bar{x}_{kj} = \dfrac{1}{n_k}\displaystyle\sum_{i\in I_k} x_{ij}$. Nous avons ainsi :

$$\sum_{i\in I_k}(x_{ij} - \bar{x}_{kj})(x_{kj'} - \bar{x}_{j'}) = \sum_{i\in I_k}(x_{kj} - \bar{x}_j)(x_{ij'} - \bar{x}_{kj'}) = 0, \tag{9.8}$$

en appliquant la formule de décomposition de Huyghens, nous obtenons :

$$t_{jj'} = d_{jj'} + e_{jj'}, \tag{9.9}$$

avec la matrice $\mathbf{D}$, matrice des covariances intra-classes, définie par :

$$d_{jj'} = \frac{1}{n}\sum_{k=1}^{q}\sum_{i\in I_k}(x_{ij} - \bar{x}_{kj})(x_{ij'} - \bar{x}_{kj'}), \tag{9.10}$$

et $\mathbf{E}$, la matrice des covariances inter-classes :

$$e_{jj'} = \frac{1}{n}\sum_{k=1}^{q} n_k(\bar{x}_{kj} - \bar{x}_j)(\bar{x}_{kj'} - \bar{x}_{j'}). \tag{9.11}$$

La variance d'une combinaison linéaire $\mathbf{a}$ se décompose donc en variance interne et externe :

$$\mathbf{a}^*\mathbf{T}\mathbf{a} = \mathbf{a}^*\mathbf{D}\mathbf{a} + \mathbf{a}^*\mathbf{E}\mathbf{a}. \tag{9.12}$$

Nous cherchons donc les combinaisons linéaires $\mathbf{a}$ telles que $\mathbf{a}^*\mathbf{Da}$ soit minimale et $\mathbf{a}^*\mathbf{Ea}$ soit maximale. C'est-à-dire telle que :

$$f(a) = \frac{\mathbf{a}^*\mathbf{Ea}}{\mathbf{a}^*\mathbf{Ta}}, \tag{9.13}$$

soit maximale. Ceci revient à résoudre (*cf.* [Celeux *et al.*, 1989]) :

$$\mathbf{T}^{-1}\mathbf{Ea} = \lambda\mathbf{a}, \tag{9.14}$$

sous la condition $\mathbf{a}^*\mathbf{Ta} = 1$. Dans notre cas le problème de l'inversion de $\mathbf{T}$ ne se pose pas puisque nous pouvons la supposer définie positive. $\mathbf{a}$ est donc le vecteur propre de $\mathbf{T}^{-1}\mathbf{E}$ associé à la plus grande valeur propre λ. Il faut donc diagonaliser $\mathbf{T}^{-1}\mathbf{E}$ qui n'est pas *a priori* symétrique. Posons :

$$\mathbf{E} = \mathbf{CC}^*, \tag{9.15}$$

avec :

$$c_{jk} = \sqrt{\frac{n_k}{n}}(\bar{x}_{kj} - \bar{x}_j). \tag{9.16}$$

Et posons :

$$\mathbf{a} = \mathbf{T}^{-1}\mathbf{Cv}. \tag{9.17}$$

L'équation $\mathbf{Ea} = \lambda\mathbf{Ta}$ s'écrit alors :

$$\mathbf{CC}^*\mathbf{T}^{-1}\mathbf{Cv} = \lambda\mathbf{Cv}. \tag{9.18}$$

Il suffit donc de diagonaliser la matrice symétrique $\mathbf{C}^*\mathbf{T}^{-1}\mathbf{C}$ d'ordre q, puis de déduire $\mathbf{a}$ à l'aide de $\mathbf{v}$.

Cas où $q = 2$: Dans le cas où il n'y a que deux classes, des simplifications apparaissent. En fait, l'analyse factorielle discriminante est alors équivalente à la régression multiple (*cf.* [Lebart *et al.*, 1995]). Notons par les indices 1 et 2, les classes représentant, dans notre cas le bruit et la parole. La matrice des covariances inter-classes $\mathbf{E}$ a pour terme général :

$$e_{jj'} = \frac{n_1}{n}(\bar{x}_{1j} - \bar{x}_j)(\bar{x}_{1j'} - \bar{x}_{j'}) + \frac{n_2}{n}(\bar{x}_{2j} - \bar{x}_j)(\bar{x}_{2j'} - \bar{x}_{j'}), \tag{9.19}$$

avec

$$\bar{x}_j = \frac{n_1}{n}\bar{x}_{1j} + \frac{n_2}{n}\bar{x}_{2j}. \tag{9.20}$$

En remplaçant $\bar{x}_j$ par sa valeur et en tenant compte du fait que $n_1 + n_2 = n$, nous avons :

$$e_{jj'} = \frac{n_1 n_2}{n^2}(\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1j'} - \bar{x}_{2j'}). \tag{9.21}$$

Ainsi la matrice symétrique $\mathbf{E}$ de rang 1, peut être considérée comme le produit d'une matrice colonne $\mathbf{c}$ par sa transposée :

$$\mathbf{E} = \mathbf{c}^*\mathbf{c}, \tag{9.22}$$

avec :

$$c_j = \frac{\sqrt{n_1 n_2}}{n}(\bar{x}_{1j} - \bar{x}_{2j}). \tag{9.23}$$

Ainsi la relation (9.14) s'écrit :

$$\mathbf{T}^{-1}\mathbf{c}^*\mathbf{c}\mathbf{a} = \lambda\mathbf{a}. \tag{9.24}$$

Nous avons donc $\lambda = \mathbf{c}^*\mathbf{T}^{-1}\mathbf{c}$ qui est l'unique valeur propre, car $\mathbf{E}$ est de rang 1. Le vecteur propre correspondant est donc $\mathbf{a} = \mathbf{T}^{-1}\mathbf{c}$, qui est l'unique fonction discriminante.
À notre connaissance cette approche n'a jamais été employée pour une détection de parole. Cependant, elle est principalement fondée sur la matrice de covariance, qui a été souvent utilisée pour définir des distances dans des systèmes de détection (*cf.* paragraphe 4.5). L'analyse factorielle discriminante permet de considérer cette matrice de façon plus globale.

B. Remarque comparative

L'ACP et l'analyse factorielle discriminante sont deux méthodes très proches. Cependant certains auteurs ont montré qu'elles ne sont pas équivalentes, plus ou moins bien adaptées à des problèmes, elles peuvent être complémentaires.

Les performances de l'analyse factorielle discriminante, avec l'ACP et l'utilisation de méthodes de filtres sur les fréquences ainsi que de transformées en cosinus sont comparées dans [Batlle *et al.*, 1998]. Il en ressort que d'une part les coefficients issus de l'ACP et de la transformé en cosinus (*i.e.* les MFCC) sont équivalents (les MFCC étant une approximation des coefficients de l'ACP, *cf.* [Charlet, 1997]), d'autre part l'analyse factorielle discriminante est la plus performante pour la décorrélation des caractéristiques des coefficients en sortie du banc de filtres, en vue de la reconnaissance de phonèmes. Dans [Wark et Sridharan, 1998] l'ACP est utilisée pour réduire le nombre de caractéristiques de l'image des lèvres et du signal de la parole, puis l'analyse factorielle discriminante est employée pour discriminer ces caractéristiques. L'utilisation des deux méthodes s'avère plus efficace pour la reconnaissance du locuteur que l'ACP seule.

C. Méthodes d'affectation

Plusieurs méthodes d'affectation d'une nouvelle trame se présentent pour les analyses factorielles :

- La première consiste à utiliser la combinaison linéaire la plus discriminante trouvée par l'analyse. Il suffit de calculer la projection de la trame sur la droite définie par

cette combinaison, qui est unique dans le cas de deux classes. Cette projection est comparée à un seuil qui détermine l'appartenance à l'une des deux classes.

- Une autre approche, permettant de traiter plus facilement le cas de plusieurs classes, consiste à affecter une nouvelle trame $\mathbf{x}$ dans une des classes du nouvel espace, en considérant les centres de gravité et les matrices de covariance inter-classes ou/et intra-classes. Dans ce cas plusieurs distances sont envisageables. Nous décrivons ci-dessous quelques unes de ces distances au centre de gravité.
- L'approche bayesienne permet de prendre en compte les probabilités *a priori* de chaque classe.
- Une autre règle d'affectation est la règle des plus proches voisins.

Les méthodes d'affectation présentées ici ne sont pas exclusivement employées pour les analyses factorielles.

C.1. Distance du centre de gravité Une règle simple est d'attribuer la trame $\mathbf{x}$ à la classe dont le centre de gravité $G_k = (G_{kj})_j$ où $G_{kj} = \bar{x}_{kj} = \dfrac{1}{n_k} \sum_{i \in I_k} x_{ij}$, est le plus proche de $\mathbf{x}$. Ainsi $\mathbf{G}_k$ est le point centre de gravité de la classe I_k dans l'espace de dimension p. Il faut alors définir une distance.

- Distance euclidienne
  La distance euclidienne usuelle dans $\mathbb{R}^p$:

$$d_e^2(\mathbf{x},\mathbf{G}_k) = \sum_{j=1}^{p} (x_j - G_{kj})^2. \tag{9.25}$$

  Exprimons cette distance dans le nouvel espace. Notons :

$$z_r = \mathbf{u}_r^*(\mathbf{x} - \bar{\mathbf{x}}), \tag{9.26}$$

  où $\bar{\mathbf{x}}$ est le vecteur $\{\bar{x}_j\}_{j=1,..,p}$, r désigne l'axe principal issu de l'analyse, et $\mathbf{u}_r$ est le r$^{\text{ième}}$ vecteur propre normalisé de $\mathbf{T}$ matrice des covariances totales, définie précédemment, correspondant à la valeur propre λ_r. La distance euclidienne s'écrit alors :

$$d_{eT}^2(\mathbf{x},\mathbf{G}_k) = \sum_{r=1}^{r_{max}} (z_r - \bar{z}_{kr})^2, \tag{9.27}$$

  où $\bar{z}_{kr} = \mathbf{u}^*(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})$, r_{max} est le nombre de valeurs propres retenues, qui peut être ici le rang de la matrice $\mathbf{X}$.
  La distance de toute trame $\mathbf{x}$ au centre de gravité $\mathbf{G}_k$ de la classe I_k dans la métrique $\mathbf{T}^{-1}$ (*i.e.* sous la condition : $\mathbf{u}^*\mathbf{T}\mathbf{u} = 1$) est :

$$d_{eT^{-1}}^2(\mathbf{x},\mathbf{G}_k) = \sum_{r=1}^{r_{max}} \frac{(z_r - \bar{z}_{kr})^2}{\lambda_r}. \tag{9.28}$$

La distance euclidienne est une distance employée dans [Haigh et Mason, 1993] sur les coefficients cepstraux, comme mesure de similarité pour une DAV. Cette distance ne semble cependant pas adaptée à ces coefficients.

– Distance de Mahalanobis globale
Si nous remplaçons les données $\mathbf{X}$ par $\hat{\mathbf{X}}$ de terme général $\hat{x}_{ij} = x_{ij} - \bar{x}_{kj}$, nous diagonalisons alors la matrice $\mathbf{D}$ au lieu de $\mathbf{T}$. Notons $\hat{\lambda}_r$ les valeurs propres de $\mathbf{D}$ et $\hat{z}_k$ les coordonnées de la trame $\mathbf{x}$ sur les nouveaux axes principaux $\hat{\mathbf{u}}_r$. La distance de $\mathbf{x}$ au centre de gravité $\mathbf{G}_k$ dans la métrique $\mathbf{D}^{-1}$ s'écrit :

$$d^2_{Mg}(\mathbf{x},\mathbf{G}_k) = \sum_{r=1}^{r_{max}} \frac{(\hat{z}_r - \bar{\hat{z}}_{kr})^2}{\hat{\lambda}_r}. \tag{9.29}$$

C'est cette distance qui est décrite dans [Hörmann et Rozinaj, 1998] à partir de deux caractéristiques calculées sur les coefficients cepstraux, pour développer un module de détection Bruit/Parole pour la reconnaissance de mots isolés.

– Distance de Mahalanobis locale
La distance de Mahalanobis locale est la distance de $\mathbf{x}$ au centre de gravité $\mathbf{G}_k$ dans la métrique $\mathbf{D}_k^{-1}$, où $\mathbf{D}_k$ est la matrice de covariance interne de la classe I_k. Notons $w_{sk} = \mathbf{v}^*_{sk}(\mathbf{x} - \bar{\mathbf{x}}_k)$, où $\bar{\mathbf{x}}_k$ est le vecteur $\{\bar{x}_{kj}\}_{j=1,..,p}$, et w_{sk} est le s$^{\text{ième}}$ vecteur propre normalisé de $\mathbf{U}^*\mathbf{D}_k\mathbf{U}$ qui correspond à la valeur propre β_{sk}. La distance s'écrit alors :

$$d^2_{Ml}(\mathbf{x},\mathbf{G}_k) = \sum_{s=1}^{s_{max}(k)} \frac{(w_{sk} - \bar{w}_{ks})^2}{\beta_{sk}}, \tag{9.30}$$

où $\bar{w}_{ks} = \mathbf{v}^*_{ks}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})$, et $s_{max}(k)$ est le nombre de valeurs propres retenues dans la classe I_k.

– Distance du χ^2
La distance du χ^2 est déterminée par :

$$d^2_{\chi^2}(\mathbf{x},\mathbf{G}_k) = \sum_{j=1}^{p} \frac{1}{s_{x_j}} \left(\frac{x_j}{s_x} - \frac{\bar{x}_{kj}}{s_{\bar{x}_k}} \right)^2, \tag{9.31}$$

où $s_{x_j} = \sum_{i=1}^{n} x_{ij}$, $s_x = \sum_{j=1}^{p} x_j$ et $s_{\bar{x}_k} = \sum_{j=1}^{p} \bar{x}_{kj}$. Dans le nouvel espace, nous avons :

$$d^2_{\chi^2}(\mathbf{x},\mathbf{G}_k) = \sum_{r=1}^{r_{max}} \frac{1}{s_{z_r}} \left(\frac{z_r}{s_z} - \frac{\bar{z}_{kr}}{s_{\bar{z}_k}} \right)^2. \tag{9.32}$$

Cependant cette distance s'applique habituellement aux tableaux de contingence, elle convient donc peu à des caractéristiques qui peuvent être binaires.

– Distance de Minkowsky
Elle dépend d'un paramètre λ positif :

$$d_M(\mathbf{x},\mathbf{G}_k) = \left(\sum_{j=1}^{p} |x_j - \bar{x}_{kj}|^{\lambda} \right)^{\frac{1}{\lambda}}. \tag{9.33}$$

Dans le nouvel espace, nous avons :

$$d_M(\mathbf{x},\mathbf{G}_k) = \left(\sum_{r=1}^{r_{max}} |z_r - \bar{z}_{kr}|^{\lambda} \right)^{\frac{1}{\lambda}}. \tag{9.34}$$

Si $\lambda = 1$, nous avons la distance des valeurs absolues, $\lambda = 2$, nous retrouvons la distance euclidienne. Lorsque $\lambda \to +\infty$, nous obtenons la distance de Tchebychev :

$$d_T(\mathbf{x},\mathbf{G}_k) = \max_r |z_r - \bar{z}_{kr}|. \tag{9.35}$$

D'autres distances sont envisageables. Ce type d'affectation ne prend cependant pas en compte les probabilités *a priori* de chaque classe.

C.2. Approche bayesienne L'approche bayesienne permet de remédier à ce manquement. Elle consiste à affecter $\mathbf{x}$ à la classe I_k pour laquelle $P(I_k/\mathbf{x})$ est maximale, or d'après la relation de Bayes :

$$P(I_k/\mathbf{x}) = \frac{P(\mathbf{x}/I_k)P(I_k)}{\sum_{k'=1}^{q} P(\mathbf{x}/I_{k'})P(I_{k'})}. \tag{9.36}$$

Il faut donc maximiser $P(\mathbf{x}/I_k)P(I_k)$. Si les classes sont équiprobables, alors $\underset{k}{\operatorname{argmax}}\, P(I_k/\mathbf{x}) = \underset{k}{\operatorname{argmax}}\, P(\mathbf{x}/I_k)$. L'hypothèse d'équiprobabilité est souvent faite, en effet les connaissances heuristiques confirment la légitimité de cette hypothèse.

Une variante de l'approche bayesienne en supposant que $\mathbf{x}$ suit une loi multinormale, est présentée dans [Lebart *et al.*, 1995]. Il est cependant bon de pouvoir éviter de telles hypothèses fortes, dans notre cas. L'ajout de caractéristiques binaires (par exemple l'existence ou non de voisement) interdit l'hypothèse de gaussianité.

Cependant cette approche est employée dans [Karray et Monné, 1998], [Arslan et Hansen, 1998] et [Singh *et al.*, 2001] (*cf.* paragraphe 4.5.2). L'hypothèse de distributions gaussiennes n'est faite que sur l'énergie du bruit et de la parole. Dans ce cas, elle est acceptable et souvent faite, même si l'estimation des statistiques de la parole pose des problèmes.

C.3. Plus proches voisins Une autre règle d'affectation, souvent utilisée en reconnaissance des formes, est la règle des plus proches voisins. Nous affectons $\mathbf{x}$ dans la classe majoritaire des m plus proches voisins. Nous sommes alors obligés d'introduire de nouveau une notion de distance. Les distances précédemment citées sont envisageables. La règle de décision pour attribuer la classe majoritaire peut être une des méthodes de fusion de décisions, dans un cas très simple, comme la méthode du vote, présentée dans [Xu *et al.*, 1992].

9.2.2 Segmentation non-paramétrique

Les méthodes de segmentation sont des méthodes de discrimination, qui construisent des arbres de décision binaire. Ces méthodes sont un cas particulier de l'application de la méthode CART (*Classification And Regression Trees*) introduite dans [Breiman *et al.*, 1993]. Elles répondent au problème de l'affectation d'une nouvelle trame, sans avoir réellement cherché au préalable un espace où les caractéristiques des trames des données d'apprentissage sont les plus discriminantes. Ces méthodes sont robustes vis-à-vis des valeurs erronées de ces données, et peu contraintes par leur nature.

principe de la méthode :

Nous considérons toujours n trames représentées par p caractéristiques, que nous cherchons à discriminer en q classes. Nous construisons un arbre de décision binaire à partir des données d'apprentissage, de manière à minimiser la variance intra-classes. Pour chaque caractéristique nous cherchons le seuil qui divise les données en deux ensembles (ou segments) de trames ayant une variance intra-classes minimale. Nous choisirons la caractéristique qui fournit les plus faibles variances intra-classes. Partant de chacun de ces ensembles, nous itérons le procédé avec les caractéristiques restantes, jusqu'à obtenir un nombre d'ensembles prédéfini. Le nombre final d'ensembles ne doit pas être trop important vis-à-vis du nombre de classes, qui donnerait une estimation trop optimiste de l'erreur théorique. À chaque segment terminal, nous associons la classe qui est la mieux représentée.

Pour affecter une nouvelle trame $\mathbf{x}$, nous testons les valeurs de ces caractéristiques en redescendant l'arbre. Ainsi $\mathbf{x}$ appartient à la classe représentée par l'ensemble final où il se trouve.

Illustrons cette méthode par un exemple, prenons 100 trames représentées par 10 caractéristiques, et 2 classes, celle du bruit et celle de la parole. Les 100 trames sont réparties équitablement en 50 trames de bruit et 50 trames de parole. Nous trouvons, par exemple, que la deuxième caractéristique est la plus discriminante, au sens de ci-dessus, pour un seuil de 120. Les deux segments ainsi formés comportent respectivement 30 trames, dont 25 de bruit et 5 de parole, et 70 trames, dont 25 de bruit et 45 de parole (*cf.* la figure 9.1). En continuant le procédé, nous obtenons, par exemple, que la première caractéristique est la plus discriminante pour le premier segment (pour un seuil de 35), tandis que ce sera la sixième (avec un seuil de 44) pour l'autre. Nous obtenons ainsi quatre segments terminaux.

En calculant le taux d'erreur de classement de l'arbre, comme la moyenne des erreurs

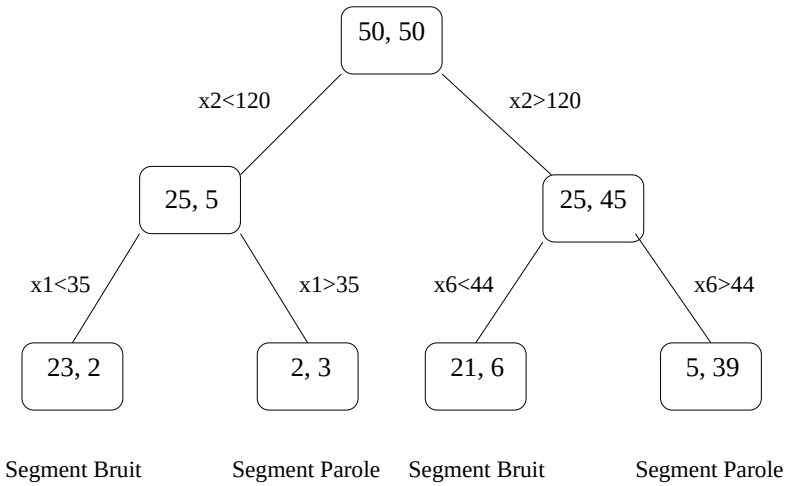


FIG. 9.1 – *Arbre de décision binaire.*

de chaque ensemble final, il est possible d'optimiser l'arbre, et de trouver le nombre final d'ensembles à déterminer. Le choix de l'arbre optimal peut néanmoins être difficile.

Cette méthode, très proche des précédentes, utilise cependant les caractéristiques conditionnellement les unes par rapport aux autres. Ce n'est donc pas une segmentation multidimensionnelle, comme l'analyse factorielle discriminante, qui résout le problème géométriquement.

L'approche de la logique floue par [Mwangi et Xydeas, 1985] et par [Cavallaro *et al.*, 1998] présentée au paragraphe 4.5.2 peut être vue comme une segmentation non-paramétrique. La différence de cette méthode est que les règles de décision se font alors automatiquement, sans connaissance heuristique des caractéristiques. La segmentation non-paramétrique semble donc plus adaptée pour un module de détection utilisant un grand nombre de caractéristiques.

Dans [Shin *et al.*, 2000], l'approche est présentée comme une fusion de décision par la méthode CART (*cf.* paragraphe 4.5.2). Cependant les différents modules de détection sont tous identiques, et n'utilisent qu'une caractéristique comparée à un seuil adapté dans les périodes de non-parole. Cette approche peut donc se ramener à la segmentation non-paramétrique pour une fusion en entrée.

9.2.3 Méthodes de classification

Les méthodes de classification visent à regrouper les trames selon leur "ressemblance". Nous pouvons espérer que les différences des caractéristiques entre les trames de bruit et de parole conduisent de proche en proche, par une des méthodes de classification, à séparer les trames en deux classes, celle de bruit et celle de parole. Ces méthodes permettent de classer des nouvelles trames, sans avoir recours à des techniques de discrimination. Il existe des algorithmes de classification conduisant directement à des partitions comme les méthodes d'agrégation autour de centres mobiles; des algorithmes ascendants qui construisent les

classes par association deux à deux; et des algorithmes descendants qui procèdent par dichotomies.

Algorithme des centres mobiles

La classification par l'algorithme des centres mobiles est particulièrement employée pour des problèmes avec beaucoup de données. Cette méthode est utilisée comme technique de partitionnement ou comme technique de réduction des données. Elle est souvent associée à d'autres méthodes de classification ou d'analyse factorielle.

Principe de la méthode :

Nous disposons toujours d'un tableau de données $\mathbf{X}$ de n lignes qui pour nous correspondent aux trames, et p colonnes qui représentent les p caractéristiques d'une trame. L'espace $\mathbb{R}^p$ est muni d'une distance d (souvent une distance euclidienne). Nous cherchons à établir q classes, pour notre problème $q = 2$.

Tout d'abord nous déterminons q centres provisoires de classes (de façon aléatoire, par exemple). Ces q centres permettent de fournir q classes, un individu appartiendra à une classe s'il est plus proche de son centre que des autres. Différentes distances comme celles présentées précédemment peuvent être employées pour déterminer les distances de l'individu aux centres des classes. Nous déterminons alors les nouveaux centres de gravité des classes. Ces nouveaux centres conduisent à une nouvelle partition des individus, suivant la même règle que précédemment. L'opération est ensuite itérée, jusqu'à ce que deux opérations successives fournissent la même partition. [Lebart *et al.*, 1995] montre que l'algorithme converge bien.

Cette méthode de classification donne des classes sphériques, dans le cas de deux classes c'est un hyperplan.

La classification hiérarchique

Ici aussi, l'algorithme fournit une partition des n trames en q classes. Nous construisons tout d'abord une matrice des distances entre les n individus. Cette distance, qui peut-être simplement une mesure de dissimilarité (*i.e.* l'inégalité triangulaire n'est pas exigée), est à définir préalablement. Une des distances présentées précédemment peut être utilisée. Une fois cette matrice calculée, nous cherchons les deux éléments les plus proches, que nous regroupons en un nouvel élément. Nous obtenons ainsi une première partition de $n-1$ éléments. Nous construisons alors une nouvelle matrice des distances, en calculant les distances du nouvel élément avec les autres. Cette distance doit elle aussi être définie. Si $\mathbf{x}$, $\mathbf{y}$ et $\mathbf{z}$ sont trois éléments, et que nous notons $\mathbf{h}$ le nouvel élément qui regroupe $\mathbf{x}$ et $\mathbf{y}$, nous pouvons définir différentes distances telles que :

$$d_g(\mathbf{h},\mathbf{z}) = min\{d(\mathbf{x},\mathbf{z}),d(\mathbf{y},\mathbf{z})\}, \tag{9.37}$$

$$d_g(\mathbf{h},\mathbf{z}) = max\{d(\mathbf{x},\mathbf{z}),d(\mathbf{y},\mathbf{z})\}, \tag{9.38}$$

$$d_g(\mathbf{h},\mathbf{z}) = \frac{n_x d(\mathbf{x},\mathbf{z}) + n_y d(\mathbf{y},\mathbf{z})}{n_x + n_y}, \tag{9.39}$$

où n_x et n_y sont les nombres d'éléments que contiennent $\mathbf{x}$ et $\mathbf{y}$ respectivement. Une fois cette nouvelle matrice calculée, nous déterminons les deux éléments les plus proches, que nous regroupons en un nouvel élément. Nous obtenons ainsi une partition de $n-2$ éléments. Nous itérons ainsi le procédé jusqu'à l'obtention d'une partition de q classes.

Remarque

De notre point de vue de telles classifications peuvent s'avérer intéressantes pour l'adaptation à l'environnement. Le problème reste la quantité trop importante des données à conserver en mémoire. Le but premier des méthodes de classification étant de trouver une répartition des données d'apprentissage, nous intéresse peu. En effet nos données d'apprentissage sont déjà classées.

9.2.4 Méthodes des réseaux de neurones

Les méthodes neuronales ont été créées avec un souci de modélisation des mécanismes de perception visuelle et auditive. Beaucoup de méthodes neuronales de discrimination ont vu le jour. L'intérêt pratique de telles méthodes tient à leur faculté d'apprentissage des détails discriminants, à leur simplicité de mise en œuvre et à leur rapidité d'exécution. Le problème reste le nombre important de données nécessaires à l'apprentissage. La dépendance de ces méthodes aux données d'apprentissage, entraîne une faible robustesse aux conditions d'observation. Le modèle le plus répandu, pour les méthodes de discrimination est le perceptron multi-couches. Il existe cependant des méthodes neuronales non-supervisés (*cf.* [Lebart *et al.*, 1995]). Ces approches de réseaux de neurones nécessitent un temps d'adaptation important qui correspond au grand nombre des données d'apprentissage pour les modèles supervisés.

Sur la figure 9.2 est représenté un perceptron à une couche cachée. Sur cet exemple nous avons pris des trames de 5 caractéristiques, pour une discrimination en deux classes, le bruit et la parole. Pour chaque trame de la base de donnée d'apprentissage, il faut optimiser les poids w et v. Une fois les poids trouvés, nous affecterons une nouvelle trame, dans une des classes en fonction de la valeur en sortie.

Dans [Héon *et al.*, 1998] un réseau de neurones multi-couches est utilisé, avec en entrée les informations cepstrales, ce qui permet de réduire le bruit. Les nouveaux coefficients permettent d'améliorer un algorithme de reconnaissance vocale.

Un perceptron à une couche cachée, avec 6 nœuds en entrée, 10 dans la couche cachée et 2 en sortie, est défini dans [Bendiksen et Steiglitz, 1990] pour une détection en segments voisés/non-voisés de la parole. Les six coefficients d'entrée sont composés de l'énergie dans différentes bandes de fréquences, et calculée sur un signal débruité. Cette classification, utilisée dans le cadre de l'analyse et synthèse de la parole, est très coûteuse en délai qui n'est pas une contrainte pour les auteurs.

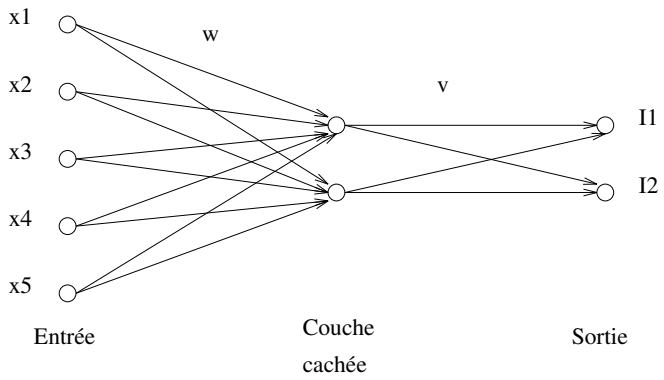


FIG. 9.2 – *Perceptron à une couche cachée.*

Dans [Cohn, 1991] un perceptron avec une couche cachée est comparé à un perceptron sans couche cachée et à une fonction linéaire discriminante. En entrée un jeu de trois ou quatre coefficients parmi cinq coefficients sur l'énergie est choisi, et en sortie il y a deux nœuds pour une classification en segments voisés/non-voisés de la parole. Cette étude montre que dans ces conditions un perceptron avec une couche cachée donne de meilleurs résultats.

Deux couches cachées composent le réseau de neurones proposé dans [Ghiselli-Cripa et El-Jaroudi, 1991]. Cependant la détection en segments voisés/non-voisés/silence obtenue n'est pas meilleure que celle de [Atal et Rabiner, 1976] qui utilise les mêmes caractéristiques : le logarithme de l'énergie du signal, le taux de passage par zéro, la corrélation des trames adjacentes de parole, le premier coefficient de prédiction linéaire et le coefficient d'autocorrélation normalisé. Dans [Atal et Rabiner, 1976] ces caractéristiques sont employées avec une distance fondée sur la matrice de covariance.

9.2.5 Discussion

À notre connaissance une étude comparative des méthodes de fusion en entrée pour des caractéristiques de la parole n'a pas été réalisée. Nous avons vu la simplicité des méthodes d'analyse factorielle discriminante et de segmentation pour notre problème à deux classes. Ces méthodes n'ont pas été employées telles quelles dans des modules de détection, cependant différents auteurs proposent des détections qui s'en approchent. Ces méthodes nécessitent néanmoins une base d'apprentissage. L'adaptation à l'environnement peut se faire, d'une part en relançant l'analyse toutes les n nouvelles trames, si ce n'est pas trop coûteux. D'autre part pour l'analyse factorielle discriminante, elle peut se faire en actualisant les coordonnées des centres de gravité de chaque classe, si la méthode d'affectation consiste à trouver le minimum de la distance entre la nouvelle trame et les deux centres de gravité, ou en actualisant la combinaison linéaire discriminante. La segmentation non-paramétrique et l'analyse factorielle discriminante permettent l'implémentation de caractéristiques numériques ou nominales. La méthode de segmentation peut paraître trop sensible aux perturbations des données, de plus le choix de l'arbre de

décision optimal n'est pas un problème aisé.

L'analyse factorielle discriminante répond donc aux trois problèmes que nous nous sommes posés, à savoir : trouver un espace qui discrimine au mieux le bruit de la parole, affecter une nouvelle trame par une des méthodes proposées et permettre une adaptation aux conditions d'appel. De plus la nature des caractéristiques n'est pas contraignante.

Notre choix pour la fusion de différentes caractéristiques se porte donc sur l'analyse factorielle discriminante.

9.3 Intégration de l'analyse factorielle discriminante

Le paragraphe 9.2 présente différentes approches d'affectation adaptées à l'analyse factorielle discriminante. L'approche la plus simple qui ne nécessite pas d'approximation par un calcul de distance, ni d'hypothèse sur la distribution des caractéristiques, est l'utilisation de la combinaison linéaire discriminante calculée par l'analyse. Cette méthode d'affectation nous paraît donc la plus adaptée à notre problème.

Dans un premier temps, nous pouvons nous limiter à l'implémentation des MFCC, le rajout d'autres caractéristiques se faisant aisément. Ces coefficients sont disponibles immédiatement, puisque les MFCC sont utilisés dans le système de reconnaissance. Le paragraphe 4.6 montre en outre que ces coefficients permettent de discriminer le bruit et la parole.

Nous utilisons ici la combinaison linéaire la plus discriminante fournie par l'analyse factorielle discriminante (AFD), qui est dans le cas de deux classes l'unique vecteur propre $\mathbf{a} = \mathbf{T}^{-1}\mathbf{c}$.

Nous avons vu au paragraphe 9.2.1, d'une part que d'après [Batlle *et al.*, 1998] l'AFD fournit de bons résultats pour la reconnaissance de phonèmes, et les MFCC sont une approximation des coefficients de l'ACP, d'autre part d'après [Wark et Sridharan, 1998] utiliser une AFD en complément de l'ACP améliore les résultats de l'ACP seul pour une reconnaissance du locuteur.

Nous cherchons ici encore à diminuer les détections de bruit qui perturbent le système de reconnaissance. Nous avons vu que diminuer les erreurs rejetables permet, de diminuer le nombre total d'erreurs (*cf.* Chapitre 3 "*Analyse des sources d'erreurs du module de détection*").

Ainsi, nous intégrons dans un premier temps la combinaison linéaire des 8 MFCC pour le passage de l'état *présomption de parole* à l'état *parole* (*cf.* figure 6.3). Cette combinaison est en fait une somme pondérée des 8 MFCC, dont les coefficients de pondération sont optimisés par l'AFD sur les bases RTC_A et GSM_A. Elle est comparée à un seuil optimisé expérimentalement sur les bases d'apprentissage à l'aide des tests de détection, noté $seuil_{AFD}$. Ainsi la condition C4 s'écrit : combinaison linéaire des MFCC $> seuil_{AFD}$. Nous obtenons ainsi une décision, si la somme est inférieure au seuil la trame est une trame de bruit, sinon c'est une trame de parole. Cette décision permet de confirmer la décision prise uniquement sur l'énergie pour le passage de l'état *présomption de parole* à l'état *parole*.

Le calcul du vecteur propre est effectué sur les bases RTC_A et GSM_A. La classe parole est constituée des segments de parole du vocabulaire et des segments de parole hors vocabulaire, issus de la segmentation manuelle. La classe bruit est constituée des segments de bruit. Ce vecteur propre permet donc de discriminer la parole des bruits segmentés, *i.e.* qui ont une énergie élevée.

9.4 Expérimentations

Pour l'évaluation du module de détection avec l'intégration des MFCC par une AFD, nous utilisons la méthodologie définit dans le Chapitre 2 "*Détection de parole pour la reconnaissance vocale*". Nous évaluons donc dans un premier temps les erreurs de détection, puis les erreurs de reconnaissance.

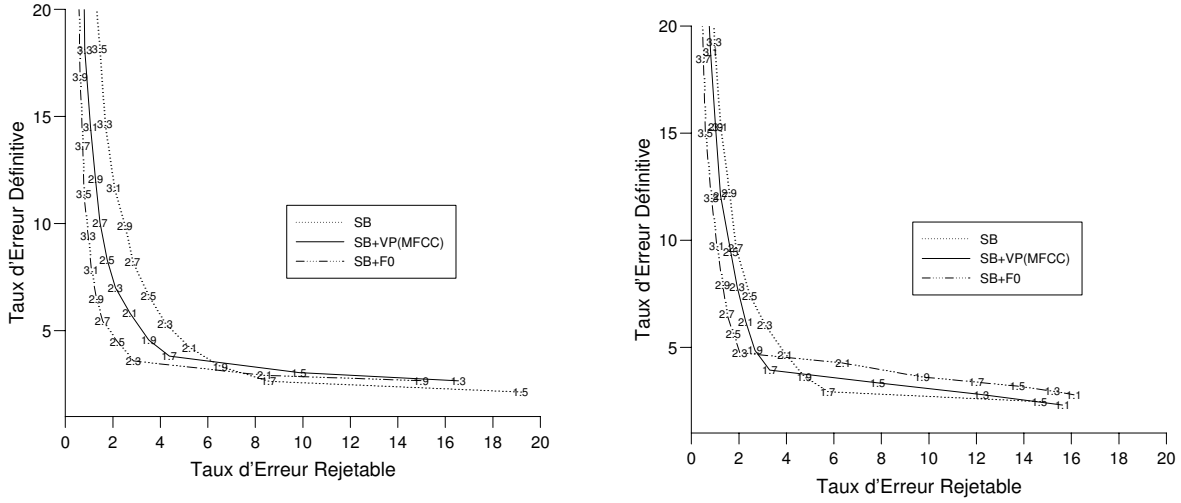
L'utilisation du vecteur propre issu de l'AFD associé avec le critère SB est appelé le critère SB+VP(MFCC). Ce critère est comparé au critère SB, sur les bases RTC_T, GSM_T et AGORA.

Nous présentons ensuite les résultats de détection avec l'intégration des coefficients de la sortie du banc de 24 filtres, avec les dérivées des MFCC associées aux MFCC, ainsi qu'avec le moment d'ordre 3 (*cf.* Chapitre 7 *"Utilisation des statistiques d'ordre supérieur"*) et le paramètre de voisement (*cf.* Chapitre 8 "*Utilisation d'un paramètre de voisement*") associés aux MFCC.

9.4.1 Résultats de détection

Nous présentons ici les résultats de la détection avec le critère SB+VP(MFCC) en comparaison du critère SB et du critère SB+F_0 présenté au Chapitre 8 "*Utilisation d'un paramètre de voisement*". La figure 9.3 donne les erreurs de détection sur la base RTC_T en fonction du mode d'enregistrement, lu ou répété. La figure 9.4 présente les résultats sur la base GSM_T en fonction du RSB, et la figure 9.5 donne les résultats sur la base AGORA. Le nouveau critère fournit de meilleurs résultats que le critère SB, en particulier sur la partie bruité de la base GSM_T, mais les résultats restent moins bons que pour le critère SB+F_0. Le tableau 9.1 donne les taux d'erreur associée pour les seuils "optimaux" de détection du critère SB+VP(MFCC) ainsi que l'intervalle de confiance des taux d'erreur associée du critère SB. Ce tableau montre que l'amélioration est significative sur toutes les bases. Nous rappelons que l'intervalle de confiance calculé est l'intervalle à 95% du nombre total de segments de parole pour les bases RTC_T, GSM_T et pour la base AGORA. Il est calculé pour le seuil de détection donnant le minimum d'erreurs totales. La différence est plus marquée sur la partie la plus bruitée de la base GSM_T et sur la base de parole continue AGORA.

Le taux d'insertion (erreurs rejetables) a été fortement diminué par rapport au critère SB, mais reste plus important que celui du critère SB+F_0.



(a) partie lue (b) partie répétée

FIG. 9.3 – *Résultats de détection des critères SB+VP(MFCC), SB et SB+F_0 sur la base RTC_T.*

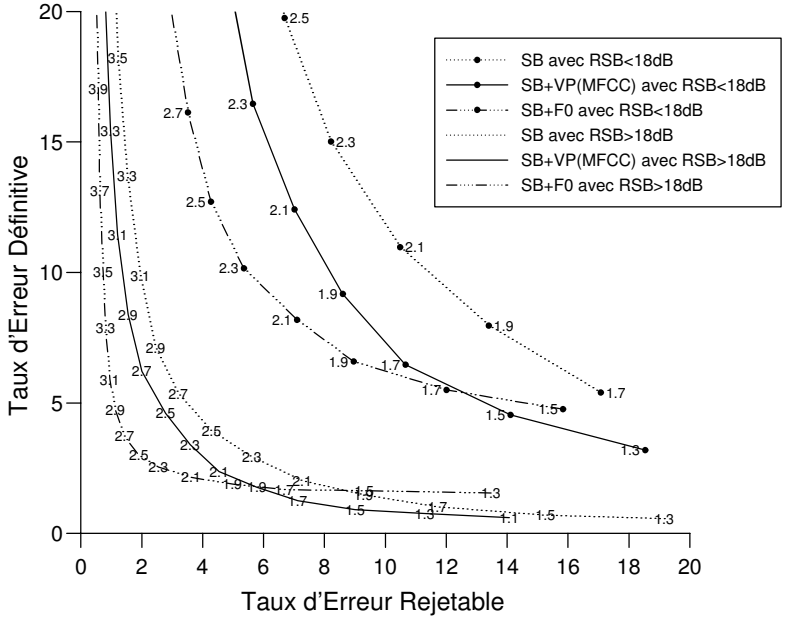


FIG. 9.4 – *Résultats de détection des critères SB+VP(MFCC), SB et SB+F_0 sur la base GSM_T.*

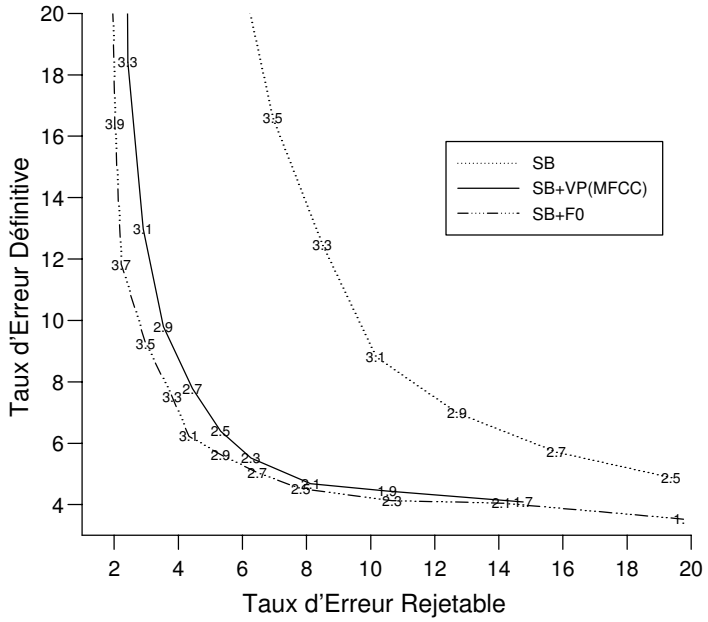


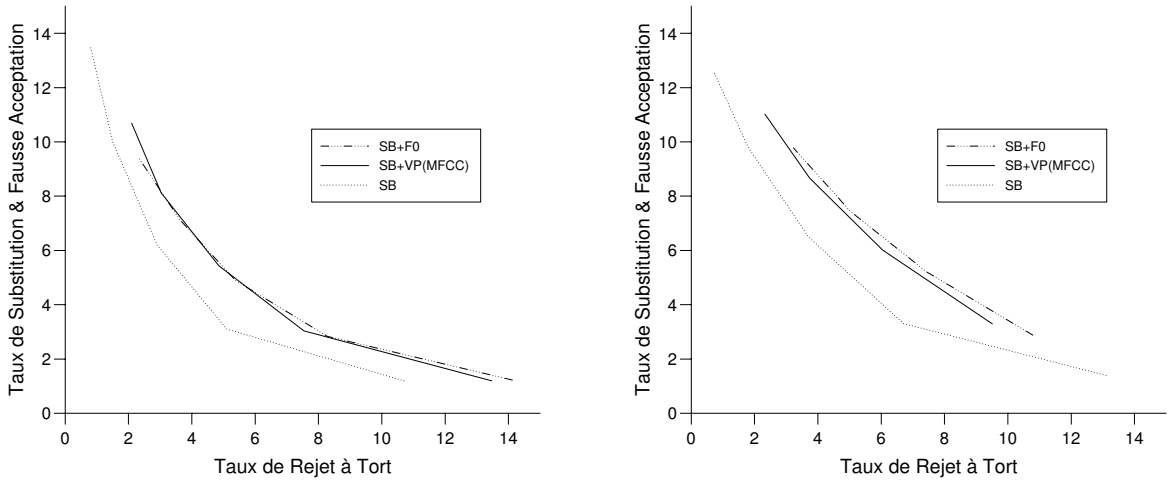
FIG. 9.5 – *Résultats de détection des critères SB+VP(MFCC), SB et SB+F_0 sur la base AGORA.*

	seuil "optimal" de SB+VP(MFCC)	taux d'erreur de SB+VP(MFCC)	seuil "optimal" de SB	intervalle de confiance de SB
RTC_T_L	1.9	8.08%	2.1	[8.74;10.17]
RTC_T_R	1.7	7.23%	1.9	[7.72;9.09]
GSM_T M18	1.7	17.12%	1.9	[20.60;22.14]
GSM_T P18	2.1	6.91%	2.5	[7.79;8.69]
AGORA	2.5	7.39	3.1%	[17.49;20.55]

TAB. 9.1 – *Taux d'erreur associée de détection du critère SB+VP(MFCC) par rapport à l'intervalle de confiance du critère SB.*

	taux d'erreur de SB+VP(MFCC)	taux d'erreur de SB	intervalle de confiance de SB
Substitution	8.2	9.2	[8.67;9.76]
Fausse Acceptation	52.7	53.5	[52.56;54.44]
Rejet à tort	8.5	13.4	[12.77;14.06]
total	22.52	27.99	[27.15;28.84]

TAB. 9.2 – *Taux d'erreur associée de reconnaissance du critère SB+VP(MFCC) par rapport à l'intervalle de confiance du critère SB sur la partie bruitée de la base GSM_T.*



(a) partie lue (b) partie répétée

FIG. 9.6 – *Résultats de reconnaissance des critères SB+VP(MFCC), SB et SB+F_0 sur la base RTC_ T.*

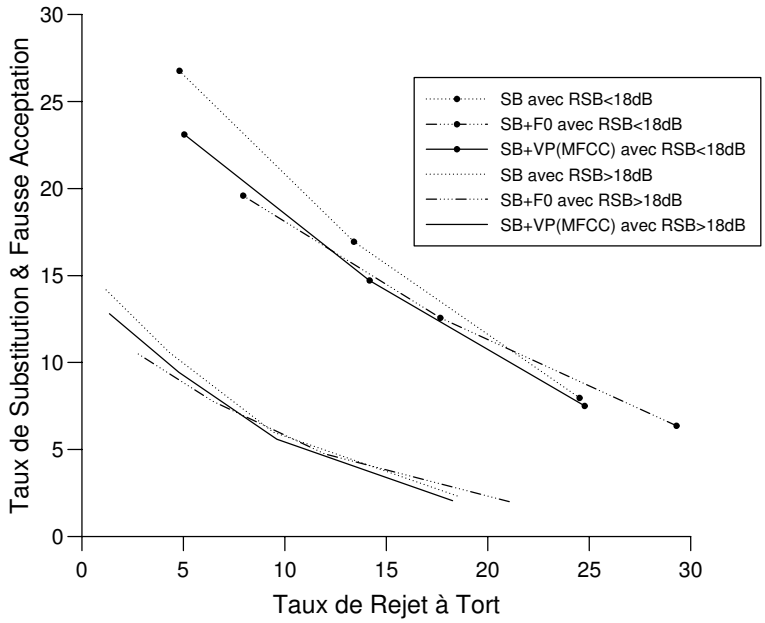
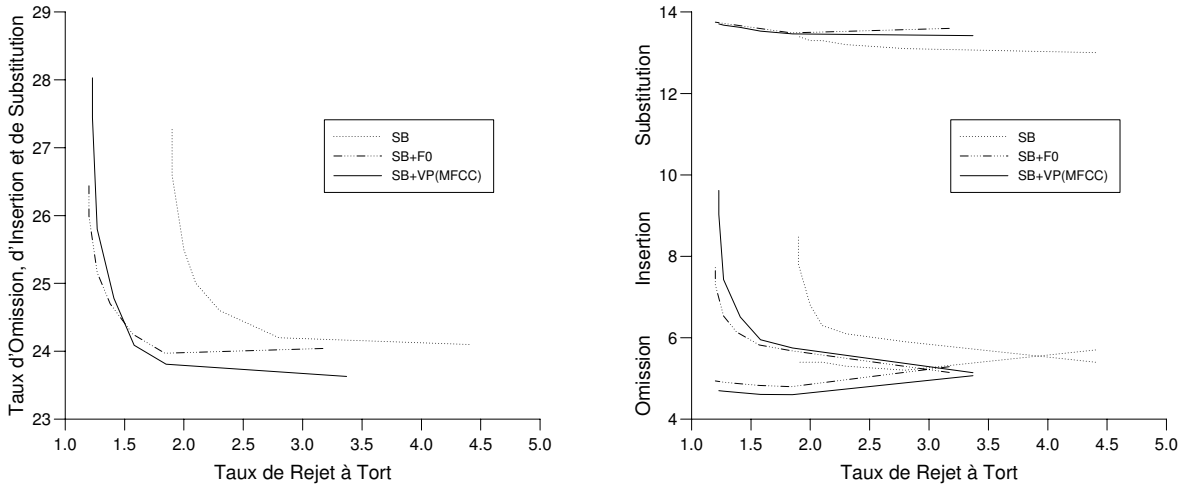


FIG. 9.7 – *Résultats de reconnaissance des critères SB+VP(MFCC), SB et SB+F_0 sur la base GSM_ T.*



(a) Les taux sont regroupés

(b) Les taux sont séparés

FIG. 9.8 – *Résultats de reconnaissance des critères SB+VP(MFCC), SB et SB+F_0 sur la base AGORA.*

	taux d'erreur de SB+VP(MFCC)	taux d'erreur de SB	intervalle de confiance de SB
Omission	4.6	5.4	[5.02;5.81]
Insertion	6.5	6.3	[5.89;6.74]
Substitution	13.6	13.3	[12.72;13.90]
Rejet à tort	1.4	2.1	[1.86;2.36]
total	26.19	27.08	[26.31;27.86]

TAB. 9.3 – *Taux d'erreur associée de reconnaissance du critère SB+VP(MFCC) par rapport à l'intervalle de confiance du critère SB sur la base AGORA.*

9.4.2 Résultats de reconnaissance

Les trois figures 9.6, 9.7 et 9.8 comparent les résultats de reconnaissance du critère SB+VP(MFCC) avec le critère SB et SB+F_0, sur les trois bases RTC_T, GSM_T et AGORA. Le seuil de détection est fixé de façon à donner les meilleurs taux de reconnaissance (*cf.* tableau F.3 en Annexe F). Il n'y a pas de différences significatives sur la base RTC_T et sur la partie calme de la base GSM_T. Pour la partie bruitée de la base GSM_T, avec un taux de rejet à tort inférieur à 20% (qui correspond ici à un poids de rejet de 400), nous obtenons une amélioration de toutes les erreurs. De plus cette amélioration est significative pour les erreurs de substitution et de rejet à tort, ainsi que de la somme des erreurs (*cf.* tableau 9.2). Ce tableau donne les taux d'erreur pour les critères SB+VP(MFCC) et SB sur la base GSM_T avec un poids de rejet de 400, ainsi que l'intervalle de confiance du critère SB. Le taux d'erreur associée est calculé en sommant le nombre de chaque erreur, divisé par le nombre total de segments références. Dans [Martin *et al.*, 2001a] ces résultats sont comparés au critère SBP. Les améliorations sont du même ordre de grandeur avec ce critère.

Pour la base AGORA, le critère SB+VP(MFCC) diminue les erreurs pour tous les types d'erreurs. Le tableau 9.3 donne les taux d'erreur pour un poids de rejet de 0. Il montre que l'amélioration est significative pour les omissions, les rejets à tort, et sur le total des erreurs. Des résultats complémentaires sont donnés dans [Martin *et al.*, 2001b].

L'amélioration des résultats de reconnaissance avec le critère SB+VP(MFCC) est donc significative sur la partie bruitée de la base GSM_T et sur la base AGORA, les deux cas les plus critiques.

9.4.3 Résultats de détection avec diverses caractéristiques

L'intégration des MFCC par une AFD améliore les performances du module de détection. Nous avons vu aux paragraphes 4.5 et 4.6 qu'il est possible d'utiliser les coefficients du vocodeur. Nous avons employé un banc de 24 filtres, les 24 coefficients permettent de calculer le vecteur propre de façon similaire à l'approche présentée pour les MFCC. La figure 9.9 montre les résultats de détections sur la base GSM_T pour ce critère SB+VP(V24) comparée aux critères SB et SB+VP(MFCC). Les résultats obtenus sont comparables au critère SB, et donc moins performants que ceux du critère SB+VP(MFCC). Ainsi le calcul des MFCC permet une utilisation plus performante de l'AFD.

Aux MFCC, il est possible d'ajouter les dérivées de chaque coefficient, calculées empiriquement par : $DMFCC_i(n) = MFCC_i(n) + 0.5MFCC_i(n-1) - 0.5MFCC_i(n-3) - MFCC_i(n-4)$, où $DMFCC_i(n)$ est la dérivée du coefficient $MFCC_i$ à la trame n. Le vecteur propre est alors calculé sur les 16 coefficients comme précédemment, et employé dans le module de détection également de la même façon. La figure 9.10 montre que ce critère SB+VP(MFCC,DMFCC) n'apporte pas d'amélioration sur la base GSM_T par rapport au critère SB+VP(MFCC).

En plus des MFCC, nous avons combiné le moment d'ordre 3 présenté au Chapitre 7 *"Utilisation des statistiques d'ordre supérieur"* et le paramètre de voisement présenté au

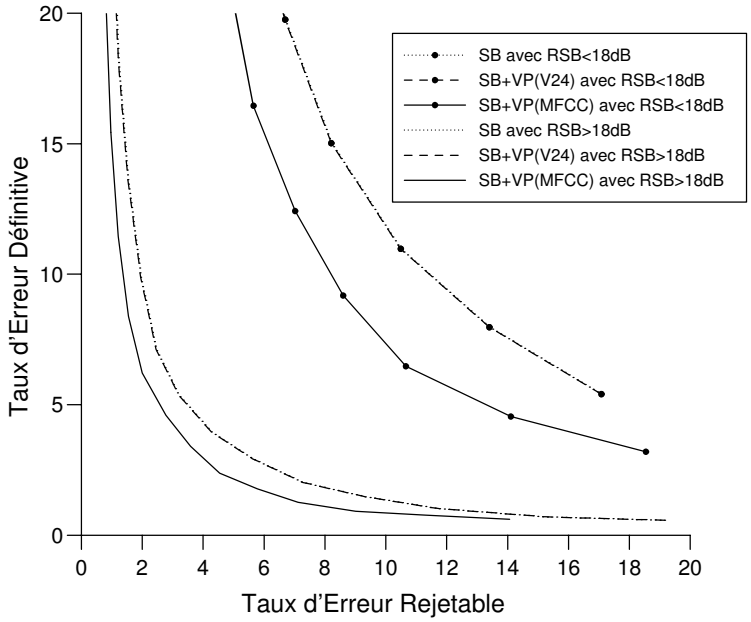


FIG. 9.9 – *Résultats de détection des critères SB+VP(V24), SB et SB+VP(MFCC) sur la base GSM_T.*

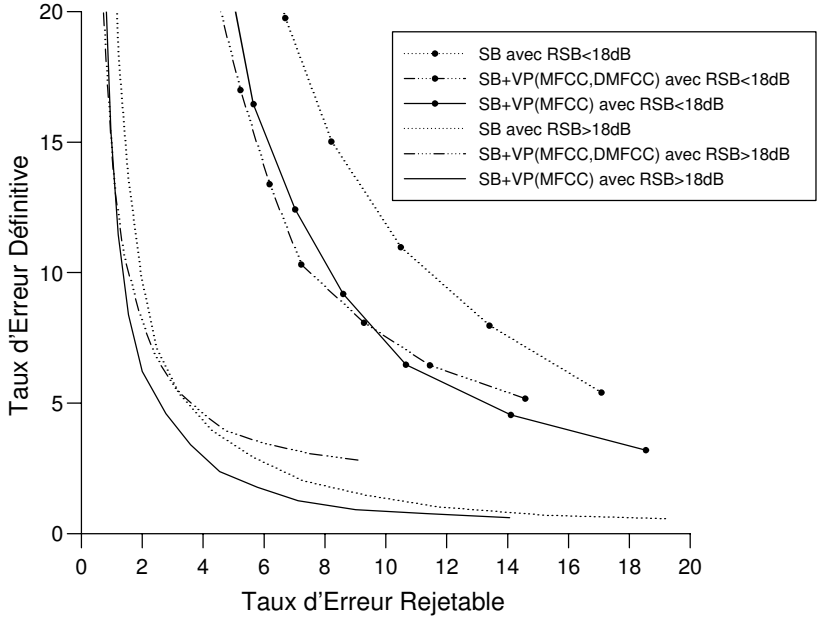


FIG. 9.10 – *Résultats de détection des critères SB+VP(MFCC,DMFCC), SB et SB+VP(MFCC) sur la base GSM_T.*

Chapitre 8 *"Utilisation d'un paramètre de voisement"*. L'intégration de ces coefficients se fait de la même façon que précédemment, par le calcul de l'unique vecteur propre **a** sur les bases GSM_A et RTC_A, qui est ensuite employé comme vecteur de pondération.

La figure 9.11 montre qu'il n'y a pas d'améliorations sur la base GSM_T avec l'ajout du moment d'ordre 3 et du paramètre de voisement. En effet, nous avons vu au Chapitre 7 *"Utilisation des statistiques d'ordre supérieur"* que le moment d'ordre 3 n'apporte pas d'amélioration au critère SB. Les résultats du Chapitre 8 *"Utilisation d'un paramètre de voisement"* montre cependant qu'un paramètre de voisement permet de détecter moins de bruits, et ainsi améliorer les performances du module de détection. Le vecteur propre sur les MFCC peut en fait être vu comme un paramètre de voisement, et ainsi le paramètre de voisement du Chapitre 8 *"Utilisation d'un paramètre de voisement"* donne une information redondante et ne permet pas d'apporter d'améliorations. En effet la représentation du vecteur propre dans l'espace fréquentiel (*cf.* figure 9.12), montre que le vecteur donne un poids important aux basses fréquences, ce qui est le principe de l'estimation d'un paramètre de voisement.

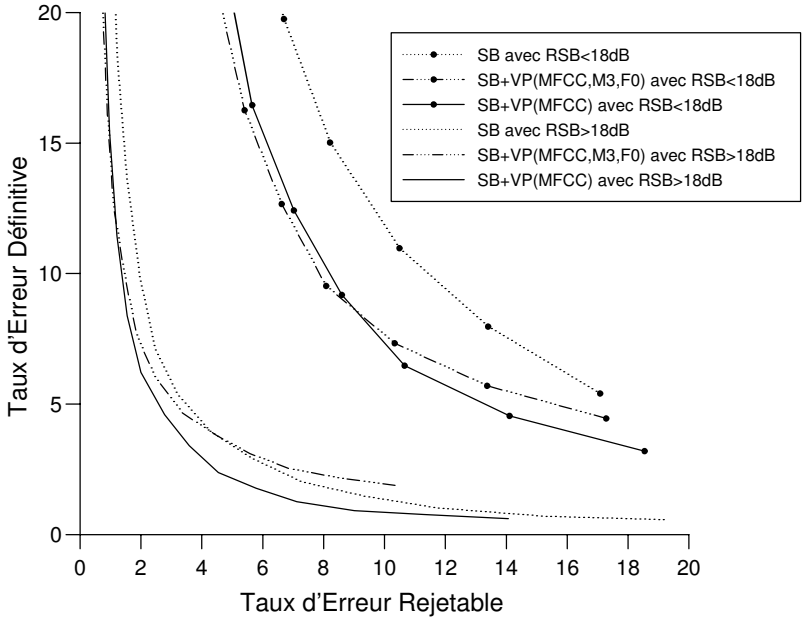


FIG. 9.11 – *Résultats de détection des critères SB+VP(MFCC,M3,F_0), SB et SB+VP(MFCC) sur la base GSM_T.*

9.5 Conclusion

Les techniques de fusion de données ici présentées semblent être intéressantes pour permettre d'intégrer un grand nombre de données dans un module de détection. Notre étude s'est portée sur la fusion en entrée. En effet la fusion de décision nécessite plusieurs

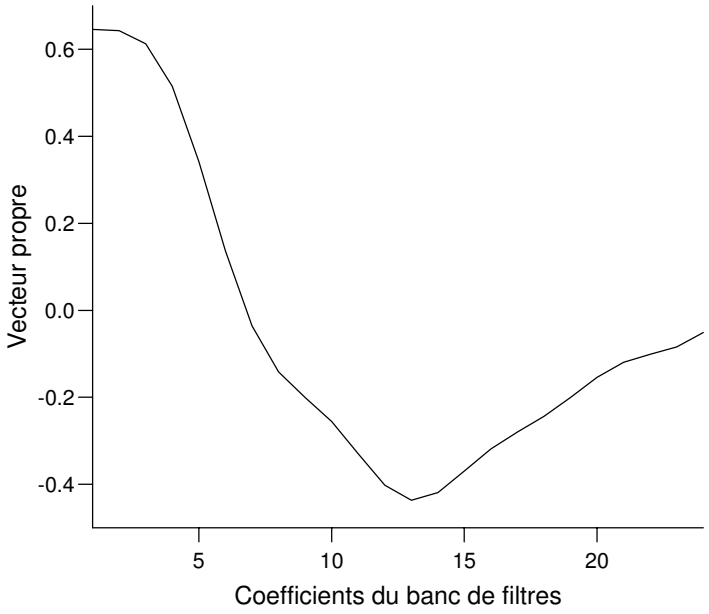


FIG. 9.12 – *Représentation du vecteur propre dans l'espace fréquentiel.*

algorithmes de détection, de préférence de nature différente, ce qui serait coûteux. Il serait cependant intéressant de comparer les méthodes de fusion en entrée et de fusion de décision pour notre problème. L'étude de l'analyse factorielle discriminante pour intégrer les MFCC dans le module de détection donne de bons résultats.

Nous avons cherché à diminuer le nombre de détections de bruit, en ajoutant une condition à l'automate. L'amélioration des résultats de détection obtenue, est significative sur la base GSM_T, sur la base AGORA et sur la partie lue de la base RTC_T. Quant aux résultats de reconnaissance, l'amélioration est significative sur la partie bruité de la base GSM_T et sur la base AGORA. Pour le reste, le fait de diminuer les détections de bruit réduit le coût de calcul du système de reconnaissance. En effet les détections de bruit peuvent être rejetées par le module de reconnaissance, cependant celui-ci est plus coûteux en temps de calcul que le module de détection de la parole. Ainsi les deux premiers objectifs sont atteints, nous avons diminuer les erreurs du module de détection pour les communications bruitées et pour la détection de parole continue. Le critère SB+F_0 reste cependant meilleur.

L'inconvénient de cette approche reste l'apprentissage nécessaire du vecteur propre des coefficients. Si les conditions d'une application sont éloignées de celles des bases d'apprentissage employées, les améliorations peuvent être moins performantes.

Ces résultats sont encourageants pour poursuivre l'étude avec l'intégration d'autres coefficients. L'intégration de la fonction linéaire discriminante peut également se faire de façon différente, en cherchant par exemple à obtenir une détection plus précise au niveau des frontières des segments. Comparer l'analyse factorielle discriminante avec une autre méthode de fusion en entrée peut être une poursuite de cette étude.

Conclusion

9.6 Bilan

Une partie importante de notre étude a été consacrée à l'évaluation du module de détection pour la reconnaissance vocale. Nous avons montré qu'une évaluation complète doit prendre en compte les résultats de la détection mais aussi les résultats du système de reconnaissance employant ce module de détection.

Cette méthode d'évaluation a permis d'effectuer une analyse détaillée des sources d'erreurs du module de détection et ainsi de déterminer les principales améliorations à apporter au module de détection.

Nous nous sommes donc fixé trois objectifs :

- Premier objectif : diminuer les erreurs du module de détection pour les communications bruitées.
- Deuxième objectif : diminuer les erreurs du module de détection de la parole continue.
- Troisième objectif : diminuer la sensibilité du seuil de détection au niveau de bruit, au changement de base de données et au réseau d'appel.

Pour atteindre ces objectifs plusieurs axes d'étude ont été abordés. Afin de réduire les erreurs dues à un bruit stationnaire, nous avons étudié les performances d'une méthode de débruitage. Les résultats du module de détection et du système de reconnaissance sont améliorés significativement pour des RSB faibles. Le premier objectif est ainsi en partie atteint. Pour réduire les détections de bruits de courte durée, nous avons étudié la meilleure façon d'intégrer une nouvelle condition. À partir de l'automate Bruit/Parole ainsi modifié avec cette nouvelle condition, nous avons étudié différentes conditions pour diminuer ces détections.

Dans un premier temps nous avons cherché à affiner l'estimation de la distribution de l'énergie, afin d'améliorer la discrimination du bruit et de la parole. Ainsi, après une étude détaillée des statistiques d'ordre supérieur, le moment d'ordre 3 du logarithme de l'énergie est apparu comme la statistique pouvant apporter le plus d'informations en complément de la moyenne et de la variance déjà employées. Différentes estimations de cette statistique ont été étudiées, cependant aucune approche n'apporte d'améliorations significatives au module de détection.

Dans un second temps nous avons donc cherché à améliorer le module de détection en apportant en complément de l'énergie une autre source d'information à l'aide d'autres

caractéristiques acoustiques. Une étude des systèmes existants de détection de parole a permis de dégager la fréquence fondamentale ainsi que les coefficients cepstraux comme étant les caractéristiques acoustiques les plus prometteuses pour répondre à notre problème.

Ainsi nous avons employé la fréquence fondamentale calculée sur tout le signal (périodes voisées et non-voisées) pour déterminer un paramètre de voisement. Le paramètre de voisement calculé de cette façon et intégré au module de détection permet une diminution significative des erreurs du module de détection aussi bien sur du signal comportant un grand nombre de bruits de courte durée que dans le cas de la détection de parole continue. Le module de débruitage employé avec ce module de détection apporte toujours une amélioration des résultats dans le cas de communications bruitées par des bruits stationnaires. Cette approche permet également de diminuer la sensibilité du réglage du seuil de détection, en particulier la sensibilité du seuil au niveau de bruit et au changement de réseau. Ainsi cette approche permet d'apporter les améliorations nécessaires au module de détection.

Nous avons cependant cherché à améliorer davantage les résultats du module de détection, en employant un grand nombre de coefficients acoustiques. Les coefficients cepstraux calculés pour le module de reconnaissance ainsi que les coefficients du vocodeur qui sont calculés pour l'obtention des coefficients cepstraux permettent une description assez complète du signal. Le problème majeur pour l'utilisation d'un grand nombre de coefficients est la fusion de ces données. Ainsi nous avons étudié les différentes approches de la fusion en entrée qui peuvent être employées pour fusionner ces caractéristiques acoustiques. L'analyse factorielle discriminante est apparue comme la plus prometteuse pour une discrimination en deux classes bruit et parole de ces coefficients. Dans un premier temps, nous avons donc intégré la combinaison linéaire des coefficients cepstraux à l'énergie. Cette approche permet une diminution des erreurs du module de détection pour des communications bruitées et pour la détection de parole continue. Cependant l'amélioration est légèrement moins bonne qu'avec l'approche du paramètre de voisement. Nous avons également étudié l'intégration des coefficients du vocodeur, des coefficients cepstraux et de leurs dérivées, ainsi que des coefficients cepstraux, du moment d'ordre 3 et du paramètre de voisement. Cependant ces approches ne semble pas apporter d'améliorations supplémentaires.

Finalement, les trois objectifs que nous nous étions fixés sont atteints par le module de détection fondé sur l'automate Bruit/Parole employant le paramètre de voisement en complément de l'énergie. Ce critère apporte de plus une amélioration lorsque le module de débruitage est associé au système de reconnaissance. Ainsi ce module de détection est robuste pour la reconnaissance de la parole en environnement bruité.

9.7 Perspectives

Le module de détection fondé sur un automate à cinq états permet de bonnes performances dans un système de reconnaissance. Une étude comparative de ce module de

détection avec d'autres approches de détection proposées dans la littérature à l'aide d'une évaluation au niveau des résultats de détection mais aussi du module de reconnaissance permettrait de situer notre approche parmi les systèmes de détection de parole.

Nous avons vu que l'approche employant un paramètre de voisement donne les meilleurs résultats. Cependant le calcul de la fréquence fondamentale pour l'obtention du paramètre de voisement reste coûteux. Un grand nombre de méthodes permettent d'estimer la fréquence fondamentale. Une méthode plus rapide que celle employée peut donner des résultats de détection semblables et ainsi diminuer le coût du système.

Ce problème de coût supplémentaire n'existe pas dans le cas de l'utilisation des coefficients cepstraux car ces coefficients sont déjà calculés pour le module de reconnaissance. Cependant les résultats obtenus sont légèrement moins bons que ceux du critère SB+F_0. Afin de chercher à améliorer davantage l'intégration de ces coefficients dans l'automate, nous pouvons effectuer une étude comparative des différentes méthodes de fusion de données dans ce contexte particulier. Une analyse qui ne nécessite pas d'apprentissage permettrait une adaptation plus aisée à toutes les applications de reconnaissance de la parole.

Le nombre important d'erreurs de segments tronqués ou élargis provoque encore beaucoup d'erreurs de reconnaissance. Il faudrait donc chercher à préciser davantage les frontières. Ceci peut se faire par une intégration différente de la condition supplémentaire. En particulier nous n'avons pas cherché à affiner la détection de fin de parole. L'intégration de cette condition au niveau de la fin de la détection proposée au paragraphe 6.4 peut permettre une meilleure détection de fin de parole sans dégrader les résultats du système de reconnaissance.

Une façon d'améliorer le système de reconnaissance est d'apporter une information en retour au module de détection à l'aide des résultats de reconnaissance. Cette approche implicite n'a pas été étudiée dans ce travail, car nous nous sommes limités à l'amélioration du module explicite. En effet le module de reconnaissance utilise les modèles de Markov qui peuvent apporter une information pour la détection de fin de parole et qui dans certains cas peuvent être plus performants pour discriminer le bruit de la parole. Cependant cette approche est aussi plus coûteuse, et le modèle de rejet a ses limites; le module de détection explicite doit donc être le plus performant possible dans la limite des moyens mis à sa disposition. Affiner le module de détection par une méthode implicite permettrait de réduire les erreurs de reconnaissance dues d 'une part aux détections de bruits, dans ce cas c'est le modèle de rejet qui doit être plus performant, d'autre part aux mauvaises détections de fin de parole (qui peuvent entraîner des erreurs de regroupement et de fragmentation).

Troisième partie

Annexes

Annexe A

Le signal de parole

A.1 Préambule sur le signal de parole

La parole est un signal émis par une suite complexe d'actions que nous produisons sans en avoir réellement conscience. Il y a tout d'abord la *génération d'une énergie ventilatoire* qui va servir à mettre en mouvement oscillatoire les cordes vocales afin de générer un son. La *vibration des cordes vocales* donne naissance à tous les sons voisés, soit 80% du temps de phonation. La *réalisation d'une disposition articulatoire* dans ce qu'il est commode de désigner sous le nom de cavités supra-glottiques, termine le processus de phonation. Les unités théoriques de base de la parole sont des *phonèmes*, ils permettent de distinguer deux sons différents. Les phonèmes se succèdent au cours du temps et durent plus ou moins longtemps. Les possibilités d'enchaînement des phonèmes confère à la parole une grande variabilité temporelle.

De plus, l'importante diversité des sons existants implique une variabilité fréquentielle du signal de parole.

Les voyelles sont classées selon différentes méthodes. Tout d'abord nous distinguons 3 groupes selon le lieu d'articulation, déterminé par la position horizontale de la masse linguale dans la cavité buccale :

- les voyelles antérieures [i,y,e,ø,œ]
- les voyelles centrales [a,ɑ]
- les voyelles postérieures [u,o,ɔ].

Il y a également la classification selon la position de la langue, haute, basse, ou moyenne. La classification des consonnes est beaucoup plus délicate. Outre la division selon la présence ou non de vibrations des cordes vocales, c'est-à-dire en voisées, ou non-voisées, nous pouvons distinguer différents modes d'articulation :

- les occlusives ou plosives [p,t,k,b,d,g]
- les fricatives [f,s,ʃ,v,z,ʒ]
- les affriquées [ts,tʃ,dz,dʒ]
- les nasales [m,n,ɳ,ŋ]
- les liquides [l,ʁ]
- les semi-voyelles [j,ʏ,w]

Le point d'articulation du son, ou la vitesse du mouvement des articulateurs permettent également de classer les consonnes.

En plus de ces classifications, un même phonème sera prononcé différemment suivant le locuteur, c'est-à-dire la personne qui parle, mais aussi suivant son état. En effet l'émission de la parole varie selon la fatigue, le stress, *etc.* Il y a donc également une variabilité du signal de parole intra-locuteur et inter-locuteur.

Le signal de parole reste donc difficile à décrire de part sa variabilité temporelle et fréquentielle et de part sa dépendance au locuteur et à son état. La production de la parole est décrite en détails dans [Calliope, 1989] et dans [Bartkova, 1999].

A.2 Analyse du signal

- **Le spectre**
  Le spectre du signal correspond au vecteur formé par le module au carré de chaque coefficient de la transformée de Fourier du signal. Il peut éventuellement être calculé en moyennant le module au carré de la transformée de Fourier sur plusieurs fenêtres. Le fait de prendre le module simplifie les calculs, puisqu'ainsi le logarithme est réel lorsque le spectre est exprimé sur une échelle en décibels. Cette simplification supprime l'information sur la phase, mais l'oreille humaine ne la perçoit que très mal (*cf.* [Bartkova, 1999]). De plus les systèmes de reconnaissance actuels n'utilisent pas cette composante du signal de parole. Plus de détails sont donnés dans [Calliope, 1989] ou dans [Rabiner et Juang, 1993]. Le spectre du signal permet de fournir de nombreuses caractéristiques du signal.
  En pratique, nous obtenons le spectre en prenant une fenêtre d'analyse de longueur $\Delta T = 32$ *ms*. Nous avons $\Delta T = n_p T_e$, où la période d'échantillonnage $T_e = 0.125$ *ms* pour un échantillonnage à 8 *kHz*, et $n_p = 256$ est le nombre de points par trame. Un tel signal échantillonné à 8 *kHz* correspond au signal de qualité téléphonique. Différents types de fenêtres d'apodisation peuvent être employés, nous utilisons des fenêtres de Hanning ($\frac{1}{2}\left[1 - \cos\left(\frac{2\pi}{n_p}(n+\frac{1}{2})\right)\right]$ avec $0 \leq n \leq n_p - 1$). Pour ne pas perdre l'information contenue au bord des fenêtres, les fenêtres se recouvrent sur un intervalle de 16 *ms*. Pour se rapprocher de la résolution fréquentielle de l'oreille humaine, le spectre peut être représenté sur une échelle Mel.
  $M = \frac{1000}{\log_{10} 2} \ln\left(1 + \frac{f}{1000}\right)$, où M est exprimé en *Mel*, et f en *Hz*.
  C'est une échelle de fréquences déduite de considérations psycho-acoustiques. Il s'agit en fait de 24 filtres passe-bande répartis sur une échelle de fréquences qui est linéaire jusqu'aux environs de 1 *kHz* et logarithmique au-delà. Le calcul de l'énergie dans chaque bande fournit ainsi 24 coefficients.
- **Les coefficients cepstraux**
  Les coefficients cepstraux sont les composants du cepstre, anagramme de spectre. Le cepstre est obtenu en prenant la transformée de Fourier inverse du logarithme du spectre. L'analyse cepstrale est une analyse qui vise à séparer les contributions respectives de la source et du conduit vocal par déconvolution. Pour cela, il est fait

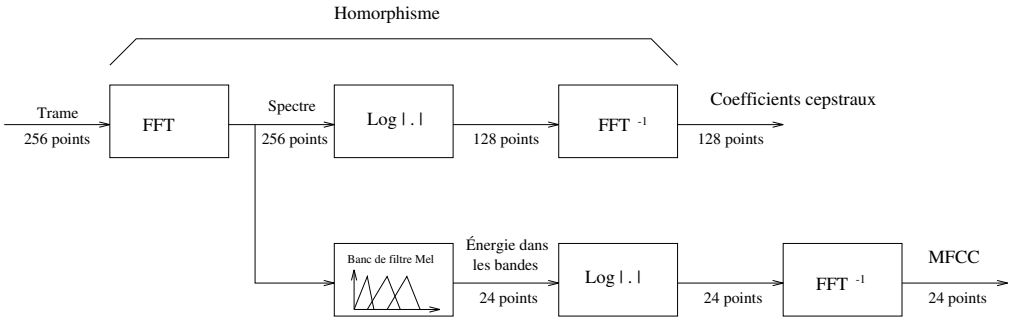


FIG. A.1 – *Calculs des coefficients cepstraux et des MFCC.*

l'hypothèse que le signal vocal s_n est produit par un signal excitateur g_n traversant un système linéaire passif de réponse impulsionnelle b_n. Nous obtenons donc $s_n = g_n * b_n$. L'homomorphisme décrit précédemment, au niveau de la définition du spectre, nous permet de déconvoluer s_n, en écrivant le produit de convolution dans un nouvel espace comme une somme. Ceci est décrit plus en détails dans [Calliope, 1989], ou [Rabiner et Juang, 1993]. En notant $S(w)$, la densité spectrale, nous écrivons :

$$\ln S(w) = \sum_{n=-\infty}^{+\infty} c_n \, e^{-inw}, \tag{A.1}$$

où $c_n = c_{-n}$ sont les coefficients cepstraux du signal. Nous avons ainsi :

$$c_n = \int_{-\pi}^{+\pi} \ln S(w) e^{inw} \frac{dw}{2\pi}. \tag{A.2}$$

Nous obtenons ainsi un ensemble de paramètres plus ou moins important, selon la précision voulue. En effet, plus nous considérons de paramètres c_n, mieux nous approchons $\ln S(w)$. Il est cependant difficile d'interpréter chaque coefficient c_n comme représentant telle ou telle caractéristique du signal (type de phonème, pitch, *etc.*) par contre il devient possible, à l'aide de ces coefficients d'estimer le pitch ou d'autres caractéristiques de la parole.

En pratique nous calculons la transformée de Fourier inverse par une transformée discrète en cosinus. Si au lieu de prendre le logarithme sur le spectre, il est calculé sur la sortie du banc de filtres de l'échelle Mel, nous obtenons les MFCC (*Mel-Frequency Cepstrum Coefficient*). La figure A.1 résume le calcul des coefficients cepstraux et des MFCC. Les énergies dans les bandes de fréquence à la sortie du banc de filtres Mel sont appelées coefficients du vocodeur. Une étude comparative des MFCC et des composantes principales de l'énergie dans chaque bande de fréquence, montre que ces coefficients sont très corrélés entre eux (*cf.* [Charlet, 1997]).

- **L'analyse LPC (Linear Predictive Coding)**
La méthode LPC est fondée sur les connaissances de la production de la parole et suppose que le modèle de production soit linéaire. Ce modèle se décompose en deux parties : la source, active, et le conduit, passif. L'onde vocale est modélisée comme la sortie d'un filtre qui put être un filtre passe-bas. Le conduit vocal est représenté par un filtre tout pôle autorégressif. Dans un premier temps, nous calculons les valeurs d'autocorrélation :

$$r(m) = \sum_{n=0}^{N-1-m} \tilde{x}(n)\,\tilde{x}(n+m) \quad m = 0,...,p, \tag{A.3}$$

où p est l'ordre de l'analyse LPC, N le nombre d'échantillons pris en compte (selon la taille de la fenêtre), et $\tilde{x}$ le signal sur une fenêtre donnée. La méthode de Durbin nous donne :

$$E^{(0)} = r(0), \tag{A.4}$$

$$k_i = \frac{1}{E^{(i-1)}} \left\{ r(i) - \sum_{j=i}^{L-1} \alpha_j^{i-1} r(|i-j|) \right\} , \tag{A.5}$$

avec $1 \leq i \leq p$,

$$\alpha_i^{(i)} = k_i, \tag{A.6}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{j-1}^{(i-1)}, \text{ pour } j < i \tag{A.7}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)}, \tag{A.8}$$

où $E^{(i)}$ est la variance de l'erreur de prédiction d'ordre i et L est le nombre de trames du signal. Nous obtenons ainsi les coefficients LPC : $a_m = \alpha_m^{(p)}$, les coefficients PARCOR (*PARtial CORrelation*) : k_m et les coefficients du logarithme du rapport des aires :

$$g_m = \ln \left(\frac{1 - k_m}{1 + k_m} \right). \tag{A.9}$$

Annexe B

Principe de la reconnaissance

Il existe différentes méthodes de reconnaissance. Il est possible de distinguer les méthodes de comparaison de formes acoustiques et les méthodes statistiques. Nous rappelons ici le module de reconnaissance utilisé dans cette étude, présenté dans [Jouvet, 1988].

Cette méthode est fondée sur une approche statistique. À partir d'une observation acoustique X, nous cherchons la séquence de phonèmes (ou de mots), W, telle que la probabilité $P(W/X)$ soit maximale. C'est-à-dire que nous cherchons la séquence $\hat{W} = \underset{W}{\text{argmax}}\, P(W/X)$. Or d'après la formule de Bayes, nous avons :

$$\hat{W} = \underset{W}{\text{argmax}} \frac{P(X/W)P(W)}{P(X)} = \underset{W}{\text{argmax}}\, P(X/W)P(W). \tag{B.1}$$

L'estimation de la probabilité *a priori* $P(W)$ relève des modèles de langage lorsque W est un mot ou une suite de mots. Nous avons $P(W) = \prod_{i=1}^{N} P(W_i/W_0, \cdots, W_{i-1})$, où W_i sont les N phonèmes (ou mots) de la séquence W. Pour la reconnaissance de parole continue cette probabilité est estimée dans notre étude à l'aide d'un *bigramme* (*cf.* [Damnati, 2000]). Le bigramme définit la probabilité de chaque mot, connaissant le mot qui le précède. C'est-à-dire que $P(W)$ est approximée à $\prod_{i=1}^{N} P(W_i/W_{i-1})$, qui est plus facile à estimer.

Pour estimer $P(X/W)$ plusieurs méthodes sont envisageables. Les modèles de Markov cachés sont très utilisés, et c'est de cette manière que la probabilité $P(X/W)$ est estimée dans le module de reconnaissance. Nous nous contentons de rappeler les principes généraux de cette méthode fondée sur les modèles de Markov cachés.

B.1 Définition

Un *modèle de Markov caché* est l'association d'une chaîne de Markov et d'un processus d'observation (l'abréviation en anglais est HMM pour *Hidden Markov Model*). Ceci

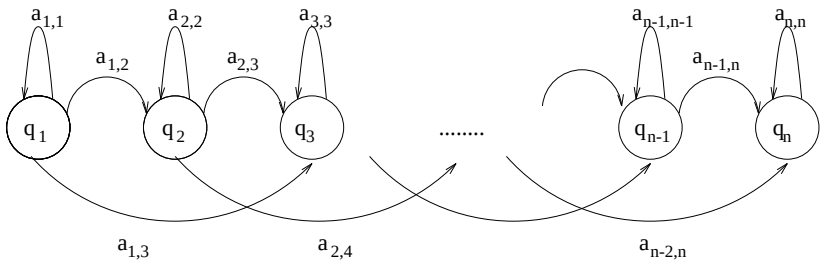


FIG. B.1 – *Modèle de Bakis.*

correspond, dans notre cas à un modèle de production de la parole, et nous permet de calculer la probabilité qu'une occurrence soit réalisée.

Rappelons qu'une chaîne de Markov est définie par un ensemble d'états, des probabilités de transitions entre ces états et par une probabilité initiale. Nous noterons :

$\{q_i, i = 1, \cdots, N_Q\}$, l'ensemble des N_Q états,

$\{\pi_i, i = 1, \cdots, N_Q\}$, les probabilités initiales, et

$\{a_{ij}, i,j = 1, \cdots, N_Q\}$, les probabilités de transition de la chaîne.

Nous avons donc $\sum_{i=1}^{N_Q} \pi_i = 1$ et $\sum_{j=1}^{N_Q} a_{ij} = 1$, $\forall i \in \{1, \cdots, N_Q\}$.

Le processus d'observation est constitué d'un ensemble de fonctions de densité de probabilité associées aux transitions entre ces états. La chaîne de Markov sert de support à ces fonctions de densité de probabilité qui fournissent des probabilités sur l'espace acoustique, c'est-à-dire des probabilités d'obtenir des trames acoustiques.

Considérant l'aspect temporel du déroulement de la parole, nous utilisons des modèles qui n'autorisent que des transitions de q_i vers q_j uniquement si $i < j$. Les modèles utilisés dans le module de reconnaissance sont de différents types, par exemple, les modèles de Bakis (*cf.* figure B.1).

Nous noterons B_{ij} la distribution acoustique associée à la transition t_{ij} de l'état q_i vers l'état q_j. Ainsi la probabilité que la trame acoustique X_t soit émise par la distribution B_{ij} est définie par :

$$B_{ij}(X_t) = P(X_t/t_{ij}). \tag{B.2}$$

Dans le cas de notre module de reconnaissance, ces fonctions de densité de probabilité sont des gaussiennes multivariées :

$$B_{ij}(X_t) = \sum_{k=1}^{NG} c_{ijk} \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma_{ijk}|^{\frac{1}{2}}} exp\{-\frac{1}{2}(X_t - \mu_{ijk})^t \Sigma_{ijk}^{-1}(X_t - \mu_{ijk})\}, \tag{B.3}$$

où NG est le nombre de gaussiennes, N le nombre de paramètres acoustiques, μ_{ijk} et Σ_{ijk} sont la moyenne et la matrice de covariance de la gaussienne k, et c_{ijk} est un coefficient

tel que :

$$\sum_{k=1}^{NG} c_{ijk} = 1, \forall i,j \in \{1, \cdots, N_Q\}. \tag{B.4}$$

Pour des raisons de simplicité et de coût de calculs, les matrices de covariance sont supposées diagonales. Cette approximation est justifiée lorsque les paramètres acoustiques utilisés sont non corrélés.

B.2 Probabilité d'émission des trames acoustiques

Nous appelons *chemin s de longueur T* dans la chaîne de Markov une suite de T transitions consécutives entre les états du modèle. Un chemin est donc défini par une suite de transitions :

$$s = t_{[0][1]} t_{[1][2]} \cdots t_{[T-1][T]}, \tag{B.5}$$

où $[\tau]$ désigne le numéro de l'état occupé à l'instant τ. Ainsi pour un modèle donné M, la probabilité d'observer l'occurrence acoustique X le long du chemin s avec le modèle M est :

$$P(X/s,M) = \prod_{\tau=1}^{T} B_{[\tau-1][\tau]}(X_\tau). \tag{B.6}$$

Nous obtenons la probabilité conjointe d'émission de l'observation et du chemin :

$$\begin{aligned} P(X,s/M) &= P(X/s,M).P(s/M) \\ &= \pi_{[0]} \prod_{\tau=1}^{T} a_{[\tau-1][\tau]} B_{[\tau-1][\tau]}(X_\tau). \end{aligned}$$

Nous appelons chemin *optimal*, le chemin $\hat{s}$ qui fournit la probabilité conjointe d'émission de l'observation et du chemin, la plus élevée sur tous les chemins possibles de longueur T.

$$P(X,\hat{s}/M) = \max_{s} P(X,s/M). \tag{B.7}$$

La probabilité d'observation de X pour le modèle M est la somme des probabilités d'émission sur tous les chemins possibles de longueur adéquate. L'algorithme de Baum-Welch permet le calcul de cette probabilité. En fait, le module de reconnaissance ne considère que la probabilité d'émission le long du chemin optimal, qui peut être calculé par l'algorithme de Viterbi. Cet algorithme est décrit dans le tableau B.1, où :

- $\Phi(\tau,q_j)$ est la probabilité d'émettre les τ premières trames et d'aboutir dans l'état q_j à l'instant τ.
- $\Psi(\tau,q_j)$ est le pointeur vers l'état précédent le long du meilleur chemin.
- $\hat{q}_\tau$ est l'état occupé après l'émission de la trame X_τ le long du chemin optimal et q_F est l'état final.

Initialisation	Pour $i = 1, \cdots, N_Q$ faire $\Phi(0,q_i) = \pi_i$
Itérations	Pour $\tau = 1, \cdots, T$ faire Pour $j = 1, \cdots, N_Q$ faire $\Phi(\tau,q_j) = \max_{q_i}\{\Phi(\tau-1,q_i)a_{ij}B_{ij}(X_\tau)\}$ $\Psi(\tau,q_j) = \hat{q}_i = \operatorname{argmax}_{q_i}\{\Phi(\tau-1,q_i)a_{ij}B_{ij}(X_\tau)\}$
Résultat	$P(X,\hat{s}/M) = \Phi(T,q_F)$
Décodage	$\hat{q}_T = q_F$ Pour $\tau = T-1, \cdots, 1$ faire $\hat{q}_\tau = \Psi(\tau+1,\hat{q}_{\tau+1})$

TAB. B.1 – *Algorithme de Viterbi.*

B.3 Apprentissage

L'apprentissage est la tâche qui consiste à estimer l'ensemble Θ des paramètres du modèle de Markov selon un certain critère. Cet ensemble comprend l'ensemble des probabilités initiales, l'ensemble des probabilités de transition, et l'ensemble des paramètres des densités de probabilité d'émission.

Différentes approches sont proposées pour estimer Θ dans [Jouvet, 1988]. C'est une approche fondée sur les statistiques d'utilisation des paramètres des modèles, pour maximiser la probabilité d'émission des données suivant les chemins optimaux, qui a été retenue.

B.4 Modèles utilisés

Différents types de modèles acoustiques correspondant aux unités de base sont considérés.

Tout d'abord, nous considérons un modèle par mots, où les mots sont pris comme unités de base. Ce modèle est construit par un apprentissage, et ne convient donc que pour la description de petits vocabulaires.

Le modèle par phonèmes est une approche pour modéliser toutes les réalisations acoustiques possibles d'un phonème.

Le modèle par allophones (qui sont des phonèmes en contexte) modélise également les réalisations acoustiques des phonèmes, mais en prenant en compte les différents contextes. L'unité de base est donc l'allophone. Ce modèle convient mieux que le modèle par mots pour de grands vocabulaires, il est cependant moins performant lorsqu'il y a peu de mots, si l'apprentissage ne dépend pas du vocabulaire.

C'est ce dernier modèle qui sera utilisé pour les applications. Un exemple de modèle par allophones est proposé sur la figure B.2 dans le cas d'un vocabulaire constitué de chiffres. Nous disposons d'une bibliothèque d'allophones, constituée de tous les phonèmes dans tous les contextes possibles de la langue française. Les mots sont décrits phonéti-

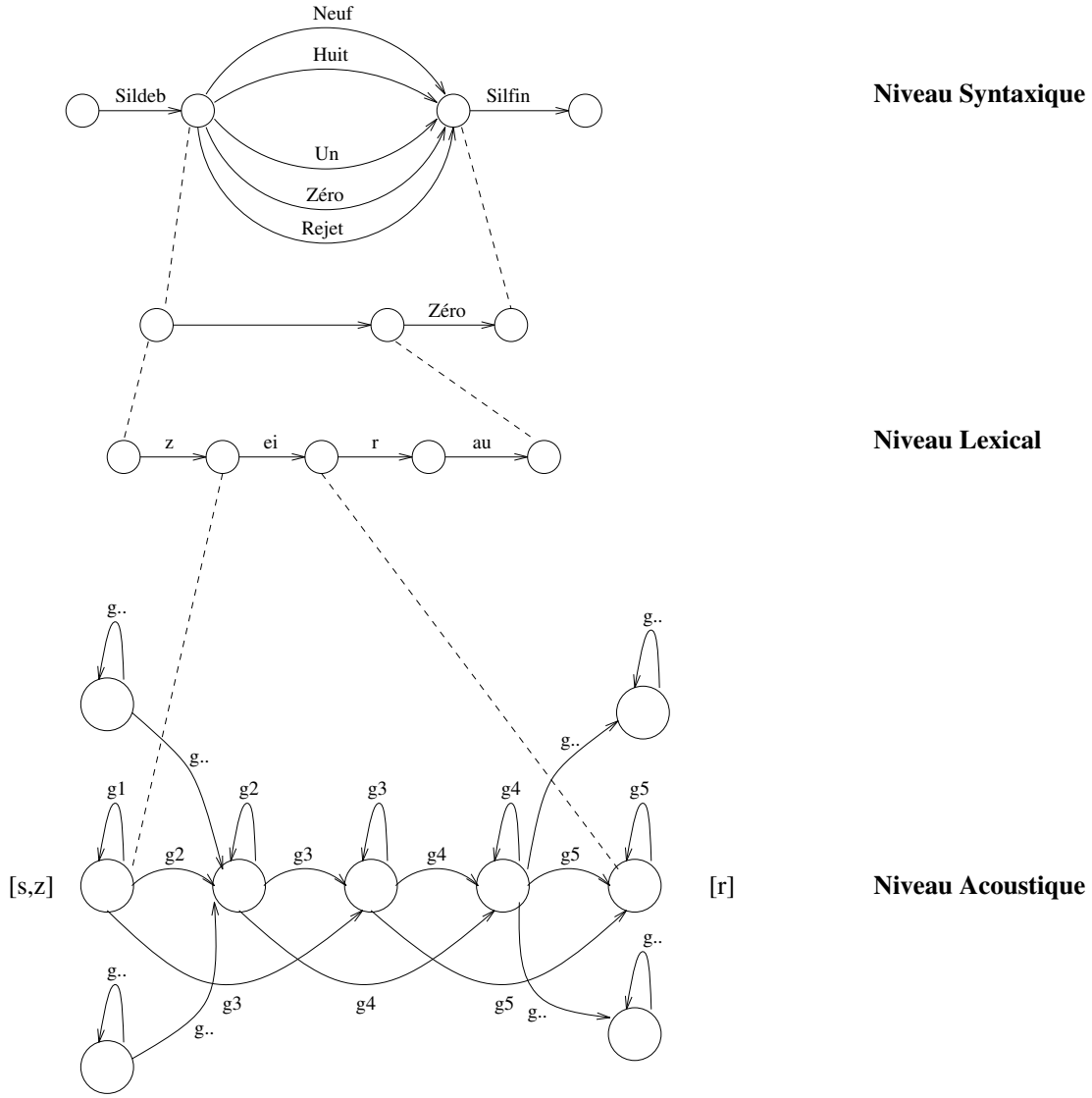


FIG. B.2 – *Modèle par allophones pour les mots isolés.*

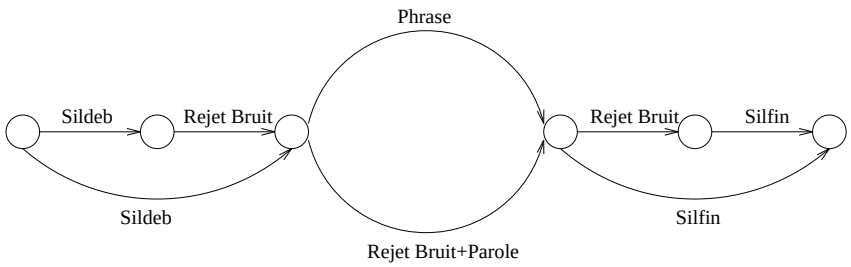


FIG. B.3 – *Modèle par allophones avec ajout d'un modèle de bruits pour la reconnaissance de parole continue.*

quement, à l'aide des allophones, tenant ainsi compte des phénomènes de coarticulation. Pour modéliser le silence éventuel avant et après le mot (ou la phrase), des modèles de silence début (Sildeb) et de silence fin (Silfin) sont rajoutés au niveau syntaxique. Le niveau syntaxique représente les différentes possibilités offertent par le vocabulaire. Un modèle de rejet permet de modéliser les détections de bruits et la parole hors vocabulaire. Ce modèle est construit à partir d'une base de bruits. Le niveau lexical permet de modéliser les différentes suites de phonèmes de la prononciation de chaque mot. Et enfin le niveau acoustique est la modélisation des prononciations possibles de chaque phonème.

Dans le cas de la reconnaissance de parole continue, le modèle utilisé est un modèle par allophones. Le modèle par mot n'est pas envisageable ici, compte tenu de l'importance du vocabulaire. Quelques modifications ont été apportées au modèle précédent (*cf.* figure B.3). L'insertion de petits mots dus aux bruits, ou à un silence trop important en début et en fin de phrase, nous a conduit à rajouter un modèle de bruits en début et en fin de phrase. Les modèles de bruit en début, en fin de phrase et au niveau de la phrase sont identiques. Ces modèles permettent de rejeter uniquement des bruits de courtes durées.

Annexe C

Signification des résultats

Il ne s'agit pas ici de présenter toutes les approches permettant la validation des résultats. Les principales méthodes dans le cadre des tests de systèmes de reconnaissance sont décrites dans [Mokbel, 1992], un point de vue plus général est proposé dans [Saporta, 1990]. Une première approche consiste à calculer l'intervalle de confiance du taux d'erreur estimé pour chacun des deux modules, puis de déterminer si ces intervalles de confiance ne se recoupent pas. Une seconde approche est un test d'hypothèse : les deux modules sont équivalents, ou ne le sont pas.

C.1 Intervalle de confiance

Soit $\hat{\tau}$ un estimateur du taux d'erreur τ. L'intervalle de confiance est déterminé par δ avec une probabilité $1-\alpha$ telle que :

$$1-\alpha = P(\hat{\tau}-\delta < \tau < \hat{\tau}+\delta). \tag{C.1}$$

L'intervalle de confiance avec la probabilité $1-\alpha$ est donc $[\hat{\tau}-\delta;\hat{\tau}+\delta]$. Nous ne connaissons pas en général τ, ainsi l'intervalle de confiance est lui-même une estimation.

Pour cette estimation deux cas se présentent : nous connaissons la loi de distribution de la mesure d'erreurs dont le paramètre τ est la moyenne, et auquel cas des techniques dites paramétriques répondent au problème, ou nous ne connaissons pas cette loi, dans ce cas des techniques non-paramétriques permettent cette estimation.

Citons quelques exemples de techniques paramétriques.

- Cas de la loi gaussienne de paramètres (τ,σ) :
  Si $\hat{\tau}$ est l'estimateur arithmétique de τ, il suit une loi gaussienne de paramètres $(\tau,\frac{\sigma}{\sqrt{n}})$, où n est la taille de l'échantillon.
  - cas σ connu :
    L'intervalle de confiance est :

$$\hat{\tau} - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \tau < \hat{\tau} + u_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{C.2}$$

si $1-\alpha = 0.95$, nous avons $u_{\alpha/2} = 1.96$, d'après la table de la fonction de répartition de la loi normale réduite.

- cas σ inconnu :
  En posant $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tau_i - \hat{\tau})^2}$, où les τ_i sont les n observations de τ. La variable aléatoire $T = \frac{\hat{\tau}-\tau}{S}\sqrt{n-1}$ suit une loi de Student à $(n-1)$ degré de liberté. L'intervalle de confiance est :

$$\hat{\tau} - t_{\alpha/2}\frac{s}{\sqrt{n-1}} < \tau < \hat{\tau} + t_{\alpha/2}\frac{s}{\sqrt{n-1}}, \tag{C.3}$$

  où $t_{\alpha/2}$ est déterminé à partir des tables de Student.

- Cas de la loi binomiale de paramètres (n,τ) quand n est grand :
  C'est aussi le problème nommé intervalle de confiance pour une proportion τ inconnue. Dans notre cas, nous pouvons supposer que pour chaque segment observé, il y a erreur ou pas, que ce soit pour les tests de détection ou de reconnaissance. Nous nous ramenons ainsi à un espace de Bernouilli. La proportion $\hat{\tau}$ pour un échantillon de taille n grand suit une loi gaussienne de paramètres $(\tau, \sqrt{\frac{\tau(1-\tau)}{n}})$. L'intervalle de probabilité symétrique est donc :

$$\tau - u_{\alpha/2}\sqrt{\frac{\tau(1-\tau)}{n}} < \hat{\tau} < \tau + u_{\alpha/2}\sqrt{\frac{\tau(1-\tau)}{n}}. \tag{C.4}$$

  Si n est grand, nous avons l'expression approchée de l'intervalle de confiance :

$$\hat{\tau} - u_{\alpha/2}\sqrt{\frac{\hat{\tau}(1-\hat{\tau})}{n}} < \tau < \hat{\tau} + u_{\alpha/2}\sqrt{\frac{\hat{\tau}(1-\hat{\tau})}{n}}. \tag{C.5}$$

Nous nous contentons de citer la technique de Jackknife, comme technique non-paramétrique.

- Le principe de la méthode consiste à diminuer le biais d'un estimateur, et fournir un estimation de l'intervalle de confiance. $\hat{\tau}$ étant l'estimateur de τ calculé sur un échantillon de taille n, notons $\hat{\tau}_{-i}$ celui calculé sur le $(n-1)$ échantillon obtenu en enlevant l'observation i, et nous appelons pseudo-valeur $\hat{\tau}_i^*$:

$$\hat{\tau}_i^* = n\hat{\tau} - (n-1)\hat{\tau}_{-i}. \tag{C.6}$$

  L'estimateur de Jackknife est l'estimateur des moyennes des pseudo-valeurs :

$$\hat{\tau}_J = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_i^* = n\hat{\tau} - \frac{n-1}{n}\sum_{i=1}^{n}\hat{\tau}_{-i}. \tag{C.7}$$

La variance de cet estimateur est :

$$\sigma_J^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{\tau}_i^* - \hat{\tau}_J)^2}{n-1}. \tag{C.8}$$

L'estimateur $\hat{\tau}_J$ a un biais plus petit que $\hat{\tau}$, il est possible de montrer que $E[\hat{\tau}] = E[\hat{\tau}_J] + \frac{a}{n}$. De plus l'estimateur $\hat{\tau}_{n-1} = \frac{\hat{\tau}_J - E[\hat{\tau}_J]}{\sigma_J}$ suit une loi de Student à $n-1$ degré de liberté. Cet estimateur permet le calcul de l'intervalle de confiance sans hypothèse sur la loi de l'échantillon.

La méthode du *bootstrap* permet également l'estimation d'un intervalle de confiance sans connaissance *a priori* de la loi.

C.2 Tests d'hypothèses

Rappelons que nous cherchons à évaluer l'estimation, pour comparer deux tests de détection ou de reconnaissance. Nous cherchons donc à savoir si nous sommes dans l'une des deux hypothèses suivantes :

H_0 : Les deux modules de détection sont équivalents.

H_1 : Les deux modules de détection ne sont pas équivalents.

L'hypothèse H_0 équivaut à "le taux d'erreur du premier module est égal à celui du second", et l'hypothèse H_1 en est la contre-opposée. C'est un test unilatère, plusieurs formalisations sont envisageables.

Si nous considérons de nouveau que l'espace des observations est un espace de Bernouilli, ce qui est une modélisation vraisemblable pour notre problème. Les estimations des taux d'erreur $\hat{\tau}_1$ et $\hat{\tau}_2$ de τ_1 et de τ_2 de chacun des modules, sont telles que $n\hat{\tau}_1$ et $n\hat{\tau}_2$ suivent une loi binomiale de paramètres respectifs (n,τ_1) et (n,τ_2).

Dans le cas de l'hypothèse H_0, nous avons $\tau_1 = \tau_2 = \tau$. Si n est grand, $\hat{\tau}_1$ et $\hat{\tau}_2$ suivent alors une loi gaussienne de paramètres $(\tau, \sqrt{\frac{\tau(1-\tau)}{n}})$. L'estimateur $\hat{\tau} = \frac{\hat{\tau}_1 + \hat{\tau}_2}{2}$ est un autre estimateur de τ. Et $\hat{\tau}_1 - \hat{\tau}_2$ suit une loi gaussienne de paramètre $(0, \sqrt{\frac{2\hat{\tau}(1-\hat{\tau})}{n}})$. Ainsi nous pouvons obtenir un intervalle de confiance pour accepter l'hypothèse H_0 :

$$-u_{\alpha/2}\sqrt{\frac{2\hat{\tau}(1-\hat{\tau})}{n}} < \tau_1 - \tau_2 < u_{\alpha/2}\sqrt{\frac{2\hat{\tau}(1-\hat{\tau})}{n}}. \tag{C.9}$$

C.3 Conclusion

Il ne s'agit pas ici de détailler toutes les méthodes d'évaluation des résultats, mais d'en présenter celles qui nous paraissent les principales dans le cadre de notre problème. Le fait de considérer l'espace d'observation comme un espace de Bernouilli semble raisonnable et permet de donner un intervalle de confiance simple à calculer. Ainsi pour

valider les résultats obtenus, nous utiliserons le calcul de l'intervalle de confiance sous cette hypothèse.

Annexe D

Bases de données

Pour le réglage des algorithmes et la validation des tests il est important d'utiliser plusieurs bases de données. Les apprentissages (réglages des algorithmes) et les tests ont été effectués sur différentes bases, l'une enregistrée sur le réseau RTC dans un contexte applicatif (*cf.* [Mauuary, 1994]), deux autres sont des bases de laboratoire, et la dernière est une base d'expérimentation. Une base est enregistrée sur le réseau GSM (*cf.* [Karray, 1998b]), une autre comporte des enregistrements GSM et RTC (notée GSM_T et RTC_T) avec un vocabulaire en commun, utilisée pour les tests, et enfin une base de parole continue enregistrée sur le réseau RTC (notée AGORA), qui est une base d'expérimentation. La segmentation manuelle de ces bases permet une évaluation par rapport à une détection supposée idéale. La segmentation manuelle a été réalisée par différentes personnes. Dans toute segmentation reste une part de subjectivité. En effet, les frontières des mots ou phrases ont été placées au plus près selon la vision et l'écoute de la personne. Des parties de signal ont pu être considérées comme inaudibles, ou reconnues comme de la parole. Des bruits ont pu être considérés comme trop faibles pour devoir être pris en compte. Dans le cas de la parole continue, corpus AGORA, le temps de pause entre les mots d'une requête peut être plus ou moins long, ce qui entraîne la possibilité de deux requêtes lorsque la pause est longue. Une requête peut être composée d'un seul mot de commande. Bref, cette segmentation "idéale", est loin de toute critique.

D.1 Les Baladins - Le corpus RTC_A

La base de données des Baladins est constituée de 999 appels à un Service Vocal Interactif (SVI) en exploitation qui fournit les programmes de cinéma de la région du Trégor (*cf.* figure D.1). Les appels enregistrés en continuité sur le réseau RTC ont une durée maximum de 2 min 30 s. Ils contiennent les mots de commande au SVI (soit un vocabulaire de 25 mots), de la parole non destinée au SVI et du bruit. Cette base de données de 999 appels dure 32 h 25 min.
Les 25 mots du vocabulaire sont :
Annulation, Art et essai, Aujourd'hui, Autre rubrique, Demain, Guide, Guingamp, Lannion, Message, Mode d'emploi, Non, Oui, Perros, Perros Guirec, Prochainement, Pro-

gramme, Précédent, Renseignement, Résumé, Répéter, Scolaire, Semaine, Suivant, Tarif, Trebeurden .
Cette base étant une base d'exploitation, le nombre de répétitions de chacun des mots dépend de l'utilisation du SVI faite par les personnes ayant appelé le service.

Pour l'évaluation des modules de détection, les segments issus de la segmentation manuelle sont regroupés en *Parole* et *Non-Parole*. Les segments *Parole* sont composés de la parole hors-vocabulaire, étiquetée [PAROLE] et également référencée par *Parole-Hors-Voc* et les mots du vocabulaire constituent les segments *Parole-Voc*. Les segments *Non-Parole* sont étiquetés [BRUIT], pour tous types de bruits, [RTRC], pour rires, toux, respirations (sauf la respiration qui suit ou précède un mot), cris, [ÉCHO], pour l'écho, [SFL], pour la respiration qui suit ou précède un mot, [BC], pour les bruits de combiné, [TIERS], pour la parole prononcée par une tierce personne, de faible intensité, et donc considérée comme de la non-parole. Pour une homogénéité entre les bases, les étiquettes ont été regroupées en [BRUIT] pour [RTRC], [BC] et [BRUIT], en [BF] (Bruits de Fond) pour [SFL] et [TIERS], qui sont considérées comme de la non-parole ainsi que l'étiquette [ÉCHO].

Les enregistrements ayant un niveau sonore différent, nous avons séparé l'ensemble des enregistrements en deux parties : les enregistrements ayant un RSB inférieur à 20 *dB*, et ceux ayant un RSB supérieur à 20 *dB* (*cf.* paragraphe D.5).

Nous avons ainsi sur l'ensemble des appels (en nombre de segments étiquetés manuellement) :

Segments	RTC_A	RTC_A M20	RTC_A P20
nb. de fichiers	999	465	534
RÉFÉRENCES	10021	4895	5126
Parole	5815	2504	3311
[PAROLE]	1329	705	624
[BRUIT]	2177	1304	873
[BF]	635	368	267
[ÉCHO]	65	14	51

Les enregistrements initiaux comportant le bruit DTMF de fin de communication et étiqueté raccrochage dans [Mauuary, 1994], ont été raccourcis pour ne plus contenir ces parties de signal. Ce qui revient à ne considérer que la partie du signal avant le raccrochage. Il y a 58.03% de Parole, 13.26% de [PAROLE] et 28.71% de bruit, regroupant toutes les étiquettes de bruit.

D.2 Le corpus GSM_A

C'est une base de données de laboratoire, où le locuteur (homme ou femme et d'âges variables) répète une liste prédéfinie de mots dans un ordre aléatoire. La base de donnée est constituée de 53 mots : *Ancien, Annulation, Annuler, Archivé, Autre choix, Autre*

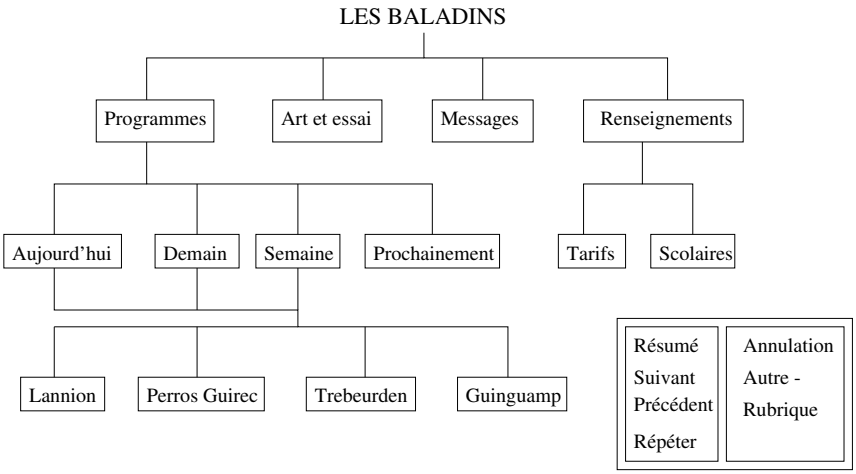


FIG. D.1 – *Arborescence du serveur "les Baladins".*

rubrique, Caractère, Cinq, Compléter, Confirmer, Conserver, Consultation, Consulter, Correction, Deux, Début, Écouter, Effacer, Enregistrer, Fin, Guide, Huit, Information, Message, Messagerie, Modifier, Neuf, Non, Nouveau, Ouais, Oui, Pause, Précédent, Quarante quatre, Quatre, Quitter, Reprise, Retour, Répéter, Répétition, Sept, Six, Sommaire, Stop, Suite, Suivant, Supprimer, Terminer, Trois, Un, Validation, Valider, Zéro. Ces mots sont répétés à peu près autant de fois chacun, normalement une fois par locuteur.

Ce corpus est composé de 67.80% de mots du vocabulaire, 6.28% de bruit de fond, étiquetés [BF], 9.39% de bruit GSM (bruits métalliques, trous), étiquetés [BGSM] , 3.97% d'écho, étiquetés [ÉCHO], 2.40% de bruit divers, étiquetés [BRUIT], 6.25% de signal inaudible, étiquetés [INAUDIBLE], et 3.91% de parole hors vocabulaire, étiquetés [PAROLE]. Il y a donc 28.29% de bruit total.

Les appels, au nombre de 389, ont été effectués dans différents types de conditions : à l'intérieur (25.39%), à l'extérieur (22.87%), dans un véhicule à l'arrêt (28.01%) ou roulant (24.02%). Ce corpus contient ainsi différents types de bruits selon l'environnement d'appel. Ces différents environnements d'appel ne reflètent pas exactement les différents niveaux de bruits. Nous avons donc séparé l'ensemble des enregistrements en deux parties selon le RSB, ceux inférieurs à 18 *dB*, et ceux supérieur à 18 *dB* (*cf.* paragraphe D.5).

	environnements				RSB	
Segments	int.	ext.	véh. arrêt	véh. roulant	RSB M18	RSB P18
nb. de fichiers	100	83	111	95	195	194
RÉFÉRENCES	8134	7237	8976	7695	17433	14609
Parole	5407	4714	6170	5432	11205	10518
[PAROLE]	335	338	287	294	644	610
[BRUIT]	306	284	106	74	445	325
[BF]	164	407	1042	400	1464	549
[BGSM]	870	648	839	652	1570	1439
[ÉCHO]	174	162	287	648	847	424
[INAUDIBLE]	878	684	245	195	1258	744

D.3 Les corpus GSM_T et RTC_T

Cette nouvelle base est constituée de 389 enregistrements d'appels sur le réseau GSM, et de 180 enregistrements sur le réseau RTC divisés en 90 enregistrements pour les mots de commande lus et de 90 enregistrements pour ceux qui sont répétés. Cette base est utilisée dans cette étude pour l'évaluation des modules de détection.

D.3.1 Le corpus GSM_T

Les mots de commande sont répétés. Les enregistrements ont été effectués dans différents environnements : intérieur, extérieur, dans un véhicule roulant à l'arrêt avec ou sans le moteur en marche. Les véhicules utilisés sont de différentes marques : Ford Escort, Fiat Brava, Opel Vectra, Alpha 146, Punto. Le main libre a parfois été utilisé. Les personnes, hommes et femmes, peuvent avoir des accents régionaux et étrangers, et sont d'âge très variable. La majorité des personnes était non abonnés lors de l'acquisition de cette base en 1997. Les appels ont été effectués avec des puissances de 2W ou 8W, et avec des combinés de différentes marques : Siemens, Nokia, Ericson, Panasonic, Motorola, Sagem.

La base est constituée de 65 mots de vocabulaire :

0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 44, Aide, Ajouter, Annuler, Avancer, Caractère, Compléter, Composer, Continuer, Corriger, D'accord, Début, Déconnecter, Détruire, Écouter, Effacer, Enregistrer, Faxer, Fin, Guidage piéton, Guide, Imprimer, Info bourse, Info finance, Info route, Info trafic, Liste, Location de voitures, Loisir, Mode d'emploi, Modifier, Non, Notifier, Ok, Opérateur, Opératrice, Oui, Pause, Précédent, Reculer, Reprendre, Reprise, Retour, Répéter, Réservation, Réécouter, Sommaire, Spectacle, Stop, Stopper, Suivant, Supprimer, Terminer, Télécopier, Valider.

D.3.2 Le corpus RTC_T

Les mots de commande sont soit répétés, soit lus. L'écho sur cette partie de la base est très important.

La base est constituée de 68 mots de vocabulaire :

0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 44, Aide, Ajouter, Annuler, Avancer, Bateau, Caractère, Circulation, Compléter, Conseils boursiers, Continuer, Corriger, Cours des valeurs, Devise, Début, Déconnecter, Écouter, Effacer, Enneigement, Enregistrer, Et demi, Et quart, Euro, Festival, Fin, Guide, Itinéraire, Livraison à domicile, Magasin, Mode d'emploi, Modifier, Moins le quart, Muse, Non, Opérateur, Opératrice, Oui, Parc, Pause, Point, Pronostic des courses, Pronostic hippiques, Précédent, Prévision météo, Rapport des courses, Reculer, Reprise, Retour, Répéter, Résultats des courses, Réécouter, Ski, Sommaire, Stop, Suivant, Supprimer, Terminer, Valider.

D.3.3 Segmentation manuelle de la base

Pour 769 enregistrements, les frontières des mots répétés ou lus et des bruits de la communication ont été marquées, puis libellées. Les étiquettes sont divisés en :

[PAROLE] qui correspond à toute la parole enregistrée qui ne fait pas partie du vocabulaire.

[BF] qui correspond aux bruits de fond (voitures, souffles, respirations, bruits divers, *etc.*)

[BRUIT] qui correspond à des bruits impulsifs (combiné, sonnerie, *etc.*)

[ÉCHO] qui correspond à l'écho des instructions demandées aux personnes.

[INAUDIBLE] qui correspond à de la parole très faible ou incompréhensible.

[BGSM] qui correspond à des bruits jugés provenir de la transmission GSM.

Partie GSM_T Les appels téléphoniques sont au nombre de 389. Ces enregistrements proviennent pour 41.13% de femmes, et pour 58.87% d'hommes, et selon les environnements : 20.31% en intérieur, 22.37% en extérieur, 33.16% pour les véhicules roulants, et 24.16% pour les véhicules à l'arrêt, qui comporte 20.57% avec le moteur en marche et 3.60% sinon.

Remarque : L'environnement le plus calme est celui du véhicule à l'arrêt, puis l'intérieur, le plus bruyant étant celui du véhicule roulant. Dans le même souci que pour la base GSM_A, les enregistrements ont été classé selon leur RSB inférieur et supérieur à 18 *dB* (*cf.* paragraphe D.5).

	environnements				RSB	
Segments	int.	ext.	véh. arrêt	véh. roulant	RSB M18	RSB P18
nb. de fichiers	79	87	94	129	167	222
RÉFÉRENCES	5969	6997	6647	9979	13110	16448
Parole	5124	5911	6083	8203	10809	14487
[PAROLE]	213	340	171	347	519	551
[BRUIT]	95	40	60	236	246	184
[BF]	435	561	267	949	1168	1038
[BGSM]	34	54	29	23	107	33
[ÉCHO]	8	16	19	86	82	46
[INAUDIBLE]	60	75	18	135	179	109

Au total, il y a 85% de Parole, 3% de [PAROLE] et 11% de bruit, regroupant les différentes étiquettes de bruit de la segmentation manuelle.

Partie RTC_T Cette partie est divisée en deux, en fonction du mode d'enregistrement, la partie comportant les mots lus est notée RTC_T_L et celle comportant les mots répétés est noté RTC_T_R, chaque sous-partie comporte 90 enregistrements.

Segments	RTC_T_R	RTC_T_L
nb. de fichiers	90	90
RÉFÉRENCES	12246	11909
Parole	6277	6437
[PAROLE]	203	223
[BRUIT]	19	28
[BF]	275	368
[BGSM]	0	0
[ÉCHO]	5467	4838
[INAUDIBLE]	5	15

Remarque : L'écho est très important en nombre, mais aussi en intensité. Il est souvent d'une énergie comparable à celle de la parole. Nous avons donc supprimé les segments d'écho pour obtenir des résultats plus fiables et proches de la réalité. Ainsi, nous obtenons sur l'ensemble de la base, 91% de Parole, 3% de [PAROLE] et 6% de bruit, comprenant l'ensemble des étiquettes de bruit.

D.4 Le corpus de parole continue - AGORA

Cette base acquise en 1999 contient 98 enregistrements sur le réseau RTC, de parole continue, qui sont des commandes à un serveur pour la gestion d'un service personnalisé. Cette base d'expérimentation a été enregistrée pour mettre en œuvre ce service. Il y a 64 enregistrements de voix d'homme, et 34 enregistrements de voix de femme.

Le modèle de reconnaissance utilisé comprend 1633 mots, qui sont des mots courants pour l'utilisation. Il y a 2520 segments de *Parole* (ou phrases), qui comprennent 12635 mots, et 1018 segments de *Non-Parole* (ou bruit).

Pour tous les enregistrements, les frontières des mots et des bruits de la communication ont été marquées, puis libellées. Les bruits sont divisés en :

[TIERS] qui correspond à toute la parole enregistrée d'une tierce personne.

[BC] qui correspond aux bruits de combiné.

[BB] qui correspond aux bruits de bouche.

[BP] qui correspond aux bruits de papier.

[TX] qui correspond aux bruits de toux.

[RI] qui correspond aux bruits de rire.

[SFL] qui correspond aux bruits de souffle.

[BRUIT] qui correspond à des bruits impulsifs non identifiés.

[ÉCHO] qui correspond à l'écho des instructions demandées aux personnes.

[RACCROCHAGE] qui correspond aux bruits de raccrochage.

Pour un homogénéité des bases ces segments ont été regroupés en [BRUIT] pour [BC], [BB], [TX], [RI], [RACCROCHAGE] et [BRUIT], en [BF] pour [TIERS] et [SFL], qui sont considérées comme de la non-parole ainsi que l'étiquette [ÉCHO].

Cet environnement comporte ainsi (en nombre de segments) :

Segments	AGORA
RÉFÉRENCES	3115
Parole	2520
[BRUIT]	331
[BF]	301
[ÉCHO]	386

D.5 Le rapport signal à bruit

Les différentes bases comportent des enregistrements plus ou moins bruités. Le partitionnement des bases en fonction de l'environnement d'appel (*intérieur*, *extérieur*, *véhicule roulant*, *véhicule à l'arrêt*) n'est pas représentatif du niveau du rapport de signal à bruit des appels.

L'adaptation des seuils de l'algorithme se faisant dans les périodes de silence, le rapport signal à bruit (RSB) ici calculé est la différence logarithmique de l'énergie sur les segments étiquetés comme de la parole à laquelle nous avons retranché l'énergie sur la partie du signal non étiquetée (pour ne pas considérer le bruit existant sur les segments de *Parole*) et l'énergie des segments étiquetés comme du bruit. Ce choix de calcul est fait en vue d'une implémentation en ligne aisée dans le module de détection, et pour pouvoir considérer le bruit existant sur les segments de *Parole*. En effet, le RSB "segmental" (*cf.* [Faucon *et al.*,

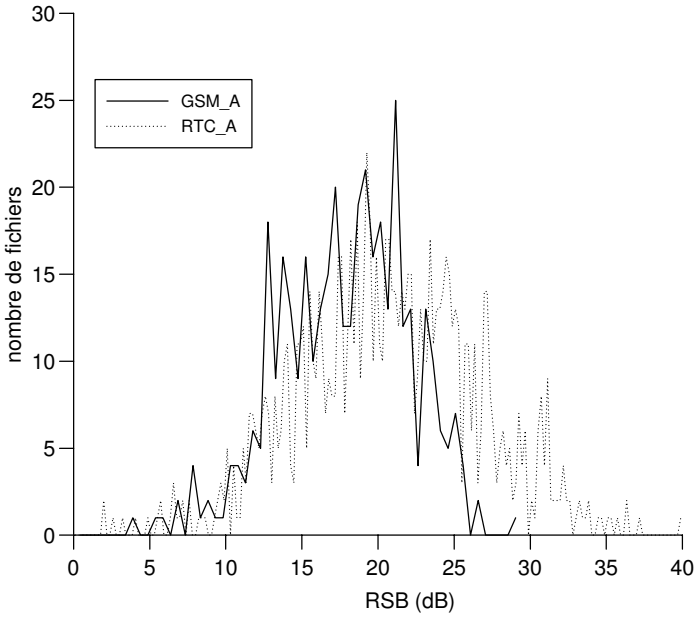
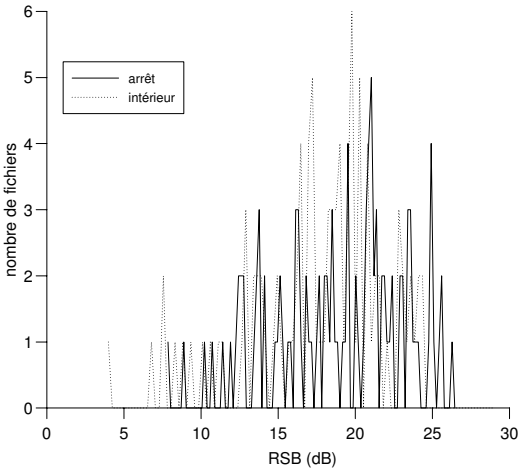


FIG. D.2 – *Rapport Signal à Bruit sur les bases RTC_A et GSM_A.*

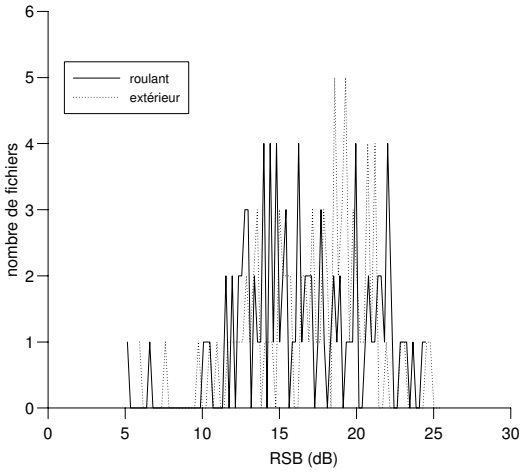
1993]) ne permet pas de considérer ce bruit par un calcul simple. Le RSB "segmental" est mieux corrélé avec les tests d'écoute, mais nous ne cherchons ici qu'un critère de séparation des fichiers selon le niveau de bruit.

La figure D.2 présente les histogrammes cumulés des fichiers de la base GSM_A et de la base RTC_A en fonction du RSB des fichiers. Il est normal que sur la base RTC_A les enregistrements soient plus calmes. Il y a cependant un recouvrement important avec des enregistrements de la base GSM_A, pour les RSB inférieurs à 20 *dB*. Pour les fichiers des bases GSM_A nous pouvons les séparer en deux parties à peu près égales, ceux inférieurs à 18 *dB* et ceux supérieurs. En effet, sur les environnements calmes (*véhicule à l'arrêt* et *intérieur*) il y a peu d'enregistrements inférieurs à 18 *dB* (*cf.* figures D.3(a) et D.3(c)), tandis que sur les environnements *extérieur* et *véhicules roulants*, nous notons une nette séparation. En effet si la fenêtre du véhicule est ouverte, ou si l'appel est effectué à partir d'un milieu très bruyant, le RSB sera plus faible (*cf.* figures D.3(b) et D.3(d)). Séparer les fichiers selon les RSB supérieurs ou inférieurs à 18 *dB* semble donc plus intéressant qu'une séparation par environnement d'appel. Cette séparation divise la base RTC_A en 465 fichiers de RSB inférieurs à 20 *dB*, notée *RTC_A M20*, et 534 fichiers supérieurs à 20 *dB*, notée *RTC_A P20*. La base GSM_A est divisée en 195 fichiers inférieurs à 18 *dB*, notée *GSM_A M18*, et 194 fichiers supérieurs à 18 *dB*, notée *GSM_A P18*. La base GSM_T est moins bruitée que la base GSM_A, elle est divisée en 167 fichiers inférieurs à 18 *dB*, notée *GSM_T M18*, et 222 fichiers supérieurs à 18 *dB*, notée *GSM_T P18*.

Base GSM_A

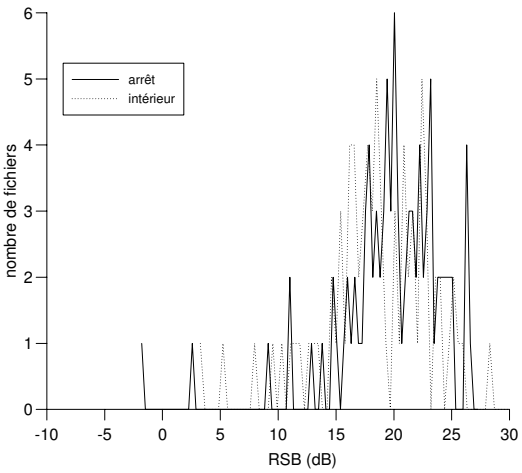


(a) véhicule à l'arrêt et intérieur

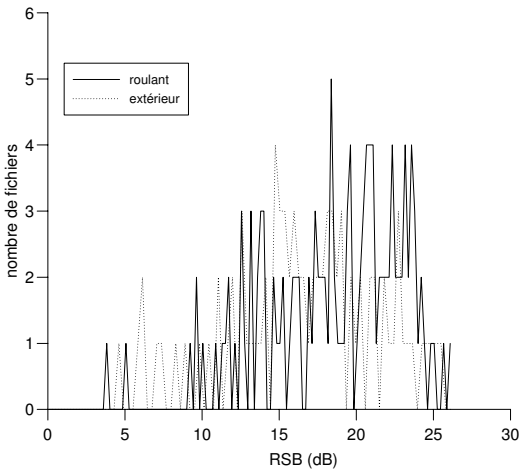


(b) véhicule roulant et extérieur

Base GSM_T



(c) véhicule à l'arrêt et intérieur



(d) véhicule roulant et extérieur

FIG. D.3 – *Rapport Signal à Bruit sur les bases GSM_A et GSM_T.*

Annexe E

Résultats du module de détection par type de mots

Nous présentons dans cette annexe le détail des erreurs étudiées au paragraphe 3.5 du critère LCT du module de détection. L'étude est faite sur la partie calme de la base GSM_A afin d'éviter l'influence du bruit de la partie bruitée de la base.

Description des tableaux

- Le tableau E.1 présente les erreurs de détection par mot pour la base GSM_A avec un RSB inférieur à 18 *dB* pour un seuil "optimal" de 9.4 *dB*. Ce tableau comprend une colonne *%total* qui donne le pourcentage d'occurrences des mots dans la partie de la base considérée. Les colonnes *Omis* et *Frag* fournissent respectivement les pourcentages de détections de mots omis et fragmentés par le module de détection. Ces pourcentages sont exprimés en fonction des occurrences des mots. Les quatre colonnes *B.P.*, *T.G.*, *T.D.* et *T.G.D.* précisent le positionnement des frontières (tronquées ou non) des détections correctement reliées, également exprimé en fonction des occurrences des mots.
- Le tableau E.2 donne pour quelques mots de la base le pourcentage de mots bien reconnus, rejetés à tort, substitués à un autre mot du vocabulaire et omis lors de la détection, en fonction de la détection :
  - si le segment issu de la segmentation manuelle correspond à la détection automatique (Cor.),
  - si le segment a été regroupé avec une autre (Reg.),
  - ou s'il a été fragmenté (Frag.).

  Ces pourcentages sont calculés en fonction du nombre d'occurrences des mots, pour un seuil "optimal" de 9.4 *dB* et avec un rejet de 200.
- Pour un même seuil "optimal" de 9.4 *dB* et avec un rejet de 200, le tableau E.3 détaille pour chaque mot, ceux qui sont bien reconnu et ceux rejetés à tort en fonction du positionnement de la détection. Cette détection est bien placée, ou tronquée à gauche, à droite ou à gauche et à droite. Les pourcentages sont calculés en fonction des occurrences des détections correctement reliées.

mot	%total	Omis	Frag.	B.P.	T.G.	T.D.	T.G.D.
Cinq	1.91	1.49	**0.00**	37.81	21.89	**19.90**	**17.91**
Consultation	1.85	9.23	**10.77**	56.41	**10.77**	5.64	4.10
Consulter	1.83	4.17	**13.54**	61.98	13.54	4.69	0.00
Début	1.79	2.66	0.00	78.19	5.32	11.17	0.53
Écouter	1.82	6.28	**21.47**	44.50	**22.51**	3.14	1.57
Enregistrer	1.90	0.50	**19.00**	62.00	0.50	**13.00**	1.50
Fin	1.76	1.62	**0.00**	51.35	**42.70**	1.08	0.00
Guide	1.73	6.59	**0.00**	70.88	11.54	7.14	1.10
Huit	1.86	5.61	1.53	37.76	3.06	**47.45**	2.04
Information	1.97	1.93	3.86	71.50	3.38	**15.46**	2.42
Modifier	1.83	5.18	**10.36**	70.98	5.18	4.66	0.52
Neuf	1.85	**0.51**	**0.00**	47.69	2.56	**44.10**	2.05
Non	1.83	2.60	**0.00**	94.79	1.56	0.00	0.00
Oui	1.98	7.21	0.48	85.10	0.96	1.44	0.00
Pause	1.83	**1.04**	**0.00**	79.27	0.52	**17.62**	0.00
Quarante quatre	2.26	0.00	0.84	46.22	0.42	**45.80**	2.10
Quatre	1.95	**0.00**	1.46	51.22	0.98	**44.39**	0.00
Quitter	1.76	4.86	0.00	37.84	**53.51**	0.00	0.54
Reprise	1.83	4.15	0.00	60.10	9.84	**17.62**	3.63
Répéter	1.89	2.01	**11.56**	66.33	9.05	7.54	0.50
Répétition	1.79	5.32	**33.51**	39.36	4.26	**12.23**	1.60
Sept	1.90	**1.50**	1.00	24.00	21.00	**24.50**	**27.50**
Six	2.04	**19.07**	0.93	26.51	16.74	**17.67**	**18.14**
Stop	1.92	5.45	0.50	23.27	**27.23**	9.90	**32.67**
Suite	1.72	**11.60**	2.21	28.18	14.92	**20.44**	**20.99**
Supprimer	1.92	7.92	0.99	45.05	**41.09**	1.49	2.48
Trois	2.02	**0.00**	**0.00**	83.02	15.09	0.94	0.00
Un	1.94	**0.00**	**0.00**	93.63	0.49	0.49	0.00
Validation	1.86	1.02	1.53	70.92	5.10	**17.35**	2.55

TAB. E.1 – *Erreurs de détection par mot.*

	Bien Reconnu		Rejeté		Substitué		Omis
mot	Cor.	Frag.	Cor.	Frag	Cor.	Frag	
Cinq	89.6	0.0	6.5	0.0	1.0	0.0	1.5
Consultation	84.6	0.0	4.6	0.5	2.1	0.5	9.2
Consulter	86.5	0.0	5.2	**2.6**	0.5	1.0	4.2
Début	88.3	0.0	3.7	0.0	2.1	0.0	2.7
Écouter	80.1	0.0	8.4	**4.2**	0.5	**4.7**	6.3
Enregistrer	85.0	0.0	5.5	**2.0**	0.0	1.0	0.5
Fin	76.8	0.0	6.5	0.0	**9.2**	0.0	1.6
Guide	84.6	0.0	6.6	0.0	2.2	0.0	6.6
Huit	79.1	1.0	8.2	0.0	**6.6**	1.0	5.6
Information	89.9	0.0	4.8	0.5	0.0	0.0	1.9
Modifier	87.0	0.0	3.1	3.1	0.0	1.0	5.2
Neuf	87.7	0.0	6.7	0.0	0.5	0.0	0.5
Non	89.1	0.0	6.8	0.0	1.0	0.0	2.6
Oui	84.1	0.0	2.4	0.0	**6.7**	0.5	7.2
Pause	92.2	0.0	4.1	0.0	0.0	0.0	1.4
Quarante quatre	85.3	0.0	6.3	0.4	0.0	0.0	0.0
Quatre	88.3	1.0	6.3	0.0	0.5	0.0	0.0
Quitter	**68.1**	0.0	**19.5**	0.0	**4.9**	0.0	4.9
Reprise	85.0	0.5	8.8	0.0	0.5	0.0	4.1
Répéter	83.9	0.0	5.5	**1.5**	0.5	1.0	2.0
Répétition	79.8	0.0	3.2	**3.7**	1.1	**6.4**	5.3
Sept	82.5	0.5	9.5	1.0	1.5	1.0	1.5
Six	81.9	0.0	7.9	0.5	**3.3**	0.5	**19.1**
Stop	77.2	0.0	**12.9**	0.0	**5.0**	1.0	5.4
Suite	85.1	1.7	6.6	0.6	**2.2**	0.0	**11.6**
Supprimer	85.6	0.0	10.9	0.0	0.5	0.0	7.9
Trois	92.0	0.0	3.3	0.0	0.5	0.0	0.0
Un	71.1	0.0	15.2	0.5	**6.4**	0.0	0.0
Validation	93.4	0.0	3.6	0.0	1.0	0.0	1.0

TAB. E.2 – *Erreurs de reconnaissance par mot.*

	Rejeté				Substitué			
mot	B.P.	T.G.	T.D.	T.D.G.	B.P.	T.G.	T.D.	T.G.D.
Cinq	5.9	1.1	0.6	0.0	1.1	0.0	0.0	0.0
Consultation	3.0	2.4	0.0	0.0	0.0	0.0	**2.4**	0.0
Consulter	4.1	1.2	0.0	0.6	0.6	0.0	0.0	0.0
Début	2.9	1.2	0.0	0.0	2.4	0.0	0.0	0.0
Écouter	0.7	**9.2**	0.0	0.7	0.0	0.7	0.0	0.0
Enregistrer	5.6	-	0.6	0.0	0.0	-	0.0	0.0
Fin	5.7	2.1	-	0.0	**9.2**	**2.8**	-	0.0
Guide	6.0	1.3	0.0	0.0	2.6	0.0	0.0	0.0
Huit	6.6	0.6	**3.2**	0.0	**5.8**	0.0	**2.6**	0.0
Information	3.7	0.5	0.5	0.5	0.0	0.0	0.0	0.0
Modifier	3.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Neuf	7.1	0.0	0.6	0.0	0.0	0.0	0.6	0.0
Non	7.0	0.0	-	-	0.6	0.6	-	-
Oui	2.6	0.0	0.0	-	**7.4**	0.0	0.6	-
Pause	4.0	0.0	0.6	-	0.0	0.0	0.0	-
Quarante quatre	7.4	0.0	0.5	0.0	0.0	0.0	0.0	0.0
Quatre	6.4	0.0	1.1	0.0	0.6	0.0	0.0	0.0
Quitter	9.6	**17.5**	0.0	0.8	1.6	**5.6**	0.0	0.0
Reprise	7.2	1.8	0.6	0.6	0.0	0.6	0.0	0.0
Répéter	5.2	1.2	0.0	0.0	0.0	0.6	0.0	0.0
Répétition	1.3	2.7	0.0	0.0	0.0	0.0	**1.3**	0.0
Sept	10.9	1.8	0.6	0.0	0.6	0.6	0.6	0.0
Six	6.8	0.6	1.1	1.1	2.8	0.0	1.1	0.0
Stop	10.1	**7.1**	0.6	0.6	3.8	2.6	0.0	0.0
Suite	5.3	1.9	0.6	0.6	2.6	0.0	0.0	0.0
Supprimer	4.3	**8.7**	0.0	0.0	0.6	0.0	0.0	0.0
Trois	3.0	0.5	0.0	0.0	0.5	0.0	0.0	0.0
Un	16.4	0.0	0.0	0.0	**9.0**	0.0	0.0	0.0
Validation	2.7	0.5	0.5	0.0	0.5	0.0	0.5	0.0

TAB. E.3 – *Erreurs de reconnaissance selon les frontières des mots détectés.*

Annexe F

Seuils optimaux sur les différentes bases

Nous donnons ici les seuils optimaux de détection et de reconnaissance sur les bases RTC_A, GSM_A et GSM_T en fonction du RSB, sur la base RTC_T en fonction du mode d'enregistrement, et sur la base AGORA.

F.1 Seuils optimaux de la détection

Le tableau F.1 donne les seuils présentant le minimum des taux d'erreur associée (erreurs rejetables et définitives), selon les bases et les critères.

Le critère SB+VP(MFCC) n'est évalué que sur les bases RTC_T, GSM_T et AGORA, puisque les bases RTC_A et GSM_A sont utilisées pour l'apprentissage.

Le tableau F.2 donne les seuils de détection optimaux pour les trois critères LCT, SB et SBP, après ajout des deux bruits *car* et *babble* sur la partie calme de la base GSM_A.

F.2 Seuils optimaux de la reconnaissance

Le tableau F.3 est obtenu par des figures représentant les taux de substitution et de fausse acceptation en fonction du taux de rejet à tort, en fonction des seuils de détection. La courbe la plus basse permet de trouver le meilleur seuil de reconnaissance pour un critère et une base données. Par exemple, sur les figures F.1(a) et F.1(b), sont représentés les courbes des erreurs du critère LCT sur la base GSM_A selon le RSB, pour les différents seuils de détection. Ainsi le seuil optimal de détection est 9.4 *dB* pour la partie bruitée et 11.3 *dB* pour la partie plus calme.

Le tableau F.3 donne les seuils, selon les critères et sur les différentes bases, qui fournissent les meilleurs résultats de reconnaissance. Ainsi, nous faisons l'étude détaillée des erreurs de reconnaissance uniquement pour ces seuils.

De même le tableau F.4 donne les seuils pour les trois critères LCT, SB et SBP avec ajout des deux bruits *car* et *babble* sur la partie calme de la base GSM_A, le tableau F.5 après le débruitage de la partie calme ayant subi l'ajout des deux bruits, et le tableau F.6 après le débruitage de la base GSM_A selon le RSB.

	Critères					
Bases	LCT	SB	SBP	SB+M3	SB+F_0	SB+VP(MFCC)
RTC_A M20	22.6	3.3	1.35	3.3	2.3	
RTC_A P20	22.6	3.1	1.35	3.1	2.5	
GSM_A M18	11.3	2.1	1.17	1.5	1.3	
GSM_A P18	13.2	2.3	1.17	1.5	1.7	
RTC_T_L	13.2	2.1	1.13	2.1	1.7	1.9
RTC_T_R	15.0	1.9	1.11	1.9	1.7	1.7
GSM_T M18	11.3	1.9	1.15	1.9	1.3	1.7
GSM_T P18	15.0	2.5	1.17	2.5	1.7	2.1
AGORA	22.6	3.1	1.19	3.1	2.1-2.3	2.5

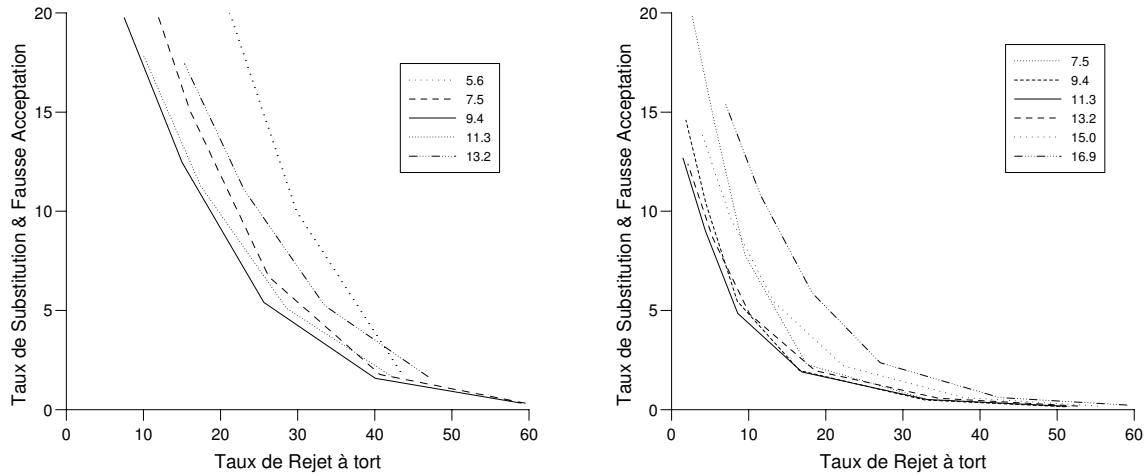
TAB. F.1 – *Seuils optimaux pour la détection sur les différentes bases.*

		Critères		
Bruits	RSB	LCT	SB	SBP
car	10 *dB*	11.3	0.7	1.11
	12.5 *dB*	9.4	1.3	1.09
	15 *dB*	9.4	1.3	1.09
babble	10 *dB*	9.4	1.1	1.09
	12.5 *dB*	9.4	1.1	1.11
	15 *dB*	9.4	1.3	1.11

TAB. F.2 – *Seuils optimaux pour la détection sur la base GSM_A bruitée.*

	Critères					
Bases	LCT	SB	SBP	SB+M3	SB+F_0	SB+VP(MFCC)
RTC_A M20	15.0	2.3	1.13	2.3	2.1	
RTC_A P20	18.8	2.7	1.23	2.7	2.5	
GSM_A M18	9.4	1.7	1.11	1.7	1.3	
GSM_A P18	11.3	1.9	1.09	1.9	1.5	
RTC_T_L	11.3	1.7	1.07	1.7	1.7	1.7
RTC_T_R	13.2	1.7	1.07	1.7	1.7	1.7
GSM_T M18	9.4	1.7	1.09	1.7	1.3	1.5
GSM_T P18	9.4	2.3	1.11	2.1	1.5	1.7
AGORA	16.9	2.3	1.11	2.1	1.5	1.5

TAB. F.3 – *Seuils optimaux pour la reconnaissance sur les différentes bases.*



(a) RSB inférieur à 18 *dB*

(b) RSB supérieur à 18 *dB*

FIG. F.1 – *Résultats de reconnaissance détaillés pour le critère LCT sur la base GSM_A.*

		Critères			
Bruits	RSB	LCT	SB	SBP	SB+F_0
car	10 *dB*	11.3	0.7	1.03	0.5
	12.5 *dB*	9.4	0.7	1.03	0.5
	15 *dB*	7.5	1.1	1.03	0.7
babble	10 *dB*	9.4	1.1	1.07	0.9
	12.5 *dB*	7.5	1.1	1.07	0.9
	15 *dB*	7.5	1.1	1.05	0.9

TAB. F.4 – *Seuils optimaux pour la reconnaissance sur la base GSM_A bruitée.*

		Critères			
Bruits	RSB	LCT	SB	SBP	SB+F_0
car	10 *dB*	9.4	3.5	1.07	1.3
	12.5 *dB*	11.3	3.5	1.09	1.3
	15 *dB*	11.3	3.3	1.07	1.3
babble	10 *dB*	13.2	2.7	1.17	2.1
	12.5 *dB*	13.2	2.9	1.15	2.1
	15 *dB*	13.2	2.9	1.13	2.1

TAB. F.5 – *Seuils optimaux pour la reconnaissance avec réduction de bruit sur la base GSM_A bruitée.*

	Critères			
Bases	LCT	SB	SBP	SB+F_0
GSM_A M18	16.9	2.1	1.03	1.5
GSM_A P18	20.7	2.1	1.01	1.5

TAB. F.6 – *Seuils optimaux pour la reconnaissance avec réduction de bruits sur la base GSM_A.*

Annexe G

Taux d'erreur associée sur les différentes bases

Nous présentons ici les taux d'erreur associée atteints pour les seuils optimaux précédemment donnés. Pour la détection, il s'agit de la somme des taux d'erreur rejetable et définitive, pour la reconnaissance, de la somme du taux d'erreur de rejet à tort et du taux de substitution et fausse acceptation pour les mots isolés, et de la somme des taux de rejet à tort, d' omissions, d'insertion et de substitution de mots pour la parole continue.

Le meilleur critère est donné pour chaque base selon le RSB. Il est également précisé par rapport à quels critères il est significativement meilleur.

G.1 Taux d'erreur associée de détection

L'intervalle de confiance est calculé par rapport au nombre de segments de *Parole-Voc*.

	Taux d'erreur par critère			Intervalle de confiance par critère		
Bases	LCT	SB	SBP	LCT	SB	SBP
RTC_A M20	50.12	53.59	56.11	48.16;52.08	51.63;55.54	54.16;58.04
RTC_A P20	19.72	23.20	24.74	18.40;21.11	21.79;24.67	23.30;26.24
GSM_A M18	29.67	28.77	26.45	28.83;30.52	27.94;29.62	25.64;27.27
GSM_A P18	13.20	12.92	13.22	12.57;13.86	12.29;13.57	12.59;13.88
RTC_T_L	8.84	9.43	9.43	08.17;09.56	08.74;10.17	08.74;10.17
RTC_T_R	9.18	8.38	8.54	08.49;09.92	07.72;09.09	07.87;09.26
GSM_T M18	21.7	21.36	20.29	20.93;22.49	20.60;22.14	19.54;21.06
GSM_T P18	8.48	8.23	8.33	08.04;08.94	07.79;08.69	07.89;08.79
AGORA	16.99	18.97	18.45	15.53;18.45	17.49;20.55	16.98;20.01

TAB. G.1 – *Taux d'erreur associée de détection et intervalle de confiance sur les différentes bases.*

Bases	Meilleur critère	Significatif par rapport à
RTC_A M20	LCT	SB, SBP
RTC_A P20	LCT	SB, SBP
GSM_A M18	SBP	LCT, SB
GSM_A P18	SB	
RTC_T_L	LCT	
RTC_T_R	SB	LCT
GSM_T M18	SBP	LCT, SB
GSM_T P18	SB	
AGORA	LCT	SB, SBP

TAB. G.2 – *Meilleurs critères pour les taux d'erreur associée de détection sur les différentes bases.*

		Taux d'erreur par critère			Intervalle de confiance par critère		
Bases	RSB	LCT	SB	SBP	LCT	SB	SBP
car	10 *dB*	63.10	24.96	29.12	62.17;64.02	24.14;25.80	28.26;30.00
	12.5 *dB*	34.22	18.25	21.82	33.32;35.13	17.52;19.00	21.04;22.62
	15 *dB*	17.72	14.41	17.03	17.00;18.46	13.75;15.09	16.32;17.76
babble	10 *dB*	47.97	20.74	28.07	47.02;48.93	19.98;21.53	27.22;28.94
	12.5 *dB*	25.58	17.21	22.61	24.76;26.42	16.50;17.94	21.82;23.42
	15 *dB*	15.93	15.40	17.93	15.24;16.64	14.72;16.10	17.21;18.67

TAB. G.3 – *Taux d'erreur associée de détection et intervalle de confiance sur la base GSM_A bruitée.*

Bruits	RSB	Meilleur critère	Significatif par rapport à
car	10 *dB*	SB	LCT, SBP
	12.5 *dB*	SB	LCT, SBP
	15 *dB*	SB	LCT, SBP
babble	10 *dB*	SB	LCT, SBP
	12.5 *dB*	SB	LCT, SBP
	15 *dB*	SB	SBP

TAB. G.4 – *Meilleurs critères pour les taux d'erreur associée de détection sur la base GSM_A bruitée.*

G.2 Taux d'erreur associée de la reconnaissance

Le meilleur taux de reconnaissance est donné à poids de rejet constant, qui est de 400 pour les bases de mots isolés, de 0 pour la base AGORA de parole continue, et de 800 pour l'étude des deux bruits *car* et *babble* ajoutés. L'intervalle de confiance est toujours calculé par rapport aux segments de *Parole-Voc* pour les bases de mots isolés, et par rapport au nombre total de mots pour la base AGORA.

	Taux d'erreur par critère			Intervalle de confiance par critère		
Bases	LCT	SB	SBP	LCT	SB	SBP
RTC_A M20	18.00	18.24	21.33	16.54;19.55	16.78;19.80	19.77;22.98
RTC_A P20	10.94	12.54	13.01	9.92;10.05	11.46;13.71	11.91;14.20
GSM_A M18	22.11	21.36	20.70	21.35;22.89	20.61;22.13	19.96;21.46
GSM_A P18	12.18	11.84	11.54	11.57;12.82	11.24;12.47	10.94;12.16
RTC_T_L	13.30	11.40	11.54	12.49;14.15	10.65;12.20	10.78;12.34
RTC_T_R	14.10	11.50	12.17	13.26;14.98	10.73;12.31	11.38;13.00
GSM_T M18	28.23	27.99	27.60	27.39;29.09	27.15;28.84	26.77;28.45
GSM_T P18	16.21	14.38	14.13	15.62;16.82	13.82;14.96	13.57;14.71
AGORA	26.78	27.08	26.81	26.02;27.56	26.31;27.86	26.04;27.59

TAB. G.5 – *Taux d'erreur associée de reconnaissance et intervalle de confiance sur les différentes bases.*

Bases	Meilleur critère	Significatif par rapport à
RTC_A M20	LCT	SBP
RTC_A P20	LCT	SB, SBP
GSM_A M18	SBP	LCT
GSM_A P18	SBP	LCT
RTC_T_L	SB	LCT
RTC_T_R	SB	LCT
GSM_T M18	SBP	
GSM_T P18	SBP	LCT
AGORA	LCT	

TAB. G.6 – *Meilleurs critères pour les taux d'erreur associée de reconnaissance sur les différentes bases.*

		Taux d'erreur par critère			Intervalle de confiance par critère		
Bases	RSB	LCT	SB	SBP	LCT	SB	SBP
car	10 *dB*	54.79	32.53	35.39	53.84;55.74	31.64;33.43	34.48;36.31
	12.5 *dB*	39.02	24.72	26.68	38.09;39.96	23.90;25.55	25.84;27.53
	15 *dB*	23.18	21.23	23.59	22.38;24.00	20.46;22.02	22.79;24.41
babble	10 *dB*	47.66	34.21	41.02	46.71;48.62	33.31;35.12	40.08;41.96
	12.5 *dB*	33.05	27.39	30.24	32.16;33.96	26.55;28.55	29.37;31.12
	15 *dB*	20.51	18.82	23.27	19.75;21.29	18.08;19.58	22.47;24.09

TAB. G.7 – *Taux d'erreur associée de reconnaissance et intervalle de confiance sur la base GSM_A bruitée.*

		Meilleur	Significatif
Bruits	RSB	critère	par rapport à
car	10 *dB*	SB	LCT, SBP
	12.5 *dB*	SB	LCT, SBP
	15 *dB*	SB	LCT, SBP
babble	10 *dB*	SB	LCT, SBP
	12.5 *dB*	SB	LCT, SBP
	15 *dB*	SB	LCT,SBP

TAB. G.8 – *Meilleurs critères pour les taux d'erreur associée de reconnaissance sur la base GSM_A bruitée.*

Annexe H

Sensibilité des différents critères

Nous présentons ici, pour les différents critères du module de détection ABP, la sensibilité du réglage du seuil de détection au changement de base (avec les bases GSM_A et GSM_T), au niveau de bruit (avec les bases GSM_A pour les RSB inférieurs et supérieurs à 18 *dB* et RTC_A pour les RSB inférieurs et supérieurs à 20 *dB*), et au changement de réseau d'appel (avec les bases GSM_T pour un RSB supérieur à 18 *dB* et RTC_T_R). Le choix de ces deux dernières bases a été fait en raison du recouvrement important du vocabulaire, le mode répété d'enregistrement et le RSB supérieur à 18 *dB* présente un niveau de bruit comparable à la base RTC_T_R.

H.1 Sensibilité au changement de base

Nous présentons les résultats de reconnaissance sur la base GSM_T avec un seuil de rejet de 400. L'appartenance du taux d'erreur associée de reconnaissance obtenu avec le seuil "optimal" de la base GSM_T à l'intervalle de confiance à 95% du taux d'erreur associée de reconnaissance obtenu avec le seuil "optimal" de la base GSM_A, fournit un critère de non-sensibilité au changement de base.

		Critères			
		LCT	SB	SBP	SB+F_0
GSM_T M18	Intervalle	27.39;29.09	27.15;28.84	27.02;28.71	26.29;27.97
	Taux d'erreur	28.23	27.99	27.60	27.12
	Sensibilité	Non	Non	Non	Non
GSM_T P18	Intervalle	14.07;15.23	14.12;15.28	13.96;15.10	12.90;14.02
	Taux d'erreur	16.21	14.38	14.13	13.45
	Sensibilité	Oui	Non	Non	Non

TAB. H.1 – *Sensibilité des différents critères au changement de base de la base GSM_A à la base GSM_T.*

H.2 Sensibilité au niveau de bruit

Nous présentons les résultats de reconnaissance d'une part sur la base GSM_A d'autre part sur la base RTC_A, avec un seuil de rejet de 400. L'appartenance du taux d'erreur associée de reconnaissance obtenu avec le seuil "optimal" de la partie bruitée de la base GSM_A à l'intervalle de confiance du taux d'erreur associée de reconnaissance obtenu avec le seuil "optimal" de la partie calme base GSM_A, et inversement, fournit un critère de non-sensibilité au niveau de bruit.

De même pour la base RTC_A, sont comparés les résultats obtenus pour la partie de la base ayant un RSB inférieur à 20 *dB*, et pour la partie de la base ayant un RSB supérieur à 20 *dB*.

		Critères			
		LCT	SB	SBP	SB+F_0
GSM_A P18	Intervalle	12.95;14.26	11.78;13.04	10.90;12.12	10.82;12.83
	Taux d'erreur	12.18	11.84	11.54	11.22
	Sensibilité	Oui	Non	Non	Non
GSM_A M18	Intervalle	21.71;23.25	20.95;22.47	20.24;21.74	19.55;21.03
	Taux d'erreur	22.11	21.36	20.70	19.91
	Sensibilité	Non	Non	Non	Non
RTC_A P20	Intervalle	13.60;16.02	13.77;16.19	17.42;20.07	10.78;12.98
	Taux d'erreur	10.94	12.54	13.01	10.73
	Sensibilité	Oui	Oui	Oui	Oui
RTC_A M20	Intervalle	12.53;15.24	14.60;17.47	15.38;18.30	11.83;14.47
	Taux d'erreur	18.00	18.24	21.33	13.34
	Sensibilité	Oui	Oui	Oui	Non

TAB. H.2 – *Sensibilité des différents critères au niveau de bruit sur les bases GSM_A et RTC_A.*

H.3 Sensibilité au réseau d'appel

Nous présentons ici les résultats de reconnaissance sur la base GSM_T avec un RSB supérieur à 18 *dB* et sur la base RTC_T_R, avec un seuil de rejet de 400. L'appartenance du taux d'erreur associée de reconnaissance obtenu avec le seuil "optimal" de la partie calme de la base GSM_T et de la base RTC_T_R, à l'intervalle de confiance du taux d'erreur associée de reconnaissance obtenu avec le seuil "optimal" respectivement de la base RTC_T_R et de la partie calme de la base GSM_T, fournit un critère de non-sensibilité au changement de réseau d'appel (RTC et GSM).

		Critères			
		LCT	SB	SBP	SB+F_0
GSM_T P18	Intervalle	13.66;14.80	14.91;16.09	14.63;15.80	13.05;14.17
	Taux d'erreur	16.21	14.38	14.13	13.45
	Sensibilité	Oui	Oui	Oui	Non
RTC_T_R	Intervalle	19.72;21.72	14.04;15.80	11.89;13.54	13.52;15.26
	Taux d'erreur	14.10	11.50	12.17	12.10
	Sensibilité	Oui	Oui	Non	Oui

TAB. H.3 – *Sensibilité des différents critères au réseau d'appel sur la partie calme de la base GSM_ T et sur la base RTC_ T_ R.*

Annexe I

Résultats de différentes intégrations d'une nouvelle condition

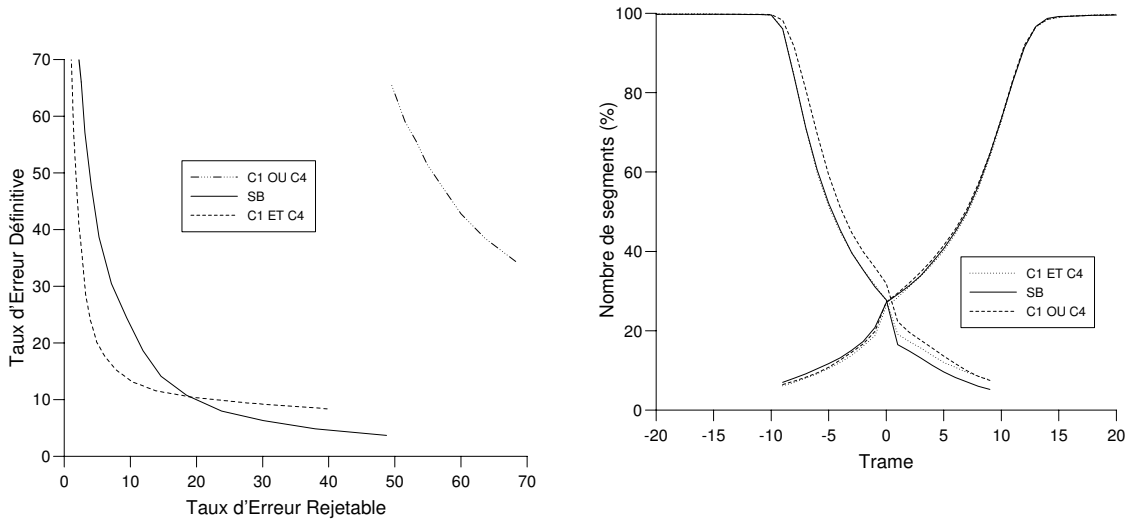
Nous présentons ici les résultats obtenus avec les différentes intégrations d'une nouvelle condition C4 du Chapitre 6 "*Intégration d'une nouvelle condition dans l'automate*". Les résultats sont présentés pour le critère SB+$F0$ présenté au Chapitre 8 "*Utilisation d'un paramètre de voisement*". Nous étudions ici uniquement l'influence que l'intégration de la condition C4 a sur les résultats de détection sur la base GSM_A qui comprend des fichiers ayant un RSB inférieur à 18 *dB*. Pour ce faire, nous représentons d'une part les erreurs de détection (les erreurs définitives en fonction des erreurs rejetables), d'autre part le placement des frontières pour le seuil qui donne le minimum des taux d'erreur associée (somme du taux d'erreur définitive et du taux d'erreur rejetable).

La figure I.1 montre que l'intégration de la condition C4 avec l'opérateur *et*, avec la règle C1 *et* C4, améliore le critère SB au niveau des erreurs définitives et rejetables (*cf.* figure I.1(a)). En revanche, il n'y a pas de différences avec le critère SB au niveau du positionnement des frontières.

L'opérateur *ou* avec le règle C1 *ou* C4 dégrade fortement les résultats. En effet, l'information du paramètre de voisement ne doit être utilisée que pour des périodes énergétiques.

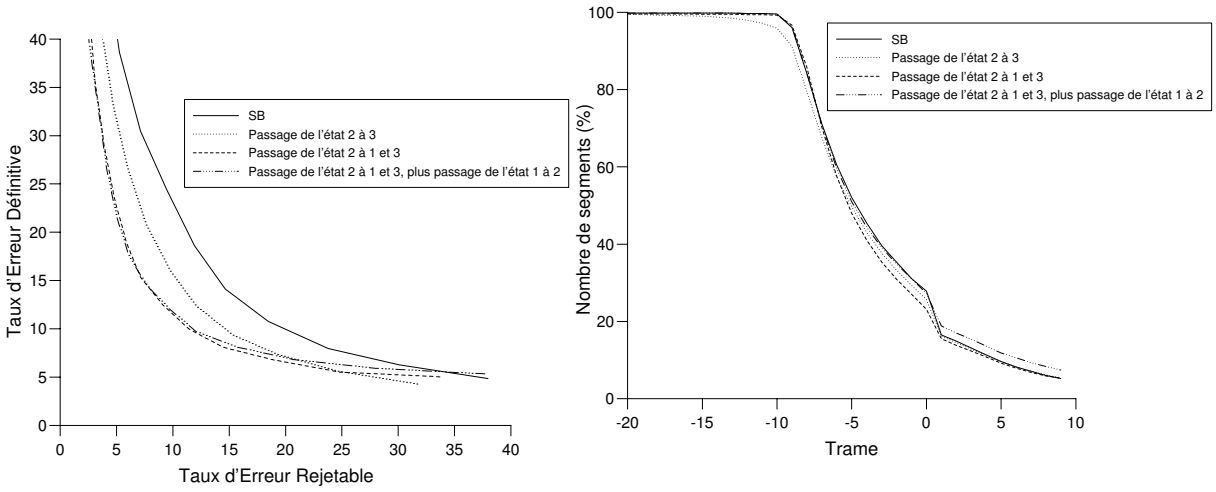
La figure I.2 présente les résultats de l'intégration de la nouvelle condition au niveau de l'état *présomption de parole*. Le passage uniquement de l'état *présomption de parole* (état 2) à l'état *parole* (état 3) améliore les résultats du critère SB. Cependant en rajoutant la nouvelle condition pour le passage de l'état *présomption de parole* à l'état *bruit ou silence* (état 1), les résultats de détection (erreurs rejetables et définitives) sont encore améliorés, et de plus la détection de la frontière gauche est légèrement meilleure. L'ajout de la condition pour le passage de l'état *bruit ou silence* à l'état *présomption de parole* ne fait dégrader que légèrement les résultats précédents.

Pour diminuer les erreurs de détection de fin de mot ou requête (*cf.* figure I.3), l'intégration peut se faire au niveau de l'état *reprise possible de parole*. L'ajout de la nouvelle condition pour toutes les transitions au niveau de cet état, avec les opérateurs logiques *et* et *ou* par les règles C4 *et* C1, et non C4 *ou* non C1, n'apporte pas d'amélioration, et dégrade même pour les taux d'erreur rejetable supérieurs à 10%. Par contre, si la condi-



(a) Résultats de détection (b) Positionnement des frontières

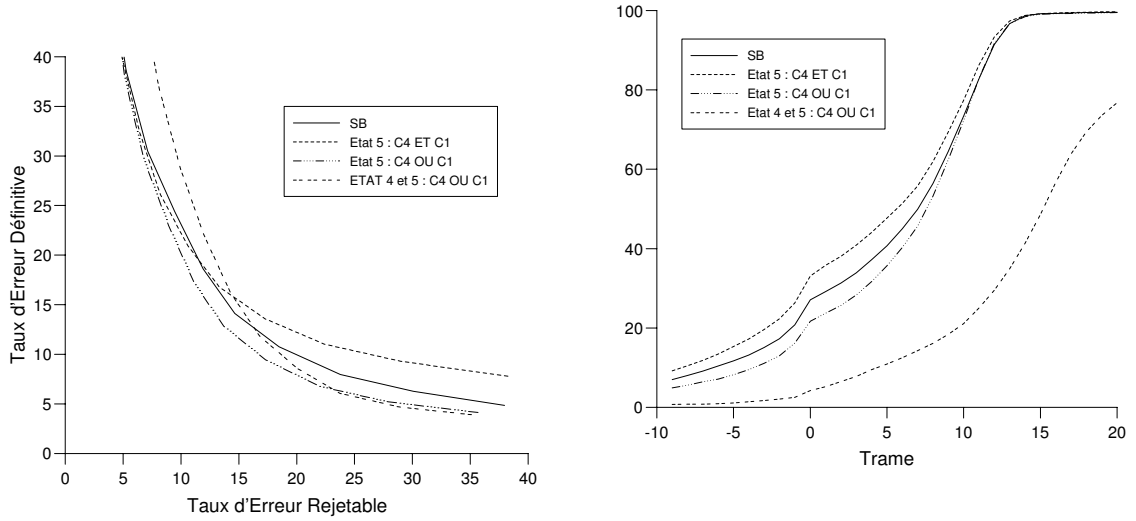
FIG. I.1 – *Condition du critère SB+F0 sur toutes les transitions avec les opérateurs logiques "et" et "ou".*



(a) Résultats de détection (b) Positionnement des frontières droites

FIG. I.2 – *Condition du critère SB+F0 au niveau de l'état "présomption de parole".*

tion est considérée au niveau du passage de l'état *reprise possible de parole* à l'état *parole* par la règle C4 *ou* C1, une amélioration par rapport au critère SB est observée. De plus les détections à droite sont légèrement moins tronquées. Si la nouvelle condition est de plus ajoutée pour le passage de l'état *plosive non voisée ou silence* à l'état *reprise possible de parole* avec la règle C4 *ou* C1, les détections à droite sont beaucoup moins tronquées, mais il y a alors plus de détections élargies à droite qui peuvent entraîner des erreurs de reconnaissance. De plus les erreurs de détection sont plus importantes que pour le critère SB.



(a) Résultats de détection

(b) Positionnement des frontières gauches

FIG. I.3 – *Condition du critère SB+F0 au niveau de l'état "reprise possible de parole" et de l'état "plosive non voisée ou silence.*

La figure I.4 permet de comparer les meilleures intégrations du critères SB+F_0 au niveau des différents passages d'un état à l'autre.

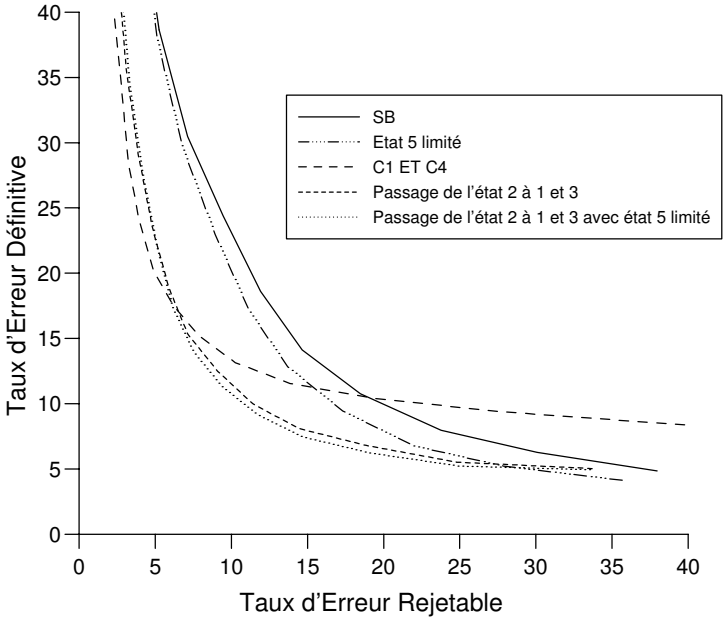


FIG. I.4 – *Comparaison des meilleures intégrations du critère SB+F0.*

Annexe J

Résultats du critère SB+F_0 avec débruitage

Les résultats du critère SB+F_0 ont montré que ce critère est le plus performant parmi ceux présentés dans ce travail. Afin de confirmer ces résultats, nous présentons ici les résultats du critère SB+F_0 avec le module de débruitage présenté au Chapitre 5 "*Méthode de débruitage*".

Nous comparons ce critère au critère SB, sur la base GSM_A débruitée, ainsi que sur cette même base avec l'ajout des deux bruits *car* et *babble*. Le protocole d'évaluation reste inchangé. Nous décrivons tout d'abord au paragraphe J.1 les résultats de détection puis les résultats de reconnaissance au paragraphe J.2.

J.1 Résultats de détection

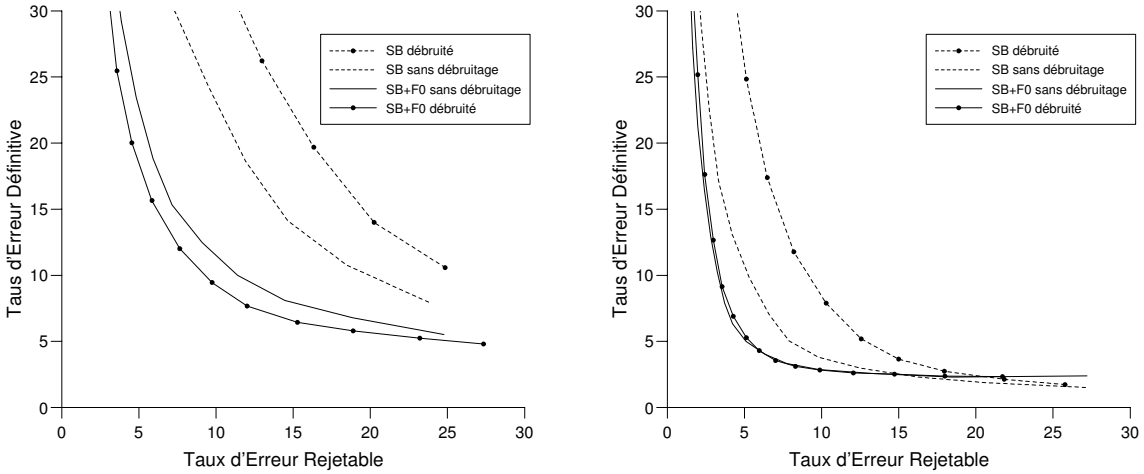
Nous présentons dans un premier temps au paragraphe J.1.1, les résultats de détection après débruitage de la base GSM_A, selon le RSB, puis dans un second temps nous décrivons au paragraphe J.1.2 les résultats de détection après débruitage sur la partie calme de la base GSM_A avec un ajout préalable de bruit. L'ajout de bruit et le module de débruitage sont similaires à l'étude réalisée au Chapitre 5 "*Méthode de débruitage*".

J.1.1 Débruitage de la base GSM_A

Les figures J.1(a) et J.1(b) donnent les résultats de détection du critère SB+F_0 et du critère SB avec et sans le module de débruitage sur la base GSM_A, pour un RSB inférieur à 18 *dB* et supérieur à 18 *dB* respectivement.

Nous constatons que le critère SB+F_0 est toujours meilleur que le critère SB, avec ou sans débruitage. Nous avons constaté que le critère SB est moins performant sur le signal débruité que sur le signal original. Dans le cas du critère SB+F_0 les résultats sont meilleurs sur le signal débruité que sur le signal original, particulièrement pour la partie de la base la plus bruitée. Ceci s'explique par le fait que les bruits impulsifs accentués par le module de débruitage ne sont pas détectés par le critère SB+F_0 qui diminue les

détections de bruits. L'amélioration du critère SB+F_0 est donc toujours significative sur les deux parties de la base GSM_A débruitée.



(a) RSB inférieur à 18 *dB*

(b) RSB supérieur à 18 *dB*

FIG. J.1 – *Résultats de détection des critères SB+F_0 et SB sur la base GSM_A avec et sans débruitage.*

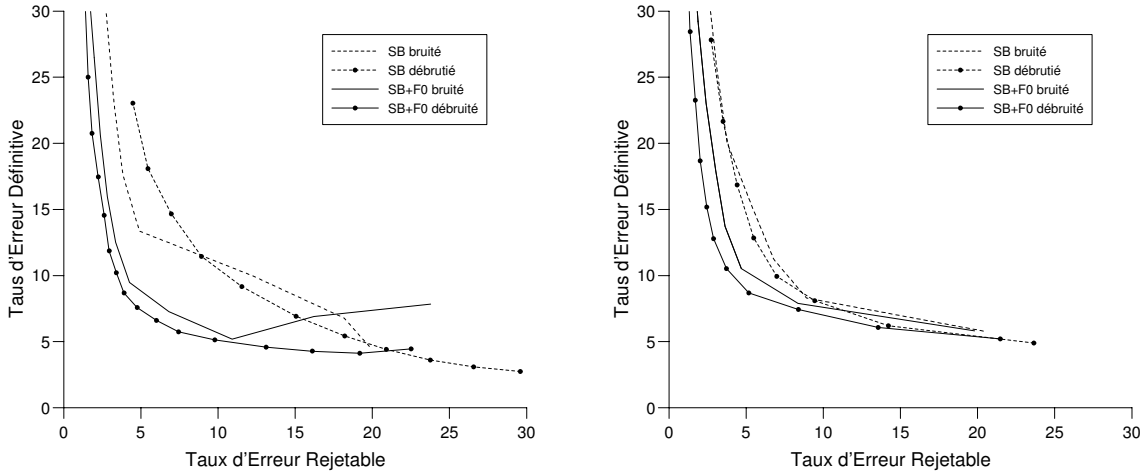
J.1.2 Débruitage de la base GSM_A après ajout de bruits

Les figures J.2(a) et J.2(b) présentent les résultats de détection avec un ajout des bruits *car* et *babble* à $12.5dB$ sur la partie calme de la base GSM_A, puis débruitée. Le module de débruitage est conçu pour de tels bruits stationnaires. Le critère SB+F_0 donne de meilleurs résultats sur le signal débruité que sur le signal bruité, et l'amélioration est toujours significative par rapport au critère SB.

Ainsi, les résultats de détection montrent qu'il peut être intéressant d'employer le module de débruitage avec le critère SB+F_0 plutôt qu'avec le critère SB. En effet les bruits impulsifs détectés par la DAV du débruitage et ensuite amplifiés, ne sont pas détectés par le critère SB+F_0. De plus la suppression du bruit de fond apporte toujours une amélioration de la détection de la parole.

J.2 Résultats de reconnaissance

Nous comparons dans ce paragraphe les résultats du système de reconnaissance avec les deux critères SB et SB+F_0, dans un premier temps après débruitage de la base GSM_A,



(a) Bruit *car*

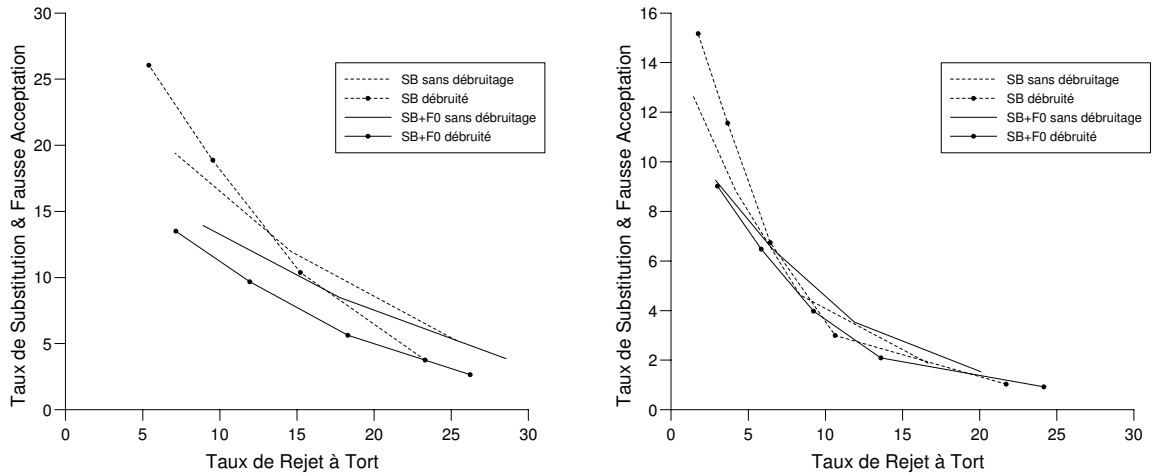
(b) Bruit *babble*

FIG. J.2 – *Résultats de détection des critères SB+F_0 et SB sur la base GSM_A bruitée avec et sans débruitage.*

selon le RSB, et dans un second temps après débruitage de la base GSM_A avec ajout des bruits *car* et *babble*.

J.2.1 Débruitage de la base GSM_A

Les figures J.3(a) et J.3(b) présentent les résultats de reconnaissance pour les critères SB et SB+F_0 avant et après débruitage de la base GSM_A, pour les enregistrements avec un RSB respectivement inférieur à 18 *dB* et supérieur à 18 *dB*. Nous constatons que les résultats avec le critère SB+F_0 sont meilleurs sur le signal débruité que sur le signal original. Cependant l'amélioration sur le signal débruité n'est pas significative. De plus, l'amélioration du critère SB+F_0 par rapport au critère SB sur le signal débruité reste très faible et n'est pas significative. Nous retrouvons les mêmes types de résultats que sur la base GSM_T du Chapitre 8 "*Utilisation d'un paramètre de voisement*". C'est-à-dire une amélioration significative des résultats de détection qui n'est pas suivie d'une amélioration importante des résultats de reconnaissance. Ceci s'explique toujours par le fait que les détections de bruits impulsifs qui ne sont plus détectés par le critère SB+F_0, le sont par critère SB, mais sont ensuite rejetés par le module de reconnaissance. Rappelons que cette baisse des détections des bruits impulsifs reste intéressante pour réduire le coût du système de reconnaissance.



(a) RSB inférieur à 18 *dB*

(b) RSB supérieur à 18 *dB*

FIG. J.3 – *Résultats de reconnaissance des critères SB+F_0 et SB sur la base GSM_A avec et sans débruitage.*

J.2.2 Débruitage de la base GSM_A après ajout de bruits

Nous comparons dans ce paragraphe les résultats du système de reconnaissance du critère SB+F_0 avec ceux du critère SB sur la partie calme de la base GSM_A avec les bruits *car* et *babble* ajoutés à 12.5 *dB*, avec et sans débruitage (*cf.* figures J.4(a) et J.4(b)).

Nous remarquons que les résultats de reconnaissance avec le critère SB+F_0 et le critère SB sont similaires avec et sans le débruitage pour ces deux bruits stationnaires, alors que les résultats de détection sont significativement meilleurs avec le critère SB+F_0. Ceci s'explique comme précédemment par le fait que les bruits impulsifs contenus dans la partie calme de la base GSM_A sont rejetés par le module de reconnaissance.

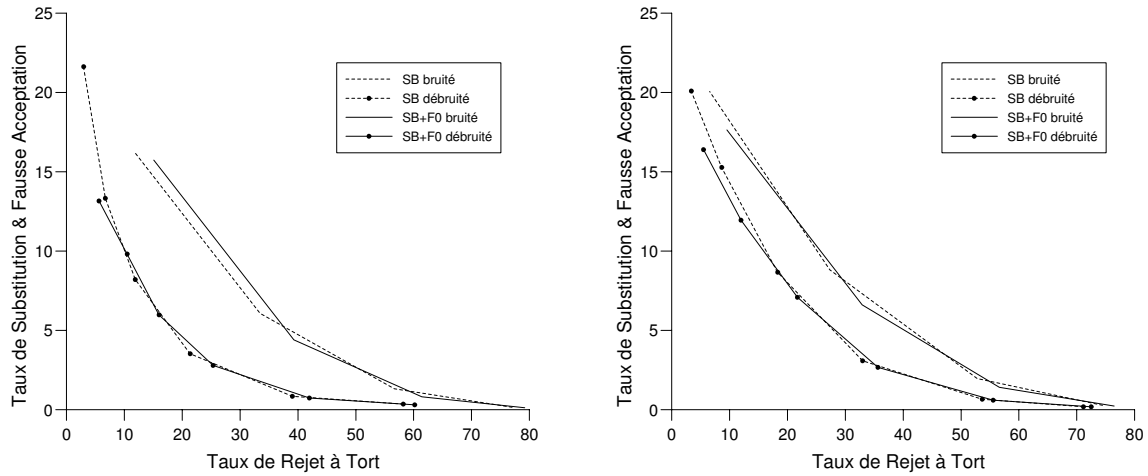
L'amélioration donnée par le module de débruitage avec le critère SB+F_0 est significative.

Ainsi, dans le cas de bruits stationnaires tels que ceux étudiés ici, le module de débruitage apporte aussi une amélioration des performances pour le critère SB+F_0.

J.3 Conclusion

Le module de débruitage apporte avec le critère SB+F_0 des résultats intéressants tant au niveau du module de détection qu'au niveau du système de reconnaissance.

Les résultats de détection avec le critère SB+F_0 restent meilleurs que ceux du critère SB sur le signal débruité. De plus les résultats de détection ne sont pas dégradés avec



(a) Bruit *car* (b) Bruit *babble*

FIG. J.4 – *Résultats de reconnaissance des critères SB+F_0 et SB sur la base GSM_A bruitée avec et sans débruitage.*

le critère SB+F_0 sur le signal débruité, alors que c'est le cas avec le critère SB pour des bruits impulsifs.

Les résultats de reconnaissance du critère SB+F_0 ne sont pas significativement meilleurs que ceux du critère SB sur le signal débruité. Cependant, l'amélioration avec le critère SB+F_0 sur le signal débruité est significative pour les bruits stationnaires.

Le critère SB+F_0 présente donc les meilleures performances avec et sans le module de débruitage.

Bibliographie

[Abdallah *et al.*, 1997a] Abdallah (I.), Montrésor (S.) et Baudry (M.). – Speech Signal Detection in Noisy Environment Using a Locoal Entropic Criterion. *European Conference on Speech Communication and Technology*, vol. 5, pp. 2595–2598. – Rhodes, Grèce, septembre 1997.

[Abdallah *et al.*, 1997b] Abdallah (I.), Montrésor (S.) et Baudry (M.). – Un algorithme Récursif pour la Segmentation des Signaux de Parole Basé sur un Critère Entropique Local. $4^{ième}$ *Congrès de la Société Française d'Acoustique*, pp. 85–88. – Marseille, France, avril 1997.

[Acero *et al.*, 1993] Acero (a.), Crespo (C.), De La Torre (C.) et Torrecilla (J.C.). – Robust HMM-Based Endpoint Detector. *European Conference on Speech Communication and Technology*, vol. 3, pp. 1551–1554. – Berlin, Allemagne, septembre 1993.

[Agaiby et Moir, 1997] Agaiby (H.) et Moir (T.J.). – Knowing the Wheat from the Weeds in Noisy Speech. *European Conference on Speech Communication and Technology*, vol. 3, pp. 1119–1122. – Rhodes, Grèce, septembre 1997.

[André-Obrecht *et al.*, 1993] André-Obrecht (R.), Puel (J.B.) et Eichene (S.). – Détection des débuts et fin de parole en environnement difficile. $14^{ième}$ *Colloque GRETSI*, pp. 157–160. – Juan-Les-Pins, France, septembre 1993.

[Ariyoshi, 2000] Ariyoshi (T.). – *Integrated Endpoint Detection for Improved Speech Recognition Method and System.* – Brevet Américain, n°US6029130, février 2000.

[Arslan et Hansen, 1998] Arslan (L.M.) et Hansen (J.H.L.). – Likelihood Decision Boundary Estimation Between HMM Pairs in Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, vol. 6, n° 4, juillet 1998, pp. 410–414.

[Atal et Rabiner, 1976] Atal (B.S.) et Rabiner (L.R.). – A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classsification with Applications to Speech Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, n° 3, juin 1976, pp. 201–212.

[Bagshaw, 1994] Bagshaw (P.). – *Automatic Prosodic Analysis for Computer Aided Pronunciation Teaching.* – Thèse de doctorat, Université d'Edinburgh, 1994.

[Bartkova, 1999] Bartkova (K.). – *Production, Description, Perception du signal vocal.* – Rapport technique, cours DEA, 1999.

[Batlle *et al.*, 1998] Batlle (E.), Nadeu (C.) et Fonollosa (J.A.R.). – Feature Decorrelation Methods In Speech Recognition A Comparative Study. *International Conference on Spoken Language Processing*, vol. 3, pp. 951–954. – Sydney, Australie, décembre 1998.

[Bendiksen et Steiglitz, 1990] Bendiksen (A.) et Steiglitz (K.). – Neural Networks for Voiced/Unvoiced Speech Classification. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 521–524. – Albuquerque, New-Mexico, États-Unis, avril 1990.

[Beritelli *et al.*, 1999] Beritelli (F.), Casale (S.) et Cavallaro (A.). – A Multi-Chanel Speech/Silence Detector based on Time Delay Estimation and Fuzzy Classification. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 93–96. – Phoenix, Arizona, États-Unis, mai 1999.

[Bouteille, 1999] Bouteille (F.). – *Etude et mise en œuvre de dispositifs de contrôle de l'écho acoustique pour terminaux mains-libres et applications multimédia.* – Diplôme de Recherche Technologique, Université de Rennes 1, 1999.

[Braun *et al.*, 1990] Braun (H.J.), Cosier (G.), Freeman (D.), Gilloire (A.), Sereno (D.), Southcott (C.B.) et Van der Krogt (A.). – *Voice control of the Pan-European digital mobile radio system.* – Rapport technique n° 3, CSELT, juin 1990.

[Breiman *et al.*, 1993] Breiman (L.), Friedman (J.H.), Ohlsen (R.A.) et Stone (C.J.). – *Classification And Regression Trees.* – Chapman & Hall, 1993.

[Bruno *et al.*, 1987] Bruno (G.), Di Benedetto (M.D.), Gilio (A.) et Mandarini (P.). – A Bayesian-Adaptive Decision Method for the V/UV/S Classification of Segments of a Speech Signal. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, n° 4, avril 1987, pp. 556–559.

[Calliope, 1989] Calliope. – *La parole et son traitement automatique.* – Masson, 1989.

[Cavallaro *et al.*, 1998] Cavallaro (A.), Beritelli (F.) et Casale (S.). – A Fuzzy Logic-Based Speech Detection Algorithm for Communications in Noisy Environments. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 565–568. – Seattle, Washington, États-Unis, mai 1998.

[Celeux *et al.*, 1989] Celeux (G.), Diday (E.), Govaert (G.), Lechevallier (Y.) et Ralambondrainy (H.). – *Classification automatique des données.* – Dunod, 1989.

[Charlet, 1997] Charlet (D.). – *Authentification Vocale par Téléphone en Mode Dépendant du Texte.* – Thèse de doctorat, École Nationale Supérieure des Télécommunications, 1997.

[Cho et Un, 1982] Cho (D.H.) et Un (C.K.). – Hybrid Companding Delta Modulation with Silence Detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1982, pp. 1340–1344.

[Cohn, 1991] Cohn (R.P.). – Robust Voiced/Unvoiced Speech Classification Using a Neural Net. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 437–440. – Toronto, Canada, mai 1991.

[Cox et Timothy, 1980] Cox (B.V.) et Timothy (L.M.K.). – Nonparametric Rank-Order Statistics Applied to Robust Voiced-Unvoiced-Silence Classification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, n° 5, octobre 1980, pp. 550–561.

[Daaboul et Adoul, 1977] Daaboul (F.) et Adoul (J.P.). – Parameter Segmentation of Speech into Voiced-Unvoiced-Silence Intervals. *International Conference on Acoustics, Speech, and Signal Processing*, pp. 327–331. – 1977.

[Damnati, 2000] Damnati (G.). – *Modèles de Langage et Classification Automatique pour la Reconnaissance de la Parole Continue dans un contexte de Dialogue Oral Homme-Machine.* – Thèse de doctorat, Université d'Avignon et des Pays du Vaucluse, 2000.

[De Souza, 1983] De Souza (P.). – A Statistical Approach to the Design of an Adaptative Self-Normalizing Silence Detector. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, n° 3, juin 1983, pp. 678–684.

[Delphin-Poulat, 1999] Delphin-Poulat (L.). – *Utilisation de Modèles de Markov Cachés pour une Compensation Synchrone à la Trame, dans un Contexte de Reconnaissance de la Parole.* – Thèse de doctorat, Université de Rennes 1, 1999.

[Depambour *et al.*, 1997] Depambour (P.), André-Obrecht (R.) et Delyon (B.). – On the Use of Phone Duration and Segmental Processing to Label Speech Signal. *European Conference on Speech Communication and Technology*, vol. 3, pp. 1627–1630. – Rhodes, Grèce, septembre 1997.

[Dermatas *et al.*, 1991] Dermatas (E.S.), Fakotakis (N.D.) et Kokkinakis (G.K.). – Fast Endpoint Detection Algorithm for Isolated Word Recognition Detector. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 733–736. – Toronto, Canada, mai 1991.

[Di Francesco, 1990] Di Francesco (R.J.). – Real-Time Speech Segmentation Using Pitch and Convexity Jump Models: Application to Variable Rate Speech Coding. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, n° 5, mai 1990, pp. 741–748.

[Doukas *et al.*, 1997] Doukas (N.), Naylor (P.) et Stathaki (T). – Voice Activity Detection Using Source Separation Techniques. *European Conference on Speech Communication and Technology*, vol. 3, pp. 1099–1102. – Rhodes, Grèce, septembre 1997.

[Ephraim et Malah, 1984] Ephraim (Y.) et Malah (D.). – Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, n° 6, décembre 1984, pp. 1109–1121.

[Faucon *et al.*, 1993] Faucon (G.), Le Bouquin (R.) et Akbari Azirani (A.). – Mesures Objectives de la Réduction de Bruit. $14^{ième}$ *Colloque GRETSI*, pp. 187–590. – Juan-Les-Pins, France, septembre 1993.

[Freeman *et al.*, 1989] Freeman (D.K.), Cosier (G.) et Southcott, C.B. Boyd (I.). – The Voice Activity Detector for the Pan-European Digital Cellular Mobile Telephone service. *International Conference on Acoustics, Speech, and Signal Processing*, pp. 369–373. – Glasgow, Royaume-Uni, mai 1989.

[Ganapathiraju *et al.*, 1996] Ganapathiraju (A.), Webster (L.), Trimble (J.), Bush (K.) et Kornman (P.). – Comparison of Energy-Based Endpoint Detection for Speech Signal Processing. *IEEE Southeastcon*, pp. 500–503. – Floride, États-Unis, avril 1996.

[Ghiselli-Cripa et El-Jaroudi, 1991] Ghiselli-Cripa (T.) et El-Jaroudi (A.). – A Fast Net Training to Voiced-Unvoiced-Silence Classification of Speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1991, pp. 441–444.

[Hahn et Park, 1992] Hahn (M.) et Park (C.K.). – An Improved Speech Detection Algorythm for Isolated Korean Utterances. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 525–528. – San Francisco, Californie, États-Unis, mars 1992.

[Haigh et Mason, 1993] Haigh (J.A.) et Mason (J.S.). – A Voice Activity Detector Based on Cepstral Analysis. *European Conference on Speech Communication and Technology*, vol. 2, pp. 1103–1106. – Berlin, Allemagne, septembre 1993.

[Haltsonen, 1984] Haltsonen (S.). – An Endpoint Relaxation Method for Dynamic Time Warping Algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984, pp. 9.8.1–9.8.4.

[Hamada *et al.*, 1990] Hamada (M.), Takizawa (Y.) et Norimatsu (T.). – A Noise Robust Speech Recognition System. *International Conference on Spoken Language Processing*, vol. 2, pp. 893–896. – 1990.

[Hanel et Jouvet, 2000] Hanel (S.) et Jouvet (D.). – Detecting the End of Spellings using Statistics on Recognized Letter Sequences for Spelled Names Recognition. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1755–1758. – Istambul, Turquie, mai 2000.

[Hariharan *et al.*, 2001] Hariharan (R.), Häkkinen (J.) et Laurila (K.). – Robust End-Of-Utterance Detection for Real-Time Speech Recognition Applications. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 249–252. – Salt Lake City, Utah, États-Unis, mai 2001.

[Héon *et al.*, 1998] Héon (M.), Tolba (H.) et O'Shaughnessy (D.). – Robust Automatic Speech Recognition by the Application of a Temporal-correlation-based Recurrent Multilayer Neural Network to the Mel-based Cepstral Coefficients. *International Conference on Spoken Language Processing*, vol. 4, pp. 1459–1462. – Sydney, Australie, décembre 1998.

[Hess, 1983] Hess (W.). – *Pitch Determination of Speech Signal.* – Springer-Verlag, 1983.

[Hirose et Iwano, 1997] Hirose (K.) et Iwano (K.). – A Method of Representing Fundamental Frequency Contours of Japanese Using Statistical Models of Moraic Transition. *European Conference on Speech Communication and Technology*, vol. 3, pp. 1431–1439. – Rhodes, Grèce, septembre 1997.

[Hörmann et Rozinaj, 1998] Hörmann (T.) et Rozinaj (G.). – *Start/End Point Detection for Word Recognition.* – Brevet Américain, n°US5794195, août 1998.

[Hsieh, 1998] Hsieh (C.-K.). – *Endpoint Detection in a Stand-Alone Real-Time Voice Recognition System.* – Brevet Américain, n°US5845092, décembre 1998.

[Huang et Yang, 2000] Huang (l.-S.) et Yang (C.-H.). – A Novel Approach to Robust Speech Endpoint Detection in Car Environments. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1751–1754. – Istambul, Turquie, mai 2000.

[ITU Recommendation, 1996] ITU Recommendation (G.729). – *Coding of speech at 8kbit/s using conjugate structure algebraic-code-excited linear-prediction (CS-ACELP).* – International Telecommunication Union, 1996.

[Iwano et Hirose, 1998] Iwano (K.) et Hirose (K.). – Representing Prosodic Words Using Statistical Models of Moraic Transition of Fondamental Frequency Contours of Japanese. *International Conference on Spoken Language Processing*, vol. 3, pp. 599–602. – Sydney, Australie, décembre 1998.

[Iwano et Hirose, 1999] Iwano (K.) et Hirose (K.). – Prosodic Word Boundary Detection Using Statistical Modeling of Moraic Fundamental Frequency Contours and its Use for Continuous Speech Recognition. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 133–136. – Phoenix, Arizona, États-Unis, mai 1999.

[Jacovitti *et al.*, 1991] Jacovitti (G.), Pierucci (P.) et Falaschi (A.). – Speech Segmentation and Classification Using Higer Order Moments. *European Conference on Speech Communication and Technology*, pp. 1371–1374. – Gènes, Italie, septembre 1991.

[Jouvet, 1988] Jouvet (D.). – *Reconnaissance de Mots Connectés Indépendamment du Locuteur par des Méthodes Statistiques.* – Thèse de doctorat, École Nationale Supérieure des Télécommunications, 1988.

[Junqua *et al.*, 1991] Junqua (J.-C.), Reaves (B.) et Mak (B.). – A Study of Endpoint Detection Algorithms in Adverse Conditions: Incidence on a DTW and HMM Recognizer. *European Conference on Speech Communication and Technology*, vol. 3, pp. 1371–1374. – Gènes, Italie, septembre 1991.

[Junqua *et al.*, 1994] Junqua (J-C), Mak (B.) et Reaves (B.). – A Robust Algorithm for Word Boundary Detection in the Presence of Noise. *IEEE Transactions on Speech and Audio Processing*, vol. 2, n° 3, juillet 1994, pp. 406–412.

[Karray et Martin, 2001] Karray (L.) et Martin (A.). – Towards Improving Speech Detection Robustness for Speech Recognition in Adverse Conditions. *soumis à Speech Communications*, 2001.

[Karray et Monné, 1998] Karray (L.) et Monné (J.). – Robust Speech/Non-Speech Detection in Adverse Conditions Based on Speech Statistics. *International Conference on Spoken Language Processing*, vol. 4, pp. 1471–1474. – Sydney, Australie, décembre 1998.

[Karray, 1998a] Karray (L.). – *Estimation des Statistiques du Bruit et de la Parole pour une Détection Bruit/Parole plus Robuste.* – Rapport technique n° 8, DT/DIH/DIPS/285, avril 1998.

[Karray, 1998b] Karray (L.). – *Nouveau Critère pour l'Automate de Détection Bruit/Parole.* – Rapport technique n° 3, DT/DIH/DIPS/48, janvier 1998.

[Kobatake *et al.*, 1989] Kobatake (H.), Tawa (K.) et Ishida (A.). – Speech/Non-Speech Discrimination for Speech Recognition System under Real Life Noise Environments. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 365–368. – Glasgow, Royaume-Uni, mai 1989.

[Kuroiwa *et al.*, 1999] Kuroiwa (S.), Naito (M.), Yamamoto (S.) et Higuchi (N.). – Robust Speech Detection Method for Telephone Speech Recognition System. *Speech Communication*, vol. 27, n° 2, 1999, pp. 135–148.

[Lacoume *et al.*, 1997] Lacoume (J.-L.), Amblard (P.-O.) et Comon (P.). – *Statistiques d'ordre supérieur pour le traitement du signal.* – Masson, 1997.

[Lamel *et al.*, 1981] Lamel (L.F.), Rabiner (L.R.), Rosenberg (A.E.) et Wilpon (J.G.). – An Improved Endpoint Detector for Isolated Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, n° 4, août 1981, pp. 777–785.

[Le Bouquin-Jeannès et Faucon, 1995] Le Bouquin-Jeannès (R.) et Faucon (G.). – Study of a Voice Activity Detector and its Influence on a Noise Reduction System. *Speech Communication*, vol. 16, n° 3, 1995, pp. 245–254.

[Lebart *et al.*, 1995] Lebart (L.), Morineau (A.) et Piron (M.). – *Statistique exploratoire multidimensionnelle.* – Dunod, 1995.

[Li *et al.*, 2001] Li (Q.), Zheng (J.), Zhou (Q.) et Lee (C.-H.). – A Robust Real-Time Endpoint Detector with Energy Normalization for ASR in Adverse Environments. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 233–236. – Salt Lake City, Utah, États-Unis, mai 2001.

[Lynch *et al.*, 1987] Lynch (J.F.), Josenhans (J.G.) et Crochiere (R.E.). – Speech/Silence Segmentation for Real-Time Coding Via Rule Based Adaptive Endpoint Detection. *International Conference on Acoustics, Speech, and Signal Processing*, pp. 1348–1351. – Dallas, États-Unis, avril 1987.

[Mak *et al.*, 1992] Mak (B.), Junqua (J.-C.) et Reaves (B.). – A Robust Speech/Non-Speech Detection Algorithm using Time and Frequency-Based Feature. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 269–272. – San Francisco, Californie, États-Unis, mars 1992.

[Martin *et al.*, 1989] Martin (T.B.), Rabiner (L.R.) et Wilpon (J.G.). – *Endpoint Detector.* – Brevet Américain, n°US4821325, avril 1989.

[Martin *et al.*, 2000] Martin (A.), Karray (L.) et Gilloire (A.). – High Order Statistics for Robust Speech/Non-Speech Detection. *European Signal Processing Conference*, pp. 469–472. – Tampere, Finlande, septembre 2000.

[Martin *et al.*, 2001a] Martin (A.), Charlet (D.) et Mauuary (L.). – Robust Speech/Non-Speech Detection Using LDA Applied to MFCC. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 237–240. – Salt Lake City, Utah, États-Unis, mai 2001.

[Martin *et al.*, 2001b] Martin (A.), Damnati (G.) et Mauuary (L.). – Robust Speech/Non-Speech Detection Using LDA Applied to MFCC for Continuous Speech Recognition. *European Conference on Speech Communication and Technology*, vol. 2, pp. 885–888. – Aalborg, Danemark, septembre 2001.

[Martin, 1982] Martin (Ph.). – Comparison of pitch detection by cepstrum and spectral combination analysis. *International Conference on Acoustics, Speech, and Signal Processing*, pp. 180–183. – 1982.

[Martin, 2000] Martin (A.). – Utilisation des moments d'ordre 3 pour une détection parole/non-parole robuste. *Journées d'Étude sur la Parole*, pp. 53–56. – Aussois, France, juin 2000.

[Martinez *et al.*, 1997] Martinez (R.), Àlvarez (A.), Gòmez (P.), Pèrez (M.), Nieto (V.) et Rodellar (V.). – A Speech Pre-Processing Technique for End-Point Detection in Highly

Non-Stationary Environments. *European Conference on Speech Communication and Technology*, vol. 3, pp. 1111–1114. – Rhodes, Grèce, septembre 1997.

[Mauuary et Karray, 1997] Mauuary (L.) et Karray (L.). – The Tuning of Speech Detection in the Context of a Global Evaluation of a Voice Response System. *European Conference on Speech Communication and Technology*, vol. 3, pp. 1539–1542. – Rhodes, Grèce, septembre 1997.

[Mauuary et Monné, 1993] Mauuary (L.) et Monné (J.). – Speech/Non-Speech Detection for Voice Response Systems. *European Conference on Speech Communication and Technology*, vol. 3, pp. 1097–1100. – Berlin, Allemagne, septembre 1993.

[Mauuary, 1994] Mauuary (L.). – *Amélioration des Performances des Serveurs Vocaux Interactifs.* – Thèse de doctorat, Université de Rennes 1, 1994.

[McCullagh, 1987] McCullagh (P.). – *Tensor Methods in Statistics.* – Charpman and Hall, 1987.

[Mokbel, 1992] Mokbel (C.). – *Reconnaissance de la Parole dans le Bruit: Bruitage/Débrutiatge.* – Thèse de doctorat, École Nationale Supérieure des Télécommunications, 1992.

[Montrésor et Baudry, 1990] Montrésor (S.) et Baudry (M.). – Représentation Temps-Échelle et Détection de la Fréquence Fondamentale du Signal de Parole. *Journées d'Étude sur la Parole*, pp. 170–174. – Montréal, Canada, mai 1990.

[Mwangi et Xydeas, 1985] Mwangi (E.) et Xydeas (C.). – Voiced-Unvoiced-Silence Classification of Speech Using Fuzzy Set Theory. *IEEE MELECON*, vol. 2, n° 4, 1985, pp. 123–126.

[Naito *et al.*, 1998] Naito (M.), Kuroiwa (S.), Takeda (K.) et Yamamoto (S.). – *Speech Endpoint Detection Method and Apparatus and Continuous Speech Recognition Method and Apparatus.* – Brevet Américain, n°US5740318, avril 1998.

[Nemer *et al.*, 1999] Nemer (E.), Goubran (R.) et Mahmoud (S.). – The Fourth-Order Cumulant of Speech Signals with Application to Voice Activity Detection. *European Conference on Speech Communication and Technology*, vol. 6, pp. 2391–2394. – Budapest, Hongrie, septembre 1999.

[Ney, 1981] Ney (H.). – An Optimization Algorithm for Determining the endpoints of isolated utterances. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 720–723. – Atlanta, Géorgie, États-Unis, mars 1981.

[Noé *et al.*, 2001] Noé (B.), Sienel (J.), Jouvet (D.), Mauuary (L.), Boves (L.), De Veth (J.) et De Wet (F.). – Noise Reduction for Noise Robust Feature Extraction for Distributed Speech Recognition. *European Conference on Speech Communication and Technology*, vol. 1, pp. 433–436. – Aalborg, Danemark, septembre 2001.

[Pincibono, 1993] Pincibono (B.). – *Signaux aléatoires.* – Dunod Université, 1993.

[Puel, 1997] Puel (J.B.). – *Reconnaissance Automatique de la Parole Téléphonique et Adaptation au GSM.* – Toulouse, Thèse de doctorat, Université Paul Sabatier, 1997.

[Rabiner et Juang, 1993] Rabiner (L.R.) et Juang (B.-H.). – *Fundamentals of Speech Recognition.* – Prentice Hall, 1993.

[Rabiner et Sambur, 1975] Rabiner (L.R.) et Sambur (M.R.). – An Algorithm for Determining the Endpoints of Isolated Utterances. *The bell system technical journal*, vol. 54, n° 2, février 1975, pp. 295–315.

[Rabiner et Sambur, 1977] Rabiner (L.R.) et Sambur (M.R.). – Application of an LPC Distance Measure to the Voiced-Unvoiced-Silence Detection Problem. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, n° 4, août 1977, pp. 338–343.

[Rabiner *et al.*, 1977] Rabiner (L.R.), Schmidt (C.E.) et Atal (B.S.). – Evaluation of a Statistical Approach to Voiced-Unvoiced-Silence Analysis for Telephone-Quality Speech. *The bell system technical journal*, vol. 56, n° 3, mars 1977, pp. 455–482.

[Ramana Rao et Srichand, 1996] Ramana Rao (G.V.) et Srichand (J.). – Word Boundary Detection Using Pitch Variations. *International Conference on Spoken Language Processing*, vol. 2, pp. 813–816. – Philadelphie, Pennsylvanie, États-Unis, octobre 1996.

[Rangoussi *et al.*, 1993] Rangoussi (M.), Bakamidis (S.) et Carayannis (G.). – Robust Enpoint Detection of Speech in the Presence of Noise. *European Conference on Speech Communication and Technology*, vol. 1, pp. 649–652. – Berlin, Allemagne, septembre 1993.

[Reaves, 1991] Reaves (B.). – Comments on "An Improved Endpoint Detector for Isolated Word Recognition". *IEEE Transactions on Signal Processing*, vol. 39, n° 2, février 1991, pp. 526–527.

[Reaves, 1997] Reaves (B.). – *Speech Detection Device for the Detection of Speech End Points Based on Variance of Frequency Band Limited Energy.* – Brevet Américain, n°US5617508, avril 1997.

[Renevey, 2000] Renevey (P.). – *Speech recognition in noisy conditions using missing feature approach.* – Thèse de doctorat, École Polytechnique Fédérale de Lausanne, 2000.

[Sakurai et Hirose, 1996] Sakurai (A.) et Hirose (K.). – Detection of Phrase Boundaries in Japanese by Low-Pass Filtering of Fundamental Frequency Contours. *International Conference on Spoken Language Processing*, vol. 2, pp. 817–820. – Philadelphie, États-Unis, octobre 1996.

[Saporta, 1990] Saporta (G.). – *Probabilités Analyse des Données et Statistiques.* – Editions Technip, 1990.

[Sarma et Venugopal, 1978] Sarma (V.V.S.) et Venugopal (D.). – Studies on Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1978, pp. 1–4.

[Savoji, 1989] Savoji (M.H.). – A Robust Algorithm for Accurate Endpointing of Speech Signals. *Speech Communication*, vol. 8, 1989, pp. 45–60.

[Segawa *et al.*, 2001] Segawa (O.), Takeda (K.) et Itakura (F.). – Continuous Speech Recognition Without End-Point Detection. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 245–248. – Salt Lake City, Utah, États-Unis, mai 2001.

[Seok et Bae, 1999] Seok (J.W.) et Bae (K.S.). – A Novel Endpoint Detection Using Discrete

Wavelet Transform. *IEICE Trans. Inf. & Syst.*, vol. E82-D, n° 11, novembre 1999, pp. 1489–1491.

[Shen *et al.*, 1998] Shen (J.-L.), Hung (J.-W.) et Lee (L.-S.). – Robust Entropy-Based Endpoint Detection for Speech Recognition in Noisy Environments. *International Conference on Spoken Language Processing*, vol. 3, pp. 1015–1018. – Sydney, Australie, décembre 1998.

[Shin *et al.*, 2000] Shin (W.-H.), Lee (B.-S.), Lee (Y.-K.) et Lee (J.-S.). – Speech/Non-Speech Classification Using Multiple Features for Robust Endpoint Detection. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1399–1402. – Istambul, Turquie, mai 2000.

[Shozakai *et al.*, 1998] Shozakai (M.), Namakamura (S.) et Shikano (K.). – Robust Speech in Car Environments. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 269–272. – Seattle, Washington, États-Unis, mai 1998.

[Singh *et al.*, 2001] Singh (R.), Seltzer (M.L.), Raj (B.) et Stern (R.M.). – Speech in Noisy Environments: Robust Automatic Segmentation, Feature Extraction, and Hypothesis Combination. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 273–276. – Salt Lake City, Utah, États-Unis, mai 2001.

[Smith *et al.*, 1999] Smith (D.C.), Townsend (J.), Nelson (D.J.) et D. (Richman). – A Multivariate Speech Activity Detector Based on the Syllable Rate. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 73–76. – Phoenix, Arizona, États-Unis, mai 1999.

[Sohn et Sung, 1998] Sohn (J.) et Sung (W.). – A Voice Activity Detector Employing Soft Decision Based Noise Spectrum Adaptation. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 365–368. – Seattle, Washington, États-Unis, mai 1998.

[Sokol, 1996] Sokol (R.). – *Réseaux Neuro-Flous et Reconnaissance des Traits Phonétiques pour l'Aide à la Lecture Labiale.* – Thèse de doctorat, Université de Rennes 1, 1996.

[Strom, 1995] Strom (V.). – Detection of Accents, Phrase Boundaries and Sentence Modality in German with Prosodic Features. *European Conference on Speech Communication and Technology*, vol. 3, pp. 2039–2041. – Madrid, Espagne, septembre 1995.

[Takeda *et al.*, 1995] Takeda (T.), Kuroiwa (S.), Nairo (M.) et Yamamoto (S.). – Top-Down Speech Detection and N-Best Meaning Search in a Voice Activated Telephone Extension System. *European Conference on Speech Communication and Technology*, vol. 2, pp. 1075–1078. – Madrid, Espagne, septembre 1995.

[Un et Lee, 1980] Un (C.K.) et Lee (H.H.). – Voiced-Unvoiced-Silence Discrimination of Speech by Delta Modulation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, n° 4, août 1980, pp. 398–407.

[Van Gerven et Xie, 1997] Van Gerven (S.) et Xie (F.). – A Comparative Study of Speech Detection Methods. *European Conference on Speech Communication and Technology*, vol. 3, pp. 1571–1574. – Rhodes, Grèce, septembre 1997.

[Wark et Sridharan, 1998] Wark (T.) et Sridharan (S.). – A Syntactic Approach to Automatic Lip Feature Extraction for Speaker Identification. *International Conference on*

Acoustics, Speech, and Signal Processing, vol. 6, pp. 3693–3696. – Seattle, Washington, États-Unis, mai 1998.

[Watanabe et Kimura, 1991] Watanabe (T.) et Kimura (T.). – *Method and Appatus for Speech Recognititon*. – Brevet Américain, n°US5062137, octobre 1991.

[Watson *et al.*, 1997] Watson (S.D.), Cheetham (B.M.G.), Barrett (P.A.), Wong (W.T.K.) et Lewis (A.V.). – A Voice Activity Detector for the ITU-T 8kbit/s Speech Coding Standard G.729. *European Conference on Speech Communication and Technology*, vol. 3, pp. 1571–1574. – Rhodes, Grèce, septembre 1997.

[Wilpon et Rabiner, 1987] Wilpon (J.G.) et Rabiner (L.R.). – Application of hidden markov models to automatic speech endpoint detection. *Computer Speech and Language*, n°2, 1987, pp. 321–341.

[Wu *et al.*, 1999] Wu (D.), Tanaka (M.), Chen (R.), Olorenshaw (L.), Amador (M.) et Menendez-Pidal (X.). – A Robust Speech Detection Algorithm for Speech Activated Hands-Free Applications. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 2407–2410. – Phoenix, Arizona, États-Unis, mars 1999.

[Xu *et al.*, 1992] Xu (L.), Krzyzak (A.) et Suen (C.Y.). – Methods of Combining Multiple Classifiers and Their Application to Handwriting Recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, n° 3, mai/juin 1992, pp. 418–435.

[Yamamoto *et al.*, 1997] Yamamoto (S.), Naito (M.) et Kuroiwa (S.). – Speech Detection Method for Speech Recognition System for Telecommunication Networks and its Field Trial. *European Conference on Speech Communication and Technology*, vol. 3, pp. 1535–1538. – Rhodes, Grèce, septembre 1997.

[Yang et Hsieh, 2000] Yang (C.-H.) et Hsieh (M.-S.). – Robust Endpoint Detection for In-Car Speech Recognition. *International Conference on Spoken Language Processing*, vol. 2, pp. 1061–1064. – Beijin, China, octobre 2000.

[Ying *et al.*, 1993] Ying (G.S.), Mitchell (C.D.) et Jamieson (L.H.). – Endpoint Detection of Isolated Utterances Based on a Modified Teager Energy Measurement. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 732–735. – Minneapolis, Minesota, États-Unis, avril 1993.

[Yoma *et al.*, 1996] Yoma (N.B.), McInnes (F.) et Jack (M.). – Robust Speech Pulse Detecion Using Adaptative Noise Modelling. *Electronics Letters*, vol. 32, n° 15, pp. 1350–1352. – juillet 1996.

[Zhu et Chen, 1999] Zhu (J.) et Chen (F.). – The Analysis and Application of a New Endpoint Detection Method Based on Distance of Autocorrelation Similarity. *European Conference on Speech Communication and Technology*, vol. 1, pp. 105–108. – Budapest, Hongrie, septembre 1999.

Glossaire

Automate : Dispositif qui réalise les opérations automatiques du cœur du module de détection.

Allophone : C'est un phonème en contexte dans un mot.

Bigramme : Modèle de langage, qui ne considère que le mot précédent.

Détection : C'est le signal détecté par le module de détection de parole.

Détection correcte : C'est une détection correctement reliée au segment manuel par le programme d'évaluation du module de détection.

Erreur de fausse acceptation : C'est une erreur de reconnaissance, qui correspond aux segments de bruit ou de parole hors vocabulaire pris pour de la parole du vocabulaire.

Erreur définitive : C'est un erreur de détection qui entraîne une erreur du module de reconnaissance. C'est une erreur de fragmentation, de regroupement ou d'omission

Erreur de fragmentation : C'est une erreur du module de détection. Un segment manuel est relié à plusieurs détections.

Erreur de regroupement : C'est une erreur du module de détection. Plusieurs segments manuels sont reliés à une seule détection.

Erreur de rejet à tort : C'est une erreur de reconnaissance de parole continue ou mots isolés, qui correspond aux segments non reconnus comme de la parole, et rejetés par le modèle de rejet ainsi que les segments non détectés par le module de détection. Dans le cas de la reconnaissance de mots isolés ces erreurs sont exprimées en segments, alors que dans le cas de la reconnaissance de parole continue elle sont exprimées en omission de mots.

Erreur d'insertion : C'est une erreur du module de détection, ou du module de reconnaissance de parole continue. Au niveau du module de détection, c'est une détection qui n'est pas reliée à un segment manuel de parole, par le programme d'évaluation du module de détection. Au niveau du module de reconnaissance, c'est une erreur de reconnaissance de parole continue, qui correspond aux insertions de mots par rapport aux mots contenus dans une requête.

Erreur d'omission : C'est une erreur du module de détection, ou du module de reconnaissance de parole continue. Au niveau du module de détection, c'est un segment manuel de parole qui n'est pas relié à une détection par le programme d'évaluation du module de détection. Au niveau du module de reconnaissance, c'est une erreur de reconnaissance de parole continue, qui correspond aux omissions de mots contenus dans une requête.

Erreur rejetable : C'est une erreur de détection qui peut être rejetée par le modèle de rejet du module de reconnaissance. C'est une erreur d'insertion.

Erreur de substitution : C'est une erreur de reconnaissance de parole continue ou mots isolés, qui correspond aux mots reconnus comme un autre mot du vocabulaire.

Étiquette : Label précisant le contenu d'un segment.

Locuteur : Personne qui émet la parole.

Modèle de rejet : Partie intégrée dans le module de reconnaissance qui permet de rejeter les détections de bruits ou de parole hors vocabulaire.

Parole continue : Dans ce terme est regroupé pour cette étude la parole spontanée et les phrases, qui sont ici des requêtes. La parole continue est ici en opposition avec les mots isolés.

Pitch : C'est le terme anglais qui couvre la fréquence des cordes vocales, la fréquence laryngienne si nous voulons faire référence au processus de génération articulatoire et la fréquence fondamentale si nous nous plaçons dans le domaine acoustique. Le pitch souvent confondu par abus de langage à la fréquence fondamentale, représente la périodicité du signal et sa structure harmonique.

Phonème : Unité théorique de la langue, qui permet de distinguer deux sons différents.

Requête : C'est la partie du signal issue de la segmentation manuelle et étiquetée qui contient une phrase. Ce terme est employé dans le cadre de la reconnaissance de parole continue.

Segment : C'est la partie du signal délimité par la segmentation manuelle et étiquetée. L'étiquette décrit le bruit ou le signal de parole sur cette partie du signal.

Segment élargi : C'est un segment relié à une détection correcte, mais dont la frontière gauche est détectée trop tôt, ou la frontière droite est détectée trop tard.

Segment tronqué : C'est un segment relié à une détection correcte, mais dont la frontière gauche est détectée trop tard, ou la frontière droite est détectée trop tôt.

Segment *Parole* **:** C'est un segment dont l'étiquette est de la parole.

Segment *Parole-Voc* **:** C'est un segment dont l'étiquette est de la parole du vocabulaire de l'application.

Segment *Parole-Hors-Voc* **:** C'est un segment dont l'étiquette est de la parole hors-vocabulaire de l'application.

Segment *Non-Parole* **:** C'est un segment dont l'étiquette n'est pas de la parole, mais divers bruits.

Segmentation : La segmentation ou segmentation manuelle est le procédé qui consiste à marquer et étiqueter manuellement des parties du signal. Ces parties de signal contiennent des bruits divers, de parole du locuteur, ou de parole d'une autre personne.

Seuil optimal : Le seuil optimal de la détection est le seuil qui donne le minimum des taux d'erreur associée de détection. Le seuil optimal de la reconnaissance est le seuil qui donne le minimum des taux d'erreur associée de reconnaissance.

Système de reconnaissance : Le système de reconnaissance comporte un module de reconnaissance et un module de détection.

Taux de fausse acceptation : Dans le cas de la reconnaissance de mots isolés, c'est le nombre d'erreurs de fausse acceptation divisé par le nombre de segments *Autre-Parole*

et *Non-Parole*, puis multiplié par cent. Dans le cas de la reconnaissance de parole continue, c'est le nombre d'erreurs de fausse acceptation divisé par le nombre total de mots, puis multiplié par cent.

Taux de substitution : Dans le cas de la reconnaissance de mots isolés, c'est le nombre d'erreurs de substitution divisé par le nombre de segments de *Parole*, puis multiplié par cent. Dans le cas de la reconnaissance de parole continue, c'est le nombre d'erreurs de substitution divisé par le nombre total de mots, puis multiplié par cent.

Taux de rejet à tort : Dans le cas de la reconnaissance de mots isolés, c'est le nombre d'erreurs de rejet à tort divisé par le nombre de segments de *Parole*, puis multiplié par cent. Dans le cas de la reconnaissance de parole continue, c'est le nombre d'erreurs de rejet à tort exprimées en omission de mots divisé par le nombre total de mots, puis multiplié par cent.

Taux d'erreur associée : Le taux d'erreur associée de détection est la somme des taux d'erreur rejetable et définitive. Le taux d'erreur associée de reconnaissance est la somme du taux d'erreur de rejet à tort et du taux de substitution et fausse acceptation pour la reconnaissance de mots isolés, et la somme des taux de rejet à tort, d'omission, d'insertion et de substitution pour la reconnaissance de parole continue.

Taux d'erreur définitive : C'est le nombre d'erreurs définitives divisé par le nombre de segments de parole (*Parole* et *Autre-Parole*), puis multiplié par cent.

Taux d'erreur rejetable : C'est le nombre d'erreurs rejetables divisé par le nombre de segments de parole (*Parole* et *Autre-Parole*), puis multiplié par cent.

Taux d'insertion : C'est le nombre d'erreurs d'insertion divisé par le nombre total de mots, puis multiplié par cent.

Taux d'omission : C'est le nombre d'erreurs d'omission divisé par le nombre total de mots, puis multiplié par cent.

Voisé : Ce dit d'un son émis par la vibration des cordes vocales.

Index